Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 1–9.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 5–13.
Original English Text Copyright © 2003 by Bronnikov, Meierovich.

GRAVITATION, ASTROPHYSICS

A Multidimensional Global Monopole and Nonsingular Cosmology[¶]

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

[a]Center for Gravitation and Fundamental Metrology, Russian Research Institute for Metrological Service, Moscow, 117313 Russia

[b]Institute of Gravitation and Cosmology, Peoples Friendship University of Russia, Moscow, 117198 Russia

*e-mail: kb@rgs.mccme.ru

[c]Kapitza Institute for Physical Problems, Russian Academy of Sciences, Moscow, 117334 Russia

**e-mail: meierovich@yahoo.com

Submitted January 29, 2003

Abstract—We consider a spherically symmetric global monopole in general relativity in $(D = d + 2)$-dimensional space–time. For $\gamma < d - 1$, where γ is a parameter characterizing the gravitational field strength, the monopole is shown to be asymptotically flat up to a solid angle defect. In the range $d - 1 < \gamma < 2d(d + 1)/(d + 2)$, the monopole space–time contains a cosmological horizon. Outside the horizon, the metric corresponds to a cosmological model of the Kantowski–Sachs type, where spatial sections have the topology $\mathbb{R} \times \mathbb{S}^d$. In the important case where the horizon is far from the monopole core, the temporal evolution of the Kantowski–Sachs metric is described analytically. The Kantowski–Sachs space–time contains a subspace with a $(d + 1)$-dimensional Friedmann–Robertson–Walker metric, whose possible cosmological application is discussed. Some estimates in the $d = 3$ case show that this class of nonsingular cosmologies can be viable. In particular, the symmetry-breaking potential at late times can give rise to both dark matter and dark energy. Other results, generalizing those known in 4-dimensional space–time, are derived, in particular, the existence of a large class of singular solutions with multiple zeros of the Higgs field magnitude. © *2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

In our recent paper with Podolyak [1], we considered the general properties of global monopole solutions in general relativity and developed some earlier results (see [2, 3] and references therein). It was confirmed, in particular, that the properties of these objects are governed by a single parameter γ, the squared energy of spontaneous symmetry breaking in Planck units. For $0 < \gamma < 1$, solutions with the entirely positive (or entirely negative) Higgs field are globally regular and asymptotically flat up to a solid angle deficit. In the range $1 < \gamma < 3$, the space–time of the solutions remains globally regular but contains a cosmological horizon at a finite distance from the center. Outside the horizon, the geometry corresponds to homogeneous anisotropic cosmological models of the Kantowski–Sachs type, whose spatial sections have the topology $\mathbb{R} \times \mathbb{S}^2$. The nonzero symmetry-breaking potential can be interpreted as a time-dependent cosmological constant, a kind of hidden vacuum matter. The potential tends to zero at late times, and the “hidden vacuum matter” disappears. This solution with a nonsingular static core and a cosmological metric outside the horizon drastically differs from the standard Big Bang models and conforms to the ideas advocated by Gliner and Dymnikova [4] that the standard Big Bang singularity could be replaced by a regular vacuum bounce.

The lack of isotropization at late times did not allow us to directly apply the toy model of a global monopole to the early phase of our Universe. But this circumstance does not seem to be a fatal shortcoming of the model because the anisotropy of the very early Universe could be damped by particle creation later, and the further stages with low energy densities might conform to the standard isotropic Friedmann cosmology. Another idea is to add a comparatively small positive quantity Λ to the symmetry-breaking potential (to “slightly raise the Mexican hat”). It can change nothing but the late-time asymptotic regime, which then becomes de Sitter, corresponding to the added cosmological constant Λ. These ideas deserve further study.

In this paper, we study the gravitational properties of global monopoles in multidimensional general relativity. This analysis can be of interest in view of numer-

---

[¶] This article was submitted by the authors in English.

ous attempts to construct a unified theory using the ideas of supersymmetry in higher dimensions. Objects like multidimensional monopoles, strings, and other topological defects might form due to phase transitions in the early Universe at possible stages when the present three spatial dimensions were not yet separated from others and a greater number of dimensions were equally important.

More specifically, we consider a self-gravitating hedgehog-type configuration of a multiplet of scalar fields with the Mexican-hat potential

$$V = (\lambda/4)(\phi^2 - \eta^2)^2$$

in D-dimensional space–time with a structure of $\mathbb{R}_t \times \mathbb{R}_\rho \times \mathbb{S}^d$ $(d = D - 2)$, where $\mathbb{R}_\rho$ is the range of the radial coordinate ρ and $\mathbb{R}_t$ is the time axis. The properties of such objects generalize the results obtained in [1] and earlier papers (e.g., [2, 3]) in a natural way. Thus, for small values of the parameter $\gamma = \kappa^2\eta^2$ characterizing the gravitational field strength, the solutions are asymptotically flat up to a solid angle deficit. Within a certain range $d - 1 < \gamma < \bar{\gamma}(d)$, the solutions are nonsingular but contain a Killing horizon and a cosmological metric of the Kantowski–Sachs type outside it. In the important case where the horizon is far from the monopole core, the temporal evolution of the Kantowski–Sachs metric is described analytically. The upper bound $\bar{\gamma}(d)$, beyond which there are no static solutions with a regular center, is also found analytically.

The above description applies to solutions with an entirely positive (or entirely negative) scalar field magnitude ϕ. As in [1], here we also find a class of solutions with any number n of zeros of $\phi(r)$, existing for $\gamma < \gamma_n(d)$, where the upper bounds γ_n are found analytically. All solutions with $n > 0$ describe space–times with a regular center, a horizon, and a singularity beyond this horizon.

We also discuss a possible cosmological application of multidimensional global monopoles, which can be of particular interest for a 5-dimensional space–time with 3-dimensional spheres $\mathbb{S}^d$. In this case, the Kantowski–Sachs type model has the spatial topology $\mathbb{R} \times \mathbb{S}^3$ outside the horizon. It is anisotropic in 4-dimensions, but the 3-dimensional spheres $\mathbb{S}^3$ are isotropic. The anisotropy is thus related only to the fourth coordinate t, which is spatial outside the horizon and is a cyclic variable from the dynamical viewpoint. If we identify $\mathbb{S}^3$ with the observed space, ignoring the extra coordinate, we obtain a closed cosmological model, with the Friedmann–Robertson–Walker line element in ordinary (3 + 1)-dimensional space–time.

A natural question arises: why is the fourth spatial dimension unobservable today? The answer cannot be found within our macroscopic theory without specifying the physical nature of the vacuum. The conventional Kaluza–Klein compactification of the extra dimension on a small circle is not satisfactory in our case because it leads to a singularity at the horizon (as demonstrated in Section 3). We therefore leave this question open and note that the global monopole model has the possibility of describing only the earliest phase of cosmological evolution. Its later stages should involve creation of matter and a sequence of phase transitions, possibly resulting in localization of particles across the t direction. We then obtain a model with a large but unobservable extra dimension, similar in spirit to the widely discussed brane world models (see reviews [5–7] and references therein).

The solutions of interest appear when the symmetry-breaking scale η is sufficiently large, and one can suspect that quantum gravity effects are already important on this energy scale. We show in Section 2.3 that this is not the case if the monopole core radius is much greater than the Planck length: the curvature and energy scales in the whole space are then much smaller than their Planckian values.

The existence of nonsingular models of the early Universe on the basis of classical gravity supports the opinion that our Universe had never undergone a stage described by full quantum gravity. In addition to those discussed here, such models are rather numerous now ([1, 4, 8–10], see also references therein). All of them are evidently free of the long-standing problems of the standard Big Bang cosmology related to the existence of multiple causally disconnected regions [11, 12].

This paper is organized as follows. In Section 2, we analyze the properties of a global monopole in $D = d + 2$ dimensions (one time coordinate and $d + 1$ spatial coordinates). It is a generalization of our previous results [1]. In Section 3, the particular case where $d = 3$ is studied in more detail along with its possible cosmological application. Unless otherwise indicated, we use the natural units $\hbar = c = 1$.

2. MULTIDIMENSIONAL GLOBAL MONOPOLE

2.1. General Characteristics

The most general form of a static, spherically symmetric metric in $D = d + 2$ dimensions is

$$ds^2 = e^{2F_0}dt^2 - e^{2F_1}d\rho^2 - e^{2F_\Omega}d\Omega^2, \tag{1}$$

where $d\Omega^2 = d\Omega_d^2$ is a linear element on a d-dimensional unit sphere parameterized by the angles φ_1, …, φ_d,

$$d\Omega_d^2 = d\varphi_d^2 + \sin^2\varphi_d(d\varphi_{d-1}^2 + \sin^2\varphi_{d-1}$$
$$\times(d\varphi_{d-2}^2 + \ldots + \sin^2\varphi_3(d\varphi_2^2 + \sin^2\varphi_2 d\varphi_1^2)\ldots)),$$

and F_0, F_1, and F_Ω are functions of the radial coordinate ρ not yet specified. The nonzero components of the

Ricci tensor are (the prime denotes $d/d\rho$)

$$\begin{aligned} R_0^0 &= e^{-2F_1}[F_0'' + F_0'(F_0' + dF_\Omega' - F_1')], \\ R_\rho^\rho &= e^{-2F_1}[dF_\Omega'' + F_0'' + dF_\Omega'^2 + F_0'^2 - F_1'(F_0' + dF_\Omega')], \\ R_2^2 &= \ldots = R_{d+1}^{d+1} = -(d-1)e^{-2F_\Omega} \\ &\quad + e^{-2F_1}[F_\Omega'' + F_\Omega'(F_0' + dF_\Omega' - F_1')]. \end{aligned} \qquad (2)$$

A global monopole with a nonzero topological charge can be constructed with a multiplet of real scalar fields ϕ^a ($a = 1, 2, \ldots, d + 1$) comprising a hedgehog configuration in $d + 1$ spatial dimensions,[1]

$$\phi^a = \phi(\rho)n^a(\varphi_1, \ldots, \varphi_d),$$

where $n^a(\varphi_1, \ldots, \varphi_d)$ is a unit vector ($n^a n^a = 1$) in the ($d + 1$)-dimensional Euclidean target space, with the components

$$\begin{aligned} n^{d+1} &= \cos\varphi_d, \\ n^d &= \sin\varphi_d\cos\varphi_{d-1}, \\ n^{d-1} &= \sin\varphi_d\sin\varphi_{d-1}\cos\varphi_{d-2}, \\ &\ldots \\ n^{d-k} &= \sin\varphi_d\sin\varphi_{d-1}\ldots\sin\varphi_{d-k}\cos\varphi_{d-k-1}, \\ &\ldots \\ n^2 &= \sin\varphi_d\ldots\sin\varphi_2\cos\varphi_1, \\ n^1 &= \sin\varphi_d\ldots\sin\varphi_2\sin\varphi_1. \end{aligned}$$

The Lagrangian of a multidimensional global monopole in general relativity is given by

$$L = \frac{1}{2}\partial_\mu\phi^a\partial^\mu\phi^a - V(\phi) + \frac{R}{2\kappa^2},$$

where R is the scalar curvature, $\kappa = \kappa_D$ is the D-dimensional gravitational constant, and $V(\phi)$ is a symmetry-breaking potential depending on $\phi = \pm\sqrt{\phi^a\phi^a}$; it is natural to choose V as the Mexican-hat potential,

$$V = \frac{\lambda}{4}(\phi^2 - \eta^2)^2 = \frac{\lambda\eta^4}{4}(f^2 - 1)^2. \qquad (3)$$

We have introduced the normalized field magnitude $f = \phi(\rho)/\eta$ playing the role of the order parameter. The model has a global $SO(d + 1)$ symmetry, which can be spontaneously broken to $SO(d)$; $\eta^{2/d}$ is the energy of symmetry breaking.

The Einstein equations can be written as

$$R_\mu^\nu = -8\pi G\tilde{T}_\mu^\nu = -8\pi G\left(T_\mu^\nu - \frac{1}{d}T\delta_\mu^\nu\right), \qquad (4)$$

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

$$\begin{aligned} \tilde{T}_0^0 &= -\frac{2}{d}V, \\ \tilde{T}_\rho^\rho &= -e^{-2F_1}f'^2 - \frac{2}{d}V, \\ \tilde{T}_2^2 &= \ldots = \tilde{T}_{d+1}^{d+1} = -e^{-2F_\Omega}f^2 - \frac{2}{d}V. \end{aligned}$$

We now use the quasiglobal coordinate ρ specified by the condition

$$F_0 + F_1 = 0,$$

which is a convenient gauge for spherically symmetric systems with Killing horizons. Introducing the functions

$$A(\rho) = e^{2F_0} = e^{-2F_1}, \quad r(\rho) = e^{F_\Omega},$$

we reduce the metric to the form

$$ds^2 = A(\rho)dt^2 - \frac{d\rho^2}{A(\rho)} - r^2(\rho)d\Omega^2 \qquad (5)$$

and obtain the equations

$$(Ar^d\phi')' - dr^{d-2}\phi = r^d\frac{\partial V}{\partial\phi}, \qquad (6)$$

$$r'' = -\frac{\kappa^2}{d}r\phi'^2, \qquad (7)$$

$$(r^dA')' = -\frac{4\kappa^2}{d}r^dV, \qquad (8)$$

$$A(r^2)'' - r^2A'' - (d-2)r^3r'\left(\frac{A}{r^2}\right)' = 2(d - 1 - \kappa^2\phi^2) \qquad (9)$$

for the unknown functions $\phi(\rho)$, $A(\rho)$, and $r(\rho)$. Only three of these four equations are independent: scalar field equation (6) follows from Einstein equations (7)–(9) because of the Bianchi identities.

Equations (6)–(8) have the same structure as Eqs. (13)–(15) in [1]. General properties of Eqs. (6)–(8) with an arbitrary value of d are the same as for $d = 2$, and the classification of their solutions is also the same. In particular, if $V(\phi) > 0$, the 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 the de Sitter space–time. Below, we focus our attention on solutions belonging to class (a1) according to [1], i.e., those with $r(\rho)$ monotonically growing from zero

---

[1] A 7D universe with a global monopole with a hedgehog configuration of scalar fields only in three extra dimensions was recently considered in [13]. Our approach is different. We consider a hedgehog configuration in all $D - 1$ space dimensions of D-dimensional space–time.

to infinity as $\rho \longrightarrow \infty$ and $A(\rho)$ changing from $A = 1$ at the regular center to $A_\infty < 0$ as $\rho \longrightarrow \infty$, and with a cosmological horizon (where $A = 0$) at some $\rho = \rho_h$.

Equation (9) is a second-order linear inhomogeneous differential equation for A. The corresponding homogeneous equation has the evident special solution

$$A(\rho) = \mathrm{const} \cdot r^2(\rho).$$

This allows expressing $A(\rho)$ in terms of $r(\rho)$ and $\phi(\rho)$ in an integral form,

$$A = C_1 r^2 - C_2 r^2 \int_\rho^\infty \frac{d\rho_1}{r^{d+2}(\rho_1)} + 2r^2 \int_\rho^\infty \frac{d\rho_1}{r^{d+2}(\rho_1)} \times \int_0^{\rho_1} d\rho_2 r^{d-2}(\rho_2)[d-1-\kappa^2\phi^2(\rho_2)]. \qquad (10)$$

We consider solutions with a large-r asymptotic behavior such that $r(\rho) \longrightarrow \infty$ and $r'(\rho) \longrightarrow \mathrm{const} > 0$ as $\rho \longrightarrow \infty$. Equation (7) gives r' as $\int [r\phi'^2] d\rho$, its convergence as $\rho \longrightarrow \infty$ implies a sufficiently rapid decay of ϕ' at large ρ, and therefore $\phi \longrightarrow \phi_\infty = \mathrm{const}$ as $\rho \longrightarrow \infty$. The potential V then tends to a constant equal to $V(\phi_\infty)$. Furthermore, Eq. (8) shows that at large r, $A(\rho)$ can grow at most as r^2 and, finally, substitution of the asymptotic form of $\phi(\rho)$, $A(\rho)$, and $r(\rho)$ in Eq. (6) leads to $dV/d\phi \longrightarrow 0$ as $\rho \longrightarrow \infty$. In application to field equations, the condition that there exists a large-r asymptotic regime implies that the scalar field then tends either to an extremum of the potential $V(\phi)$ or to an inflection point with zero derivative. For the Mexican-hat potential, it can be either the maximum at $\phi = 0$ (the trivial unstable solution for ϕ and the de Sitter metric with the cosmological constant $(1/4)\kappa^2\lambda\eta^4$) or a minimum of V where $f = 1$ and $V = 0$. For a “slightly raised Mexican hat” (potential (3) plus a small constant V_+), we have a de Sitter asymptotic behavior with $f = 1$ and $V = V_+$.

A regular center requires that $A = A_c + O(r^2)$ and $Ar'^2 \longrightarrow 1$ as $\rho \longrightarrow \rho_c$ such that $r(\rho_c) = 0$. Without loss of generality, we set $\rho_c = 0$ and $A_c = 1$.

For potential (3), regularity at $\rho = 0$ and the asymptotic condition at $\rho \longrightarrow \infty$ lead to $C_1 = C_2 = 0$ and Eq. (10) then implies that

$$A(\rho) = 2r^2(\rho) \int_\rho^\infty \frac{d\rho_1}{r^{d+2}(\rho_1)} \int_0^{\rho_1} d\rho_2 r^{d-2}(\rho_2) \times [d-1-\kappa^2\phi^2(\rho_2)]. \qquad (11)$$

Equation (8) provides another representation for $A(\rho)$ satisfying the regular center conditions,

$$A(\rho) = 1 - \frac{4\kappa^2}{d} \int_0^\infty \frac{d\rho_1}{r^d(\rho_1)} \int_0^{\rho_1} d\rho_2 r^d(\rho_2) V(\rho_2). \qquad (12)$$

From (11), we find the limiting value of A at $\rho \longrightarrow \infty$,

$$A(\infty) = \frac{d-1-\gamma}{\alpha^2(d-1)}, \quad \gamma = \kappa^2\eta^2, \qquad (13)$$

where $\alpha = dr/d\rho$ at $\rho \longrightarrow \infty$,

$$\alpha = 1 - \frac{\kappa^2}{d} \int_0^\infty r(\rho)\phi'^2(\rho) d\rho.$$

Equation (13) shows that $\gamma = d - 1$ is a critical value of γ: the large-r asymptotic behavior can be static only if $\gamma \le d - 1$; for $\gamma < d - 1$, it is flat up to a solid angle deficit, in full similarity to the conventional case $d = 2$ [1, 2]. If $\gamma > d - 1$, then $A(\infty) < 0$ and there is a horizon at some $\rho = \rho_h$ where $A = 0$. From (12),

$$\frac{4\kappa^2}{d} \int_0^{\rho_h} \frac{d\rho_1}{r^d(\rho_1)} \int_0^{\rho_1} d\rho_2 r^d(\rho_2) V(\rho_2) = 1$$

and we therefore have

$$A(\rho) = -\frac{4\kappa^2}{d} \int_{\rho_h}^{\rho} \frac{d\rho_1}{r^d(\rho_1)} \int_0^{\rho_1} d\rho_2 r^d(\rho_2) V(\rho_2). \qquad (14)$$

The γ dependence of ρ_h, with $\gamma = \kappa^2\eta^2$, can be found from the relation

$$\frac{4\kappa^2}{d} \int_{\rho_h}^{\infty} \frac{d\rho_1}{r^d(\rho_1)} \int_0^{\rho_1} d\rho_2 r^d(\rho_2) V(\rho_2) = -\frac{d-1-\gamma}{\alpha^2(d-1)}. \qquad (15)$$

2.2. Large-*r* Asymptotic Behavior

From (6), we can find the asymptotic behavior of the field $f(r)$ and the potential $V(r)$ as $r \longrightarrow \infty$. For large ρ, we have $A \longrightarrow A(\infty)$ (see (13)), and the field equation (6) reduces to

$$\frac{1}{r^d}\frac{d}{dr}\left(r^d \frac{df}{dr}\right) - \frac{d-1}{\gamma-d+1}\left[\lambda\eta^2(1-f^2) - \frac{d}{r^2}\right] f = 0,$$
$$r \longrightarrow \infty.$$

A regular solution of this equation must tend to unity as $r \longrightarrow \infty$, and for $\psi = 1 - f$, we have the linear equation

$$\psi_{,rr} + \frac{d}{r}\psi_{,r} + \frac{2\lambda\eta^2(d-1)}{\gamma-d+1}\left[\psi - \frac{d}{2\lambda\eta^2 r^2}\right] = 0, \qquad (16)$$
$$r \longrightarrow \infty.$$

The general solution of the corresponding homogeneous equation

$$\psi_{0,rr}+\frac{d}{r}\psi_{0,r}+\frac{2\lambda\eta^2(d-1)}{\gamma-d+1}\psi_0 = 0$$

can be expressed in terms of Bessel functions,

$$\psi_0(r) = r^{-(d-1)/2}\left[C_1J_{-(d-1)/2}\left(\frac{r}{r_0}\right)+C_2Y_{-(d-1)/2}\left(\frac{r}{r_0}\right)\right],$$

$$r_0^2 = \frac{\gamma-d+1}{2\lambda\eta^2(d-1)}.$$

A special solution of inhomogeneous equation (16) at $r \longrightarrow \infty$ is

$$\psi = \frac{d}{2\lambda\eta^2r^2}+O\left(\frac{1}{r^4}\right).$$

The general solution of Eq. (16) gives the asymptotic behavior for the Higgs field magnitude f as $r \longrightarrow \infty$,

$$f(r) = 1-\frac{d}{2\lambda\eta^2r^2}-\frac{C}{(\lambda\eta^2r^2)^{d/4}}\sin\left(\frac{r}{r_0}+\frac{\pi d}{4}+\varphi\right), \quad r \longrightarrow \infty. \tag{17}$$

Because of the boundary conditions imposed, the integration constants C and φ are functions of d and γ that can be found numerically. The function $C(\gamma)$ for $d = 3$ is presented in Fig. 1. From (17), we find the asymptotic behavior of V,

$$V(r) = \frac{\lambda\eta^4}{4}\left[\frac{d}{\lambda\eta^2r^2}+\frac{2C}{(\lambda\eta^2r^2)^{d/4}}\sin\left(\frac{r}{r_0}+\frac{\pi d}{4}+\varphi\right)\right]^2, \quad r \longrightarrow \infty. \tag{18}$$

2.3. *Bounds of the Classical Regime and the Monopole Core*

Of certain interest are solutions with the cosmological large-r behavior, i.e., those with $\gamma > d - 1$. The latter condition means that the scalar field, approaching η at large r, actually takes near- or trans-Planckian values. Indeed, in D dimensions, the Planck length l_D and mass m_D are expressed in terms of the gravitational constant $\kappa = \kappa_D$ as

$$l_D = \kappa^{2/d}, \quad m_D = \kappa^{-2/d}, \quad d = D-2.$$

Therefore,

$$\eta^2 = \frac{\gamma}{\kappa^2} = \gamma m_D^d,$$

and in the case of interest where $\gamma \sim d$, we have

$$\eta \sim (m_D)^{d/2}\sqrt{d}. \tag{19}$$

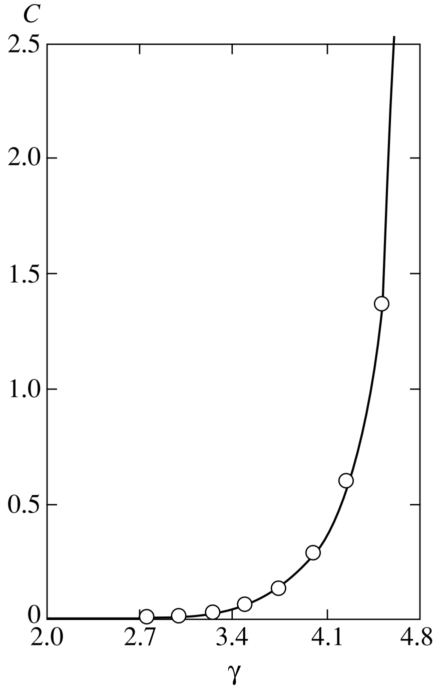


Fig. 1. Function $C(\gamma)$ found numerically for $d = 3$.

We can, however, remain at sub-Planckian curvature values, thus avoiding the necessity of invoking quantum gravity, if we require sub-Planckian values of the potential V in the entire space, i.e.,

$$\kappa^2V = \frac{1}{4}\kappa^2\lambda\eta^4 \ll m_D^2.$$

For η given by (19), this implies that

$$\lambda \ll \frac{4}{d^2}m_D^{2-d}. \tag{20}$$

We can thus preserve the classical regime even with large η by choosing sufficiently small values of λ. In terms of lengths, this condition is equivalent to the requirement that the monopole core radius

$$r_{\text{core}} = \frac{1}{\sqrt{\lambda}\eta}$$

is much greater than the Planck length,

$$\frac{1}{\sqrt{\lambda}\eta} \gg l_D. \tag{21}$$

One may note that this condition is external with respect to the theory because general relativity does not contain an internal restriction on the gravitational field strength. Moreover, in ordinary units, our dimensionless gravitational field strength parameter, expressed as $\gamma = \kappa^2c^{-4}\eta^2$, does not contain $\hbar$. We obtain restriction (20) or (21) only when we compare the characteristic length r_{core} existing in our theory with the Planck length $l_D = (\hbar\kappa^2/c^3)^{1/d}$.

We now discuss the solutions for γ slightly exceeding the critical value $d - 1$. In the case where $\gamma - (d - 1) \ll 1$, the horizon radius r_h is much greater than r_{core} and the constant C turns out to be negligibly small (this is confirmed numerically, see Fig. 1). At large ρ_2, the integrand in the inner integrals in (12), (14), and (15) is then given by

$$d\rho_2 r^d(\rho_2)V(\rho_2) \approx \frac{d^2}{4\alpha\lambda}\frac{dr}{r^{4-d}}.$$

The main contribution to the above inner integrals comes from the monopole core if $d < 3$ and from the upper limit if $d > 3$. For $d = 3$, it is a logarithmic integral. As a result, we have different behaviors of $\rho_h(\gamma)$ at $\gamma - (d - 1) \ll 1$ for $d = 2$ and $d \geq 3$.[2]

For $d = 2$ (4-dimensional general relativity),

$$\int_0^{\rho_1} d\rho_2 r^d(\rho_2)V(\rho_2) \approx \int_0^{\infty} d\rho_2 r^2(\rho_2)V(\rho_2) = \text{const}$$

and it follows from (15), in agreement with [1], that the horizon radius r_h, is inversely proportional to $\gamma - 1$,

$$r_h = \frac{\text{const}}{\gamma - 1}, \quad \gamma - 1 \ll 1, \quad d = 2.$$

For $d > 3$, we find that at $\gamma - (d - 1) \ll 1$, the horizon radius r_h is inversely proportional to the square root of $\gamma - (d - 1) \ll 1$,

$$r_h = \sqrt{\frac{\gamma d(d-1)}{2(d-3)(\gamma - d + 1)}\frac{1}{\lambda\eta^2}}, \qquad r_h^2 \gg \frac{1}{\lambda\eta^2}, \quad d > 3. \tag{22}$$

It is thus confirmed that for $\gamma - (d - 1) \ll 1$, the horizon is located far from the monopole core,

$$r_h^2 \gg \frac{1}{\lambda\eta^2}.$$

The function $A(r)$ at $r > r_h$ can then be found analytically. In this case, $r(\rho)$ is a linear function at $r > r_h$ and $dr = \alpha d\rho$. From (14) at $r > r_h$, we find

$$A(r) = -\frac{\gamma + 1 - d}{\alpha^2(d-1)}\left(1 - \frac{r_h^{d-1}}{r^{d-1}}\right) + \frac{\gamma d}{2\alpha^2(d-3)\lambda\eta^2 r^2}\left[1 - \left(\frac{r_h}{r}\right)^{d-3}\right]. \tag{23}$$

The condition of the applicability of (23) is $l_D \ll r_h$. In view of $r_{\text{core}} \ll r_h$, it is less restrictive than condition (21).

2.4. Solutions with $f(\phi)$ Changing Its Sign

As in [1], numerical integration of the field equations shows that in addition to solutions with totally positive (or totally negative) $f(u)$, there also exist solutions with a regular center such that $f(u)$ changes its sign $n \geq 1$ times. All these solutions exist for $\gamma < \gamma_n(d)$, where $\gamma_n(d)$ are some critical values of the parameter γ. For $n > 0$, all of them have a horizon, and the absolute value of f at the horizon $|f_{h,n}(\rho_h)|$ is a decreasing function of γ, vanishing as $\gamma \longrightarrow \gamma_n - 0$. Moreover, as $\gamma \longrightarrow \gamma_n(d)$, the function $f(u)$ vanishes in the whole range $\rho \leq \rho_h$ and is small inside the horizon for γ close to $\gamma_n(d)$. This allows us to find the critical values $\gamma_n(d)$ analytically: Eq. (6) reduces to a linear equation for f in a given (de Sitter) background and, combined with the boundary conditions $f(0) = 0$ and $f(\rho_h) < \infty$, leads to a linear eigenvalue problem. Its solution (see [1] for details) in the d-dimensional case gives the upper limits $\gamma_n(d)$ and the corresponding minimal horizon radii $r_h = r_{hn}$ for solutions with the Higgs field magnitude f changing its sign n times,

$$r_{hn} = \sqrt{(2n+1)(2n+d+2)/\lambda\eta^2}, \tag{24}$$

$$\gamma_n = \frac{2d(d+1)}{(2n+1)(2n+d+2)}. \tag{25}$$

For $d = 2$, Eqs. (24) and (25) reduce to Eq. (52) in [1]. Under condition (21), these solutions remain in the classical gravity regime.

3. 5-DIMENSIONAL MODELS AND NONSINGULAR COSMOLOGY

3.1. The Extra Dimension

At present, there is no evidence for the existence of more than three spatial dimensions up to achievable energies of about several hundred GeV. However, this energy is quite small on the Planck scale (on the order of 10^{19} GeV). Our solutions of possible cosmological interest correspond to $\gamma > d - 1$, i.e., the Planck energy scale. Even under condition (21), there remains an enormous range of scales in the early Universe in which the number of equally important spatial dimensions can be greater than 3.

If we try to consider our $d = 3$ solutions in the cosmological context, the extra coordinate is t in (1) and (5). The coordinate t is time inside the horizon and becomes a fourth spatial coordinate outside it, where $A(\rho) < 0$. Metric (5) takes the form

$$ds^2 = \frac{d\rho^2}{|A(\rho)|} - |A(\rho)|dt^2 - r^2(\rho)d\Omega_3^2.$$

Introducing the proper time τ of a comoving

---

[2] This is the only important qualitative difference between the general case $d \geq 3$ and the particular case $d = 2$ considered in [1].

observer outside the horizon,

$$\tau = \int_{\rho_h}^{\rho} \frac{d\rho}{\sqrt{|A(\rho)|}}, \qquad (26)$$

we obtain a 5-dimensional Kantowski–Sachs cosmology with a closed Friedmann–Robertson–Walker metric in the (3 + 1)-dimensional space–time section of constant t,

$$ds_4^2 = d\tau^2 - a^2(\tau)d\Omega_3^2 - |A(\rho(\tau))|dt^2. \qquad (27)$$

The 4-dimensional spherical radius $r(\rho)$ now plays the role of the scale factor, $a(\tau) = r(\rho(\tau))$.

It is tempting to explain the unobservability of the extra dimension parameterized by the t coordinate by compactifying t with a certain “period” T in the spirit of Kaluza–Klein models. Such a compactification would lead to a singularity at $r = r_h$, however, as is clear from Fig. 2. If $t \in \mathbb{R}$, the static region (the left quadrant in the diagram) is connected with the future cosmological region (the upper quadrant) by the horizon, crossed by photons and massive particles without problems, but if the t axis is made compact by identifying, e.g., the points t_1 and t_2 on the t axis, the static and cosmological regions in the diagram take the form of the dashed sectors, actually tubes of a variable thickness, connected at one point only, the ends (tips) of the tubes. The curvature invariants do not change due to this identification and remain finite, and the emerging singularity in the ρt plane resembles the conical singularity.

Compactification is not the only possibility of explaining why the t coordinate is invisible. It can also be assumed that at some instant of the proper cosmological time τ of 5-dimensional model (27), a phase transition occurs at a certain energy scale $1/T$, leading to localization of matter on the 3-spheres in the spirit of brane world models. Anyway, within our macroscopic theory without specifying the structure of the physical vacuum, it is impossible to explain why the extra dimension is not seen now. It is nevertheless interesting to describe some cosmological characteristics of the $d = 3$ global monopole.

3.2. Some Cosmological Estimates

For $d = 3$, the inner integrals in (14) and (15) have a logarithmic character, and instead of (22) and (23), we obtain

$$\gamma - 2 = \frac{3}{\lambda\eta^2 r_h^2}[B + \ln(\lambda\eta^2 r_h^2)], \qquad (28)$$

$$r_h^2 \gg \frac{1}{\lambda\eta^2}, \quad d = 3$$

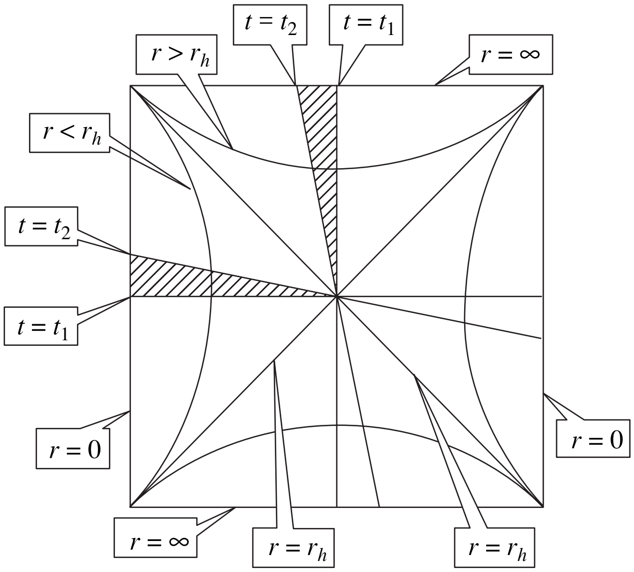


Fig. 2. Carter–Penrose diagram of a global monopole with a cosmological horizon. The diagonals of the square represent the horizons. After identification of t_1 and t_2, only the dashed regions survive.

and

$$A(a) = -\frac{\gamma - 2}{2\alpha^2}\left(1 - \frac{r_h^2}{a^2}\right) + \frac{3\gamma}{2\alpha^2\lambda\eta^2}\frac{\ln(a/r_h)}{a^2}, \qquad (29)$$

$$a > r_h, \quad d = 3.$$

The dependence $a(\tau)$ can be found from Eq. (26). In (28), B is a constant close to unity; our numerical estimate gives $B \approx 0.75$. The dimensionless radius of the horizon $\sqrt{\lambda}\eta r_h$ is presented in Fig. 3 as a function of γ for $d = 3$ (solid line). The dashed line is asymptotic dependence (28) valid for $\gamma - 2 \ll 1$ The function $A(\tau) \equiv A(a(\tau))$ is shown in Fig. 4 for $d = 3$ and $\gamma = 3$, 3.5, and 4. The numerical and analytic results are shown by solid and dashed lines, respectively. It is remarkable that only for $\gamma = 4$ is approximate analytic dependence (29)—which is, strictly speaking, valid for $\gamma - 2 \ll 1$—slightly different from the more precise dependence found numerically.

Far outside the horizon, $A(a)$ tends to a constant value,

$$A(a) \longrightarrow -\frac{\gamma - 2}{2\alpha^2}, \quad a \gg r_h,$$

and metric (27) describes a uniformly expanding world with a linear dependence $a(\tau)$ at late times,

$$a(\tau) = \alpha\sqrt{|A(\infty)|}\tau = \sqrt{\frac{\gamma - 2}{2}}\tau, \quad \tau \longrightarrow \infty. \qquad (30)$$

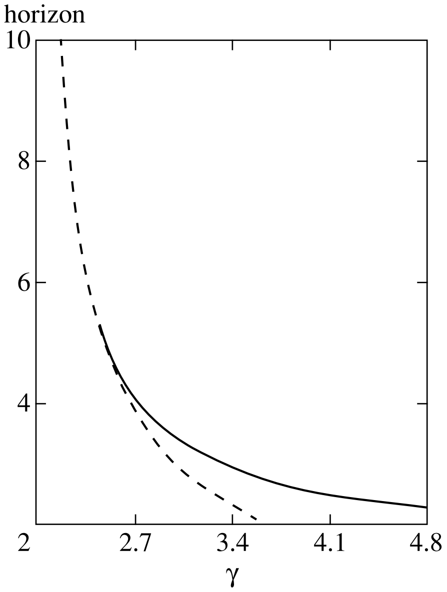


Fig. 3. The dimensionless horizon radius $\sqrt{\lambda}\eta r_h$ vs. γ for $d = 3$ (solid line). The dashed line is asymptotic dependence (28) valid for $\gamma - 2 \ll 1$.

The Hubble parameter $H = \dot a/a$, where the dot denotes $d/d\tau$, is found analytically from expression (29) for $A(a)$ ($d = 3$, $a > \rho_h \gg 1/\sqrt{\lambda}\eta$):

$$H(a) = \frac{1}{a}\sqrt{\frac{\gamma-2}{2}\left(1-\frac{r_h^2}{a^2}\right) - \frac{3\gamma}{2}\frac{\ln(a/r_h)}{\lambda\eta^2 a^2}}. \qquad (31)$$

The temporal evolution of the Hubble parameter $H(\tau)$ is shown in Fig. 5 for $\gamma = 3$, 3.5, and 4. The expansion starts from the horizon at $\tau = 0$ and rather quickly approaches the late-time behavior $H(\tau) = \tau^{-1}$. We actually have the asymptotic regime almost immediately after the beginning.

If we try to extrapolate this late-time regime to the present epoch, we can use the estimate given in [11] (Box 27.4), $\dot a \approx 0.66$; Eqs. (30) and (28) then lead to

$$\gamma = 2 + 2\dot a^2 = 2.87, \quad \sqrt{\lambda}\eta r_h \approx 3.65. \qquad (32)$$

These estimates conform to the monopole parameter values leading to a nonsingular cosmology.

Symmetry-breaking potential (18), averaged over the oscillations, $V(\tau) \equiv \overline{V(a(\tau))}$, is a decreasing function of τ,

$$V(\tau) = \frac{9}{(\gamma-2)^2\lambda\tau^4} + \frac{\lambda\eta^4 C^2}{2[(\gamma/2-1)\lambda\eta^2\tau^2]^{3/2}}, \qquad (33)$$
$$\tau \longrightarrow \infty.$$

In cosmology, scalar field potentials are often interpreted as a time-dependent effective cosmological constant. The reason is that V enters the energy-momentum tensor as a Λ term. In our case, as can be seen from (33), this term behaves as a mixture of two components, one decaying with the cosmological expansion as radiation ($\propto \tau^{-4} \propto a^{-4}$) and the other as matter without pressure ($\propto \tau^{-3} \propto a^{-3}$) in 4 dimensions. The 4-dimensional energy density corresponding to V is proportional to $V\sqrt{|A|}$. However, at late times, the extra-dimension scale factor $\sqrt{|A|}$ tends to a constant and therefore the 5- and 4-dimensional behaviors of the energy density actually coincide at large τ. We can say that the potential $V(\phi)$ in the global monopole model gives rise to

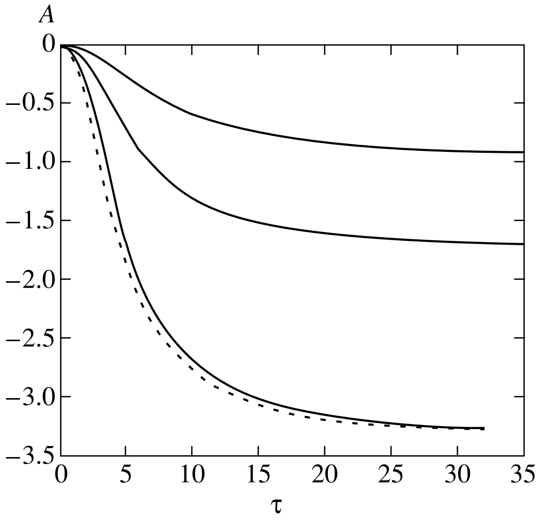


Fig. 4. Function $A(\tau) \equiv A(a(\tau))$ for $d = 3$ and $\gamma = 3$, 3.5, and 4 (from top down). Solid lines show numerical results and dashed lines show analytic dependence (29).

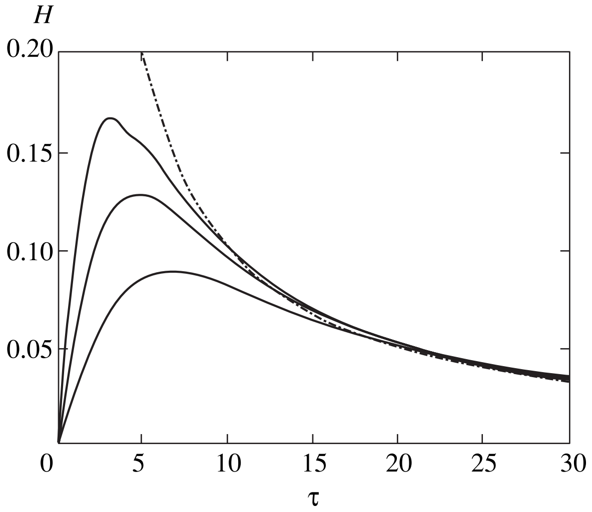


Fig. 5. Hubble parameter $H(\tau)$ for $\gamma = 3$, 3.5, and 4 (from bottom up). At late times, $H(\tau) = 1/\tau$ (dashed curve).

both dark radiation and dark matter. We recall that in accordance with modern views, both must necessarily be present in the Universe from the observational standpoint [12].

These estimates can only show that the 5-dimensional global monopole model is, in principle, able to give plausible cosmological parameters. Quantitative estimates certainly require a more complete model including further phase transitions, one of which should explain the unobservability of the fifth dimension.

The authors are grateful to A.F. Andreev, A.A. Starobinsky, V.A. Marchenko, and M.Yu. Kagan for useful discussions.

REFERENCES

1. K. A. Bronnikov, B. E. Meierovich, and E. R. Podolyak, Zh. Éksp. Teor. Fiz. **122**, 459 (2002) [JETP **95**, 392 (2002)].
2. A. Vilenkin and E. P. S. Shellard, *Cosmic Strings and Other Topological Defects* (Cambridge Univ. Press, Cambridge, 1994).
3. S. L. Liebling, Phys. Rev. D **61**, 024030 (1999).
4. E. B. Gliner, Usp. Fiz. Nauk **172**, 221 (2002) [Phys. Usp. **45**, 213 (2002)]; E. B. Gliner and I. G. Dymnikova, Usp. Fiz. Nauk **172**, 227 (2002).
5. V. A. Rubakov, Usp. Fiz. Nauk **171**, 913 (2001) [Phys. Usp. **44**, 871 (2001)]; E-print archives, hep-ph/0104152.
6. R. Maartens, E-print archives, gr-qc/0101059.
7. D. Langlois, E-print archives, gr-qc/0207047.
8. A. A. Starobinsky, Pis'ma Astron. Zh. **4**, 155 (1978) [Sov. Astron. Lett. **4**, 82 (1978)].
9. I. G. Dymnikova, A. Dobosz, M. L. Fil'chenkov, and A. A. Gromov, Phys. Lett. B **506**, 351 (2001).
10. K. A. Bronnikov and J. C. Fabris, J. High Energy Phys. **09**, 062 (2002); E-print archives, hep-th/0207213.
11. C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
12. V. Sahni and A. Starobinsky, E-print archives, astro-ph/9904398.
13. K. Benson and I. Cho, E-print archives, hep-th/0104067.

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 10–23.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 14–27.
Original English Text Copyright © 2003 by Ritus.

NUCLEI, PARTICLES, AND THEIR INTERACTION

The Symmetry, Inferable from Bogoliubov Transformation, between Processes Induced by a Mirror in Two-Dimensional and a Charge in Four-Dimensional Space–Time¶

V. I. Ritus

Tamm Department of Theoretical Physics, Lebedev Physical Institute, Russian Academy of Sciences, Moscow, 119991 Russia

e-mail: ritus@lpi.ru

Received December 30, 2002

Abstract—We consider the symmetry between creation of pairs of massless bosons or fermions by an accelerated mirror in (1 + 1)-dimensional space and emission of single photons or scalar quanta by an electric or scalar charge in (3 + 1)-dimensional space. The relation of Bogoliubov coefficients describing the processes generated by a mirror to Fourier components of the current or charge density implies that the spin of any disturbances bilinear in the scalar or spinor field coincides with the spin of quanta emitted by the electric or scalar charge. The mass and invariant momentum transfer of these disturbances are essential for the relation of Bogoliubov coefficients to invariant singular solutions and the Green functions of wave equations for both (1 + 1)- and (3 + 1)-dimensional spaces, and especially for the integral relations between these solutions. One of these relations leads to the coincidence of the self-action changes and vacuum–vacuum amplitudes for an accelerated mirror in two-dimensional space–time and a charge in four-dimensional space–time. Both invariants of the Lorentz group, spin and mass, play an essential role in the established symmetry. The symmetry embraces not only the processes of real quanta radiation, but also the processes of the mirror and charge interactions with fields carrying spacelike momenta. These fields accompany their sources and determine the Bogoliubov matrix coefficients $\alpha^{B,F}_{\omega'\omega}$. It is shown that the Lorentz-invariant traces $\pm\mathrm{tr}\,\alpha^{B,F}$ describe the vector and scalar interactions of the accelerated mirror with a uniformly moving detector. This interpretation rests essentially on the relation between propagators of the waves with spacelike momenta in two- and four-dimensional spaces. The traces $\pm\mathrm{tr}\,\alpha^{B,F}$ coincide with the products of the mass shift $\Delta m_{1,0}$ of the accelerated electric or scalar charge and the proper time of the shift formation. The symmetry fixes the value of the bare fine structure constant $\alpha_0 = 1/4\pi$. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The Hawking particle production mechanism during black hole formation is analogous to the emission from an ideal mirror accelerated in vacuum [1]. In turn, there is a close analogy between the radiation of pairs of scalar (spinor) quanta from an accelerated mirror in (1 + 1)-dimensional space and the radiation of photons (scalar quanta) by an accelerated electric (scalar) charge in (3 + 1)-dimensional space [2, 3]. All these processes are therefore interrelated. The *in* and *out* sets of the wave equation solutions typically used for a massless scalar field in problems with moving mirrors are given by

$$\phi_{\mathrm{in}\,\omega'} \propto e^{-i\omega' v} - e^{-i\omega' f(u)},\qquad \phi^*_{\mathrm{in}\,\omega'} \propto e^{i\omega' v} - e^{i\omega' f(u)}, \tag{1a}$$

$$\phi_{\mathrm{out}\,\omega} \propto e^{-i\omega g(v)} - e^{-i\omega u},\qquad \phi^*_{\mathrm{out}\,\omega} \propto e^{i\omega g(v)} - e^{i\omega u}, \tag{1b}$$

with the zero boundary condition

$$\phi|_{\mathrm{traj}} = 0$$

on the mirror's trajectory. Here, the variables $u = t - x$ and $v = t + x$ are used and the mirror (or charge) trajectory in the u, v plane is given by any of the two mutually inverse functions

$$v^{\mathrm{mir}} = f(u),\quad u^{\mathrm{mir}} = g(v).$$

We refer to [3] for the *in* and *out* sets of solutions of the massless Dirac equation. Dirac solutions differ from (1) by the presence of bispinor coefficients at the u- and v-plane waves. The current densities corresponding to these solutions have only tangential components at the boundary. The boundary condition for both scalar and spinor fields is therefore purely geometrical, and it does not contain any dimensional parameters.

The Bogoliubov coefficients $\alpha_{\omega'\omega}$ and $\beta_{\omega'\omega}$ arise as the coefficients of the expansion of the solutions of the

---

¶This article was submitted by the author in English.

out set in the solutions of the *in* set; the coefficients $\alpha^*_{\omega'\omega}$ and $\mp\beta_{\omega'\omega}$ arise similarly in the inverse expansion. The upper and lower signs correspond to the scalar (Bose) and spinor (Fermi) fields. The mean number of quanta with the frequency ω and the wave vector $\omega > 0$ radiated by the accelerated mirror to the right semispace is then given by the integral

$$d\bar{n}_\omega = \frac{d\omega}{2\pi}\int_0^\infty \frac{d\omega'}{2\pi}|\beta_{\omega'\omega}|^2. \qquad (2)$$

At the same time, the spectra of photons and scalar quanta emitted by electric and scalar charges moving along the trajectory $x_\alpha(\tau)$ in (3 + 1)-dimensional space are determined by the Fourier transforms of the electric current density 4-vector $j_\alpha(k)$ and the scalar charge density $\rho(k)$,

$$s = 1, \quad j_\alpha(k) = e\int d\tau \dot{x}_\alpha(\tau)\exp(-ik^\alpha x_\alpha(\tau)), \qquad (3)$$

$$d\bar{n}_k^{(1)} = |j_\alpha(k_+, k_-)|^2\frac{dk_+dk_-}{(4\pi)^2}, \qquad (4)$$

$$s = 0, \quad \rho(k) = e\int d\tau \exp(-ik^\alpha x_\alpha(\tau)), \qquad (5)$$

$$d\bar{n}_k^{(0)} = |\rho(k_+, k_-)|^2\frac{dk_+dk_-}{(4\pi)^2}, \qquad (6)$$

where s and k^α are the spin and 4-momentum of the quanta,

$$k^2 = k_1^2 + k_\perp^2 - k_0^2 = 0, \quad k_\perp^2 = k_0^2 - k_1^2 = k_+k_-,$$
$$k_\pm = k^0 \pm k^1,$$

and it is assumed in (4) and (6) that the trajectory $x^\alpha(\tau)$ has only x^0 and x^1 nontrivial components.

The symmetry between creation of Bose or Fermi pairs by the accelerated mirror in (1 + 1)-dimensional space and emission of single photons or scalar quanta by the electric or scalar charge in (3 + 1)-dimensional space consists, first of all, in the coincidence of the spectra. If we set

$$2\omega = k_+, \quad 2\omega' = k_-,$$

we have

$$\left|\beta^B_{\omega'\omega}\right|^2 = \frac{1}{e^2}|j_\alpha(k_+, k_-)|^2,$$
$$\left|\beta^F_{\omega'\omega}\right|^2 = \frac{1}{e^2}|\rho(k_+, k_-)|^2. \qquad (7)$$

A more refined assertion in the Bose case is

$$\beta^{B*}_{\omega'\omega} = -\sqrt{\frac{k_+}{k_-}}\frac{j_-(k)}{e} = \sqrt{\frac{k_-}{k_+}}\frac{j_+(k)}{e} = \frac{\varepsilon_{\alpha\beta}k^\alpha j^\beta(k)}{e\sqrt{k_+k_-}}, \qquad (8)$$

$$j_-(k) = e\int du \exp\left[\frac{i}{2}(k_+u + k_-f(u))\right],$$
$$j_+(k) = e\int dv \exp\left[\frac{i}{2}(k_-v + k_+g(v))\right]. \qquad (9)$$

The 2-vectors $j_\alpha(k)$ and $a_\beta(k) = \varepsilon_{\alpha\beta}k^\alpha/\sqrt{k_+k_-}$ are spacelike for timelike k^α; in a system where $k_+ = k_-$ or $\omega = \omega'$, they have only spatial components that are precisely equal to $e\beta^{B*}_{\omega'\omega}$ and 1, respectively.

In the Fermi case, we have

$$\beta^{F*}_{\omega'\omega} = \frac{1}{e}\rho(k). \qquad (10)$$

In Section 2, we underline the symmetry between analytical expressions for the Bogoliubov coefficients α and β^* and, at the same time, the physical distinction between them: $\beta^{B,F*}$ is the amplitude of the source of waves that are bilinear in massless Bose or Fermi fields and carry timelike momenta, whereas $\alpha^{B,F}$ is the amplitude of the source of similar waves that carry spacelike momenta (see (14) and (15)). In Sections 3 and 4, we show that the waves with timelike momenta emitted and absorbed by the source are involved in forming the imaginary part of the source self-action. This physical picture is naturally embodied in integral relation (20) between propagators $\Delta_2(z, m)$ of virtual pairs with masses m, $\mu \le m < \infty$ in two-dimensional space–time and the propagator $\Delta_4(z, \mu)$ of the particle in four-dimensional space–time. Analytic properties of the expressions obtained also allow us to define the real part of self-action. This leads to the coincidence of the self-actions and hence, of the vacuum–vacuum amplitudes of the mirror and the charge if we set $e^2 = 1$. In Section 5, the fields of perturbations carrying spacelike momenta are considered. These fields are defined by the matrices $\alpha^{B,F}$. Their Lorentz-invariant traces $\pm\mathrm{tr}\alpha^{B,F}$ are considered in Section 6. They describe, respectively, the vector and scalar interactions of the accelerated mirror with a uniformly moving detector in the neighborhood of the point of contact of their trajectories. In Sections 7 and 8, the traces $\pm\mathrm{tr}\alpha^{B,F}$ are found for three specific trajectories permitting analytic solutions. The general expressions for the traces are given, and their ultraviolet and infrared singularities are also considered there. In these sections, we compare the found traces $\pm\mathrm{tr}\alpha^{B,F}$ with the mass shifts $\Delta m_{1,0}$ of the electric and scalar charges moving along the same trajectory as the mirror, but in (3 + 1)-dimensional space.

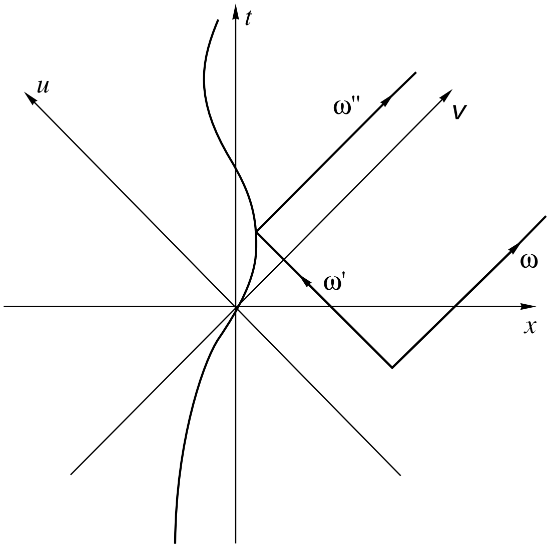


Fig. 1.

The mass shifts $\Delta m_{1,\,0}$ of the charges moving along the exponential trajectory are found in Section 9. In the Conclusions, we discuss the relation of the traces $\pm\mathrm{tr}\,\alpha^{B,\,F}$ to the general definition of the self-action accounting for interference effects and draw attention to the fact that the symmetry fixes the value of the bare charge squared, $e_0^2 = 1$, which corresponds to the bare fine structure constant $\alpha_0 = 1/4\pi$. The smallness and geometrical origin of this value may be interesting in quantum electrodynamics. In the Appendix, the even singular solutions of inhomogeneous wave equations with mass and momentum transfer parameters are considered. Integral relations (20) and (100) between these solutions for (1 + 1)- and (3 + 1)-dimensional spaces are very important for the symmetry considered.

2. PHYSICAL INTERPRETATION OF $\beta^*_{\omega'\omega}$

The absolute pair production amplitude and the single-particle scattering amplitude are related by [4]

$$\langle \mathrm{out}\,\omega''\omega|\mathrm{in}\rangle = -\sum_{\omega'}\langle \mathrm{out}\,\omega''|\omega'\mathrm{in}\rangle\beta^*_{\omega'\omega}. \qquad (11)$$

The coefficient $\beta^*_{\omega'\omega}$ was interpreted as the amplitude of a source of a pair of massless particles potentially emitted to the right and to the left with the respective frequencies ω and ω'. While the particle with the frequency ω actually escapes to the right, the particle with the frequency ω' propagates for some period of time and is then reflected by the mirror and is actually emitted to the right with an altered frequency ω'' (see Fig. 1).

In the time interval between pair creation and reflection of the left particle, we have the virtual pair with the energy k^0, momentum k^1, and mass m,

$$\begin{gathered} k^0 = \omega + \omega', \quad k^1 = \omega - \omega', \\ m = \sqrt{-k^2} = 2\sqrt{\omega\omega'}. \end{gathered} \qquad (12)$$

In addition to the polar timelike 2-vector k^α, the axial spacelike 2-vector q^α is very important,

$$\begin{gathered} q_\alpha = \varepsilon_{\alpha\beta}k^\beta, \quad q^0 = -k^1 = -\omega + \omega', \\ q^1 = -k^0 = -\omega - \omega' < 0. \end{gathered} \qquad (13)$$

In terms of k^α and q^α, the symmetry between the α and β coefficients becomes clearly expressed,

$$\begin{gathered} s = 1, \quad e\beta^{\beta*}_{\omega'\omega} = -\frac{q_\alpha j^\alpha(k)}{\sqrt{k_+k_-}}, \\ e\alpha^B_{\omega'\omega} = -\frac{k_\alpha j^\alpha(q)}{\sqrt{k_+k_-}}, \end{gathered} \qquad (14)$$

$$s = 0, \quad e\beta^{F*}_{\omega'\omega} = \rho(k), \quad e\alpha^F_{\omega'\omega} = \rho(q). \qquad (15)$$

We note that Eqs. (3) and (5) define the current density $j^\alpha(k)$ and the charge density $\rho(k)$ as functionals of the trajectory $x^\alpha(\tau)$ and functions of any 2- or 4-vector k^α. It can be shown that in (1 + 1)-dimensional space, $j^\alpha(k)$ and $j^\alpha(q)$ are spacelike and timelike polar vectors, respectively, if k^α and q^α are timelike and spacelike vectors.

The boundary condition on the mirror leads to the appearance of vector or scalar disturbance waves bilinear in massless fields in the vacuum of the massless scalar or spinor field. There are two types of these waves:

(1) waves with an amplitude of $\alpha_{\omega'\omega}$ ($\alpha^*_{\omega'\omega}$) that carry a spacelike momentum directed to the left (right);

(2) waves with an amplitude of $\beta^*_{\omega'\omega}$ ($\beta_{\omega'\omega}$) that carry a timelike momentum with a positive (negative) frequency.

The waves with spacelike momenta appear even if the mirror is at rest or moves uniformly (the Casimir effect), while the waves with timelike momenta appear only for the accelerated mirror.

The pair of Bose (Fermi) particles has spin 1 (0) because its source is a current density vector (charge density scalar) (see [5] or problem 12.15 in [6]).

3. THE APPEARANCE OF MASS IN THE MASSLESS THEORY AND OF INVARIANT SINGULAR SOLUTIONS OF THE WAVE EQUATION WITH MASS

It follows from (8) that the bilinear in massless Bose-field disturbances defined by the amplitudes $\beta^{B*}_{\omega'\omega}$ forms a positive-frequency current density vector. Its minus-component at the point *U*, *V* can be represented as

$$\iint_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}\frac{1}{e}j_-(k)\exp(-i\omega U - i\omega' V) = \frac{1}{8\pi^2} \times \int du \int_0^\infty d\rho\rho \int_{-\infty}^\infty d\theta\exp(-i\rho(z^0\cosh\theta - z^1\sinh\theta)), \quad (16)$$

if the hyperbolic variables ρ and θ are used instead of ω and ω',

$$d\omega d\omega' = \frac{1}{2}\rho d\rho d\theta, \quad \omega = \frac{1}{2}\rho e^\theta, \quad \omega' = \frac{1}{2}\rho e^{-\theta}, \quad \rho = 2\sqrt{\omega\omega'}, \quad \theta = \ln\sqrt{\frac{\omega}{\omega'}}, \quad (17)$$

and $z^\alpha = x^\alpha - x^\alpha(\tau)$ (see Fig. 2). As can be seen from (12), $\rho = m$ is the mass of the pair and θ is the rapidity. The integral over rapidity in (16) is the well-known invariant positive-frequency singular function of the wave equation for two-dimensional space–time,

$$\int_{-\infty}^\infty d\theta\exp(-im(z^0\cosh\theta - z^1\sinh\theta)) = -4\pi i\Delta_2^+(z, m) \quad (18)$$

$$= 2\theta(-z^2)K_0(i\varepsilon(z^0)m\sqrt{-z^2}) + 2\theta(z^2)K_0(m\sqrt{z^2}),$$

$$(\partial_t^2 - \partial_x^2 + m^2)\Delta_2^+(z, m) = 0. \quad (19)$$

This function describes the wave field of pairs with mass *m* and any possible positive-frequency momenta. It follows that the pairs are created, propagated, and absorbed near the mirror within a spacelike interval of the order of m^{-1}.

Using the very important integral relation between the singular functions of wave equations for *d*- and (*d* + 2)-dimensional space–times,

$$\Delta^f_{d+2}(z, \mu) = \frac{1}{4\pi}\int_{\mu^2}^\infty dm^2\Delta^f_d(z, m), \quad (20)$$

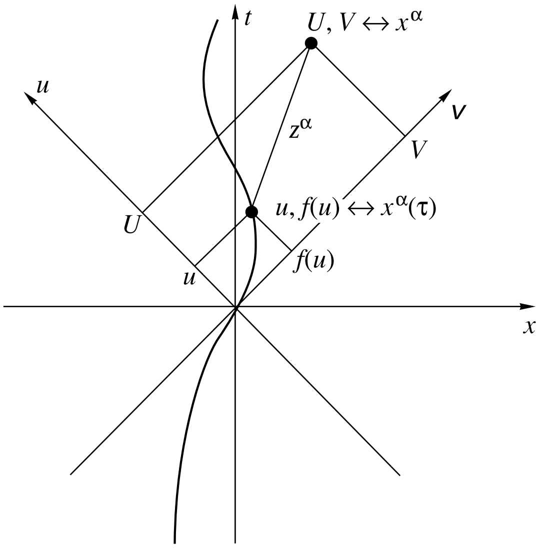


Fig. 2.

we can represent the right-hand side of (16) as

$$-\frac{i}{4\pi}\int du \int_{\mu^2\to 0}^\infty dm^2\Delta_2^+(z, m) = -i\int_{-\infty}^\infty du\Delta_4^+(z, \mu). \quad (21)$$

The small mass μ is retained to eliminate the infrared divergence in what follows.

Similarly, the positive-frequency plus-component of the current density at the point *U*, *V* can be represented as

$$\iint_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}\frac{1}{e}j_+(k)\exp(-i\omega U - i\omega' V) = -i\int_{-\infty}^\infty dv\Delta_4^+(z, \mu). \quad (22)$$

The differentials *du* and *dv* in (21) and (22) can be replaced by $d\tau\dot{x}_-(\tau)$ and $d\tau\dot{x}_+(\tau)$.

The bilinear in massless Fermi-field disturbances defined by the amplitudes $\beta^{F*}_{\omega'\omega}$ forms a positive-frequency charge density scalar. At the point *U*, *V*, it can be represented by

$$\iint_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}\frac{1}{e}\rho(k)\exp(-i\omega U - i\omega' V) = -i\int_{-\infty}^\infty d\tau\Delta_4^+(z, \mu). \quad (23)$$

If we set the point U, V on the trajectory, so that

$$U = x_-(\tau'), \quad V = x_+(\tau'), \quad z^\alpha = x^\alpha(\tau') - x^\alpha(\tau),$$

and integrate (21) over V and (22) over U, then their half-sum differs from $\mathrm{tr}\beta^+\beta$ only by the factor i,

$$\begin{aligned} \mathrm{tr}\beta^{B+}\beta^B &\equiv \int\!\!\int_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}\left|\beta^B_{\omega'\omega}\right|^2 \\ &= \frac{i}{2}\int\!\!\int(dudV + dvdU)\Delta_4^+(z, \mu) \\ &= -i\int\!\!\int d\tau d\tau' \dot{x}_\alpha(\tau)\dot{x}^\alpha(\tau')\Delta_4^+(z, \mu). \end{aligned} \tag{24}$$

The real part of the function Δ^+, which is odd in z, and its imaginary part, which is even in z, are related to the causal (Feynman) function Δ^f that is even in z,

$$\begin{gathered} \Delta^+(z, \mu) = \frac{1}{2}\Delta(z, \mu) + \frac{i}{2}\Delta^1(z, \mu), \\ \mathrm{Re}\Delta^+ = \varepsilon(z^0)\mathrm{Re}\Delta^f, \quad \mathrm{Im}\Delta^+ = \mathrm{Im}\Delta^f, \end{gathered} \tag{25}$$

and $\mathrm{tr}\,\beta^{B+}\beta^B$ can therefore be written as

$$\mathrm{tr}(\beta^+\beta)^B = \mathrm{Im}\int\!\!\int d\tau d\tau' \dot{x}_\alpha(\tau)\dot{x}^\alpha(\tau')\Delta_4^f(z, \mu). \tag{26}$$

$\beta^{F+}\beta^F$ can be obtained from the right-hand side of (26) by the substitution

$$\dot{x}_\alpha(\tau)\dot{x}^\alpha(\tau') \longrightarrow 1.$$

4. VACUUM–VACUUM AMPLITUDE $\langle \mathrm{out}|\mathrm{in}\rangle = e^{iW}$

According to DeWitt [7], Wald [8], and others (including the present author [4]),

$$\begin{gathered} 2\mathrm{Im}W^{B,F} = \pm\frac{1}{2}\mathrm{tr}\ln(1 \pm \beta^+\beta) \\ \text{or} \quad \pm\mathrm{tr}\ln(1 \pm \beta^+\beta) \end{gathered} \tag{27}$$

in the respective cases where the particle is identical or nonidentical to the antiparticle. We confine ourselves to the last case and assume that $\mathrm{tr}\beta^+\beta \ll 1$. Then

$$2\mathrm{Im}W^{B,F} = \mathrm{Im}\int\!\!\int d\tau d\tau' \begin{Bmatrix} \dot{x}_\alpha(\tau)\dot{x}^\alpha(\tau') \\ 1 \end{Bmatrix} \Delta_4^f(z, \mu). \tag{28}$$

We can omit the Im symbols from both sides of this equation and define the actions for Bose and Fermi mirrors in (1 + 1)-dimensional space as

$$W^{B,F} = \frac{1}{2}\int\!\!\int d\tau d\tau' \begin{Bmatrix} \dot{x}_\alpha(\tau)\dot{x}^\alpha(\tau') \\ 1 \end{Bmatrix} \Delta_4^f(z, \mu). \tag{29}$$

We compare this with the well-known actions for electric and scalar charges in (3 + 1)-dimensional space,

$$W^{1,0} = \frac{1}{2}e^2\int\!\!\int d\tau d\tau' \begin{Bmatrix} \dot{x}_\alpha(\tau)\dot{x}^\alpha(\tau') \\ 1 \end{Bmatrix} \Delta_4^f(z, \mu). \tag{30}$$

The symmetry would be complete if $e^2 = 1$, i.e., if the fine structure constant were $\alpha = 1/4\pi$. This “ideal” value of the fine structure constant for the charges would correspond to the ideal, geometric boundary condition on the mirror.

For the mirror trajectory with a nonzero relative velocity β_{21} of its ends (nonzero relative rapidity $\theta = \mathrm{arctanh}\beta_{21}$), the changes of the actions due to acceleration are given by

$$\begin{aligned} \mathrm{Re}\Delta W^B &= \frac{1}{8\pi}\left(\frac{\theta}{\tanh\theta} - 1\right), \\ \mathrm{Re}\Delta W^F &= \frac{1}{8\pi}\left(1 - \frac{\theta}{\sinh\theta}\right). \end{aligned} \tag{31}$$

For a uniformly accelerated mirror with the proper acceleration a, its velocity is

$$\beta(\tau) = \tanh a\tau,$$

where τ is the proper time. Then,

$$\theta = a(\tau_2 - \tau_1)$$

and as $\tau_2 - \tau_1 \longrightarrow \infty$,

$$\mathrm{Re}\Delta W^B = \frac{|a|}{8\pi}(\tau_2 - \tau_1). \tag{32}$$

By definition,

$$\mathrm{Re}\Delta m^B = -\frac{\partial \mathrm{Re}\Delta W^B}{\partial \tau_2} = -\frac{|a|}{8\pi} \tag{33}$$

is the self-energy shift of an accelerating Bose mirror. It differs from the mass shift of a uniformly accelerated electron only by the absence of the factor $e^2 = 4\pi\alpha$. The self-energy shift of a uniformly accelerated Fermi mirror is

$$\mathrm{Re}\Delta m^F = 0.$$

There are two arguments in favor of defining the action by means of the causal function $\Delta_4^f(z, \mu)$.

1. The action must represent not only the radiation of real quanta but also the self-energy and polarization effects. While the radiation effects are described by solutions of the homogeneous wave equation, the self-energy and polarization effects require solutions of the inhomogeneous wave equation, which contain information about the proper field of a source. Such solutions of the homogeneous and inhomogeneous wave equations are the functions

$$(1/2)\Delta^1 = \mathrm{Im}\Delta^f$$

and

$$\bar{\Delta} = \mathrm{Re}\Delta^f.$$

2. While the appearance of

$$(1/2)\Delta^1 \equiv \mathrm{Im}\Delta^f$$

in the imaginary part of the action is a consequence of mathematical transformations of the integral

$$\int\!\!\int_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}|\beta_{\omega'\omega}|^2$$

(transformations similar to the Plancherel theorem), the function $\bar{\Delta} \equiv \mathrm{Re}\Delta^f$ in the real part of the action is unique if it appears as the real part of the analytic continuation of $(i/2)\Delta^1(z, \mu)$ to negative z^2 that is even in z (as Δ^1 itself).

To conclude Sections 3 and 4, we note that both the function $\Delta_2(z, m)$ describing the propagation of a virtual pair with mass $m = \rho = 2\sqrt{\omega\omega'}$ in two-dimensional space–time and the mass spectrum of these pairs arise owing to the transition from the variables ω and ω' to the hyperbolic variables ρ and θ, which reflect the Lorentzian symmetry of the problem. Further integration over the mass leads to the function $\Delta_4(z, \mu)$ that coincides with the propagator of a particle moving in four-dimensional space–time with the mass μ equal to the least mass of virtual pairs. Thus, relation (20) appears in the framework of the present method and is immanent to the symmetry, relating the processes in two- and four-dimensional space–times.

In [9], relation (20) was obtained by the author independently of the processes considered and was required in proving that the integration variable involved in it coincides with the pair mass $m = 2\sqrt{\omega\omega'}$.

5. FORMATION OF TACHYON DISTURBANCES WITH THE INVARIANT MOMENTUM TRANSFER

The bilinear in massless Bose field perturbations that are defined by the amplitudes $\alpha^B_{\omega'\omega}$ and carry spacelike momenta to the left can be represented at the point U, V by the two current density components

$$\begin{aligned}&\int\!\!\int_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}\frac{1}{e}j_\pm(q)\exp(i\omega U - i\omega' V)\\ &= \frac{1}{8\pi^2}\int d\tau \dot{x}_\pm(\tau)\int_0^\infty d\rho\rho\int_{-\infty}^{\infty} d\theta \exp[i\rho(z^0\sinh\theta - z^1\cosh\theta)]\end{aligned} \tag{34}$$

if we again use the change of variables in (17) and the notation

$$z^\alpha = x^\alpha - x^\alpha(\tau).$$

The integral over θ is now given by

$$\begin{aligned}&\int_{-\infty}^{\infty} d\theta \exp[i\rho(z^0\sinh\theta - z^1\cosh\theta)] = 4\pi i\Delta^L_2(z, \rho)\\ &= 2\theta(-z^2)K_0(\rho\sqrt{-z^2}) + 2\theta(z^2)K_0(i\varepsilon(z^1)\rho\sqrt{z^2}).\end{aligned} \tag{35}$$

The integrand in the left-hand side of (35) is a wave with a spacelike 2-momentum q^α,

$$q^1 = -\omega - \omega' = -\rho\cosh\theta,$$

$$q^0 = -\omega + \omega' = -\rho\sinh\theta,$$

$$\rho = \sqrt{q^2}.$$

The function $\Delta^L_2(z, \rho)$ is a superposition of plane waves with spacelike momenta directed to the left and with a fixed invariant momentum transfer $\rho = 2\sqrt{\omega\omega'}$. It satisfies the wave equation with a negative mass squared,

$$(\partial_t^2 - \partial_x^2 - \rho^2)\Delta^L_2(z, \rho) = 0. \tag{36}$$

Using the integral relation similar to (20) (see the Appendix), we can represent the right-hand side of (34) as

$$\begin{aligned}&\frac{i}{4\pi}\int d\tau \dot{x}_\pm(\tau)\int_{\nu^2\to 0}^{\infty} d\rho^2\Delta^L_2(z, \rho)\\ &= -i\int d\tau \dot{x}_\pm(\tau)\Delta^L_4(z, \nu).\end{aligned} \tag{37}$$

The small momentum transfer ν is retained to eliminate the infrared divergence in what follows.

Similarly, the bilinear in the Fermi field disturbances that are defined by the amplitudes $\alpha^F_{\omega'\omega}$ and carry left-directed spacelike momenta form the charge density scalar. It can be represented at the point U, V by the integral

$$\begin{aligned}&\int\!\!\int_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}\frac{1}{e}\rho(q)\exp(i\omega U - i\omega' V)\\ &= -i\int d\tau\Delta^L_4(z, \nu).\end{aligned} \tag{38}$$

These representations can be useful in problems close to static ones involving another characteristic length in addition to or instead of acceleration.

6. INTERPRETATION OF THE TRACES $\pm\mathrm{tr}\alpha^{B,F}$ OF BOGOLIUBOV COEFFICIENTS

The invariant description of the mirror trajectory in the u, v plane requires that the function

$$u^{\mathrm{mir}} = g(v)$$

contains two positive parameters κ and κ' transforming as

$$x_+ = v, \quad x_- = u$$

and actually connects the invariant variables κu and $\kappa' v$ between themselves,

$$u^{\mathrm{mir}} = g(v) = \frac{1}{\kappa}G(\kappa' v). \tag{39}$$

Its expansion near the origin $u = v = 0$ on the trajectory is given by

$$g(v) = \frac{1}{\kappa}\left(\kappa' v + b\kappa'^2 v^2 + \frac{1}{3}c\kappa'^3 v^3 + \dots\right), \tag{40}$$

where b, c, ... are some numbers. Because the mirror velocity $\beta(v)$ and the proper acceleration $a(v)$ are defined by

$$\beta(v) = \frac{1-g'(v)}{1+g'(v)}, \quad a(v) = -\frac{g''(v)}{2g'^{3/2}(v)}, \tag{41}$$

the first two coefficients of expansion (40) define the mirror velocity β_0 and acceleration a_0 at the zero point,

$$\beta_0 = \frac{1-\kappa'/\kappa}{1+\kappa'/\kappa}, \quad a_0 = -b\sqrt{\kappa\kappa'}. \tag{42}$$

The absolute value of the acceleration at the zero point is denoted by

$$w_0 = |b|\sqrt{\kappa\kappa'}.$$

We define a Lorentz-invariant trace of α by the formula

$$\mathrm{tr}\alpha = \int\!\!\int_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}\alpha_{\omega'\omega}2\pi\delta\left(\sqrt{\frac{\kappa'}{\kappa}}\omega - \sqrt{\frac{\kappa}{\kappa'}}\omega'\right), \tag{43}$$

where the Lorentz-invariant argument of the δ-function is the difference of the frequencies

$$\Omega = \sqrt{\frac{\kappa'}{\kappa}}\omega, \quad \Omega' = \sqrt{\frac{\kappa}{\kappa'}}\omega' \tag{44}$$

of the reflected and incident waves in the proper system of the mirror at the moment $u = v = 0$. In accordance with (42), the multipliers $\sqrt{\kappa'/\kappa}$ and $\sqrt{\kappa/\kappa'}$ entering (44) are the Doppler factors relating the frequencies in the laboratory and proper systems. In the proper system of the mirror,

$$\Omega = \Omega' = \sqrt{\omega\omega'}.$$

In accordance with (43), $\mathrm{tr}\alpha$ is a Lorentz-invariant dimensionless quantity or, perhaps, has the dimensionality of the action because $\hbar = 1$. We now consider its physical meaning. For this, we turn to the equality of expressions (34) and (37),

$$\begin{aligned}&\int\!\!\int_0^\infty \frac{d\omega d\omega'}{(2\pi)^2}\frac{1}{e}j_\pm(q)\exp(i\omega U - i\omega' V) \\ &= -i\int d\tau \dot{x}_\pm(\tau)\Delta_4^L(z, \nu),\end{aligned} \tag{45}$$

where

$$z^\alpha = x^\alpha - x^\alpha(\tau), \quad x_- = U, \quad x_+ = V.$$

We set the point U, V on the tangent line to the mirror trajectory at zero point, so that

$$\begin{aligned}U &= X_-(\tau') = \sqrt{\frac{\kappa'}{\kappa}}\tau', \\ V &= X_+(\tau') = \sqrt{\frac{\kappa}{\kappa'}}\tau',\end{aligned} \tag{46}$$

where τ' is the proper time of the point on the tangent line, and integrate both sides of (45) over

$$dU = \dot{X}_- d\tau' \quad \text{or} \quad dV = \dot{X}_+ d\tau'$$

for the upper or lower sign in (45), respectively. Taking into account Eq. (14) and current conservation, we then obtain $\mathrm{tr}\alpha$ in the left-hand side for both the upper and lower signs in (45). In the right-hand side, we obtain the integral

$$\begin{aligned}&-i\int\!\!\int d\tau d\tau' \dot{x}_\pm(\tau)\dot{X}_\mp(\tau')\Delta_4^L(z, \nu), \\ &z^\alpha = X^\alpha(\tau') - x^\alpha(\tau),\end{aligned} \tag{47}$$

where according to the result for the left-hand side, we can replace

$$\begin{aligned}&\dot{x}_\pm(\tau)\dot{X}_\mp(\tau') \\ &= -\dot{x}_\alpha(\tau)\dot{X}^\alpha(\tau') \mp \varepsilon_{\alpha\beta}\dot{x}^\alpha(\tau)\dot{X}^\beta(\tau')\end{aligned} \tag{48}$$

with only the first term that is symmetric with respect to the permutation

$$\dot{x}_\alpha(\tau) \rightleftarrows \dot{X}_\alpha(\tau').$$

We thus obtain

$$\begin{aligned}\mathrm{tr}\alpha^B &= i\int\!\!\int d\tau d\tau' \dot{x}_\alpha(\tau)\dot{X}_\alpha(\tau')\Delta_4^L(z, \nu), \\ z^\alpha &= X^\alpha(\tau') - x^\alpha(\tau).\end{aligned} \tag{49}$$

Integrating both parts of Eq. (38) along tangent line (46) similarly and taking Eqs. (15) and (43) into account, we obtain

$$\begin{aligned} \mathrm{tr}\alpha^F &= -i\iint d\tau d\tau' \Delta_4^L(z, \nu), \\ z^\alpha &= X^\alpha(\tau') - x^\alpha(\tau). \end{aligned} \tag{50}$$

For trajectories in the Minkowsky plane on the left of their tangent line at zero point, the coordinate $z^1 \ge 0$. In this case, $\Delta_4^L(z, \nu)$ can be replaced by the function

$$\begin{aligned} \Delta_4^{LR}(z, \nu) &= \frac{1}{4\pi}\delta(z^2) - \frac{\nu}{8\pi\sqrt{z^2}}\theta(z^2) \\ &\times [J_1(\nu\sqrt{z^2}) - iN_1(\nu\sqrt{z^2})] \\ &+ i\frac{\nu}{4\pi^2\sqrt{-z^2}}\theta(-z^2)K_1(\nu\sqrt{-z^2}), \end{aligned} \tag{51}$$

which differs from the causal function $\Delta_4^f(z, \mu)$ by complex conjugation and the replacement $\mu \longrightarrow i\nu$ (or by the replacement $z^2 \longrightarrow -z^2$, $\mu \longrightarrow \nu$). Further details about this function are given in the Appendix.

For the above trajectories, we therefore have that

$$\begin{aligned} \pm\mathrm{tr}\alpha^{B, F} &= i\iint d\tau d\tau' \left\{ \begin{matrix} \dot{x}_\alpha(\tau)\dot{X}^\alpha(\tau') \\ 1 \end{matrix} \right\} \Delta_4^{LR}(z, \nu), \\ z^\alpha &= X^\alpha(\tau') - x^\alpha(\tau). \end{aligned} \tag{52}$$

The expression obtained allows us to interpret $\pm\mathrm{tr}\alpha^{B, F}$ as a functional describing the interaction of two vector or scalar sources by the exchange of vector or scalar quanta with spacelike momenta. One of the sources moves along the mirror trajectory while the other simultaneously moves along the tangent line to the trajectory at zero point. The second source can be considered as a probe or detector of the excitation created by the accelerated mirror in the vacuum.

7. TRACES OF THE BOGOLIUBOV COEFFICIENTS FOR HYPERBOLIC AND EXPONENTIAL TRAJECTORIES

We consider $\mathrm{tr}\alpha^{B, F}$ for the hyperbolic mirror trajectory

$$u^{\mathrm{mir}} = g(v) = \frac{\kappa' v}{\kappa(1 - \kappa' v)}. \tag{53}$$

Using Eqs. (14) and (4) in [3], it is not difficult to represent $\alpha_{\omega'\omega}^{B, F}$ via the Macdonald functions $K_{1, 0}$,

$$\alpha_{\omega'\omega}^{B, F} = \frac{2}{\sqrt{\kappa\kappa'}}\exp\left[i\left(\frac{\omega}{\kappa} + \frac{\omega'}{\kappa'}\right)\right]K_{1, 0}\left(2i\sqrt{\frac{\omega\omega'}{\kappa\kappa'}}\right). \tag{54}$$

In accordance with (43), we then have

$$\begin{aligned} \mathrm{tr}\alpha^{B, F} &= \frac{1}{\pi}\int_0^\infty d\left(\frac{\omega}{\kappa}\right)\exp\left(2i\frac{\omega}{\kappa}\right)K_{1, 0}\left(2i\frac{\omega}{\kappa}\right) \\ &= \frac{1}{2\pi}\int_0^\infty dz \exp(iz)K_{1, 0}(iz). \end{aligned} \tag{55}$$

The variable z in this integral has a simple physical meaning: it is equal to the ratio of the invariant momentum transfer to the invariant proper acceleration at the zero point (but for hyperbolic motion, the acceleration is the same on the entire trajectory),

$$z = \frac{\rho}{w_0}, \quad \rho = 2\sqrt{\omega\omega'}, \quad w_0 = \sqrt{\kappa\kappa'}. \tag{56}$$

The ultraviolet divergence of integral (55) is removed by subtracting the leading term of the $z \longrightarrow \infty$ expansion from the integrand. The infrared divergence (in the Bose case) is removed by introducing a nonzero lower limit $\varepsilon = \nu/w_0 \ll 1$ defined by the minimal momentum transfer ν. As a result, we obtain the integral

$$\begin{aligned} \mathrm{tr}\alpha^{B, F} &= \frac{1}{2\pi}\int_{s\varepsilon}^\infty dz\left[e^{iz}K_s(iz) - \sqrt{\frac{\pi}{2iz}}\right], \\ s &= 1, 0, \quad \varepsilon \ll 1. \end{aligned} \tag{57}$$

The integration contour can now be rotated to the negative imaginary semiaxis such that in the Bose case, it bypasses the singularity at zero along the arc of a circle with a small radius ε. Further calculation leads to the simple expressions

$$\begin{aligned} \mathrm{tr}\alpha^B &= \frac{1}{2\pi}\left[-\frac{\pi}{2} - i\left(\ln\frac{2w_0}{\gamma\nu} - 1\right)\right], \\ \nu &\ll w_0, \quad \gamma = 1, 781\ldots, \end{aligned} \tag{58}$$

$$\mathrm{tr}\alpha^F = \frac{1}{2\pi}i. \tag{59}$$

For the exponential motion of the mirror with

$$\begin{aligned} u^{\mathrm{mir}} &= -\frac{1}{\kappa}\ln(1 - \kappa' v), \\ v^{\mathrm{mir}} &= \frac{1}{\kappa'} - \frac{1}{\kappa'}\exp(-\kappa u), \end{aligned} \tag{60}$$

the same Eqs. (14) and (4) in [3] lead to the Bogoliubov coefficients

$$\alpha_{\omega'\omega}^B = \frac{1}{\kappa}\sqrt{\frac{\omega}{\omega'}}\Gamma\left(\frac{i\omega}{\kappa}\right)\exp\left(i\frac{\omega'}{\kappa'} - \frac{i\omega}{\kappa}\ln\frac{i\omega'}{\kappa'}\right), \tag{61}$$

$$\alpha_{\omega'\omega}^F = \frac{1}{\sqrt{i\kappa\omega'}}\Gamma\left(\frac{1}{2} + \frac{i\omega}{\kappa}\right)\exp\left(\frac{i\omega'}{\kappa'} - \frac{i\omega}{\kappa}\ln\frac{i\omega'}{\kappa'}\right). \tag{62}$$

The traces $\mathrm{tr}\alpha^{B,F}$ whose divergences were removed by the above prescription, are given by

$$\mathrm{tr}\alpha^B = \frac{1}{2\pi}\int_\varepsilon^\infty dx\left[\Gamma(ix)\exp(ix - ix\ln ix) - \sqrt{\frac{2\pi}{ix}}\right], \tag{63}$$

$$\mathrm{tr}\alpha^F = \frac{1}{2\pi} \times\int_0^\infty dx\left[\Gamma\left(\frac{1}{2}+ix\right)\frac{\exp(ix - ix\ln ix)}{\sqrt{ix}} - \sqrt{\frac{2\pi}{ix}}\right]. \tag{64}$$

In these integrals, the variable x is equal to one-fourth of z, which has the meaning of the momentum transfer in units of w_0 (as in (56)),

$$x = \frac{1}{4}z, \quad z = \frac{\rho}{w_0}, \quad \rho = 2\sqrt{\omega\omega'}, \quad w_0 = \frac{1}{2}\sqrt{\kappa\kappa'}. \tag{65}$$

Similarly, $\varepsilon = \nu/4w_0 \ll 1$. We note that in the course of exponential motion (60), the proper acceleration increases from zero to infinity; as a function of the proper time τ, it is given by

$$a(\tau) = -\frac{w_0}{1 - w_0\tau}. \tag{66}$$

It is now not difficult to see that the subtracted terms in integrals (63) and (64) exactly coincide with similar terms in integrals (57) if we express them through the physical variable z. In other words, up to the removal of the ultraviolet divergence from the integrals defining $\mathrm{tr}\alpha$, the asymptotic behavior of the integrands in the variable $z = \rho/w_0 \longrightarrow \infty$ is described by the universal formula

$$\frac{1}{2\pi}\sqrt{\frac{\pi}{2iz}}. \tag{67}$$

We show in the next section that this assertion is correct for any timelike trajectory in expansion (40) for which $b > 0$.

The integration contour in integrals (63) and (64) can be rotated to the negative imaginary axis bypassing the infrared singularity at zero (in the Bose case) along the arc with a radius ε. We then obtain

$$\mathrm{tr}\alpha^B = \frac{1}{2\pi}\left[-\frac{\pi}{2} - i\left(\ln\frac{4w_0}{\nu} - \int_0^\infty dt\ln tB'(t)\right)\right], \quad \nu \ll w_0, \tag{68}$$

$$\mathrm{tr}\alpha^F = -\frac{1}{2\pi}i\int_0^\infty\frac{dt}{\sqrt{t}}\left(\Gamma\left(\frac{1}{2}+t\right)\exp(t - t\ln t) - \sqrt{2\pi}\right) = \frac{1}{2\pi}i \times 0.8843\ldots \tag{69}$$

In the integral in (68), the function $B'(t)$ is the derivative of the function

$$B(t) = \Gamma(1+t)\exp(t - t\ln t) - \sqrt{2\pi t}.$$

The numerical value of this integral is 2.2194.... If we transform the imaginary part of (68) to the form of the imaginary part of (58), we obtain

$$\ln\frac{4w_0}{\nu} - 2.2194\ldots = \ln\frac{2w_0}{\gamma\nu} - 0.9491\ldots$$

Therefore, the values of $\mathrm{tr}\alpha^{B,F}$ for the exponential and hyperbolic motions are rather close to each other.

8. ULTRAVIOLET AND INFRARED SINGULARITIES OF $\mathrm{tr}\alpha^{B,F}$

It is not difficult to obtain the general expression for $\mathrm{tr}\alpha^{B,F}$ in the form of a double integral that is a functional of the mirror trajectory and is tangent to it at the point $u = v = 0$. Indeed, after substitution of the Bogoliubov coefficients

$$\alpha^B_{\omega'\omega} = \sqrt{\frac{\omega'}{\omega}}\int_{-\infty}^{\infty}dv\exp(i\omega'v - i\omega g(v)), \qquad \alpha^F_{\omega'\omega} = \int_{-\infty}^{\infty}dv\sqrt{g'(v)}\exp(i\omega'v - i\omega g(v)) \tag{70}$$

in (43) and a trivial integration over the frequency ω', we obtain

$$\mathrm{tr}\alpha^{B,F} = \frac{1}{2\pi}\int_0^\infty d\left(\frac{\omega}{\kappa}\right) \times\int_{-\infty}^{\infty}dx\{1, \sqrt{G'(x)}\}\exp\left[-i\frac{\omega}{\kappa}(G(x) - x)\right], \tag{71}$$

where 1 and $\sqrt{G'(x)}$ in the braces refer to the Bose and Fermi cases, respectively. The Lorentz invariance of these expressions is evident, but the integral over (ω/κ) diverges at the upper limit because the integrand behaves as $\sqrt{\kappa/\omega}$ at $\omega/\kappa \longrightarrow \infty$. Indeed, the condition $|x| \ll 1$ is essential in the integral over x as $\omega/\kappa \longrightarrow \infty$. The functions $G(x) - x$ and $G'(x)$ can then be replaced by the first terms of their expansions near zero, that is, by bx^2 and 1 (see (40)). Consequently, at $\omega/\kappa \longrightarrow \infty$, the integral over x is given by

$$\int_{-\infty}^{\infty}dx\exp\left(-i\frac{\omega}{\kappa}bx^2\right) = \sqrt{\frac{\pi\kappa}{ib\omega}} \tag{72}$$

in both the Bose and Fermi cases.

It is easy to show that the next term of the asymptotical expansion of the integral over x behaves as $(\kappa/\omega)^{3/2}$.

Then, after subtraction from the integral over x of the first term of its asymptotic expansion in the parameter $\omega/\kappa \longrightarrow \infty$, we make the integral over ω/κ convergent at the upper limit. If we pass from the variable ω/κ to the variable z,

$$\frac{\omega}{\kappa} = \sqrt{\frac{\omega\omega'}{\kappa\kappa'}} = \frac{b\rho}{2w_0} = \frac{1}{2}bz, \tag{73}$$

the subtracted term in $\mathrm{tr}\alpha^{B,F}$ acquires the universal form

$$\frac{1}{2\pi}\int_0^\infty dz\sqrt{\frac{\pi}{2iz}}. \tag{74}$$

We recall that $z = \rho/w_0$ has the meaning of the invariant momentum transfer in units of proper acceleration.

Although the expressions

$$\mathrm{tr}\alpha^{B,F} = \frac{1}{2\pi}\int_0^\infty ds\left[\int_{-\infty}^{\infty} dx\{1, \sqrt{G'(x)}\}\right.$$
$$\left.\times\exp(-is(G(x)-x)) - \sqrt{\frac{\pi}{ibs}}\right], \quad s = \frac{\omega}{\kappa}, \tag{75}$$

do not contain ultraviolet divergences, they can contain infrared divergences if the spectral function (the function of s in the square brackets in (75)) has the singular behavior $\propto 1/s$ as $s \longrightarrow 0$. It is clear that the behavior of the spectral function near $s = \omega/\kappa = 0$ and in the main forming region of the integral over s is determined by the behavior of the trajectory $G(x)$ far from the point of contact, where expansion (40) cannot be applied, i.e., at the distances $|x| \gtrsim 1$.

We now demonstrate the application of Eq. (75) in the example of another trajectory,

$$u^{\mathrm{mir}} = -\frac{1}{\kappa}\ln(2-e^{\kappa' v}), \quad G(x) = -\ln(2-e^x), \tag{76}$$

for which the spectral function can be expressed in terms of the well-known transcendental functions. This trajectory, as the hyperbolic one in (53), has two asymptotes but it approaches them following an exponential, not a powerlike law. Therefore, on both ends of the trajectory, the proper acceleration

$$a(v) = -\sqrt{\frac{\kappa\kappa'}{e^{\kappa' v}(2-e^{\kappa' v})}} \tag{77}$$

tends to $-\infty$ and attains the minimal value in its modulus $a_0 = -\sqrt{\kappa\kappa'}$ at the zero point.

The integral over x in (75), in which the upper limit for trajectory (76) is equal to ln2, is reduced to the tabular integral 2.2.5.1 in [10] after changing the variable x to $t = 1 - e^x$. As a result, we obtain

$$\mathrm{tr}\alpha^{B} = \frac{1}{2\pi}\int_\varepsilon^\infty ds\left[\frac{\sqrt{\pi}\Gamma(is)}{\Gamma\left(\frac{1}{2}+is\right)} - \sqrt{\frac{\pi}{is}}\right], \tag{78}$$

$$\mathrm{tr}\alpha^{F} = \frac{1}{2\pi}\int_0^\infty ds\left[\frac{\sqrt{\pi}\Gamma\left(\frac{1}{2}+is\right)}{\Gamma(1+is)} - \sqrt{\frac{\pi}{is}}\right]. \tag{79}$$

Because the spectral function has an infrared singularity in the Bose case, the corresponding divergence of the integral over s for $\mathrm{tr}\alpha^B$ is removed by introducing a small but finite lower limit $\varepsilon = \nu/w_0$. Its physical meaning is the minimum momentum transfer in units of the acceleration at the zero point.

After rotating the s-integration contour to the negative imaginary semiaxis with the singularity at zero being bypassed (in the Bose case) along the arc of a circle with a radius of ε, we obtain

$$\mathrm{tr}\alpha^{B} = \frac{1}{2\pi}\left[-\frac{\pi}{2} - i\left(\ln\frac{w_0}{\nu} - B\right)\right], \tag{80}$$

$$\mathrm{tr}\alpha^{F} = \frac{1}{2\pi}iF, \tag{81}$$

where positive constants B and F are defined by the integrals

$$B = \int_0^\infty dt\ln tB'(t) = 1.887789\ldots,$$
$$B(t) = \frac{\sqrt{\pi}\Gamma(1+t)}{\Gamma\left(\frac{1}{2}+t\right)} - \sqrt{\pi t}, \tag{82}$$

$$F = -\int_0^\infty dt\left[\frac{\sqrt{\pi}\Gamma\left(\frac{1}{2}+t\right)}{\Gamma(1+t)} - \sqrt{\frac{\pi}{t}}\right] = 1.869957\ldots \tag{83}$$

The imaginary part of (80) can be transformed to the form of the imaginary part of (57),

$$\ln\frac{w_0}{\nu} - 1.887789\ldots = \ln\frac{2w_0}{\gamma\nu} - 2.003721\ldots$$

The expressions for $\pm\mathrm{tr}\alpha^{B,F}$ obtained for the three different trajectories of the mirror are close to each other qualitatively and quantitatively (see (58), (59), (68), (69), and (80), (81)). All of them have a negative imaginary part with an infrared logarithmic singularity in the Bose case. This singularity is accompanied by the appearance of the real negative part of $\mathrm{tr}\,\alpha^B$, namely, $\mathrm{Re}\,\mathrm{tr}\,\alpha^B = -1/4$, whereas $\mathrm{Re}\,\mathrm{tr}\,\alpha^F = 0$. Similar expressions for $\pm\mathrm{tr}\,\alpha^{B,F}$ are typical of the trajectories with the $G(x)$ function increasing stronger (decreasing weaker) than x, as x tends to the upper (lower) limit.

Because the functionals $\pm\mathrm{tr}\,\alpha^{B,F}$ have the meaning of the action in accordance with (52), we compare them with the changes $\Delta W_{1,0}$ of self-actions of the electric and scalar charges in hyperbolic motion [11, 12],

$$\Delta W_{1,0} = -(\tau_2-\tau_1)\Delta m_{1,0}, \tag{84}$$

$$\Delta m_1 = \frac{e^2 w_0}{4\pi^2}\left[-\frac{\pi}{2} - i\left(\ln\frac{2w_0}{\gamma\mu} - \frac{1}{2}\right)\right],$$
$$\Delta m_0 = -i\frac{e^2 w_0}{8\pi^2}. \tag{85}$$

In this motion, the proper acceleration of the charge is constant and the square of the interval between two points on the trajectory is a function of only the length of the arc connecting them,

$$(x_\alpha(\tau) - x_\alpha(\tau'))^2 = f(\tau-\tau'). \tag{86}$$

Therefore, the change of the charge self-interaction is proportional to the time interval $\tau_2 - \tau_1$ that the charge is in hyperbolic motion multiplied by the mass shift $\Delta m_{1,0}$ of the charge. The mass shift occurs because of a change of the interaction of the charge with its own field, which is essentially modified at the distances of the order of w_0^{-1} from the charge due to acceleration. In other words, the shift is formed on the arc length

$$|\tau-\tau'| \sim w_0^{-1}$$

with the center τ_c at any point of the trajectory inside the acceleration interval (τ_1, τ_2). The independence of the shift from τ_c means that it is a constant of motion. This is not so for trajectories with a variable acceleration (see Section 9).

Unlike $\Delta W_{1,0}$, which describes the change of interaction of the charge with itself due to acceleration, the functionals $\pm\mathrm{tr}\,\alpha^{B,F}$ describe the interaction of the accelerated mirror with the probe executing the uniform motion along the tangent to the trajectory of the mirror at the point where the mirror has the acceleration w_0. This interaction is transmitted by vector or scalar perturbations created by the mirror in the Bose- or Fermi-field vacuum; these perturbations carry a spacelike momentum on the order of w_0. According to (51), at distances on the order of w_0^{-1} from the mirror, the field of these perturbations decreases exponentially in timelike directions and oscillates with a damped amplitude in spacelike directions. It can be said that such a field moves together with the mirror and is its "proper field." Hence, the probe interacts with the mirror for a time on the order of w_0^{-1}, while the charge interacts with itself the entire time and feels the change of the interaction over the entire time of acceleration. It is therefore not surprising that the expressions for $\pm\mathrm{tr}\,\alpha^{B,F}$ coincide in essence with $\Delta W_{1,0}$ if we set

$$\tau_2 - \tau_1 = 2\pi/w_0, \quad e^2 = 1$$

in the latter and change the sign. In other words, $\pm\mathrm{tr}\,\alpha^{B,F}$ are the mass shifts of the proper field of the mirror multiplied by the characteristic proper time of their formation.

9. MASS SHIFTS OF ELECTRIC AND SCALAR CHARGES IN EXPONENTIAL MOTION

To calculate the self-actions of electric and scalar charges in exponential motion, we use Eq. (30). It is convenient to use the charge trajectory (60) in the form of a function of the proper time,

$$u^{\mathrm{mir}}(\tau) = -\frac{2}{\kappa}\ln(1-w_0\tau),$$
$$v^{\mathrm{mir}}(\tau) = \frac{1}{\kappa'}(2w_0\tau - w_0^2\tau^2). \tag{87}$$

Then

$$\dot{x}_\alpha(\tau)\dot{x}^\alpha(\tau') = -\frac{1+z^2}{1-z^2},$$
$$(x-x')^2 = -(\tau-\tau')^2\frac{\operatorname{arctanh} z}{z}, \tag{88}$$
$$z = \frac{w_0(\tau-\tau')}{2-w_0(\tau+\tau')}.$$

We now introduce new variables $\xi = (\tau+\tau')/2$ and z instead of τ and τ'. At fixed ξ in the interval $-\infty < \xi < w_0^{-1}$, the variable z changes in the interval $-1 < z < 1$. Using the causal function Δ_4^f expressed via the Macdonald function, we obtain

$$\Delta W_1 = e^2\int_{-\infty}^{w_0^{-1}} d\xi\left(\frac{1}{w_0}-\xi\right)$$
$$\times\int_{-1}^{1} dz\,\dot{x}_\alpha(\tau)\dot{x}^\alpha(\tau')\Delta_4^f(x-x',\mu)\Big|_0^F$$
$$= -\frac{e^2}{2\pi^2}\int_{-\infty}^{w_0^{-1}} d\xi\int_0^\infty \frac{du\,\mu}{\sinh 2u} \tag{89}$$
$$\times\left\{\cosh 2u\frac{\tanh u}{u}K_1(i\lambda\sqrt{u\tanh u}) - K_1(i\lambda\tanh u)\right\}.$$

In the last expression, the variable $u = \operatorname{arctanh} z$ is used instead of z and λ is a function of ξ,

$$\lambda(\xi) = 2\mu(w_0^{-1} - \xi).$$

Our problem is now to find the integral over u in the region of the variable ξ where $\lambda(\xi) \ll 1$, supposing, of course, that the infrared parameter $\mu/w_0 \ll 1$. This integral coincides, in essence, with the mass shift of the electric charge,

$$\Delta m_1 = \frac{e^2}{2\pi^2}\int_0^\infty \frac{du\mu}{\sinh 2u} \times \left\{ \cosh 2u \sqrt{\frac{\tanh u}{u}} K_1(i\lambda\sqrt{u\tanh u}) - K_1(i\lambda \tanh u) \right\}. \tag{90}$$

To calculate Δm_1 with $\lambda(\xi) \ll 1$, we divide the integration interval into two intervals, $0 \le u \le u_1$ and $u_1 \le u < \infty$, by a point u_1 such that $u_1 \gg 1$, but $\lambda u_1 \ll 1$. Using the expansion of the Macdonald function at a small argument, we then obtain

$$\Delta m_1 \approx \frac{e^2 w_0}{4\pi^2(1 - w_0\xi)} \times \left\{ \frac{1}{i}\int_0^{u_1} du\left(\frac{\coth 2u}{u} - \frac{1}{2\sinh^2 u}\right) + \int_{u_1}^\infty \frac{du\lambda}{\sqrt{u}} K_1(i\lambda u) \right\} \tag{91}$$

$$= \frac{e^2 w_0}{4\pi^2(1 - w_0\xi)} \times \left\{ -\pi - i\left(2\ln\frac{w_0}{\gamma\mu(1 - w_0\xi)} + \ln\frac{2\gamma}{\pi} + \frac{1}{2}\right) \right\}.$$

The mass shift Δm_0 of the scalar charge differs from (90) by the replacement $\cosh 2u \longrightarrow -1$ in the first term in the braces and by the change of sign of the second term. Under the same condition $\lambda(\xi) \ll 1$, we then obtain

$$\Delta m_0 = -i\frac{e^2 m_0}{4\pi^2(1 - w_0\xi)}\left(\ln 2 - \frac{1}{2}\right). \tag{92}$$

It follows from (91), (92), and (66) that the mass shift depends on the absolute value

$$w(\xi) = \frac{w_0}{1 - w_0\xi}$$

of the proper acceleration of the mirror at the instant ξ, which may be considered as the center of the forming region of the shift. As the acceleration substantially changes at such an interval, the mass shifts in (91) and (92) do not coincide with the mass shifts of uniformly accelerated charges in (84) and (85) if we replace $w(\xi)$ with w_0. Nevertheless, rather close coincidence arises under the replacements $w(\xi) \longrightarrow 0.5w_0$ and $w(\xi) \longrightarrow 2.6w_0$ for Δm_1 and Δm_0, respectively.

10. CONCLUSIONS

The basis for the symmetry between processes induced by a mirror in two-dimensional space–time and by a charge in four-dimensional space–time is relation (14), (15) between the Bogoliubov coefficients $\beta^{B,F}_{\omega'\omega}$ and the current density $j^\alpha(k)$ or charge density $\rho(k)$ depending on the timelike momentum k^α. The squares of these quantities represent the spectra of real pairs and particles radiated by the accelerated mirror and the charge.

In the present paper, the symmetry is extended to the self-actions of the mirror and the charge and to the corresponding vacuum–vacuum amplitudes (cf. (29) and (30)). In essence, it is embodied in the discovered relation (20) between propagators of a massive pair in two-dimensional space and of a single particle in four-dimensional space.

Equation (29) for $W^{B,F}$ was obtained under the condition that the mean number $\mathrm{tr}\beta^+\beta$ of pairs created is small and the interference of two or more pairs is negligible. In the general case, $W^{B,F}$ is given by Eq. (27), which can also be written as

$$2\operatorname{Im} W^{B,F} = \pm\mathrm{tr}\ln(\alpha^+\alpha)^{B,F}, \tag{93}$$

because

$$\alpha^+\alpha \mp \beta^+\beta = 1$$

(see [4, 7]). As follows from (27) or (93), the imaginary part of the action differs from zero and is then positive only if $\beta \neq 0$, i.e., if the radiation of real particles indeed occurs.

For $W^{B,F}$, formula (93) allows us to choose the expression

$$W^{B,F} = \pm i\,\mathrm{tr}\ln\alpha^{B,F}, \tag{94}$$

which was called natural by DeWitt [7]. However, this expression is by no means unique. The expressions

$$W^{B,F} = \pm i\,\mathrm{tr}\ln(\alpha e^{i\gamma})^{B,F}, \quad W^{B,F} = \pm i\,\mathrm{tr}\ln\alpha^{B,F+}$$

have the same imaginary part. Nevertheless, Eq. (94) is interesting as the definition of both the real and imaginary parts of the self-actions $W^{B,F}$ via the Bogoliubov coefficients $\alpha^{B,F}_{\omega'\omega}$ only, which reduce to the current density $j^\alpha(q)$ or to the charge density $\rho(q)$ that depends on the spacelike momentum q^α in accordance with Eqs. (14) and (15). This implies that the field of the corresponding perturbations propagates in the vacuum together with the mirror, comoves it, and at the same time contains information about the radiation of real quanta.

Unfortunately, the author failed to find a simple integral representation for the matrix $\ln\alpha$. Nevertheless, if we again assume that the mean number of emitted particles is small, we can consider α or $i\alpha$ close to 1. Expanding $\ln i\alpha$ near $i\alpha = 1$ and confining ourselves to the first term, we then obtain

$$\begin{aligned} W^{B,F} &= \pm i\,\mathrm{tr}\ln i\alpha^{B,F} \approx \pm i\,\mathrm{tr}(i\alpha^{B,F} - 1) \\ &= \mp\,\mathrm{tr}\alpha^{B,F} + \dots \end{aligned} \tag{95}$$

These qualitative arguments allow us to state that the functionals $\pm\mathrm{tr}\alpha^{B,F}$ are similar to the corresponding self-actions with the opposite sign and must therefore have negative imaginary parts. This is confirmed by all examples considered in Sections 7 and 8. Nevertheless, the exact physical meaning of the quantities $\pm\mathrm{tr}\alpha^{B,F}$ is clearly defined by Eq. (52).

Here, we also want to focus attention on one prediction following from the symmetry between processes induced by the mirror in two-dimensional, and by the charge in four-dimensional, space–time. The symmetry predicts the value $e_0^2 = 1$ for the charge squared (in Heaviside units), which corresponds to the fine structure constant $\alpha_0 = 1/4\pi$. Because the radiation corrections are not taken into account in both spaces and the processes in (1 + 1)-dimensional space are due to the purely geometrical boundary condition, it is natural to think that the above-mentioned values of the charge squared and of the fine structure constant are the unrenormalized bare values of these constants. They are therefore marked with the index 0.

It is quite interesting that the bare fine structure constant has a purely geometrical origin and that its value is small,

$$\alpha_0 = 1/4\pi \ll 1.$$

The smallness of α_0 has an essential meaning for quantum electrodynamics, where it *a priori* justifies the applicability of perturbation theory and where the radiative corrections in accordance with the well-known formula [13]

$$\alpha = \frac{\alpha_0}{1 + (\alpha_0/3\pi)N\ln(\Lambda^2/m^2)} \tag{96}$$

diminish the renormalized value of α in comparison with the unrenormalized one. Here, N is the number of charged particles with masses in the interval (m, Λ) and Λ is the upper limit of the particle energy up to which quantum electrodynamics is correct.

APPENDIX

It is convenient to define the singular function $\Delta_d^{LR}(z, \nu)$ and the causal function $\Delta_d^f(z, \mu)$ in a d-dimensional space–time by the Fourier representation

$$\begin{aligned} \Delta_d^{LR}(z, \nu) &= \int \frac{d^dq}{(2\pi)^d}\frac{e^{iqz}}{q^2 - \nu^2 + i\varepsilon}, \\ \Delta_d^f(z, \mu) &= \int \frac{d^dq}{(2\pi)^d}\frac{e^{iqz}}{q^2 + \mu^2 - i\varepsilon}. \end{aligned} \tag{97}$$

These functions are the even singular solutions of the inhomogeneous wave equations

$$\begin{aligned} (-\partial^2 - \nu^2)\Delta^{LR}(z, \nu) &= \delta(z), \\ (-\partial^2 + \mu^2)\Delta^f(z, \mu) &= \delta(z), \end{aligned} \tag{98}$$

with opposite signs in front of the parameters ν^2 and μ^2, where ν and μ are the momentum transfer and the mass. Their proper time representations (in particular, for $d = 4$)

$$\begin{aligned} \Delta_4^{LR}(z, \nu) &= \frac{1}{(4\pi)^2}\int_0^\infty \frac{ds}{s^2}\exp\left(-i\nu^2 s - \frac{iz^2}{4s}\right), \\ \Delta_4^f(z, \mu) &= \frac{1}{(4\pi)^2}\int_0^\infty \frac{ds}{s^2}\exp\left(-i\mu^2 s + \frac{iz^2}{4s}\right) \end{aligned} \tag{99}$$

as well as the explicit expressions in terms of the Macdonald function differ by complex conjugation and by the replacement $\mu \longrightarrow i\nu$ or by the replacement $z^2 \longrightarrow -z^2$, $\mu \longrightarrow \nu$.

The integral relation

$$\Delta_{d+2}^{LR}(z, \nu) = -\frac{1}{4\pi}\int_{\nu^2}^\infty d\rho^2 \Delta_d^{LR}(z, \rho) \tag{100}$$

is very important for the symmetry discussed in this paper. It differs from the similar relation (20) for the causal functions not only by the sign. Being written for $z^2 < 0$, it is understood for $z^2 > 0$ in the sense of analytic continuation to the lower half-plane of complex z^2. On the other hand, relation (20), being written for $z^2 > 0$, is understood for $z^2 < 0$ as the analytic continuation to the upper half-plane of complex z^2. For the Δ^+-functions, such a continuation must be carried out in the upper half-plane if $z^0 > 0$ and in the lower one if $z^0 < 0$.

ACKNOWLEDGMENTS

The author is grateful to A.I. Nikishov and A.I. Ritus for valuable discussions and help.

The work was supported by the Russian Foundation for Basic Research (grant nos. 00-15-96566 and 02-02-16944).

REFERENCES

1. S. W. Hawking, Commun. Math. Phys. **43**, 199 (1979).
2. A. I. Nikishov and V. I. Ritus, Zh. Éksp. Teor. Fiz. **108**, 1121 (1995) [JETP **81**, 615 (1995)].
3. V. I. Ritus, Zh. Éksp. Teor. Fiz. **110**, 526 (1996) [JETP **83**, 282 (1996)].
4. V. I. Ritus, Zh. Éksp. Teor. Fiz. **114**, 46 (1998) [JETP **87**, 25 (1998)].
5. J. Schwinger, *Particles, Sources and Fields* (Addison-Wesley, Reading, Mass., 1970; Mir, Moscow, 1976), Vol. 1.
6. A. P. Lightman, W. H. Press, R. H. Price, and S. A. Teukolsky, *Problem Book in Relativity and Gravitation* (Princeton Univ. Press, Princeton, N.J., 1975).
7. B. S. DeWitt, Phys. Rep. C **19**, 295 (1975).
8. R. M. Wald, Commun. Math. Phys. **45**, 9 (1975).
9. V. I. Ritus, Zh. Éksp. Teor. Fiz. **116**, 1523 (1999) [JETP **89**, 821 (1999)]; E-print archives, hep-th/9912004.
10. A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series, Elementary Functions* (Nauka, Moscow, 1981; Gordon and Breach, New York, 1986).
11. V. I. Ritus, Zh. Éksp. Teor. Fiz. **75**, 1560 (1978) [Sov. Phys. JETP **48**, 788 (1978)].
12. V. I. Ritus, Zh. Éksp. Teor. Fiz. **80**, 1288 (1981) [Sov. Phys. JETP **53**, 659 (1981)].
13. V. B. Berestetskiĭ, E. M. Lifshitz, and L. P. Pitaevskiĭ, *Quantum Electrodynamics*, 3rd ed. (Nauka, Moscow, 1989; Pergamon Press, Oxford, 1982).

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 101–115.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 114–129.
Original Russian Text Copyright © 2003 by Vaks, Pankratov.

On the Accuracy of Various Statistical Methods for Describing Ordering Phase Transitions in BCC Alloys

V. G. Vaks* and I. R. Pankratov
Russian Research Centre Kurchatov Institute, Moscow, 123182 Russia
**e-mail: vaks@mbslab.kiae.ru*

Received February 18, 2003

Abstract—The accuracy of various statistical methods for describing $B2$- or $D0_3$-type (CuZn or Fe_3Al) ordering phase transitions characteristic of BCC alloys is examined. The temperature–concentration phase diagrams were calculated for several models of alloys discussed in the literature with the use of three methods, namely, the mean-field and pair cluster approximations and the tetrahedron cluster field method developed in this work. The calculation results were compared with each other and with Monte Carlo calculations. The accuracy of various methods was found to sharply depend on the type of interactions in the system, primarily, on the presence of strong competing interactions and short-range correlations that hinder ordering. In the absence of such correlations, in particular, in Fe–Al type alloys with extended interactions, the use of the pair cluster approximation allows the results obtained in the mean-field approximation to be substantially refined without any noticeable complication of calculations. At the same time, for systems with short-range and competing interactions, the use of this approximation could drastically distort the form of phase diagrams, whereas the tetrahedron cluster field method allows the phase diagrams of these systems to be calculated fairly accurately for temperatures and concentrations of interest to physics. *© 2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

Although the Monte Carlo method and other direct simulation methods for studying phase transitions in metallic alloys have been thoroughly elaborated [1, 2], the development of simple and fairly accurate analytic methods in this area remains a topical issue, in particular, in relation to two problems that attract much interest:

(1) The solution of “inverse” problems, that is, the determination of effective interatomic interaction parameters in alloys from experimental data on phase diagrams or diffuse scattering. In such problems, analytic methods are still more convenient and used more frequently than direct simulations, see, e.g., [3–11].

(2) Studies of the kinetics of phase transitions and the evolution of microstructures in alloys, in which the use of direct simulations often encounters difficulties, for example, related to the necessity of taking into account long-range elastic interactions, lattice defects, etc. [12–16].

The simplest analytic approximation for describing both equilibrium and nonequilibrium alloys is the mean-field approximation, see, e.g., [3–5]. However, the accuracy of this approximation is usually fairly low, and it gives qualitatively incorrect results in describing ordering in FCC lattices, in particular, the most typical $L1_2$ and $L1_0$ transitions. For this reason, quantitative studies are usually performed using various versions of cluster methods, in particular, the well-known cluster variation method [17, 18], whose accuracy is usually fairly high, see, e.g., [1, 10]. However, the use of standard cluster variation method variants to describe nonuniform and nonstationary states, leads to very cumbersome equations, which are difficult to use in actual calculations [19]. This is why a simplified version of the cluster variation method, namely, the tetrahedron cluster-field method, was suggested to describe nonuniform FCC alloys [19]. It was shown that, when applied to usual, realistic alloy models, the tetrahedron cluster-field method combines simplicity and fairly high accuracy of calculations. This method was used to study a number of problems of microstructure evolution in FCC alloys that experience ordering [14–16].

In this work, the results obtained in [19] to BCC alloys and the $B2$ and $D0_3$ ordering types characteristic of these alloys are generalize. For these systems, the accuracy of various statistical methods has been studied in much less detail than for FCC alloys [1]. In particular, in the majority of works on Fe–Al and Fe–Si alloys, which are widely discussed in the literature, only the simplest mean-field approximation was used [3–5, 12, 13]. At the same time, Monte Carlo studies [1, 2] showed that the mean-field approximation might substantially distort the phase diagrams of systems of this type, especially in the presence of strong competing interactions resulting in the “frustrated,” hindered character of the ordering transitions under consideration. A fairly high accuracy was only attained with the use of the tetrahedron [1, 18] or more complex variants of the cluster variation method. However, as has been mentioned above, the application of the cluster variation method to nonuniform and nonequilibrium systems is hardly possible. A realistic consideration of, for example, the structure of antiphase and interphase boundaries or the kinetics of ordering therefore requires developing fairly simple and accurate methods applicable to BCC alloys, similar to the tetrahedron cluster field method for FCC alloys.

The development and estimation of the accuracy of such methods for BCC alloys is the main goal of this work. In particular, we study the accuracy of the simplest cluster method (the pair cluster approximation) and show that its use considerably refines the results of mean-field calculations for many alloys without significant complications. At the same time, a comparison of the results obtained by various methods with each other and with Monte Carlo calculations gives a complete idea of the accuracy of all methods under consideration.

The main equations of the cluster methods [19, 14] and their generalization to nonuniform and nonequilibrium alloys are given in Section 2. In Section 3, the tetrahedron cluster-field approximation suggested in [19] is generalized to the BCC lattice. The mean-field, pair cluster, and tetrahedron cluster-field approximation equations for the thermodynamic potentials of the $B2$ and $D0_3$ phases are given in Section 4; these equations are then used to calculate phase diagrams. In Section 5, we give and discuss the results of such calculations for models of Fe–Al-type alloys and compare them with the available Monte Carlo calculation results [2]. A similar analysis for alloy models with short-range and competing interactions (see above) is given in Section 6. The main results are summarized in Conclusions.

2. GENERAL FORMULAS OF CLUSTER METHODS FOR EQUILIBRIUM AND QUASI-EQUILIBRIUM DISTRIBUTIONS

The general ideas and equations of cluster methods in the formulation that we use are given in [14, 19]. Below, we cite the necessary results of these works and generalize them to nonuniform and nonequilibrium alloys. For definiteness, consider an A_cB_{1-c} binary substitution alloy with a $c \le 0.5$ mean concentration. Various distributions of atoms over lattice sites i are described by their mean occupancies $c_i = \langle n_i \rangle$, where n_i is 1 if the site is occupied by atom A and zero if it is occupied by atom B, and the averaging is performed over the $P\{n_i\}$ distribution function, which can correspond to either an equilibrium or a nonequilibrium alloy state. The general problem is to calculate the generalized free energy $F\{c_i\}$ defined in [20, 21] for an arbitrary, not necessarily equilibrium, distribution of occupancies $\{c_i\}$.

As has been mentioned in [14], the most general expression for the distribution function can be written in the form

$$P\{n_i\} = \exp[\beta(\Omega - Q)], \tag{1}$$

where $\beta = 1/T$ is the inverse temperature and the "quasi-Hamiltonian" Q is given by

$$Q = -\sum_i \lambda_i n_i + \sum_{i>j} a_{ij} n_i n_j + \sum_{i>j>k} a_{ijk} n_i n_j n_k + \dots \tag{2}$$

Here, the coefficient λ_i at the operator n_i will be called "the site chemical potential i" and the $a_{i\dots j}$ coefficients of the products of operators $n_i \dots n_j$ will be called "quasi-interactions." The sum of the terms with $a_{i\dots j}$ in (2) is an analog of the interaction Hamiltonian $H_{\rm int}$ for the alloy,

$$H_{\rm int} = \sum_{i>j} v_{ij} n_i n_j + \sum_{i>j>k} v_{ijk} n_i n_j n_k + \dots, \tag{3}$$

where $v_{i\dots j}$ are the effective interaction constants. The generalized thermodynamic potential Ω in (1) is determined from the normalization condition

$$\Omega = -T \ln \mathrm{Tr} \exp(-\beta Q), \tag{4}$$

where operator Tr denotes the summation over all sets of $\{n_i\}$ values. For equilibrium distributions, (1) is the Gibbs distribution in which quasi-interactions $a_{i\dots j}$ in (2) are equal to interactions $v_{i\dots j}$ in (3) and the site chemical potentials λ_i are constant over the system; that is, $\lambda_i = \mathrm{const} = \mu$. If we describe nonequilibrium states arising during phase transitions, we can usually assume that quasi-equilibrium relations are satisfied [20, 21]; that is,

$$a_{i\dots j} = v_{i\dots j}; \text{ although, generally, } \lambda_i \ne \mathrm{const}. \tag{5}$$

There are at least two arguments in favor of the validity of these relations under usual conditions of phase transformations. First, the equality $a_{i\dots j} = v_{i\dots j}$ is strictly valid both before and after the transition, for example, both before quenching an initially equilibrium disordered alloy from a high temperature T_h to a lower temperature T_l corresponding to other equilibrium phases and after equilibrium at the new T_l temperature is attained. For this reason, there is no thermodynamic driving force that would tend to change the $a_{i\dots j}$ parameter in (1) from its equilibrium value $v_{i\dots j}$. At the same time, the high-temperature chemical potential value $\mu(T_h)$ differs from the low-temperature value $\mu(T_l)$, and the μ value must change. Because of the diffusion character of phase transitions, these changes in μ occur via the formation of local inhomogeneities, for example, antiphase or interphase boundaries, which results in local violations of the conditions $\lambda_i = \mathrm{const}$ and leads to the last inequality in (5). In addition, quasi-interactions $a_{i\dots j}$ in distribution (1) mainly describe the short-range order. After a change of the external conditions, such as temperature, this short-range order is established relatively fast, for a time on the order of the time of one interatomic exchange τ. At the same time, the complete evolution of a microstructure as a rule takes much longer times $t \gg \tau$ [12–16]. For this reason, possible fluctuative deviations from (5) at small $t \lesssim \tau$ are of no significance to this evolution.

For both equilibrium and quasi-equilibrium distributions, the free energy $F\{c_i\}$ is expressed through Ω and λ_i in a simple way [14, Eq. (8)],

$$F = \Omega + \sum_i \lambda_i c_i, \tag{6}$$

and the site chemical potential λ_i is related to F by the equation [14, Eq. (9)] $\lambda_i = \partial F/\partial c_i$.

Cluster methods deal with clusters, that is, some aggregates of lattice sites, which we for brevity denote by Greek letters, for example, $\{i, j, \ldots, k\} = \alpha$. The P_α distribution of the probabilities of various configurations in cluster α is described by its effective quasi-Hamiltonian Q_α and the corresponding potential Ω_α. In the exact formulation, they are obtained by averaging complete distribution (1) over variables $\{n_i\}$ of all sites outside the cluster,

$$P_\alpha = \exp[\beta(\Omega_a - Q_\alpha)] = \underset{i \notin \alpha}{\mathrm{Tr}} \exp[\beta(\Omega - Q)], \tag{7}$$

$$\Omega_\alpha = -T \ln Z_\alpha = -T \ln \underset{i \in \alpha}{\mathrm{Tr}} \exp(-\beta Q_\alpha). \tag{8}$$

Let us denote the operator that describes the occupancy of some subcluster $\gamma = \{i, j, \ldots, l\}$ of cluster α by $n_\gamma = n_i n_j \ldots n_l$. The Q_α quasi-Hamiltonian of the cluster can then be written in a compact form as the sum over all possible occupancies n_γ in this cluster,

$$\begin{aligned} Q_\alpha &= \sum_i (-\lambda_i + \psi_i^\alpha) n_i + \sum_{i<j} (v_{ij} + \psi_{ij}^\alpha) n_i n_j + \ldots \\ &\equiv \sum_{\gamma \subseteq \alpha} (v_\gamma + \psi_\gamma^\alpha) n_\gamma. \end{aligned} \tag{9}$$

Here, the v_γ value for "one-site" occupancy $\gamma = i$ is $(-\lambda_i)$, and the ψ_γ^α values describe the renormalization of the v_γ values, that is, the $(-\lambda_i)$ values and intracluster interactions $v_{i\ldots j}$, caused by cluster interaction with its environment. It follows from (8) and (9) that the mean occupancy $\langle n_\gamma \rangle$ for an arbitrary cluster α containing subcluster γ is related to field ψ_γ^α in this cluster as

$$\langle n_\gamma \rangle = (\partial \Omega_\alpha / \partial \psi_\gamma^\alpha)_{T, v_\gamma}. \tag{10}$$

To calculate thermodynamic potentials, for example, Ω in (4) or F in (6), we must, first, express the total quasi-Hamiltonian Q via cluster quasi-Hamiltonians Q_α in the form of an expansion that sufficiently rapidly converges as the size of clusters increases and, second, choose the type of clusters and the approximation for finding fields ψ_γ^α.

The first problem is solved by the cumulant expansion method described in [19], which leads to the main equation for all cluster methods [19, Eq. (12)], namely,

$$Q = \sum_\alpha \nu_\alpha Q_\alpha. \tag{11}$$

The choice of one or another approximation corresponds to retention in the right-hand side of Eq. (11) of only those clusters whose size $|\alpha|$ (the number of sites) does not exceed some maximum $|\alpha|_{\max}$ value. The summation in (11) is therefore only performed over retained clusters of a maximum size and all their subclusters.

The geometric coefficients ν_α and renormalizing fields ψ_γ^α in (9)–(11) for an arbitrary subcluster γ of basis clusters should satisfy the constraint equations [19, Eqs. (13), (14)]

$$\sum_{\alpha \supseteq \gamma} \nu_\alpha = 1, \quad \sum_{\alpha \supseteq \gamma} \nu_\alpha \psi_\gamma^\alpha = 0, \tag{12}$$

where the summations are over all clusters α containing subcluster γ. This, in particular, ensures equality of the coefficients of all n_γ operators, that is, of n_i, $n_i n_j$, etc., in the left- and right-hand sides of (11).

It follows from definition (4) that the mean quasi-Hamiltonian Q value over distribution (1) satisfies the usual thermodynamic equation

$$\langle Q \rangle = -\frac{\partial}{\partial \beta} \ln \mathrm{Tr} \exp(-\beta Q) = \frac{\partial}{\partial \beta}(\beta \Omega). \tag{13}$$

The integration of this equation in β with $\langle Q \rangle$ found in approximation (11) taking into account (7) yields

$$\Omega = \sum_\alpha \nu_\alpha \Omega_\alpha, \tag{14}$$

where Ω_α is given by (8). This result takes into account Eq. (17) from [19], according to which the total derivative of the right-hand side of (14) with respect to each of the independent ψ_γ^α variables is zero, that is, $d\Omega/d\psi_\gamma^\alpha = 0$.

The geometric ν_α coefficients in (11) and (12) have a number of general properties discussed, for example, in [10, 19]. Among these properties, note the "theorem about internal clusters," that is, such subclusters that enter into only one basis cluster α_m. For internal subclusters γ, the ν_γ coefficients in (11)–(14) and the corresponding $\psi_\gamma^{\alpha_m}$ fields in the basis clusters are zero. The internal clusters therefore make no contribution to the observed values and can be excluded from consideration. This theorem is used below in the method of cluster fields.

The cluster variation method contains no approximations other than those made to derive (11)–(14). Statistical properties are therefore calculated by solving system (10), (12). The dependence of fields ψ_γ^α and Ω_α values in (8) on mean occupancies $\langle n_\gamma \rangle$ is determined analytically. This dependence is given by Eq. (20) in [19]. The $\langle n_\gamma \rangle$ values themselves are then given by the equation [19, Eq. (21)]

$$\beta v_\gamma = \sum_{\alpha \supseteq \gamma} \nu_\alpha l_\gamma^\alpha \{\langle n \rangle\}, \tag{15}$$

where $l_\gamma^\alpha \{\langle n \rangle\}$ is some function of all $\langle n_\zeta \rangle$ given by Eq. (20) from [19].

Substituting $\langle n_\gamma \rangle$ found by (15) into Eq. (20) from [19] for ψ_γ^α and into Eq. (8) for Ω_α solves the problem of finding potential $\Omega = \Omega(\lambda_i)$ in (14). However, the mean site occupancies c_i and, accordingly, free energy $F\{c_i\}$ can more conveniently be used as independent variables in physical problems than the chemical potentials of sites λ_i and potential $\Omega(\lambda_i)$; $F\{c_i\}$ is related to Ω by (6). Equation (15) for one-particle occupancy $\gamma = i$ then determines the $\lambda_i\{c_j\}$ function, and (8) and (14) determine the $\Omega\{c_i\}$ function.

The cluster field method is a simplified version of the cluster variation method, in which one-site environment fields ψ_i^α are only included in cluster quasi-Hamiltonians (9), whereas interaction ψ_γ^α renormalizations with $|\gamma| \geq 2$ are neglected. Generally, Eq. (10) for many-particle averages $\langle n_\gamma \rangle$ with $|\gamma| \geq 2$ in clusters of different sizes $|\alpha|$ is then violated (in the complete cluster variation method, this equality is satisfied by introducing the corresponding renormalizing fields ψ_γ^α). For this reason, the degree of self-consistency and, accordingly, the accuracy of the cluster field method should generally be lower than those of similar cluster variation method variants. The renormalizations of intracluster interactions or their thermodynamic contributions are, however, absent for some choices of basis clusters due to the theorem about internal subclusters mentioned above; that is, fields ψ_γ^α with $|\gamma| \geq 2$ or coefficients ν_α in (11) and (14) vanish for clusters containing these fields. The cluster field method then becomes equivalent to the cluster variation method, and its self-consistency and accuracy need not be inferior to those of the latter. The tetrahedron cluster-field approximation for the FCC lattice, developed in [14, 19], corresponds exactly to this case, and its high accuracy of describing realistic alloy models can be considered natural.

System (15) in the cluster field method is drastically simplified. As variables, site "activities" y_i^α entering in Eq. (8) for Ω_α rather than ψ_i^α can conveniently be used,

$$y_i^\alpha = \exp[\beta(\lambda_i - \psi_i^\alpha)]. \tag{16}$$

The partition function Z_α of a cluster in (8) has the form of a polynomial of y_l^α of degree $|\alpha|$,

$$Z_\alpha = \sum_{\gamma \subseteq \alpha} \exp(-\beta v_\gamma) \prod_{l \in \gamma} y_l^\alpha, \tag{17}$$

where the last factor is the product of all activities y_l^α corresponding to the given occupancy $\gamma = \{i, j, \ldots, k\}$, that is, $y_i^\alpha y_j^\alpha \ldots y_k^\alpha$. Equation (10) takes the form

$$c_i = y_i^\alpha \partial \ln Z_\alpha / \partial y_i^\alpha. \tag{18}$$

According to (17), the right-hand side of equality (18) is the ratio of two polynomials in y_j^α. For each cluster α, system (18) for the $y_i^\alpha\{c_j\}$ functions is usually easily solved numerically using the conjugate-gradient method. After solving this system, (17), (8), and (14) are used to find the $\Omega\{c_i\}$ potential, and Eqs. (16) and (12) for one-site occupancy $\gamma = i$ determine the site chemical potential $\lambda_i\{c_j\}$,

$$\lambda_i = T \sum_{\alpha \supseteq i} \nu_\alpha \ln y_i^\alpha. \tag{19}$$

3. THE TETRAHEDRON CLUSTER-FIELD METHOD FOR THE BCC LATTICE

The tetrahedron cluster-field method for the BCC lattice can more conveniently be discussed if we first give the equations for the thermodynamic potential $\Omega\{c_i\}$ and the site chemical potential $\lambda_i\{c_j\}$ in the more simple mean-field and pair cluster approximations. For simplicity, consider the pair interaction case when Hamiltonian (3) contains only the first sum. Interaction v_{ij} between atoms in sites i and j will be written as $\delta_{ij,n} v_n$, where $\delta_{ij,n}$ is nonzero and equals 1 only if sites i and j are nth neighbors in the lattice, and v_n is the interaction value. The mean-field equations for $\Omega\{c_i\}$ and $\lambda_i\{c_j\}$ then take the form (see, e.g., [21])

$$\begin{aligned} \Omega &= T\sum_i \ln c_i' - \frac{1}{2}\sum_{i,j,n}\delta_{ij,n}v_n c_i c_j, \\ \lambda_i &= T\ln\frac{c_i}{c_i'} + \sum_{j,n}\delta_{ij,n}v_n c_j, \end{aligned} \tag{20}$$

where $c_i' = 1 - c_i$. In the pair cluster approximation, all pairs of sites i and j for which v_{ij} interactions are nonzero are considered. The $\Omega\{c_i\}$ and $\lambda_i\{c_j\}$ values are then given by [19, Eqs. (32)–(37)]

$$\begin{aligned} \Omega &= T\sum_i \ln c_i' + \frac{1}{2}\sum_{i,j,n}\delta_{ij,n}\Omega_n^{ij}, \\ \lambda_i &= T\ln\frac{c_i}{c_i'} + \sum_{j,n}\delta_{ij,n}\lambda_{ni}^{ij}. \end{aligned} \tag{21}$$

Here,

$$\Omega_n^{ij} = -T\ln(1 - c_i c_j g_n^{ij}), \quad \lambda_{ni}^{ij} = T\ln(1 - g_n^{ij}c_j),$$

and the g_n^{ij} function describes interaction $v_{ij} = v_n$ in terms of the Mayer function $f_n = \exp(-\beta v_n) - 1$,

$$\begin{aligned} g_n^{ij} &= 2f_n/[R_n^{ij} + 1 + f_n(c_i + c_j)], \\ R_n^{ij} &= \{[1 + (c_i + c_j)f_n]^2 - 4c_i c_j f_n(f_n + 1)\}^{1/2}. \end{aligned} \tag{22}$$

If interaction is weak, $\beta v_n \ll 1$, the g_n^{ij} function transforms into $(-\beta v_n)$, and Eq. (21) of the pair cluster approximation transforms into Eq. (20) of the mean-field approximation.

As follows from the discussion given below, simple mean-field (Eq. (20)) and, especially, pair cluster approximation (Eq. (21)) are usually sufficient for describing systems with extended interactions such as the Fe–Al-type alloys considered in Section 5. Strong correlations between nearest and next-nearest neighbors, however, arise in the systems with short-range and competing interactions discussed in Section 6. To correctly describe such systems, these correlations should be taken into account, at least, within tetrahedra of the type of the set of sites (*a*, *b*, *c*, *d*) shown in Fig. 1. For example, the study of the model with $v_2/v_1 = 0.5$ and $v_{n>2} = 0$ (suggested in [4] to describe Fe–Si alloys) with the use of the mean-field approximation, the tetrahedron cluster variation method, and the Monte Carlo method [1] showed that the accuracy of the mean-field approximation for models of this type is fairly low (see Fig. 9), while the tetrahedron cluster variation method gives rather accurate results, just as for analogous models in the FCC lattice [10]. At the same time, a similar high accuracy for the FCC lattice is also attained with the use of the much simpler tetrahedron cluster-field method described in [19]. Therefore, it seems desirable to develop a similar simplified version of the tetrahedron cluster variation method for the BCC lattice.

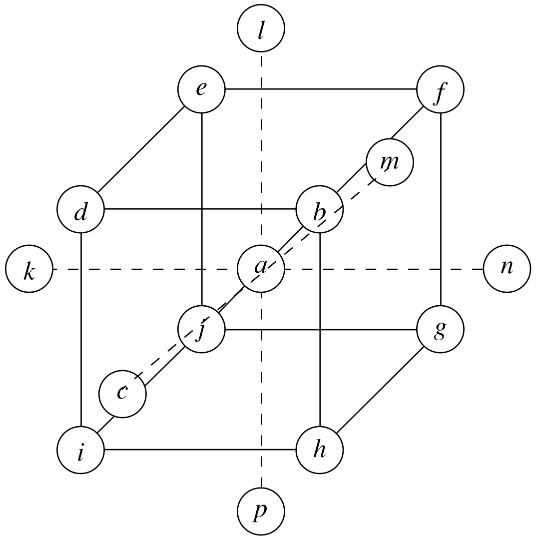


Fig. 1. BCC lattice sites discussed in the text in relation to the choice of tetrahedral clusters and "residual" next-nearest neighbor interactions v_2 in the tetrahedron cluster-field method.

When the usual tetrahedron cluster variation method is used, we must consider all possible tetrahedra of neighboring sites of the (*a*, *b*, *c*, *d*) type (Fig. 1), including, for example, the (*a*, *b*, *d*, *l*), (*a*, *b*, *f*, *l*) clusters, etc., as basis clusters. Each three-site subcluster (*a*, *b*, *d*) then enters into two basis tetrahedra, and pair subclusters of the (*a*, *b*) or (*a*, *c*) type, into six or four basis tetrahedra, respectively. Therefore, all these subclusters are not internal, and so we must take into account all renormalization fields ψ_γ^α for all subclusters γ in Eqs. (9) and (10) of the general cluster variation method. This makes the solution of these equations for nonuniform distributions $\{c_i\}$ a very complex task [19].

In the cluster field method, all subclusters except one-site ones should be internal in order that interaction renormalizations in them might be neglected. For this reason, basis clusters should have common vertices but not common edges; for the FCC lattice, this is illustrated by Fig. 1 in [19]. Therefore, that only four basis tetrahedra should pass through each BCC lattice site in the tetrahedron cluster-field method. These tetrahedra include all eight nearest-neighbor bonds which correspond to interaction v_1 (for shortness, v_1 bonds) and four of six bonds corresponding to next-nearest neighbor interactions v_2 (v_2 bonds). Interactions v_2 that do not enter into basis tetrahedra can be taken into account using either mean-field or pair cluster approximation, as in applying the tetrahedron cluster-field method to the FCC lattice [19].

Note, however, that when the v_2/v_1 ratio between the competing interactions under consideration is not small, and the more distant interactions $v_{n>2}$ favoring ordering are small or absent, such a simplified treatment of the "residual" (not included into basis clusters) contribution of interactions v_2 can cause difficulties at low temperatures, especially if the pair cluster approximation is used. This manifests itself by the formal arising of "anti-Curie" points in calculations and the disappearance of the B_2 and $D0_3$ orderings at low temperatures, which is similar to the appearance of anti-Curie points in describing Heisenberg antiferromagnets in the pair cluster approximation [22]. These difficulties arise because the ground state of the pair cluster under consideration then differs from the ground state of the whole system. For this reason, in constructing the tetrahedron cluster-field method for the BCC lattice, we will describe this residual contribution of v_2 interactions in the simple mean-field rather than pair cluster approximation. The difficulties with the anti-Curie point for realistic interaction models will then arise only at very low temperatures, when the phase transformations are already "frozen" by kinetic factors. As concerns physically interesting temperatures, it is shown below that the accuracy of the tetrahedron cluster-field approximation at these temperatures usually remains fairly high.

The contributions of interactions of more distant neighbors $v_{n>2}$ can be taken into account in both pair cluster and mean-field approximations. For the usual interaction models the results virtually coincide, and for definiteness, we will describe these contributions in the pair cluster approximation. Also note that the choice of basis tetrahedra mentioned above and "residual" interactions v_2 described in the mean-field approximation is generally not unique. For example, if we choose the (*a*, *b*, *c*, *d*), (*a*, *e*, *f*, *m*), (*a*, *i*, *j*, *k*), and (*a*, *g*, *h*, *n*) or (*a*, *d*, *e*, *k*), (*a*, *b*, *f*, *n*), (*a*, *h*, *i*, *d*), and (*a*, *g*, *j*, *m*) tetrahedra as four basis tetrahedra that pass through site *a* in Fig. 1, the (*a*, *l*) and (*a*, *p*) v_2 bonds and also all other v_2 bonds parallel to the *z* axis will be described in the mean-field approximation. Four other similar choices of basis tetrahedra are possible, in which all v_2 bonds parallel to the *x* or *y* axis are described in the mean-field approximation, and there is a total of six variants for choosing tetrahedron basis clusters. If the distribution of site occupancies $\{c_i\}$ has cubic symmetry (as, for example, in a homogeneous phase with *B*2 or $D0_3$ ordering or in a disordered phase), the results obtained in thermodynamic potential $\Omega\{c_i\}$ and site chemical potential $\lambda_i\{c_i\}$ calculations by (14) and (19) do not depend on the variant that is used. However, such calculations for an arbitrary nonuniform distribution $\{c_i\}$ may depend on the selected set of basis tetrahedra. Generally, calculation results should therefore be averaged over all six possible variants of choosing basis clusters.

All things considered, we can rewrite (14) and (19) for the Ω and λ_i values in the tetrahedron cluster field method for the BCC lattice in a form similar to Eqs. (13)–(21) from [14] obtained for the FCC lattice,

$$\begin{aligned}\Omega = {} & T\sum_i \ln c'_i - T\frac{1}{6}\sum_{s=1}^{6}\sum_{\{i,j,k,l\}\in t_s}\ln(Z_{ijkl}c'_ic'_jc'_kc'_l) \\ & -\frac{1}{6}\sum_{i,j}\delta_{ij,2}v_2c_ic_j + \frac{1}{2}\sum_{i,j,n>2}\delta_{ij,n}\Omega_n^{ij},\end{aligned} \tag{23}$$

$$\begin{aligned}\lambda_i = {} & T\ln\frac{c_i}{c'_i} + T\frac{1}{6}\sum_{s=1}^{6}\sum_{\{j,k,l\}\in t_s^i}\ln\left(\frac{y_i^{ijkl}c'_i}{c_i}\right) \\ & +\frac{1}{3}\sum_j\delta_{ij,2}v_2c_j + \sum_{j,n>2}\delta_{ij,n}\lambda_{ni}^{ij}.\end{aligned} \tag{24}$$

Here, Ω_n^{ij} and λ_{ni}^{ij} in the last sums are the same as in (21), and the denotation t_s or t_s^i in the second term in (23) or (24) refers to the basis tetrahedron corresponding to one of six variants *s* of choosing basis clusters described above. At a given choice *s*, the summation in (23) is over all tetrahedra t_s in the lattice corresponding to this choice, and the summation in (24) is over the four t_s^i tetrahedra that contain site *i*. The results obtained by (23) and (24) are then averaged over six possible cluster choice variants *s*. The Z_α value in (23) with $\alpha = \{i, j, k, l\}$ is the partition function of the tetrahedral cluster [Eq. (17)], and y_i^α in (24) is the activity [Eq. (16)] of site *i* in this cluster. If, when writing $\alpha = \{i, j, k, l\}$, we suppose that the neighboring pairs of indices (*i*, *j*), (*j*, *k*), (*k*, *l*), and (*l*, *i*) correspond to the nearest lattice sites, and "nonneighboring" pairs (*i*, *k*) and (*j*, *l*), to next-nearest sites, the partition function Z_α takes the form similar to Eq. (21) from [21] for the FCC lattice,

$$\begin{aligned}Z_\alpha = {} & 1 + \sum_p y_p^\alpha + \zeta_1(y_i^\alpha y_j^\alpha + y_j^\alpha y_k^\alpha + y_k^\alpha y_l^\alpha + y_l^\alpha y_i^\alpha) \\ & + \zeta_2(y_i^\alpha y_k^\alpha + y_j^\alpha y_l^\alpha) + \zeta_1^2\zeta_2(y_i^\alpha y_j^\alpha y_k^\alpha + y_j^\alpha y_k^\alpha y_l^\alpha \\ & + y_k^\alpha y_l^\alpha y_i^\alpha + y_l^\alpha y_i^\alpha y_j^\alpha) + \zeta_1^4\zeta_2^2 y_i^\alpha y_j^\alpha y_k^\alpha y_l^\alpha.\end{aligned} \tag{25}$$

Here, $\zeta_1 = \exp(-\beta v_1)$; $\zeta_2 = \exp(-\beta v_2)$; index *p* is *i*, *j*, *k*, or *l*; and the dependence of y_p^α on site occupancies $c_{p'}$ in the cluster is found from system (18). As with FCC lattices (see [19] and [14]), this system of four algebraic equations for each tetrahedron α is easily solved numerically by the conjugate-gradient method. If interaction is weak, $\beta v_n \ll 1$, Eqs. (23) and (24) of the tetrahedron cluster-field method and also Eq. (21) of the pair cluster approximation transform into Eq. (20) of the mean-field approximation.

4. THERMODYNAMIC POTENTIALS IN THE *B*2 AND $D0_3$ PHASES

In this section, we apply general equations (20)–(24) to calculate the thermodynamic potentials of the *B*2 and $D0_3$ phases; the results are used below to construct phase diagrams. The structure of these phases and the mean-field approximation equations for their thermodynamic potentials are described, for example, in [1–5, 13]. The necessary relations are as follows. For the *B*2 homogeneous phase, the mean occupancy c_i of site *i* with lattice vector $\mathbf{R}_i$ can be written through the mean concentration *c* and order parameter η as

$$c_i = c + \eta\cos(\mathbf{g}_1 \cdot \mathbf{R}_i), \tag{26}$$

where $\mathbf{g}_1 = [111]2\pi/a$ is the *B*2 superstructure vector and *a* is the BCC lattice constant. Equation (26) describes the presence of two sublattices, *a* and *b*, in the *B*2 phase. These sublattices are shifted with respect to each other by vector $[111]a/2$ (see Fig. 1) and have the following site occupancies:

$$c_a = c + \eta, \quad c_b = c - \eta. \tag{27}$$

The $D0_3$ structure corresponds to a further splitting of the "enriched" sublattice of the *B*2 phase (if $\eta > 0$, this is sublattice *a*) into two different sublattices and to the presence of two order parameters, η and ζ. The occupancies c_i can then be written as [13]

$$\begin{aligned} c_i &= c + \eta\cos(\mathbf{g}_1 \cdot \mathbf{R}_i) \\ &+ 2\zeta[\theta(\eta)\cos(\mathbf{g}_2 \cdot \mathbf{R}_i) + \theta(-\eta)\sin(\mathbf{g}_2 \cdot \mathbf{R}_i)], \end{aligned} \tag{28}$$

where $\mathbf{g}_2 = [111]\pi/a$ is the $D0_3$ superstructure vector and $\theta(x)$ is the step function, equal to unity at $x > 0$ and zero at $x < 0$. Equation (28) describes the presence of three sublattices, *a*, *b*, and *c*, with the occupancies

$$c_a = c + \eta + 2\zeta, \quad c_b = c - \eta, \quad c_c = c + \eta - 2\zeta. \tag{29}$$

For example, if site *a* in Fig. 1 corresponds to the first sublattice in (29), then the *b*, *d*, *e*, … sites correspond to the second sublattice, and the *c*, *k*, *l*, … sites, to the third one. If $D0_3$ ordering is complete, we have $\eta = \zeta = c$, $c_a = 4c$, and $c_b = c_c = 0$ in (29); that is, the first sublattice is only occupied. In the mean-field approximation, the expression for the free energy per atom $f = F/N$ (where *N* is the total number of atoms, that is, lattice sites) in the *B*2 phase according to Eqs. (6), (20), and (26) can be written as

$$\begin{aligned} f_{B2}^{MFA} &= T\frac{1}{2}(c_a\ln c_a + c'_a\ln c'_a + c_b\ln c_b + c'_b\ln c'_b) \\ &+ \frac{1}{2}V_0c^2 + \frac{1}{2}V_{\mathbf{g}_1}\eta^2, \end{aligned} \tag{30}$$

where $c'_p = 1 - c_p$ and V_0 or $V_{\mathbf{g}_1}$ is the Fourier component of the interaction $v_{ij} = v(\mathbf{R}_i - \mathbf{R}_j)$,

$$V_{\mathbf{k}} = \sum_{\mathbf{r}} v(\mathbf{r})\cos(\mathbf{k} \cdot \mathbf{r}), \tag{31}$$

for the wave vector $\mathbf{k} = 0$ or $\mathbf{k} = \mathbf{g}_1$. The equilibrium order parameter $\eta = \eta(c, T)$ for the given concentration *c* and temperature *T* can then be found either from the condition of the equality of the chemical potentials of the sublattices $\lambda_a = \lambda_b$ calculated by (19) or from the equivalent condition of minimum (with respect to η) free energy (30) at given *c* and *T*. Similarly, for the $D0_3$ phase, the free energy in the mean-field approximation is given by the equation

$$\begin{aligned} f_{D0_3}^{MFA} &= T\frac{1}{4}[c_a\ln c_a + c'_a\ln c'_a + 2(c_b\ln c_b + c'_b\ln c'_b) \\ &+ c_c\ln c_c + c'_c\ln c'_c] + \frac{1}{2}V_0c^2 + \frac{1}{2}V_{\mathbf{g}_1}\eta^2 + V_{\mathbf{g}_2}\zeta^2. \end{aligned} \tag{32}$$

Here, occupancies c_a, c_b, and c_c are related to the η and ζ order parameters by (29) and $V_{\mathbf{g}_2}$ is the Fourier component of interaction (31) for $\mathbf{k} = \mathbf{g}_2$. The equilibrium $\eta(c, T)$ and $\zeta(c, T)$ values are found either from the equality of the chemical potentials of the sublattices $\lambda_a = \lambda_b = \lambda_c$ or from the equivalent conditions of minimum free energy (32) with respect to η and ζ.

In the pair cluster approximation (PCA), the general equations for free energy *f* and chemical potential λ_p in sublattice *p* for an arbitrary homogeneous phase follow from (6) and (21),

$$\begin{aligned} f^{PCA} &= \omega + \sum_p \xi_p\lambda_p, \\ \omega &= \sum_p \nu_p\left(\ln c'_p + \frac{1}{2}\sum_{q,n} z_{np}^q\Omega_n^{pq}\right), \\ \lambda_p^{PCA} &= T\ln\frac{c_p}{c'_p} + \sum_{q,n} z_{np}^q\lambda_{np}^{pq}. \end{aligned} \tag{33}$$

Here, $\omega = \Omega/N$ is the thermodynamic potential per atom, ν_p is the fraction of sites of type *p* in the given phase, z_{np}^q is the number of sites of type *q* in the *n*th coordination sphere of site p, and Ω_n^{pq} and λ_{np}^{pq} are given as Ω_n^{ij} and λ_{ni}^{ij} in (21) at $c_i = c_p$ and $c_j = c_q$. For the $B2$ phase, the p index in (33) in the notation of (27) is a or b; $\nu_a = \nu_b = 1/2$; and the nonzero z_{np}^q values, for example, for the first four neighbors n, are given by

$$\begin{gathered} z_{1a}^b = z_{1b}^a = 8, \quad z_{2a}^a = z_{2b}^b = 6, \\ z_{3a}^a = z_{3b}^b = 12, \quad z_{4a}^b = z_{4b}^a = 24. \end{gathered} \tag{34}$$

Similarly, for the $D0_3$ phase, the ν_p and z_{np}^q parameters in (33) in the notation of (29) are

$$\begin{gathered} \nu_a = \nu_c = 1/4, \quad \nu_b = 1/2, \\ z_{1a}^b = z_{1c}^b = 8, \quad z_{1b}^a = z_{1b}^c = 4, \\ z_{2a}^c = z_{2c}^a = z_{2b}^b = 6, \quad z_{3a}^a = z_{3c}^c = z_{3b}^b = 12, \\ z_{4a}^b = z_{ac}^b = 24, \quad z_{4b}^a = z_{4b}^c = 12. \end{gathered} \tag{35}$$

In the tetrahedron cluster-field approximation (TCA), the contributions of interactions v_1 and v_2 between the nearest and next-nearest neighbors in (21) are described by the second and third terms of (23) and (24). For the $B2$ and $D0_3$ homogeneous phases, this leads to the replacement of the terms with $n = 1$ and $n = 2$ in Eqs. (33) for ω and λ_p by the terms

$$\begin{aligned} \omega^{TCA}(v_1, v_2) &= -T\ln(Z_\alpha c'_i c'_j c'_k c'_l) \\ &\quad - \frac{1}{6}\sum_{p,q} \nu_p z_{2p}^q v_2 c_p c_q, \end{aligned} \tag{36}$$

$$\lambda_p^{TCA}(v_1, v_2) = 4T\ln\frac{y_p^\alpha c'_p}{c_p} + \frac{1}{3}\sum_q z_{2p}^q v_2 c_q, \tag{37}$$

where the notation is the same as in (23) and (24). If, in addition, site i in tetrahedral cluster α in (36) and (37) corresponds to sublattice a in (27) and (29), the c_i, c_j, c_k, and c_l occupancies in the $B2$ phase are related to c_a and c_b as $c_i = c_k = c_a$ and $c_j = c_l = c_b$. For the $D0_3$ phase, similar relations have the form $c_i = c_a$, $c_j = c_l = c_b$, and $c_k = c_c$.

Next, consider spinodals for the $B2$ and $D0_3$ type ordering (for brevity, the $B2$ and $D0_3$ spinodals), that is, the $T = \theta(c)$ curves in the c, T plane that are boundaries of alloy stability with respect to these orderings. The general condition of the stability of the distribution of occupancies $\{c_{i0}\}$ with respect to the arising of a concentration wave with the superstructure vector $\mathbf{k}_s$, that is, with respect to small fluctuations δc_i of the form

$$\delta c_i = c_i - c_{i0} = \eta_s \cos(\mathbf{k}_s \cdot \mathbf{R}_i), \tag{38}$$

is a positive value of the coefficient of η_s^2 in the expansion of free energy $F\{c_i\}$ in powers of δc_i at small η_s values. The condition of vanishing of this coefficient gives the equation for the ordering spinodal

$$\sum_{i,j} S_{ij} \cos(\mathbf{k}_s \cdot \mathbf{R}_i) \cos(\mathbf{k}_s \cdot \mathbf{R}_j) = 0, \tag{39}$$

where the $S_{ij} = \partial^2 F/\partial c_i \partial c_j$ derivatives, called below the "thermodynamic stiffnesses" matrix or simply "stiffnesses," are calculated for the initial state $\{c_i\} = \{c_{i0}\}$. For the $B2$ spinodal, this state is the disordered BCC phase $A2$, in which S_{ij} value only depends on the relative distance $\mathbf{r}_{ij} = \mathbf{R}_i - \mathbf{R}_j$; $\mathbf{k}_s$ is the $\mathbf{g}_1$ vector from (26); the $\cos(\mathbf{k}_s \cdot \mathbf{R}_i)$ values are 1 or –1; and spinodal equation (39) reduces to the condition of vanishing of the Fourier component $S_\mathbf{k}$ of the $S_{ij} = S(\mathbf{r}_{ij})$ function for $\mathbf{k} = \mathbf{g}_1$,

$$\sum_{\mathbf{r}} S(\mathbf{r}) \cos(\mathbf{g}_1 \cdot \mathbf{r}) = 0. \tag{40}$$

For the $D0_3$ spinodal, the initial state is the $B2$ phase, $\mathbf{k}_s$ is the vector $\mathbf{g}_2$ in (28), and the factor $\cos(\mathbf{k}_s \cdot \mathbf{R}_i)$ in (38) and (39) vanishes in all sites of "depleted" sublattice b in (26) and (27). Equation (39) can therefore be written as

$$\sum_{\mathbf{r}_a} S(\mathbf{r}_a) \cos(\mathbf{g}_2 \cdot \mathbf{r}_a) = 0, \tag{41}$$

where the summation is taken only over vectors $\mathbf{r}_a = \mathbf{R}_j^a - \mathbf{R}_0^a$ of the enriched sublattice a.

Let us write the explicit equations for the spinodals in the approximations under consideration. For the mean-field approximation, it follows from Eqs. (6) and (20) that the stiffness S_{ij} has the form

$$S_{ij}^{MFA} = \delta_{ij}\left(\frac{T}{c_i c'_i}\right) + \delta_{ij,n} v_n. \tag{42}$$

Substituting this S_{ij} value into (40) and (41) yields the well-known mean-field approximation equations for ordering spinodals θ(c), see, e.g., [23],

$$\begin{aligned}\theta_{B2}^{MFA} &= (-V_{\mathbf{g}_1})cc' \\ &= (8v_1 - 6v_2 - 12v_3 + 24v_4)c(1-c), \\ \theta_{D0_3}^{MFA} &= (-V_{\mathbf{g}_2})c_a c_a' \\ &= (6v_2 - 12v_3)(c+\eta)(1-c-\eta),\end{aligned} \tag{43}$$

where $V_{\mathbf{g}_1}$, $V_{\mathbf{g}_2}$, and c_a are the same as in (30), (32), and (27). Here and below, explicit equations for the spinodals for illustration purposes are given for models with interactions v_n up to the fourth neighbors (for shortness: four-neighbor-interaction models). Equations (43), in particular, show that increasing the v_2/v_1 ratio suppresses $B2$ ordering while the negative v_3 values (characteristic of the Fe–Al alloy models discussed below) promote both the $B2$ and $D0_3$ ordering.

The general equations for stiffnesses S_{ij} in the cluster field method were considered in [6] (see Eqs. (10) and (11) in [6]). In the pair cluster approximation, "reduced" stiffnesses βS_{ij} has the form [6, Eq. (23)]

$$\beta S_{ij}^{PCA} = \delta_{ij}\left(\frac{1}{c_i c_i'} + \sum_{k,n} a_n^{ik}\right) - \frac{\delta_{ij,n} f_n}{R_n^{ij}}. \tag{44}$$

Here, f_n and R_n^{ij} are the same as in (22), and a_n^{ij} can be written in terms of these quantities as follows:

$$a_n^{ij} = 4c_j c_j' f_n^2 / R_n^{ij}[(R_n^{ij}+1)^2 - f_n^2(c_i - c_j)^2]. \tag{45}$$

Substituting (44) into general equations (40) and (41) yields the equations of the pair cluster approximation for the $B2$ and $D0_3$ spinodals. For example, for models with interactions up to the fourth neighbors, the $B2$ spinodal θ_{B2}^{PCA} is determined by the equation

$$\begin{aligned}&\frac{1}{cc'} + (8a_1 + 6a_2 + 12a_3 + 24a_4) \\ &+ 8\frac{f_1}{R_1} - 6\frac{f_2}{R_2} - 12\frac{f_3}{R_3} + 24\frac{f_4}{R_4} = 0,\end{aligned} \tag{46}$$

where $a_n = 4cc' f_n^2/R_n(R_n+1)^2$ and $R_n = (1 + 4cc'f_n)^{1/2}$. According to (41), the equation for the $D0_3$ spinodal $\theta_{D0_3}^{PCA}$ in the pair cluster approximation for the same model has the form

$$\begin{aligned}&\frac{1}{c_a c_a'} + (8a_1^{ab} + 6a_2^{aa} + 12a_3^{aa} + 24a_4^{ab}) \\ &+ 6\frac{f_2}{R_2^{aa}} - 12\frac{f_3}{R_3^{aa}} = 0,\end{aligned} \tag{47}$$

where the a_n^{ii} and R_n^{ij} values with i and j equal to a or b are given by (45), (22), and (27). If interactions are weak, $\beta v_n \ll 1$, Eqs. (46) and (47) of the pair cluster approximation transform into (43) of the mean-field approximation.

In the tetrahedron cluster-field approximation, the contributions to the stiffnesses made by the nearest and next-nearest neighbor interactions v_1 and v_2, that is, the terms with $n = 1$ and $n = 2$ in Eq. (44) of the pair cluster approximation, should be replaced by the contribution $S_{ij}(v_1, v_2) = \partial\lambda_i(v_1, v_2)/\partial c_j$ corresponding to the derivatives of the second and third terms in (24). As both the $B2$ and $D0_3$ phases that we are considering have cubic symmetry, the form of (40) and (41) is independent of the choice of variant s of division into clusters in (24), and both index s and averaging over s can be omitted. Using Eq. (11) from [6] to determine the $\partial\lambda_i(v_1, v_2)/\partial c_j$ derivatives, we can write this contribution in the form

$$\begin{aligned}\beta S_{ij}^{TCA}(v_1, v_2) &= \delta_{ij}\left(\sum_{k,l,m\in t^i} S_{ii}^{iklm} - \frac{4}{c_i c_i'}\right) \\ &+ \delta_{ij,1} S_{ij}^{ijkl} + \delta_{ij,2}\left(S_{ij}^{ikjl} + \frac{\beta v_2}{3}\right),\end{aligned} \tag{48}$$

where the matrix S_{ij}^{α} for cluster α is inverse to the matrix $A_{ij} = \partial c_i/\partial \ln y_j^{\alpha}$,

$$\begin{aligned}S_{ij}^{\alpha} &= (A^{-1})_{ij}, \\ A_{ij} &= \delta_{ij}c_i c_i' + (1-\delta_{ij})\left(\frac{y_i^{\alpha} y_j^{\alpha}}{Z_{\alpha}} - c_i c_j\right).\end{aligned} \tag{49}$$

Substituting (48) into (40) and (41) gives the tetrahedron cluster-field approximation equations for the $B2$ and $D0_3$ spinodals. For example, in the fourth-neighbor-interaction model, Eq. (46) for θ_{B2}^{TCA} transforms into

$$\begin{aligned}&\frac{1}{cc'} + 4\left(S_{11}^{\alpha} - \frac{1}{cc'} - 2S_{12}^{\alpha} + S_{13}^{\alpha}\right) + (12a_3 + 24a_4) \\ &+ 2v_2 - 12\frac{f_3}{R_3} + 24\frac{f_4}{R_4} = 0.\end{aligned} \tag{50}$$

Here, the S_{ij}^{α} matrix elements are calculated according to (49), where the i, j, k, and l sites of cluster α are numbered as 1, 2, 3, and 4 and their occupancies $c_1 = c_2 = c_3 = c_4 = c$ correspond to the disordered phase $A2$. Similarly, Eq. (47) of the tetrahedron cluster-field approxi-

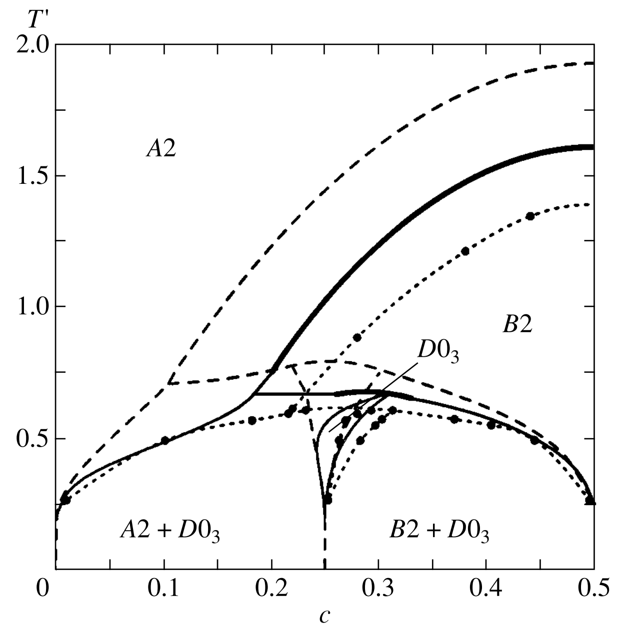


Fig. 2. Phase diagrams c, T' ($T' = T/v_1$ is the reduced temperature) of a binary A_cB_{1-c} BCC alloy calculated for model I in table. Here and below, the dashed line corresponds to the mean-field approximation; the dotted line, to the pair cluster approximation; and the solid line, to the tetrahedron cluster-field method. Solid circles correspond to Monte Carlo calculations [2]. The results obtained by the pair cluster and tetrahedron cluster-field approximations for models I, II, III, IV, and V (see table) and shown in Figs. 2, 3, 4, 5, and 6, respectively, coincide to within the accuracy of graphic representation.

mation for the $D0_3$ spinodal in the same model transforms into

$$\frac{1}{c_a c'_a} + 4\left(S^{\alpha}_{11} - \frac{1}{c_a c'_a} - S^{\alpha}_{13}\right) + (12a_3^{aa} + 24a_4^{ab}) - 2v_2 - 12\frac{f_3}{R_3^{aa}} = 0, \tag{51}$$

where the 1, 2, 3, and 4 sites of cluster α have occupancies $c_1 = c_3 = c_a$ and $c_2 = c_4 = c_b$ with the c_a and c_b values from (27) corresponding to the $B2$ phase.

Reduced interactions $v'_n = v_n/v_1$ for Fe–Al models under consideration

Model	I	II	III	IV	V
References	[2, 7]	[2, 8]	[2, 9]	[5]	[12]
v'_2	0.07	0.081	0.167	0.184	–0.8
v'_3	–0.23	–0.305	–0.208	–0.844	–0.5
v'_4	–0.11	–	–	–	–

5. CALCULATIONS OF PHASE DIAGRAMS FOR Fe–Al ALLOY MODELS

Below, we study the accuracy of the mean-field, pair cluster, and tetrahedron cluster-field approximations used to calculate phase diagrams with $B2$- and $D0_3$-type orderings in the BCC lattice. For this purpose, we will compare the results of such calculations with each other and with the available Monte Carlo studies [1, 2].

As has been mentioned, the accuracy of analytic methods sharply depends on the type of interactions in the system under consideration. In this section, we consider the Fe–Al type alloy models for which interactions that determine ordering are fairly long-range. The table contains the literature estimates of the $v'_n = v_n/v_1$ reduced interaction constants for these alloys. Models I, II, and III were suggested by Schmid and Binder [2] based on an analysis of the data on diffuse scattering in Fe–Al alloys obtained in [7], [8], and [9], respectively. Model IV was suggested by Hasaka [5] based on fitting the observed Fe–Al phase diagram to the mean-field approximation calculations. We will not take into account magnetic interactions discussed by Schmid and Binder [2] because their presence is of no importance for the problems under consideration. Comparisons with the calculations performed in [2] will only be made for the region of low magnetic atom concentrations (that is, for $x < 0.5$ in Fe_xAl_{1-x}), in which the magnetic interaction contributions considered in [2] appear to be negligibly small. In addition, we consider model V used in [12], which corresponds to the presence of a broad two-phase $A2$–$B2$ region in the phase diagram, much broader than in real Al–Fe alloys, because this model illustrates the special features of phase transitions in systems of this type [12].

The phase diagrams calculated for models I–V are shown in Figs. 2–6. First, consider the phase diagrams for “realistic” models I–III obtained from direct experimental estimates of interaction constants [7–9]. The results shown in Figs. 2, 3, and 4 for these three models are qualitatively similar. The conclusions that can be drawn from their comparison are as follows.

(1) Mean-field calculations, even for systems with fairly long-range interactions, can involve substantial errors. First of all, they systematically overestimate the $A2 \longrightarrow B2$ second-order phase transition temperatures; that is, they shift the $\theta_{B2}(c)$ $B2$ spinodal in the (c, T) phase diagram upward. This overestimation can be characterized by the relative error δ_c in the critical temperature T_c (that is, the θ_{B2} spinodal maximum) with respect to the T_c^{MC} value found by the Monte Carlo method: $\delta_c^{MFA} = (T_c^{MFA} - T_c^{MC})/T_c^{MC}$. The errors δ_c^{MFA} for models I, II, and III amount to 40, 16, and 20%, respectively.

(2) The use of the pair cluster approximation allows the description to be considerably refined compared with the mean-field calculations. The error in the criti-

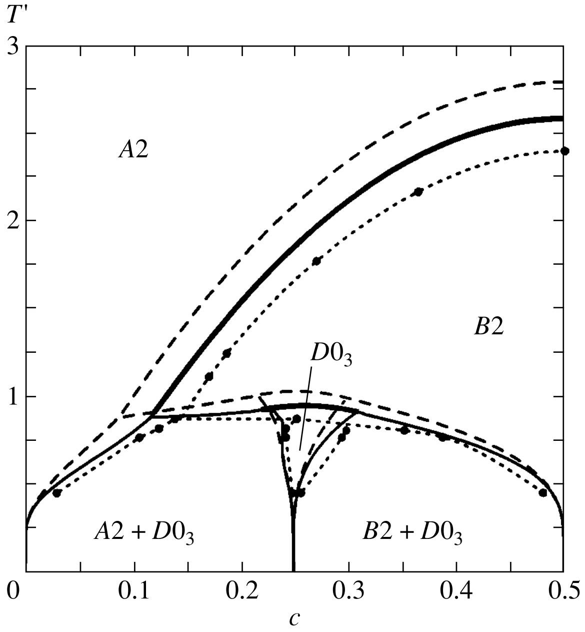


Fig. 3. The same as in Fig. 2 for model II in table.

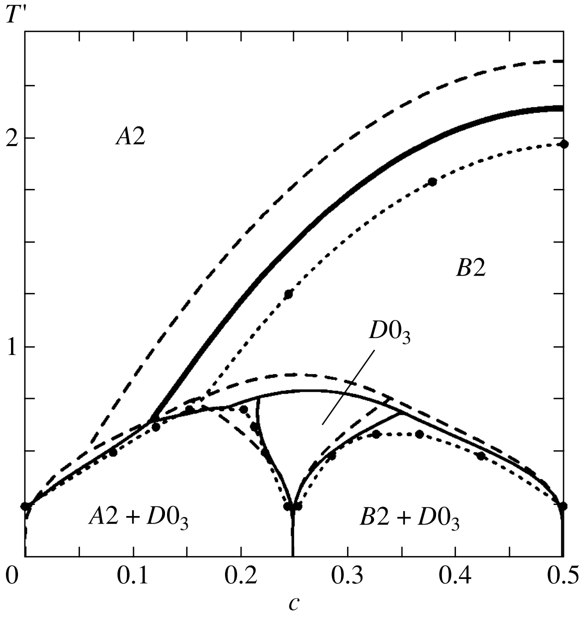


Fig. 4. The same as in Fig. 2 for model III in table.

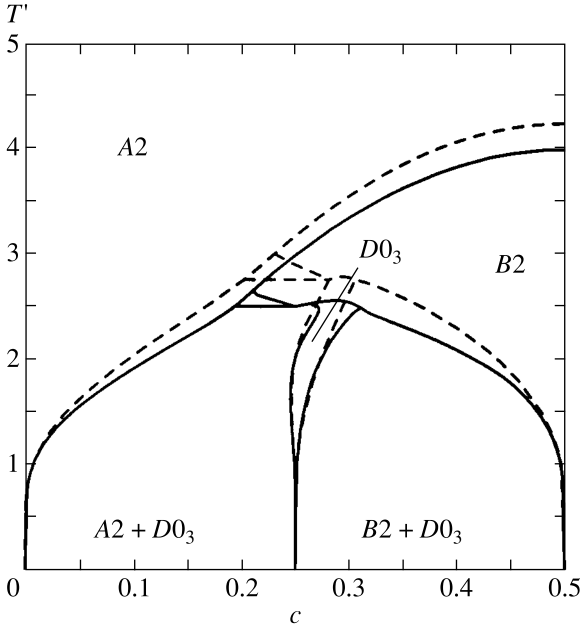


Fig. 5. The same as in Fig. 2 for model IV in table without Monte Carlo calculations.

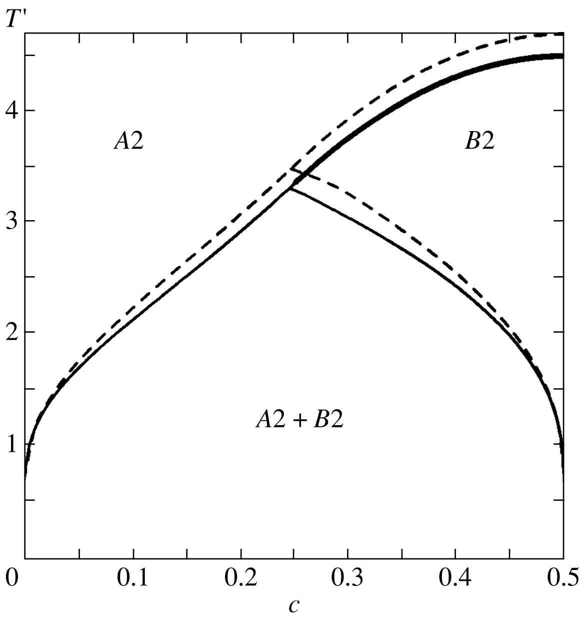


Fig. 6. The same as in Fig. 5 for model V in table.

cal temperature mentioned see above, that is, the $\delta_c^{PCA} = (T_c^{PCA} - T_c^{MC})/T_c^{MC}$ value, obtained in calculations by the pair cluster method is 17, 8, and 8% for models I, II, and III, respectively; that is, this error is 2–2.5 times lower than that of the mean-field calculations. Therefore, the use of the pair cluster approximation considerably increases the accuracy of calculations compared to the mean-field approximation without noticeably complicating them.

(3) The results obtained in the pair cluster and tetrahedron cluster-field approximations for models I–III virtually coincide because of the smallness of v_2 interactions, that is, the v_2/T_c ratios, for these models. The contribution of v_2 is therefore effectively described by the pair cluster approximation, and the equations of the tetrahedron cluster-field approximation effectively transform into the equations of the pair cluster approximation.

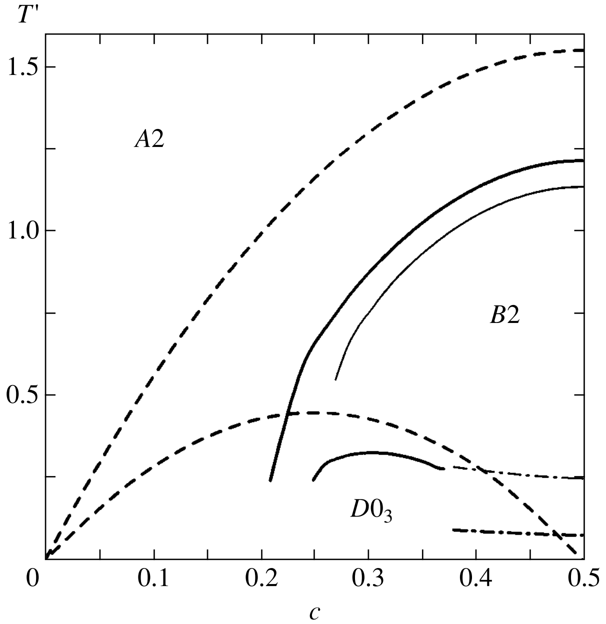


Fig. 7. The same as in Fig. 5 for the model with $v'_2 = v_2/v_1 = 0.3$ and $v_{n>2} = 0$. The dash-and-dot lines indicate the anti-Curie point positions.

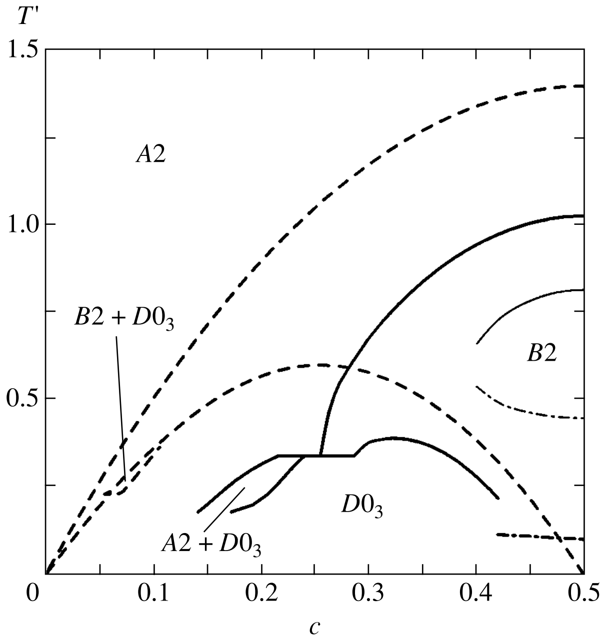


Fig. 8. The same as in Fig. 7 for the model with $v'_2 = 0.4$, $v_{n>2} = 0$.

(4) The accuracy of the approximate methods under consideration usually increases for models I–III as temperature decreases. The phase equilibria that include the $D0_3$ phase are therefore described more accurately than the $A2 \longrightarrow B2$ transitions, especially when the pair cluster or tetrahedron cluster-field approximation is used.

For model IV, the "attraction" of the third neighbors ($-v_3$), which favors ordering, is much stronger than for models I–III. For this reason, the accuracy of even the simple mean-field approximation (much more, of the pair cluster approximation) appears to be fairly high, and the difference of the δ_c^{MFA} and δ_c^{PCA} errors is two to three times smaller than for models I–III. Unlike from models I–IV, model V is characterized by non-nearest-neighbor interactions v_2 and v_3 that are not small and are negative; that is, these interactions favor $B2$ ordering. For this reason, the accuracy of both the mean-field and pair cluster approximations is fairly high and the difference of the δ_c^{MFA} and δ_c^{PCA} errors is as low as 4%; that is, it is three to five times smaller than for models I–III. The results obtained for models I–III (see above) allow us to expect that the exact transition temperatures for models IV and V should differ from the T_{PCA} values calculated in the pair cluster approximation by no more than how much T_{PCA} differs from T_{MFA}, that is, $(T_{PCA} - T_{ex}) \lesssim (T_{MFA} - T_{PCA})$.

6. PHASE DIAGRAMS FOR SYSTEMS WITH SHORT-RANGE AND COMPETING INTERACTIONS

In this section, we consider the BCC alloy models in which "reduced" repulsive interactions between next-nearest neighbors $v'_2 = v_2/v_1$ are comparatively strong and the other interactions v'_n that favor ordering are absent or weak. As has been mentioned, the competition between the v_1 and v_2 interactions in such systems hinders the transition and decreases reduced ordering temperatures $T' = T/v_1$. As a result, correlations between the occupancies of the nearest- and next-nearest-neighbor sites are of significance in the transition region.

Note that, as distinguished from the models of Fe–Al-type alloys considered above, there is no experimental evidence for the existence of BCC alloys of the type under consideration. The model with $v'_2 = 0.5$ and $v'_n = 0$ was suggested in its time from considerations of simplicity of calculations in the simplest mean-field approximation [4]; the phase diagram obtained was similar to that observed for Fe–Si alloys. Subsequent calculations by the cluster variation and Monte Carlo methods [18, 1], however, showed that the exact phase diagram for this model did not resemble the experimental one too closely. In addition, the available experimental data on interactions in metallic alloys show that these alloys as a rule involve fairly strong interactions between the third and more distant neighbors, which substantially influence the form of the phase diagrams [6–11, 14]. For this reason, the scale of correlations in the models described below may be substantially exaggerated compared with real alloys.

For brevity, models of interactions between two nearest neighbors with $v_1 > v_2 \geq 0$, $v_{n>2} = 0$ will be called "v'_2 models." Our calculations of the phase diagrams of v'_2 models in the mean-field, pair cluster, and tetrahedron cluster-field approximations (Figs. 7–9) allow the following conclusions to be drawn.

(1) At small $v'_2 \lesssim 0.15$ values, the phase diagrams of v'_2 models in the pair cluster and tetrahedron cluster-field approximations virtually coincide, as for the Fe–Al alloy models discussed above. The use of the pair cluster (or tetrahedron cluster) approximation considerably refines the results of simple mean-field calculations. For example, for $v'_2 = 0$ (that is, in the presence of only nearest neighbor interactions), the errors $\delta_c = (T_c - T_c^{MC})/T_c^{MC}$ for the mean-field and pair cluster approximations are 0.26 and 0.09, respectively; that is, the pair cluster approximation gives a twice smaller error than the mean-field approximation.

(2) At larger $v'_2 \gtrsim 0.25$, the accuracy of the pair cluster approximation begins to decrease sharply and becomes substantially lower than the accuracy of the tetrahedron cluster-field approximation. This is seen from Figs. 3 and 4. The pair cluster approximation underestimates the region of ordered phase stability on the side of high temperatures (the calculated θ_{B2}^{PCA} spinodals lie below the θ_{B2}^{TCA} spinodals) and on the side of low temperatures, where the above-mentioned anti-Curie points T_{ac} corresponding to the (fictitious) disappearance of the ordered phase become quite distinct.

Note that anti-Curie points are present in v'_2 models at all $v'_2 > 0$ in both pair cluster and tetrahedron cluster-field approximations. At small v'_2, the T_{ac} values are, however, very small compared with the T_c critical temperature. For example, at $v'_2 = 0.1$, the reduced critical temperature $T'_c = T_c/v \approx 1.6$, and the reduced $T'_{ac} = T_{ac}/v_1$ values at concentrations c of 0.2 to 0.5 vary from $T'_{ac} \approx 0.2$ to $T'_{ac} = 0.1$ according to the pair cluster calculations and from $T'_{ac} \approx 0.1$ to $T'_{ac} = 0.03$ according to the tetrahedron cluster-field approximation. For this reason, the presence of anti-Curie points for small v'_2 values is not significant for the physically interesting temperature range $T/T_c \gtrsim 0.3$.

If $v'_2 \gtrsim 0.25$, anti-Curie T_{ac}^{PCA} temperatures, however, sharply increase and the pair cluster approximation becomes inapplicable for describing v'_2 models. For example, at $v'_2 \geq 0.421$, the $B2$ ordered phase does not appear at all in the pair cluster approximation. This is illustrated by the calculations shown in Figs. 7–9.

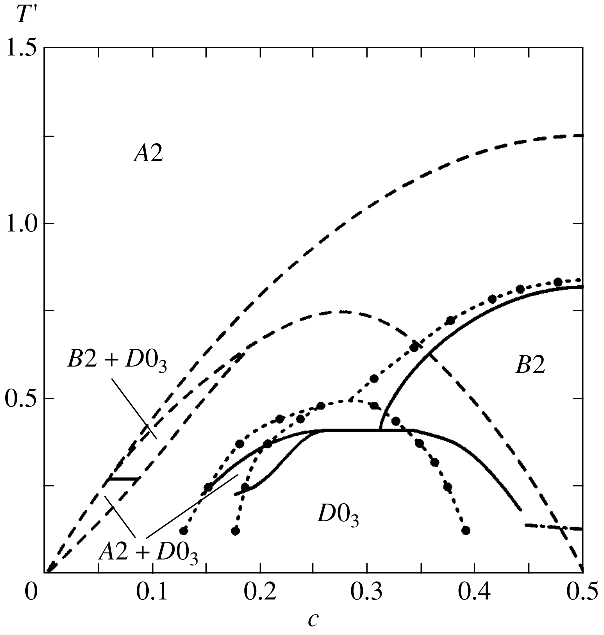


Fig. 9. The same as in Fig. 7 for model with $v'_2 = 0.5$, $v_{n>2} = 0$. Solid circles correspond to Monte Carlo calculations [1].

(3) Anti-Curie points are also present in the tetrahedron cluster calculations for v'_2 models. The corresponding T_{ac} values are, however, much lower than in the pair cluster approximation, and, for all $v'_2 \leq 0.5$ values under consideration, the T_{ac} values are small compared with the T_c critical temperature. This is illustrated by Figs. 7–9. Figure 9 shows that the phase diagram obtained for $v'_2 = 0.5$ in the tetrahedron cluster-field approximation is in satisfactory agreement with the results of Monte Carlo calculations [1] at all $T \gtrsim 0.3T_c$ that are not too low, although errors caused by the presence of anti-Curie points become noticeable at small $T \lesssim 0.2T_c$. Figure 9 also shows that mean-field calculations for this model not only quantitatively, but also qualitatively, distort the form of the phase diagram, and, as mentioned, the pair cluster approximation is completely inapplicable to such models.

Next, note that at large v'_2, the presence of even comparatively weak interactions between more distant neighbors that favor ordering can substantially change phase diagrams compared with "pure" v'_2 models. For example, Fig. 10 shows that the addition of a very weak attraction corresponding to $v'_3 = -0.05$ to the v'_2 model with $v'_2 = 0.5$ noticeably broadens the region of ordered phase stability and decreases the difficulties of

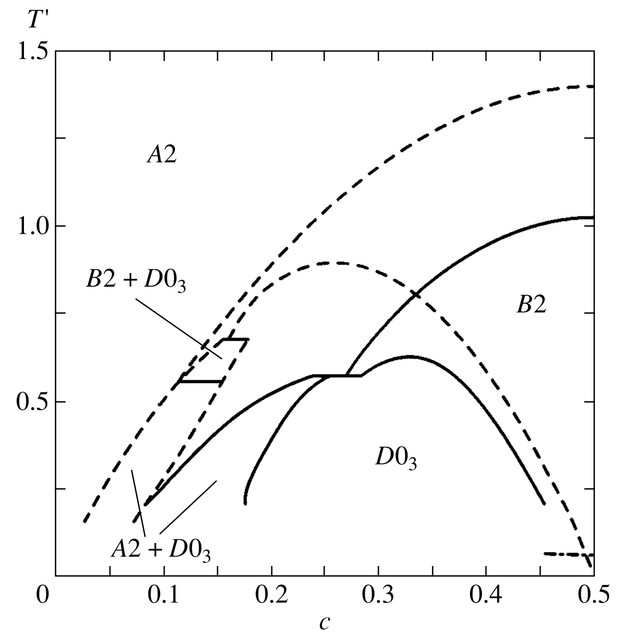


Fig. 10. The same as in Fig. 7 for model with v'_2 = 0.5, v'_3 = –0.05, and $v_{n>3}$ = 0.

the method related to the presence of the anti-Curie point by decreasing the T_{ac}/T_c ratio more than twice. As has been mentioned above (and as illustrated by the v'_n values in the table), the presence of interactions between distant neighbors that favor ordering is typical of all alloys studied experimentally. Therefore, there is good reason to believe that, if alloys with large v'_2 do exist in nature, they are also characterized by noticeable “attraction” of more distant neighbors and the accuracy of the tetrahedron cluster-field approximation for them should be substantially higher than for the v'_2 model shown in Fig. 9.

7. CONCLUSIONS

The main results of this work are as follows. The cluster methods for describing the thermodynamic properties of alloys developed earlier were generalized to nonuniform and nonequilibrium atomic distributions. The tetrahedron cluster-fields method for BCC lattices was suggested; this method could also be used to describe alloys with strong short-range and competing interactions that hinder ordering. The accuracy of various statistical methods (the mean-field and pair cluster approximations and the tetrahedron cluster-field method developed in this work) for describing phase diagrams of BCC alloys with *B*2- and $D0_3$-type orderings was studied. The results of calculations by various methods were compared with each other and with the available Monte Carlo results. The accuracy of various methods was found to sharply depend on the type of interactions in the system, first of all, on the presence of competing interactions and correlations mentioned above, which impede ordering. In the absence of such correlations, in particular, for the Fe–Al-type alloys with long-range interactions, the use of the pair cluster approximation substantially refines the results obtained by the simple mean-field method without any noticeable complication of calculations. A comparison with the available Monte Carlo calculations allowed realistic estimates to be obtained for exact transition temperatures based on the temperature values calculated in the mean-field and pair cluster approximations. Employing both the mean-field and pair cluster approximations to systems with short-range and competing interactions may sharply distort the form of phase diagrams. At the same time, the use of the tetrahedron cluster field approximation for such systems enables one to fairly accurately calculate their phase diagrams for physically interesting temperature and concentration values. It is likely that taking into account interactions between the third and more distant neighbors, which are usually fairly noticeable in real alloys, should increase the accuracy of tetrahedron cluster calculations compared with the calculations described in this work.

ACKNOWLEDGMENTS

The authors thank V.Yu. Dobretsov and K.D. Belashchenko for useful discussions. This work was financially supported by the Russian Foundation for Basic Research (project nos. 00-02-17692 and 00-15-96709).

REFERENCES

1. B. Dünweg and K. Binder, Phys. Rev. B **36**, 6935 (1987).
2. F. Schmid and K. Binder, Phys. Rev. B **46**, 13553 (1992).
3. S. V. Semenovskaya, Phys. Status Solidi B **64**, 291 (1974).
4. G. Inden, Acta Metall. **22**, 945 (1974).
5. M. Hasaka, Trans. Jpn. Inst. Met. **21**, 660 (1980).
6. V. G. Vaks, N. E. Zein, and V. V. Kamyshenko, J. Phys. F **18**, 1641 (1988).
7. W. Schweika, Mater. Res. Soc. Symp. Proc. **166**, 249 (1990).
8. V. Pierron-Bohnes, M. C. Cadeville, A. Finel, and O. Schaerpf, J. Phys. (Paris) **11**, 247 (1991).
9. V. Pierron-Bohnes, S. Lefevbre, M. Bessiere, and A. Finel, Acta Metall. Mater. **38**, 2701 (1991).
10. A. Finel, in *Statics and Dynamics of Alloy Phase Transformations*, Ed. by A. Gonis and P. E. A. Turchi (Plenum, New York, 1994), p. 495, NATO ASI Ser., Ser. B, Vol. 319.

11. R. V. Chepulskii and V. N. Bugaev, J. Phys.: Condens. Matter **10**, 8771 (1998).
12. V. Yu. Dobretsov, V. G. Vaks, and G. Martin, Phys. Rev. B **54**, 3227 (1996).
13. K. D. Belashchenko, G. D. Samolyuk, and V. G. Vaks, J. Phys.: Condens. Matter **11**, 10567 (1999).
14. K. D. Belashchenko, V. Yu. Dobretsov, I. R. Pankratov, *et al.*, J. Phys.: Condens. Matter **11**, 10593 (1999).
15. I. R. Pankratov and V. G. Vaks, J. Phys.: Condens. Matter **13**, 6031 (2001).
16. K. D. Belashchenko, I. R. Pankratov, G. D. Samolyuk, and V. G. Vaks, J. Phys.: Condens. Matter **14**, 565 (2002).
17. R. Kikuchi, Phys. Rev. **81**, 988 (1951).
18. R. Kikuchi and C. M. van Baal, Scr. Metall. **8**, 425 (1974).
19. V. G. Vaks and G. D. Samolyuk, Zh. Éksp. Teor. Fiz. **115**, 158 (1999) [JETP **88**, 89 (1999)].
20. V. G. Vaks, Pis'ma Zh. Éksp. Teor. Fiz. **63**, 447 (1996) [JETP Lett. **63**, 471 (1996)].
21. K. D. Belashchenko and V. G. Vaks, J. Phys.: Condens. Matter **10**, 1965 (1998).
22. V. G. Vaks and N. E. Zein, Zh. Éksp. Teor. Fiz. **67**, 1082 (1974) [Sov. Phys. JETP **40**, 537 (1974)].
23. M. A. Krivoglaz, Zh. Éksp. Teor. Fiz. **32**, 1368 (1957) [Sov. Phys. JETP **5**, 1115 (1957)].

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 116–126.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 130–141.
Original Russian Text Copyright © 2003 by Red'ko, Kagan, Rodionov, Pol'shin.

SOLIDS
Electronic Properties

Magnetic Field Effects on the Phonon Thermal Electromotive Force in *n*-Bi–Sb Semiconducting Alloys

N. A. Red'ko*, V. D. Kagan, N. A. Rodionov, and V. I. Pol'shin

Ioffe Physicotechnical Institute, Russian Academy of Sciences,
ul. Politekhnicheskaya 26, St. Petersburg, 194021 Russia
**e-mail: nikolaj.a.redko@mail.ioffe.ru*

Received December 15, 2002

Abstract—The phonon thermal electromotive force component α_{22} ($\nabla T \parallel C_1$) prevails in n-Bi$_{1-x}$Sb$_x$ ($0.07 \le x \le 0.16$) semiconducting alloys at low temperatures. This component increases by almost an order of magnitude in a classically strong transverse magnetic field $\mathbf{H}$ with $\mathbf{H} \parallel C_3$, which results in an increase in thermoelectric efficiency. The transverse Nernst–Ettingshausen coefficient $Q_{12,3}$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) changes sign from negative at $T > 10$ K to positive at $T < 10$ K. The observed characteristics of the phonon thermal electromotive force and the phonon transverse Nernst–Ettingshausen coefficient are explained in terms of the theory of electron–phonon drag for electrons with a strongly anisotropic spectrum. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The interaction of electrons with phonons in semiconductors causes not only their mutual scattering but also the drag effect, which can be observed for a nonequilibrium distribution of phonons, for instance, when there is a temperature gradient along a sample. The low-temperature thermal electromotive force (EMF) of semiconductors comprises two components, namely, the diffusion component caused by the redistribution of charge carriers under the action of a temperature gradient and the phonon component caused by the redistribution of charge carriers as a result of their collisions with phonons, which are in a nonequilibrium state in the presence of a temperature gradient. It follows that the contribution to the phonon thermal EMF is only made by the flux part of the phonon distribution function, which, in addition, is proportional to the temperature gradient. At low temperatures, the phonon thermal EMF component exceeds the diffusion component several or several dozen times depending on the degree of electron gas degeneracy.

The mechanism of carrier–phonon drag in solids in the presence of a temperature gradient was first predicted by Gurevich [1]. Currently, a large number of works have been published in which experimental data and the theory of the phonon drag of thermal EMF and thermomagnetic coefficients in semiconductors with isotropic electron energy spectra are described. A review of these works can be found, for instance, in monographs [2–5].

According to phonon drag theory, thermal EMF and thermomagnetic coefficients in semiconductors with isotropic electron energy spectra do not contain the small factor kT/ζ present in diffusion coefficients [2]. Secondly, phonon thermal EMF does not depend on transverse magnetic fields, and the phonon transverse Nernst–Ettingshausen coefficient is zero. Here, ζ is the chemical potential of electrons in semiconductors.

However, according to theoretical work [6], thermal EMF in anisotropic semimetal bismuth does depend on the transverse magnetic field and the phonon transverse Nernst–Ettingshausen coefficient is nonzero.

The purpose of this work was to analyze the behavior of the phonon thermal EMF and the phonon transverse Nernst–Ettingshausen coefficient of Bi–Sb semiconducting alloys depending on magnetic field and temperature. Previously [7, 8], we discussed the behavior of the diffusion thermal EMF and the diffusion transverse Nernst–Ettingshausen coefficient under temperature and magnetic field variations. These data were used to study the mechanisms of electron relaxation in alloys and to explain the maximum, which was observed in the magnetic field dependence of diffusion thermal EMF.

2. SAMPLES AND PROCEDURE FOR MEASUREMENTS

The temperature dependences of thermal EMF $\alpha_{22}(0)$ ($\nabla T \parallel C_1$) in the absence of a magnetic field and thermomagnetic effects $Q_{12,3}(H, T)$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) and changes in the thermal EMF in a magnetic field $\Delta\alpha_{22}(H) = \alpha_{22}(H) - \alpha_{22}(0)$ were measured for single crystals of Bi$_{1-x}$Sb$_x$ ($0.07 \le x \le 0.16$) semiconducting alloys doped with tellurium as a donor impurity. The samples were cut on an electroerosion unit from the middle part of a single crystalline ingot grown by the method of horizontal zone recrystallization. The samples were cut as $3 \times 3 \times 30$ mm^3 rectangular parallelepipeds, whose faces were perpendicular to the crystallographic axes C_1, C_2, and C_3. The largest sample dimension was along the C_1 bisector axis. The samples were etched with a 1 : 1 C_2H_5OH–HNO_3 solution. The main sample parameters are listed in the table.

The sample to be studied with heaters at its ends was soldered to the bottom of a vacuum chamber 18 mm in diameter placed into a thermostatic liquid (helium, hydrogen, or nitrogen). The heater at the end soldered to the chamber bottom served to control the mean temperature of the sample, and the heater at the opposite end, to create a temperature gradient ∇T in the sample. The temperature was measured for two cross sections spaced $l_T \approx 15$ mm at a constant heat flux. The measurements were taken with carbon resistance thermometers at $1.5 < T < 40$ K and copper–constantan thermocouples at $30 < T < 80$ K. The thermal EMF was measured for sample–copper thermoelectric couples; the absolute thermal EMF of copper did not exceed 1 μV/K in the temperature range of measurements. The experimental thermal EMF data are given below without taking this circumstance into account. The thermomagnetic effects were measured for Bi–Sb alloys in magnetic fields of $0 \leq H < 18$ kOe at temperatures of $1.4 \leq T < 40$ K.

3. MEASUREMENT RESULTS

The electron gas in the region of impurity conductivity ($T < 40$ K) is degenerate in the semiconducting alloys under consideration, and the Fermi surface of the alloys consists of three electronic ellipsoids centered at the L Brillouin zone points, which are situated in the mirror reflection planes [8]. One of the minor ellipsoid axes coincides with the C_2 binary axis of the crystal; the ellipsoids are turned about this axis through a small angle φ, whose values are listed in the table. As a result of these turns, the two other ellipsoid axes form angles φ with the C_1 and C_3 crystallographic axes. The electronic ellipsoids are equivalent in accordance with crystal symmetry and have strongly anisotropic effective masses, whose values are also listed in the table.

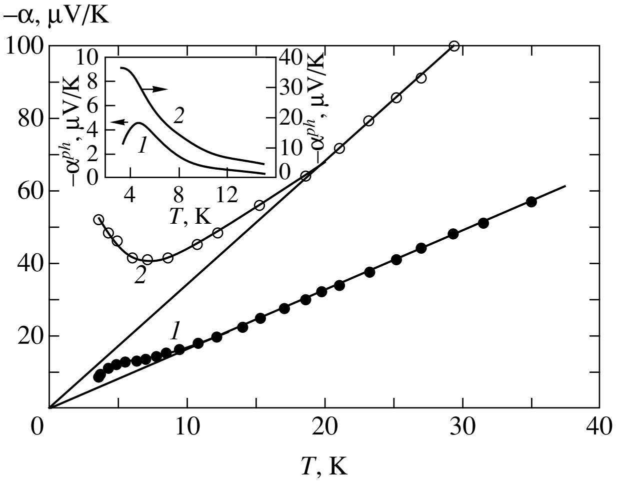


Fig. 1. Temperature dependence of the thermal EMF of the n-$Bi_{0.93}Sb_{0.07}$ semiconducting alloy (*1*) in zero magnetic field, $\alpha_{22}(0)$ ($\nabla T \parallel C_1$) and (*2*) in a transverse classically strong magnetic field, α_∞ ($\mathbf{H} \parallel C_3$). Solid straight lines pass through the experimental diffusion thermal EMF values at $T > 20$ K and are extrapolated to $T = 0$ K. The same dependences for the phonon component are shown in the inset.

We measured the temperature dependence of thermal EMF α_{22} ($\nabla T \parallel C_1$) for semiconducting $Bi_{1-x}Sb_x$ ($0.07 \leq x \leq 0.16$) alloys, for which a linear temperature dependence was observed at $T > 18$ K. Such a dependence is characteristic of semiconductors with a degenerate electron gas, $\alpha_{22}(T) \propto kT/\zeta$; this was discussed in [7, 8]. The thermal EMF of Bi–Sb alloys deviates from linearity below $T = 18$ K. By way of example, these thermal EMF deviations for $Bi_{0.93}Sb_{0.07}$ are shown in Fig. 1 (curve *1*). The experimental thermal EMF val-

Main parameters of the studied n-$Bi_{1-x}Sb_x$ ($0.07 \leq x \leq 0.16$) semiconducting alloys

No.	x	φ	ε_{gL}, meV	n, 10^{17} cm^{-3}	ζ, meV	m_1/m_0, 10^{-3}	m_2/m_0	m_3/m_0, 10^{-3}	δ
1	0.07	5.7°	7.5	1.37	18.6	0.91	0.32	1.2	24.5
2	0.09	5.55°	11	1.95	19	1.43	0.48	1.74	24
3	0.12	5.3°	16.4	1.57	15.5	2.42	0.72	2.58	22.6
4	0.13	5.14°	19	1.58	14.3	3.1	0.83	3.0	21.6
5	0.15	4.97°	22	1.61	13.7	3.86	0.94	3.45	20.5
6	0.16	4.88°	23.6	1.71	13.5	4.37	1.0	3.7	19.7

Note: x, ε_{gL}, n, ζ, m_i, and $\delta = (m_1 + m_4)^2/4m_1m_4$ are the concentration of antimony in the alloy, the forbidden band energy in the semiconducting alloy, the concentration of electrons in the sample, the chemical potential of electrons in the sample, the effective masses of electrons in the ellipsoid at the band bottom in the alloy, and the parameter taking into account the anisotropy of the energy spectrum of electrons in the alloy, respectively.

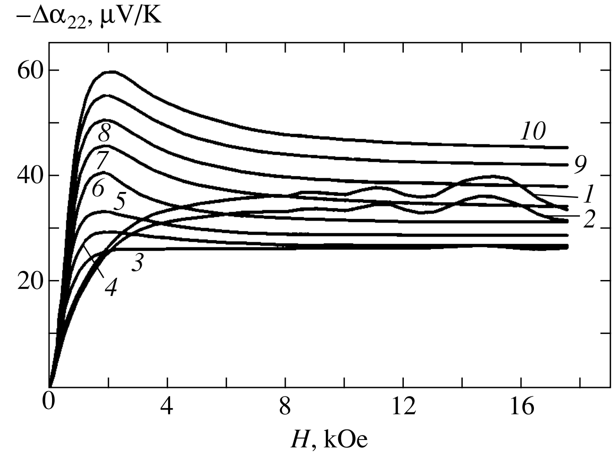


Fig. 2. Transverse magnetic field dependences of thermal EMF changes $\Delta\alpha_{22}(H) = \alpha_{22}(H) - \alpha_{22}(0)$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) for the n-$Bi_{0.93}Sb_{0.07}$ alloy at various temperatures: (*1*) 4.2, (*2*) 4.8, (*3*) 8.5, (*4*) 10.6, (*5*) 12.3, (*6*) 15.4, (*7*) 18.6, (*8*) 21.0, (*9*) 23.2, and (*10*) 25.1 K.

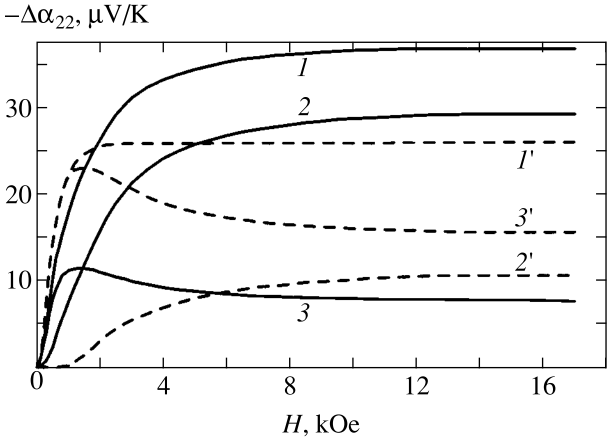


Fig. 3. Transverse magnetic field dependences of changes in the thermal EMF $\Delta\alpha_{22}(H)$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$): (*1*, *1'*) experimental dependences and their (*2*, *2'*) phonon and (*3*, *3'*) diffusion components for the n-$Bi_{0.93}Sb_{0.07}$ semiconducting alloy at (*1*, *2*, *3*) 4.2 and (*1'*, *2'*, *3'*) 8.5 K.

ues at $T < 18$ K exceed the diffusion thermal EMF, which is linearly continued in Fig. 1 from the region of high temperatures $T > 18$ K to the lower temperatures. This difference is natural to relate to the phonon thermal EMF component. The temperature dependence of the phonon thermal EMF component is shown in the inset in Fig. 1 (curve *1*); this dependence was obtained by subtracting the diffusion thermal EMF from the total experimental value. The phonon thermal EMF increases by the power law $\alpha^{ph}(T) \propto T^{-2.8}$ as temperature lowers from $T = 18$ K and reaches a maximum at $T \approx 4.5$ K, whereas the temperature corresponding to the phonon heat conductivity maximum is somewhat lower ($T \approx 4$ K).

We also measured the thermal EMF $\alpha_{22}(H)$ ($\nabla T \parallel C_1$) in transverse magnetic fields up to 18 kOe with $\mathbf{H} \parallel C_3$. These measurements were made by two methods. First, direct thermal EMF measurements for the Bi–Sb alloy under study were performed at a given temperature and a certain magnetic field value, that is, point-by-point, to obtain $\alpha(H) = V/\Delta T$ dependences, where V is the voltage drop between sample contacts for thermal EMF measurements and ΔT is the temperature difference between these contacts. The second procedure was based on recording ΔV, that is, changes in the voltage drop between sample contacts for thermal EMF measurements, in various magnetic fields. Further, the equation

$$\Delta\alpha(H) = \alpha(H) - \alpha(0) = \Delta V/\Delta T$$

was used to calculate the change in the thermal EMF of the Bi–Sb sample under study at a given temperature as a function of the magnetic field value. A family of $\Delta\alpha_{22}(H) = \alpha_{22}(H) - \alpha_{22}(0)$ curves versus magnetic field with $\mathbf{H} \parallel C_3$ recorded at various temperatures is shown in Fig. 2 for the $Bi_{0.93}Sb_{0.07}$ semiconducting alloy.

In classically strong magnetic fields ($\Omega\tau \gg 1$), we have $\Delta\alpha_{22}(H) = \Delta\alpha_\infty$. Here, $\Omega = eH/(c\sqrt{M_1M_4})$ is the cyclotron frequency; τ is the total relaxation time of alloy electrons; $M_i = m_i(1 + 2\zeta/e_g)$; m_i and M_i are the effective masses of electrons in Bi–Sb alloys at the conduction band bottom and at the Fermi level, respectively; and ε_g is the forbidden band energy. The experimental thermal EMF values $\alpha_\infty = \Delta\alpha_\infty + \alpha_{22}(0)$ in a classically strong magnetic field ($\Omega\tau \gg 1$) are shown in Fig. 1 versus temperature (curve *2*). Above $T = 20$ K, the temperature dependence of thermal EMF α_∞ is linear and corresponds to the diffusion thermal EMF component in a classically strong magnetic field for semiconductors with a degenerate electron gas [2, 7]. The experimental thermal EMF changes $\Delta\alpha_\infty^{\text{expt}}$ in a classically strong magnetic field coincide with the diffusion component $\Delta\alpha_\infty^{\text{diff}}$ at $T > 20$ K and linearly depend on temperature. At low temperatures of $T < 20$ K, the $\Delta\alpha_\infty^{\text{diff}}$ value exceeds $\Delta\alpha_\infty^{\text{diff}}$ obtained by linear extrapolation from the temperature region $T > 20$ K to lower temperatures.

The temperature dependence of the phonon thermal EMF $\alpha_\infty^{ph} = \Delta\alpha_\infty^{ph} + \alpha^{ph}(0)$ in a classically strong magnetic field ($\Omega\tau \gg 1$) is shown in the inset in Fig. 1 (curve *2*). Phonon thermal EMF changes were found by subtracting the $\Delta\alpha_\infty^{\text{diff}}$ diffusion component from the experimental $\Delta\alpha_\infty^{\text{expt}}$ value, $\Delta\alpha_\infty^{ph} = \Delta\alpha_\infty^{\text{expt}} - \Delta\alpha_\infty^{\text{diff}}$.

The phonon components of thermal EMF changes (curves *2*, *2'*) as a function of transverse magnetic field

(**H** || C_3) measured at 4.2 and 8.5 K for the $Bi_{0.93}Sb_{0.07}$ alloy is shown in Fig. 3. This dependence was obtained by subtracting the diffusion thermal EMF components (curves *3*, *3*') from the experimental $\Delta\alpha_{22}$ ($\nabla T \parallel C_1$) values (curves *1*, *1*').

The diffusion component of thermal EMF changes in a transverse magnetic field was calculated by the equation [8]

$$\Delta\alpha(H) = -\frac{\pi^2 k^2 T}{3e\zeta}\beta\left\{1 + \frac{\Omega^2\tau^2(\delta - 2) - \delta}{(1 + \Omega^2\tau^2)(\delta + \Omega^2\tau^2)}\right\}. \quad (1)$$

The $\Delta\alpha(H)$ value was found using the τ relaxation time of electrons in Bi–Sb alloys obtained by linearly extrapolating the temperature dependence of the inverse electron relaxation time from high temperatures of $T > 20$ K to the low-temperature region. The inverse electron relaxation times for the alloys at $T > 20$ K were taken from [8]. The total electron relaxation time τ in Bi–Sb semiconducting alloys depends on temperature, and this dependence is caused by the contribution of electron scattering by acoustic phonons. The inverse electron relaxation time linearly depends on temperature ($\tau^{-1} \propto T$) at $T > \Theta_e = 19$ K, because, starting with this temperature, the number of phonons that interact with electrons increases linearly. Here, $\Theta_e = 2p_F s/k$ is the electron Debye temperature, where p_F and s are the Fermi momentum of the electron and the velocity of sound in the alloy. The Debye temperature varies in the interval $\Theta_{\min} \approx 1\ \mathrm{K} \le \Theta_e \le \Theta_{\max} \approx 19$ K for the electron ellipsoid in the alloys under consideration.

Note that the linear extrapolation of the temperature dependence of the electron relaxation time performed in this work is inaccurate at $T < \Theta_e$. At the same time, electron scattering by impurities rather than phonons prevails in the alloys to a substantial extent. The error introduced by this extrapolation should therefore be small.

The β parameter that we used to calculate the $\Delta\alpha(H)$ dependence by (1) was found from the experimental data on the diffusion thermal EMF $\Delta\alpha_\infty$ by the equation [8]:

$$\Delta\alpha_\infty = -\frac{\pi^2 k^2 T}{3e\zeta}\beta. \quad (2)$$

Equation (2) gives $\beta = 1.35$ for the $Bi_{0.93}Sb_{0.07}$ alloy. The chemical potential of electrons ζ used in the calculations and the δ parameter for the samples studied in this work are given in the table.

The field dependences of the diffusion and phonon thermal EMF components are different in the alloys (Fig. 3). The phonon thermal EMF component exhibits a monotonic magnetic field dependence, whereas the dependence of the diffusion component is nonmonotonic. The magnetic field (**H** || C_3) dependence of the diffusion thermal EMF component in Bi–Sb semiconducting alloys has a maximum in intermediate magnetic fields ($\Omega\tau \approx 1$) caused by the multivalley character and strong anisotropy of the electron energy spectrum [7, 8]. Also note that the diffusion thermal EMF component is differential in character, whereas the phonon thermal EMF component is an integral characteristic corresponding to averaged features of electron energy and phonon spectrum anisotropy and electron and phonon relaxation processes [8]. For this reason, the phonon thermal EMF component exhibits a monotonic magnetic field dependence in the alloys.

The temperature dependence of the phonon thermal EMF in a classically strong magnetic field, $\alpha_\infty^{ph} = \Delta\alpha_\infty^{ph} + \alpha^{ph}(0)$, is shown in Fig. 1 (inset, curve *2*); the dependence follows the $T^{-1.7}$ law. The exponent is smaller than for the phonon thermal EMF in a zero magnetic field and slightly larger than for the temperature dependence of the phonon heat conductivity of Bi–Sb semiconducting alloys ($\kappa \propto T^{-1.3}$). The $\alpha_\infty^{ph}(T)$ curve has a maximum at $T \approx 3.5$ K, which is shifted downward from the temperature of maximum photon thermal EMF in the absence of a magnetic field.

The transverse Nernst–Ettingshausen coefficient

$$Q = -\frac{Vl}{hH\Delta T}$$

was measured for Bi–Sb semiconducting alloys in the same temperature range. Here, V, h, H, and ΔT are the voltage drop between the measuring Nernst–Ettingshausen probes on the sample, the distance between the measuring probes, the transverse magnetic field applied to the sample, and the temperature difference between points spaced l along the sample, respectively.

The magnetic field dependence of the transverse Nernst–Ettingshausen coefficient $Q_{12,3}$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) at various temperatures is shown in Fig. 4 for the $Bi_{0.93}Sb_{0.07}$ semiconducting alloy. The $Q_{12,3}$ value is negative at $T > 10$ K because of predominant electron scattering by acoustic phonons and by the fluctuation potential related to fluctuations of the concentration of Bi and Sb in the alloy (the “alloy” mechanism of electron scattering) [2, 7, 8]. According to the theory of semiconductors with isotropic electron energy spectra, positive $Q_{12,3}$ values might be evidence of predominant electron scattering by ionized impurities [2].

In [8], we analyzed the contributions of all electron scattering mechanisms in *n*-Bi–Sb semiconducting alloys at $T > 20$ K; the important role played by the alloy mechanism was noted. The sign of the transverse Nernst–Ettingshausen coefficient is negative for the alloy electron scattering mechanism, as for scattering by acoustic phonons according to experimental $Q_{12,3}$ measurements at $T > 10$ K. In Bi–Sb alloys, a decrease in the temperature has no effect on the predominance of

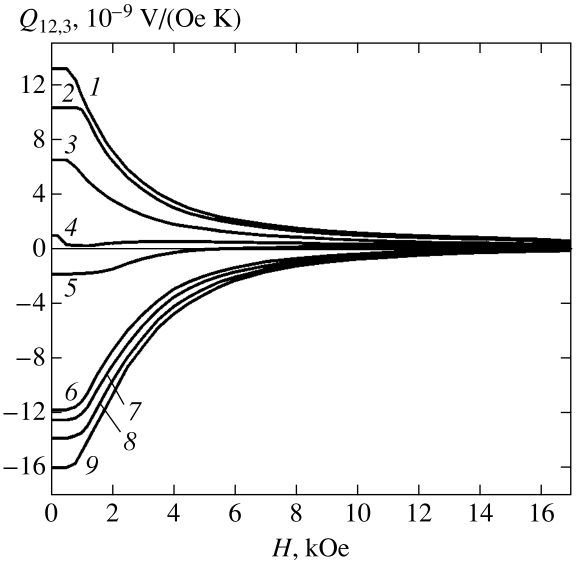


Fig. 4. Magnetic field dependences of the transverse Nernst–Ettingshausen coefficient $Q_{12,\,3}$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) for the *n*-$Bi_{0.93}Sb_{0.07}$ semiconducting alloy at various temperatures: (*1*) 4.2, (*2*) 4.8, (*3*) 6.2, (*4*) 8.5, (*5*) 10.6, (*6*) 18.6, (*7*) 21.0, (*8*) 23.2, and (*9*) 25.1 K.

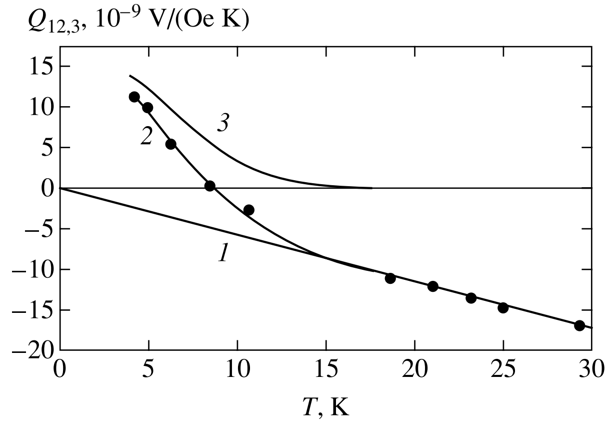


Fig. 5. Temperature dependences of the transverse Nernst–Ettingshausen coefficient: (*2*) experimental $Q_{12,\,3}$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) dependence and its (*1*) diffusion and (*3*) phonon components in a constant magnetic field $H = 1$ kOe for the *n*-$Bi_{0.93}Sb_{0.07}$ semiconducting alloy.

the alloy electron scattering mechanism and therefore cannot cause the reversal of the sign of the transverse Nernst–Ettingshausen diffusion coefficient.

Like the thermal EMF, the transverse Nernst–Ettingshausen coefficient of the alloys at low temperatures contains two (phonon and diffusion) components. The temperature dependence of the transverse Nernst–Ettingshausen coefficient $Q_{12,\,3}$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) measured at $H = 1$ kOe for the $Bi_{0.93}Sb_{0.07}$ semiconducting alloy is shown in Fig. 5. Curve *2* is the experimental $Q_{12,\,3}(T)$ dependence. At temperatures above $T = 18$ K, the experimental values of $Q_{12,\,3}(T)$ fall on a straight line (straight line *1*) that corresponds to the diffusion component. The diffusion coefficient components at $T > 18$ K, like the diffusion thermal EMF components, are discussed and compared with theoretical results in [7, 8].

A characteristic feature of the experimental $Q_{12,\,3}(T)$ curve is a change in sign from negative at $T > 10$ K to positive at $T < 10$ K. The difference between the total experimental value (Fig. 5, curve *2*) and the diffusion component (straight line *1*) is shown in Fig. 5 (curve *3*); this difference corresponds to the phonon component. According to the theory described in [2], the phonon contribution to $Q_{12,\,3}(T)$ is zero in semiconductors with an isotropic electron energy spectrum, which does not explain the dependences observed experimentally.

The temperature dependences of the $Q_{12,\,3}$ coefficient measured for $Bi_{0.93}Sb_{0.07}$ in various magnetic fields are shown in Fig. 6. The contribution of the phonon component compared with the diffusion component increases as the magnetic field grows stronger. This causes a change in the sign of the coefficient at higher temperatures.

The magnetic field dependences of the phonon contribution to the transverse Nernst–Ettingshausen coefficient measured for $Bi_{0.93}Sb_{0.07}$ at 4.2 and 8.5 K are shown in Fig. 7 (curves *3* and *3*', respectively). These dependences were obtained by subtracting the diffusion contribution to the transverse Nernst–Ettingshausen coefficient (curves *1*, *1*') from the experimental $Q_{12,\,3}(T)$ dependences (curves *2*, *2*'). The diffusion contribution to the transverse Nernst–Ettingshausen coefficient was calculated by the equation [8]

$$Q_{12,\,3} = -\frac{\pi^2 k^2 T}{6\zeta}\beta\frac{(M_1 + M_4)\tau}{M_1 M_4 c(\delta + \Omega^2\tau^2)}. \qquad (3)$$

In these calculations, we used the same electron relaxation times τ as in the calculations of the diffusion component of $\alpha_{22}(H)$ by (1); these τ values were taken from [8]. The diffusion component of $Q_{12,\,3}$ for $Bi_{0.93}Sb_{0.07}$ was calculated using the $\beta = 1.35$ parameter and the tabulated chemical potential ζ and δ values.

We come to the conclusion that the results of processing the thermal EMF data given above for the $Bi_{0.93}Sb_{0.07}$ alloy show that the phonon thermal EMF at low temperatures (Fig. 1, inset) depends fairly strongly on the transverse magnetic field. The phonon thermal EMF in a classically strong magnetic field is almost an order of magnitude larger than the phonon thermal EMF in the absence of a magnetic field. This is at variance with the kinetic theory of semiconductors with isotropic electron energy spectra, according to which the phonon thermal EMF is independent of magnetic field, and the Nernst–Ettingshausen effect is absent under phonon drag conditions [2].

We undertook a theoretical analysis to elucidate the reasons for the observed features of the phonon thermal EMF in magnetic fields and the phonon transverse Nernst–Ettingshausen coefficient in semiconducting Bi–Sb alloys.

4. PHONON DRAG THEORY FOR SEMICONDUCTING ALLOYS

The magnetic field dependence of the tensor of phonon thermal EMF was found by solving the kinetic equation for a strongly anisotropic nonparabolic dispersion law of L electrons in semiconducting Bi–Sb alloys within the framework of the Lax model,

$$\varepsilon_{\mathbf{p}} = \sqrt{\left(\frac{\varepsilon_g}{2}\right)^2 + \frac{\varepsilon_g}{2}\left(\frac{p_1^2}{m_1} + \frac{p_2^2}{m_2} + \frac{p_2^3}{m_3}\right)} - \frac{\varepsilon_g}{2}, \qquad (4)$$

where $\varepsilon_{\mathbf{p}}$ is the energy of the electron with momentum $\mathbf{p}$ (its components along the ellipsoid axes are p_1, p_2, and p_3) and m_i are the effective masses of the electrons in the ellipsoid.

The distribution function of the electrons was found by solving the linearized kinetic equation

$$\begin{aligned}\frac{f_{\mathbf{p}}}{\tau_e(\varepsilon_{\mathbf{p}})} &= \sum_{s=1}^{3}\int\frac{d^3q}{(2\pi)^3}\frac{2\pi}{\hbar}|C_{\mathbf{q}}|^2 n_{\mathbf{q}} \\ &\times\{[f_0(\varepsilon_{\mathbf{p}} - \hbar\omega_{\mathbf{q}}^{(s)}) - f_0(\varepsilon_{\mathbf{p}})]\delta(\varepsilon_{\mathbf{p}} - \varepsilon_{\mathbf{p}-\hbar\mathbf{q}} - \hbar\omega_{\mathbf{q}}^{(s)}) \\ &+ [f_0(\varepsilon_{\mathbf{p}} + \hbar\omega_{\mathbf{q}}^{(s)}) - f_0(\varepsilon_{\mathbf{p}})] \\ &\times\delta(\varepsilon_{\mathbf{p}} - \varepsilon_{\mathbf{p}+\hbar\mathbf{q}} + \hbar\omega_{\mathbf{q}}^{(s)})\},\end{aligned} \qquad (5)$$

where $f_{\mathbf{p}}$ and $n_{\mathbf{q}}$ are the nonequilibrium additions to the distribution functions of the electrons and phonons, which depend on the electronic momentum $\mathbf{p}$ and the phonon wave vector $\mathbf{q}$;

$$f_0(\varepsilon_{\mathbf{p}}) = \frac{1}{\exp[(\varepsilon_{\mathbf{p}} - \zeta)/kT + 1]}$$

is the equilibrium Fermi distribution function for electrons with the chemical potential ζ; $\omega_{\mathbf{q}}^{(s)}$ is the frequency of the phonon with the wave vector $\mathbf{q}$ for the sth branch of vibrations; $\tau_e(\varepsilon_{\mathbf{p}})$ is the total electron relaxation time under scattering by phonons, ionized impurities, and fluctuations of the concentrations of Bi and Sb in the alloy; and $C_{\mathbf{q}}$ is the electron–phonon coupling constant.

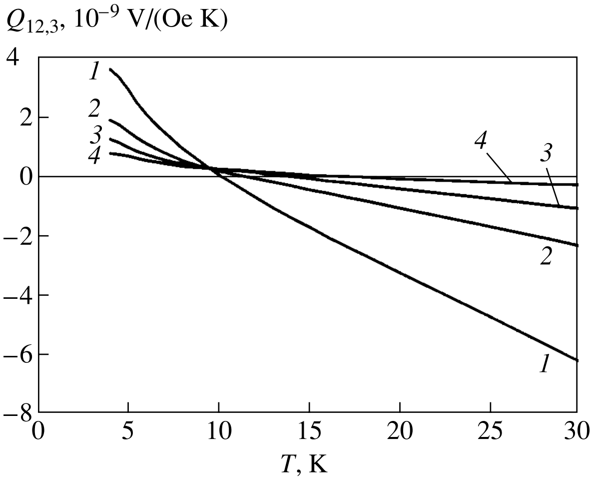


Fig. 6. Temperature dependences of the experimental transverse Nernst–Ettingshausen coefficient $Q_{12,3}$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) for the n-$Bi_{0.93}Sb_{0.07}$ semiconducting alloy in various constant magnetic fields: (*1*) 4, (*2*) 7, (*3*) 10, and (*4*) 15 kOe.

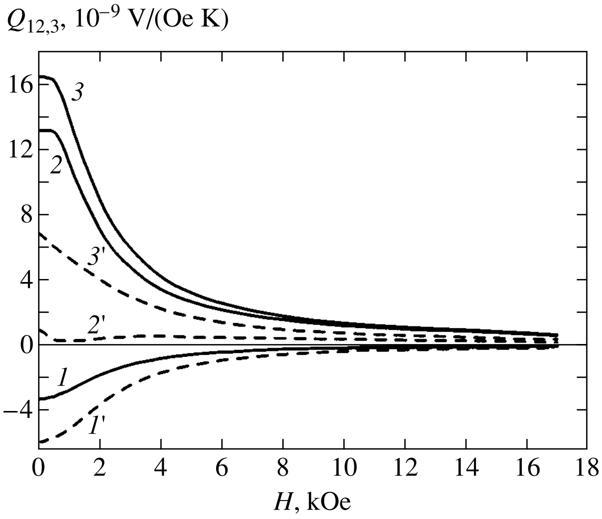


Fig. 7. Magnetic field dependences of the transverse Nernst–Ettingshausen coefficient: (*2*, *2*') experimental $Q_{12,3}$ ($\nabla T \parallel C_1$, $\mathbf{H} \parallel C_3$) dependences and their (*1*, *1*') diffusion and (*3*, *3*') phonon components for the n-$Bi_{0.93}Sb_{0.07}$ semiconducting alloy at (*1*, *2*, *3*) 4.2 and (*1*', *2*', *3*') 8.5 K.

The phonon distribution function is found from the kinetic equation for the sth-branch phonons,

$$\begin{aligned}&\frac{\partial\omega_{\mathbf{q}}}{\partial\mathbf{q}}\frac{\partial N_0}{\partial T}\nabla T + \left(\frac{1}{\tau_{ph-im}(\omega_{\mathbf{q}})} + \frac{1}{\tau_{ph-e}(\omega_{\mathbf{q}})}\right)n_{\mathbf{q}} \\ &+ \frac{1}{\tau_{ph-ph}(\omega_{\mathbf{q}})}\left(n_{\mathbf{q}} + \mathbf{q}\cdot\mathbf{V}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}}\right) = 0,\end{aligned} \qquad (6)$$

where $N_0 = [\exp(\hbar\omega_{\mathbf{q}}/kT) - 1]^{-1}$ is the equilibrium Planck function.

The kinetic equation for phonons was solved with the inclusion of scattering of phonons by impurities τ_{ph-im} and electrons τ_{ph-e} and phonon–phonon scattering τ_{ph-ph}. The Callaway method for solving the kinetic equation makes it possible to take into account the total momentum conservation in phonon–phonon collisions; this is attained by using a term containing the mean phonon drift velocity **V**. This term is determined from the condition of the conservation of the total momentum in normal phonon–phonon collisions,

$$\sum_{s=1}^{3}\int d^3q\frac{\mathbf{q}}{\tau_{ph-ph}(\omega_{\mathbf{q}}^{(s)})}\left(n_{\mathbf{q}}+\mathbf{q}\cdot\mathbf{V}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right) = 0. \qquad (7)$$

The analysis performed above does not take into account the influence of the nonequilibrium electron distribution function on the nonequilibrium phonon distribution function; that is, it ignores mutual drag of electrons and phonons. This effect can only become significant when mutual scattering of electrons and phonons in the electronic and phonon systems is the major scattering mechanism. Such a situation can hardly arise in alloys. According to the detailed analysis of the experimental data performed in [8], this is not the situation with the electronic system. In this work, we are interested in the experimental thermal EMF data obtained at temperatures lower than those in [8], when scattering of electrons by phonons is negligibly small compared with alloy and impurity scattering. It follows that we can ignore mutual drag.

Equation (6) is completely independent of (5); we therefore solve it first taking into account condition (7). The obtained solution is substituted into (5), where the phonon energy is on the order of kT, which is much lower than the mean electron energy equal to the Fermi energy. This allows us to expand the right-hand side of (5) in powers of $\hbar\omega_{\mathbf{q}}/\varepsilon_{\mathbf{p}}$. Upon solving the kinetic Eqs. (5) and (6), we obtain the electron distribution function $f_{\mathbf{p}}$ antisymmetric with respect to momentum and proportional to ∇T. We can then calculate the thermoelectric tensor, which is the proportionality factor between thermoelectric current and temperature gradient. As for the diffusion component, we must sum the contributions to this tensor of three equivalent ellipsoids with the tilt angle φ between them and the C_3 axis. Multiplying the calculated tensor and the resistance tensor [2] gives the tensor of thermal EMF in a magnetic field with $\mathbf{H} \parallel C_3$. In this way, we find the α_{ii} diagonal and $\alpha_{ij} = Q_{ij}H$ off-diagonal components of the phonon tensor; that is, we determine the phonon transverse Nernst–Ettingshausen coefficient. The equations for the diagonal component of the tensor of phonon thermal EMF, $\alpha_{11}(H) = \alpha_{22}(H)$, in a transverse magnetic field ($\mathbf{H} \parallel C_3$) for semiconducting Bi–Sb alloys are

$$\alpha_{11}(H) = -\frac{1}{e}\frac{3\hbar(1+2\zeta/\varepsilon_g)}{8\rho\sqrt{[2\zeta(1+\zeta/\varepsilon_g)]^3}}$$

$$\times\sum_{s=1}^{3}\int\frac{d^3q}{(2\pi)^2}\frac{\tau_{ph}(\Lambda_{ik}q_i l_k^{(s)})^2}{\sqrt{q_1^2/m_1+q_2^2/m_2+q_3^2/m_3}}$$

$$\times\left[\left(1+\frac{\eta}{\delta+\Omega^2\tau_e^2}\right)\left(q_1\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_1}\frac{\partial N_0}{\partial T}-\frac{b_1}{a_1}\frac{q_1^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)\right.$$

$$+\left(1-\frac{\eta}{\delta+\Omega^2\tau_e^2}\right)\frac{m_4\cos^2\varphi}{m_2} \qquad (8)$$

$$\times\left(q_2\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_2}\frac{\partial N_0}{\partial T}-\frac{b_2}{a_2}\frac{q_2^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)$$

$$+\left(1-\frac{\eta}{\delta+\Omega^2\tau_e^2}\right)\frac{m_4\sin^2\varphi}{m_3}$$

$$\left.\times\left(q_3\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_3}\frac{\partial N_0}{\partial T}-\frac{b_3}{a_3}\frac{q_3^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)\right],$$

where

$$\delta = \frac{(m_1+m_4)^2}{4m_1m_4}, \qquad (9)$$

$$\eta = \frac{m_4^2-m_1^2}{4m_1m_4}, \qquad (10)$$

$$a_i = \int d^3q\frac{q_i^2\tau_{ph}}{\tau_{ph-ph}\tau_R}\left(-\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right), \quad i = 1, 2, 3, \qquad (11)$$

$$b_i = \int d^3q\frac{\tau_{ph}}{\tau_{ph-ph}}\left(q_i\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_i}\right)\left(\frac{\partial N_0}{\partial T}\right), \quad i = 1, 2, 3, \qquad (12)$$

$$\frac{1}{\tau_R(q)} = \frac{1}{\tau_{ph-im}(q)}+\frac{1}{\tau_{ph-e}(q)}, \qquad (13)$$

$$\frac{1}{\tau_{ph}(q)} = \frac{1}{\tau_R(q)}+\frac{1}{\tau_{ph-ph}(q, T)}. \qquad (14)$$

Here, e is the absolute charge of the electron; ρ is the density of the Bi–Sb alloy; ζ is the chemical potential of electrons; $\Omega = eH/(c\sqrt{M_1M_4})$ is the cyclotron frequency of electrons, the same for all three ellipsoids if $\mathbf{H} \parallel C_3$; $M_i = m_i(1 + 2\zeta/e_g)$ are the effective masses of electrons at the Fermi level; Λ_{ik} is the deformation potential tensor; $\mathbf{l}$ is the polarization vector of the sth phonon branch; $\tau_R(q)$ is the resistive phonon relaxation time corresponding to momentum relaxation in the phonon system; and $\tau_{ph}(q)$ is the total phonon relaxation time. The effective mass m_4 is related to the tilt of the ellipsoids with respect to the C_3 and C_1 axes as

$$\frac{1}{m_4} = \frac{\cos^2\varphi}{m_2} + \frac{\sin^2\varphi}{m_3}. \qquad (15)$$

Note that the integration in (8) with respect to wave vectors $\mathbf{q}$ is performed in a limited volume of momenta. This volume is determined by the laws of conservation of energy and momentum for interacting electrons and phonons,

$$\frac{\hbar^2 q_1^2}{m_1} + \frac{\hbar^2 q_2^2}{m_2} + \frac{\hbar^2 q_3^2}{m_3} \le 8\zeta\left(1 + \frac{\zeta}{\varepsilon_g}\right). \qquad (16)$$

The integrand in (8) contains terms with b_i/a_i coefficients. These terms take into account contributions of phonon–phonon collisions when the phonon kinetic equation is solved by the Callaway method. For very pure substances with $\tau_R(q) \gg \tau_{ph}(q)$, the term proportional to $\tau_R(q)$ becomes very large, which corresponds to two-step drag [9]. In semiconducting alloys, resistive scattering resulting from scattering of phonons by impurities is not weak, and, parametrically, the second term has the same order of magnitude as the first one.

The integrals that determine the coefficients a_i (11) and b_i (12) have no bearing on electron–phonon interactions. For this reason, restriction (16) on the volume of momenta for the integration does not apply to them; that is, the integration in (11) and (12) is performed over the entire volume of phonon wave vectors.

The δ (9) and η (10) parameters are large in magnitude and comparable because $m_4 \gg m_1$. They equal $m_4/4m_1$, and their difference is 1/2. As a consequence, the major contribution to the phonon thermal EMF in a zero magnetic field is made by the first term present in (8), and the equation for $\alpha_{11}(0)$ can be written in the form

$$\begin{aligned}\alpha_{11}(0) &= -\frac{1}{e}\frac{3\hbar(1 + 2\zeta/\varepsilon_g)}{8\rho\sqrt{[2\zeta(1 + \zeta/\varepsilon_g)]^3}} \\ &\times \sum_{s=1}^{3}\int\frac{d^3q}{(2\pi)^2}\frac{\tau_{ph}(\Lambda_{ik}q_i l_k^{(s)})^2}{\sqrt{q_1^2/m_1 + q_2^2/m_2 + q_3^2/m_3}} \\ &\times\left[2\left(q_1\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_1}\frac{\partial N_0}{\partial T} - \frac{b_1}{a_1}\frac{q_1^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right) + \frac{2m_1}{m_4}\frac{m_4\cos^2\varphi}{m_2}\right. \qquad (17) \\ &\times\left(q_2\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_2}\frac{\partial N_0}{\partial T} - \frac{b_2}{a_2}\frac{q_2^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(2)}}\right) + \frac{2m_1}{m_4}\frac{m_4\sin^2\varphi}{m_3} \\ &\left.\times\left(q_3\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_3}\frac{\partial N_0}{\partial T} - \frac{b_3}{a_3}\frac{q_3^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)\right].\end{aligned}$$

In (8), the coefficients $1 \pm \eta/(\delta + \Omega^2\tau^2)$ of three terms in square brackets in the integrand all equal one in a classically strong magnetic field. As a result, the phonon thermal EMF of the alloys in a classically strong field with $\mathbf{H} \parallel C_3$ is

$$\begin{aligned}\alpha_\infty &= -\frac{1}{e}\frac{3\hbar(1 + 2\zeta/\varepsilon_g)}{8\rho(2\zeta)^{3/2}(1 + \zeta/\varepsilon_g)^{3/2}} \\ &\times \sum_{s=1}^{3}\int\frac{d^3q}{(2\pi)^2}\frac{\tau_{ph}(\Lambda_{ik}q_i l_k^{(s)})^2}{\sqrt{q_1^2/m_1 + q_2^2/m_2 + q_3^2/m_3}} \\ &\times\left[\left(q_1\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_1}\frac{\partial N_0}{\partial T} - \frac{b_1}{a_1}\frac{q_1^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)\right. \qquad (18) \\ &+ \frac{m_4\cos^2\varphi}{m_2}\left(q_2\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_2}\frac{\partial N_0}{\partial T} - \frac{b_2}{a_2}\frac{q_2^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right) \\ &\left.+ \frac{m_4\sin^2\varphi}{m_3}\left(q_3\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_3}\frac{\partial N_0}{\partial T} - \frac{b_3}{a_3}\frac{q_3^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)\right].\end{aligned}$$

In an isotropic degenerate semiconductor, limitation (16) for the region of phonon–electron interactions determines the maximum phonon momentum $\sqrt{8m\zeta}$ in electron–phonon interaction. This corresponds to the electronic Debye temperature $\Theta_e = s\sqrt{8m\zeta}/k$, where s is the velocity of sound. An anisotropic semiconducting alloy can be treated in terms of only two different electronic Debye temperatures with the effective masses $m_1 \approx m_3$ and m_2. This gives $\Theta_{1e} \approx 1$ K and $\Theta_{2e} \approx 19$ K. Nevertheless, the use of some average electronic Debye temperature to estimate integrals (17) and (18) is, we believe, incorrect. The calculations will therefore be performed differently. We ignore the small term with the large effective mass m_2 in condition (16). The restriction is then only imposed on the wave vectors q_1 and q_3. The integration in q_2 is performed over an unbounded region of phonon wave vector values, but, because of the Planck distribution function for phonons, the integral is determined by the heat wave vector $kT/\hbar s$. This value is larger than the maximum value for $q_1 = \sqrt{8m_1\zeta}/\hbar$. For this reason, everywhere in the integral except the square root in the denominator we ignore q_1 and q_3 in comparison with q_2. This integration procedure takes into account that the temperature is within the interval between Θ_{1e} and Θ_{2e}. For phonon relaxation times in scattering of phonons by impurities, phonons, and electrons, the following equations for the mean heat momenta are used:

$$1/\tau_{ph-im} = d_1T^4, \quad 1/\tau_{ph-ph} = d_2T^5, \quad 1/\tau_{ph-e} = d_3T, \tag{19}$$

where d_i are constant coefficients.

Contributions to electron relaxation come from electron scattering by impurities and electron–phonon scattering. The electron relaxation times on both impurities and phonons in Bi–Sb alloys were determined in [8] for temperatures above Θ_{2e}. In semiconducting alloys, only electron–phonon scattering depends on temperature, but, as mentioned, this scattering is insignificant at low temperatures compared with electron scattering by impurities. For this reason, electron–phonon scattering can be ignored in analyzing the temperature dependence of the phonon thermal EMF. Consider the temperature dependence of $\alpha_{11}(0)$ [Eq. (17)] using the temperature dependence of phonon relaxation times given by (19) and the temperature dependence of the values present in the integrand in (17), namely, $\partial N_0/\partial\omega_{\mathbf{q}} \propto T^{-1}$ and $q \propto T$ for thermal phonons. The temperature dependence of the phonon thermal EMF to the right of the maximum is determined by phonon–impurity and phonon–phonon scattering; it is given by the equation

$$\alpha_{11}(0) = \frac{1}{c_1T^2 + c_2T^3}. \tag{20}$$

The thermal EMF to the left of the maximum is determined by phonon–electron scattering,

$$\alpha_{11}(0) = c_3T^1. \tag{21}$$

The intersection of these curves forms the temperature maximum.

The integral of the first term in Eq. (18) for α_∞ is two times smaller than that in Eq. (19) for $\alpha_{11}(0)$, but the integral of the third term with the same temperature dependence is added. In addition, there is the integral of the second term, which differs from the first term by the replacement of the wave vector q_1 with the heat vector q_2. This increases α_∞ and changes its temperature dependence. The temperature dependence of the phonon thermal EMF $\alpha_\infty(T)$ in a classically strong magnetic field is determined by the predominant second term in (18); to the right of the maximum, this dependence is described by the equation

$$\alpha_\infty = \frac{1}{c_4T + c_5T^2} + \frac{1}{c_6 + c_7T}. \tag{22}$$

Equation (18) for the $\alpha_\infty(T)$ dependence contains two terms, because two terms in the integrand give different temperature dependences. The temperature dependence of the phonon thermal EMF α_∞ to the left of its maximum is also different than that for $\alpha_{11}(0)$, namely,

$$\alpha_\infty = c_8T^2 + c_9T^3. \tag{23}$$

A comparison of the temperature dependences of $\alpha_{11}(0)$ [Eqs. (20), (21)] and α_∞ [Eqs. (22), (23)] shows that the temperature maximum of α_∞ is shifted to lower temperatures compared with $\alpha_{11}(0)$.

A comparison of Eqs. (17) for $\alpha_{11}(0)$ and (18) for α_∞ provides evidence that the thermal EMF value in classically strong magnetic fields can be increased. To quantitatively estimate this increase requires the integrals in (17) and (18) for $\alpha_{11}(0)$ and α_∞ to be calculated numerically. The transition from temperature dependence (20) for $\alpha_{11}(0)$ to dependence (22) for α_∞ and the shift of the temperature maximum to lower temperatures are in qualitative agreement with the experimental temperature dependence of the phonon thermal EMF observed in classically strong magnetic fields (Fig. 1, inset).

In a weak magnetic field ($\Omega\tau \ll 1$), the diffusion transverse Nernst–Ettingshausen coefficient for a degenerate semiconductor does not depend on the magnetic field and its sign is determined by the predominant (unique) electron scattering mechanism [2, 7],

$$Q = \left(r - \frac{1}{2} - \gamma\right)\frac{\pi^2 ku}{3ec}\frac{1 + 2\zeta/\varepsilon_g}{1 + \zeta/\varepsilon_g}. \tag{24}$$

Here, r is the parameter of the electron scattering mechanism ($r = 0$ for scattering by acoustic phonons or point impurities and $r = 2$ for scattering by ionized impurities), γ is the parameter characterizing the degree to which the band is nonparabolic, and u is the mobility of electrons. The temperature-induced change in the sign of the transverse Nernst–Ettingshausen coefficient in semiconductors with a degenerate electron gas is usually related to the theory of the diffusion contribution to this coefficient, according to which electron scattering by ionized impurities predominates at low temperatures, and scattering by acoustic phonons, at high temperatures [2, 7].

Note that the contributions of all scattering mechanisms to the diffusion transverse Nernst–Ettingshausen coefficient in n-Bi–Sb semiconducting alloys were analyzed in [8]. It was shown that the alloy mechanism of scattering played an important role. The sign of the diffusion transverse Nernst–Ettingshausen coefficient is negative for the alloy mechanism of electron scattering and for scattering by acoustic phonons, as observed in the alloys at $T > 10$ K. Temperature decrease has no effect on the predominance of the alloy scattering mechanism in Bi–Sb alloys and, therefore, cannot cause a change in the sign of the diffusion transverse Nernst–Ettingshausen coefficient. For this reason, the change in the sign of the transverse Nernst–Ettingshausen coefficient at $T < 10$ K can only be related to electron–phonon drag, which was observed for the thermal EMF of Bi–Sb alloys and has been discussed above. At $T < 10$ K, the phonon thermal EMF component prevails over the diffusion component.

Note that the transverse Nernst–Ettingshausen coefficient caused by electron–phonon drag in an isotropic semiconductor with a degenerate electron gas is zero [2]. The corresponding physical picture is clear without calculations. As the additional force in the kinetic equation for electrons in the right-hand side of (5) is proportional to $[(\mathbf{p}\cdot\nabla T)/m][\partial f_0/\partial\varepsilon_{\mathbf{p}}]$ by virtue of the problem symmetry, the energy dependence of the proportionality factor can be ignored, because in all terms, the energy equals the Fermi energy. This force is fully analogous to the $[(\mathbf{p}\cdot\mathbf{E})/m][\partial f_0/\partial\varepsilon_{\mathbf{p}}]$ force in the kinetic equation for calculating the response of electrons to electric field $\mathbf{E}$. It follows that the thermoelectric tensor, which relates thermal current to temperature gradient, is proportional to the conductivity tensor. In a strong magnetic field, both tensors are essentially nondiagonal. However, multiplying the thermoelectric tensor and the resistance tensor, we obtain a diagonal unit tensor. The zero off-diagonal components of the thermoelectric tensor are indicative of the equality to zero of the transverse Nernst–Ettingshausen coefficient.

Calculations of the phonon transverse Nernst–Ettingshausen coefficient in a semiconductor with an anisotropic electron energy spectrum at $\mathbf{H} \parallel C_3$ gives an equation which is in many respects similar to that for the phonon thermal EMF, namely,

$$Q_{12,3}(H) = \frac{3\hbar(m_4 - m_1)\tau_e}{16\rho c\sqrt{[2\zeta(1+\zeta/\varepsilon_g)]^3}\,m_4 m_1(\delta + \Omega^2\tau^2)}$$
$$\times \sum_{s=1}^{3}\int\frac{d^3q}{(2\pi)^2}\frac{\tau_{ph}(\Lambda_{ik}q_i l_k^{(s)})^2}{\sqrt{q_1^2/m_1 + q_2^2/m_2 + q_3^2/m_3}}$$
$$\times\Bigg[-\left(q_1\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_1}\frac{\partial N_0}{\partial T} - \frac{b_1}{a_1}\frac{q_1^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right) \tag{25}$$
$$+\frac{m_4\cos^2\varphi}{m_2}\left(q_2\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_2}\frac{\partial N_0}{\partial T} - \frac{b_2}{a_2}\frac{q_2^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)$$
$$+\frac{m_4\sin^2\varphi}{m_3}\left(q_3\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_3}\frac{\partial N_0}{\partial T} - \frac{b_3}{a_3}\frac{q_3^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)\Bigg].$$

In a low magnetic field, the equation for the phonon transverse Nernst–Ettingshausen coefficient takes the form

$$Q_{12,3}(H \longrightarrow 0) = \frac{3\hbar(m_4 - m_1)\tau_e}{4\rho c\sqrt{[2\zeta(1+\zeta/\varepsilon_g)]^3}(m_1 + m_4)^2}$$
$$\times \sum_{s=1}^{3}\int\frac{d^3q}{(2\pi)^2}\frac{\tau_{ph}(\Lambda_{ik}q_i l_k^{(s)})^2}{\sqrt{q_1^2/m_1 + q_2^2/m_2 + q_3^2/m_3}}$$
$$\times\Bigg[-\left(q_1\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_1}\frac{\partial N_0}{\partial T} - \frac{b_1}{a_1}\frac{q_1^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right) \tag{26}$$
$$+\frac{m_4\cos^2\varphi}{m_2}\left(q_2\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_2}\frac{\partial N_0}{\partial T} - \frac{b_2}{a_2}\frac{q_2^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)$$
$$+\frac{m_4\sin^2\varphi}{m_3}\left(q_3\frac{\partial\omega_{\mathbf{q}}^{(s)}}{\partial q_3}\frac{\partial N_0}{\partial T} - \frac{b_3}{a_3}\frac{q_3^2}{\tau_{ph-ph}}\frac{\partial N_0}{\partial\omega_{\mathbf{q}}^{(s)}}\right)\Bigg].$$

The equations for a_i, b_i, and τ_{ph} are given in (11), (12), and (14). The phonon transverse Nernst–Ettingshausen coefficient is proportional to the difference of m_4 and m_1, which is zero in semiconductors with an isotropic electron energy spectrum. This result conforms to the theory of semiconductors with isotropic electron energy spectra.

The integrals that determine the phonon transverse Nernst–Ettingshausen coefficient are the same as in Eq. (8) for the phonon thermal EMF but have different signs. It can be shown that, if α_∞ increases compared with $\alpha_{11}(0)$, then (25) has a positive value. The phonon transverse Nernst–Ettingshausen coefficient is then positive, and its value, on the order of $|\alpha_{11}(H)|/H$, ensures its predominance over the negative diffusion coefficient.

We reach the conclusion that the change in the sign of the transverse Nernst–Ettingshausen coefficient in semiconducting Bi–Sb alloys at low temperatures shown in Figs. 4–6 can be explained by the predominance of the positive phonon contribution to $Q_{12,3}$ over the negative diffusion contribution.

Also note that, according to (3) and (25), the diffusion and phonon transverse Nernst–Ettingshausen coefficients similarly depend on the magnetic field. The absolute value of their ratio therefore does not depend on the magnetic field and is a function of temperature alone. For this reason, the temperature at which the transverse Nernst–Ettingshausen coefficient changes sign should not depend on the magnetic field either by virtue of the same equations. This property of the model is at variance with experiment. Experimentally (Figs. 4–6), we observe a magnetic field dependence of the temperature at which the sign of $Q_{12,3}$ changes. An increase in the magnetic field increases the sign reversal temperature. It can be suggested that the shortcoming of the model is the introduction of the electron relaxation time. We need a theory that would more accurately take into account the action of the collision operator on the nonequilibrium electron distribution function and would be capable of explaining the magnetic field dependence of the temperature at which $Q_{12,3}$ changes sign.

5. CONCLUSIONS

Note that the reason for the strong dependence of the phonon thermal EMF on transverse magnetic field in Bi–Sb semiconducting alloys is the anisotropy of the electron energy spectrum. An increase in the phonon thermal EMF in a transverse magnetic field causes an increase in the thermoelectric efficiency $Z = \alpha^2/\kappa\rho$ of the semiconducting alloys, where κ is the thermal conductivity and ρ is the resistivity. For instance, in the $Bi_{0.93}Sb_{0.07}$ alloy, Z increases by a factor of 1.5 in a magnetic field of $H = 0.3$ kOe (as compared to the case of $H = 0$) at $T = 4.2$ K. Electron energy spectrum anisotropy in semiconducting alloys results in nonzero phonon transverse Nernst–Ettingshausen coefficients. The reason for the change in the sign of the transverse Nernst–Ettingshausen coefficient from negative to positive at $T < 10$ K in Bi–Sb semiconducting alloys is the presence of a substantial phonon transverse Nernst–Ettingshausen coefficient rather than a change in the mechanism of electron scattering in passing from high to low temperatures.

REFERENCES

1. L. É. Gurevich, Zh. Éksp. Teor. Fiz. **16**, 193 (1946); Zh. Éksp. Teor. Fiz. **16**, 416 (1946).
2. B. M. Askerov, *Electronic Transport Phenomena in Semiconductors* (Nauka, Moscow, 1985).
3. A. I. Ansel’m, *Introduction to the Theory of Semiconductors* (Nauka, Moscow, 1978).
4. F. J. Blatt, P. A. Schroeder, C. L. Foiles, and P. Greig, *Thermoelectric Power of Metals* (Plenum, New York, 1976; Metallurgiya, Moscow, 1980).
5. P. S. Zyryanov and M. I. Klinger, *Quantum Theory of Electron Transport Phenomena in Crystalline Semiconductors* (Nauka, Moscow, 1976).
6. I. Ya. Korenblit, Fiz. Tekh. Poluprovodn. (Leningrad) **2**, 1425 (1968) [Sov. Phys. Semicond. **2**, 1185 (1968)].
7. V. D. Kagan, N. A. Red’ko, N. A. Rodionov, and V. I. Pol’shin, Fiz. Tverd. Tela (St. Petersburg) **42**, 1376 (2000) [Phys. Solid State **42**, 1414 (2000)].
8. V. D. Kagan, N. A. Red’ko, N. A. Rodionov, and V. I. Pol’shin, Zh. Éksp. Teor. Fiz. **122**, 377 (2002) [JETP **95**, 325 (2002)].
9. V. A. Kozlov and É. L. Nagaev, Pis’ma Zh. Éksp. Teor. Fiz. **13**, 639 (1971) [JETP Lett. **13**, 455 (1971)]; V. A. Kozlov, N. S. Lidorenko, and É. L. Nagaev, Fiz. Tverd. Tela (Leningrad) **15**, 1458 (1973) [Sov. Phys. Solid State **15**, 856 (1973)].

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 127–137.
Translated from Zhurnal Éksperimental’noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 142–153.
Original Russian Text Copyright © 2003 by Dodin, Zharov.

SOLIDS
Electronic Properties

Terahertz Radiation of Bloch Oscillators Excited by an Electromagnetic Field in Lateral Semiconductor Superlattices

E. P. Dodin* and A. A. Zharov

Institute of Physics of Microstructures, Russian Academy of Sciences, Nizhni Novgorod, 603950 Russia
**e-mail: dodin@ipm.sci-nnov.ru*

Received December 23, 2002

Abstract—The action of a strong high-frequency electromagnetic field on a lateral semiconductor superlattice is considered based on the quasi-classical electron transport theory in the self-consistent wave formulation. The theory predicts that a lateral superlattice can emit terahertz radiation wave trains, which are associated with periodic excitation of Bloch oscillations in the superlattice arising because of the development of transient processes in it in a variable self-consistent electric field. The conditions necessary for observing Bloch oscillator radiation were found. The spectral composition of radiation transmitted through the superlattice and the energy efficiency of frequency multiplication related to Bloch oscillator excitation were calculated. *© 2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

According to Bloch and Zener [1, 2], electrons in a periodic potential oscillate in real space in static electric fields. The frequency of oscillations is proportional to a biased electric field, $\omega_B \sim eE_0d/\hbar$, where E_0 is the applied electric field strength, e is the charge of the electron, and d is the potential period. The nature of these oscillations, known as Bloch oscillations, is determined by limited electron energies in the energy band and is related to Bragg reflections from Brillouin zone boundaries. The Bloch oscillations manifest themselves under the conditions of a quasi-ballistic electron flight through the Brillouin zone, in other words, when the inverse mean free time between collisions is short compared with the Bloch frequency ω_B. In bulk crystals, this condition cannot be satisfied because of the small value of d = 0.1–0.3 nm, which prevents the observation (and excitation) of Bloch oscillations in them.

Real prospects for observing Bloch oscillations (and, simultaneously, hope for constructing a tunable generator of terahertz frequencies) arose with the appearance of semiconductor superlattices, that is, artificially grown heterostructures with chemical compositions periodically changing on a nanometer scale [3]. The presence of an additional periodic potential determined by sample chemical composition changes splits the energy spectrum of electrons into a system of allowed and forbidden minibands with a characteristic width of $\Delta \sim$ 10–100 meV [4, 5].

The necessity of achieving a high mobility of charge carriers in the conduction miniband for providing ballistic transport in the Brillouin zone imposes very severe requirements on the perfection of semiconductor superlattice structures. Lateral superlattices based on AlGaAs/GaAs appear to be most promising in this respect. In these superlattices (which are ordered structures on the surface of a dielectric substrate, such as systems of coupled quantum dots or wires), an additional potential is created for the two-dimensional electron gas localized close to the surface of the semiconducting substrate, and the spectrum of minibands is formed when electrons move over the surface of their localization. A noticeable increase in the mobility of carriers can be achieved in lateral superstructures by creating superstructures whose geometry allows electron scattering by optical phonons to be essentially suppressed between minibands and within them by appropriately selecting the widths of the first allowed and forbidden minibands [6, 7].

The specified and expected advances in technologies have actually brought lateral superstructures to the fore as candidates for creating tunable terahertz radiation sources. Several fundamental obstacles, however, crop up on the way to designing such sources. First, even under the conditions of rare collisions, macroscopic current oscillations at the Bloch frequency relax; as a result, a stationary current–voltage characteristic is established [8, 9]. Secondly, this current–voltage characteristic possesses negative differential conductivity in the region of electric fields corresponding to Bloch frequencies exceeding the frequency of collisions. The presence of a decreasing current–voltage characteristic portion results in the low-frequency instability of a uniform quasi-neutral spatial charge distribution. This instability splits the initially uniform electric field in strongly doped samples into separate domains, whose

1063-7761/03/9701-0127$24.00 © 2003 MAIK “Nauka/Interperiodica”

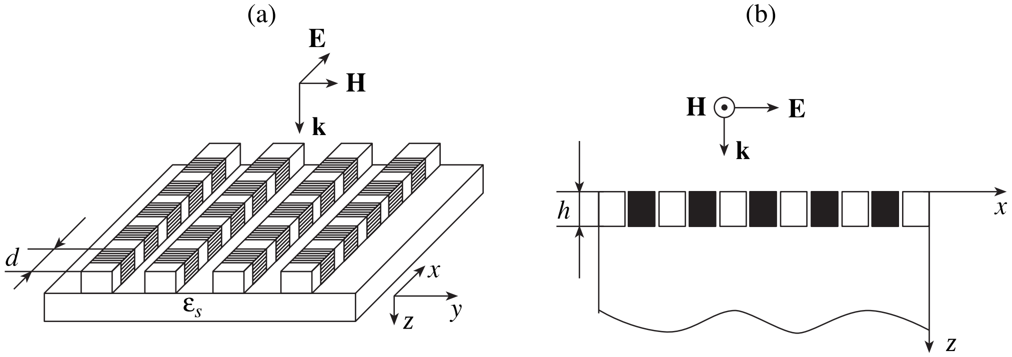


Fig. 1. Geometry of the problem: (a) an example of a GaAs/AlGaAs lateral superlattice (a quasi-one-dimensional system of coupled quantum dots) and (b) scheme of problem statement.

spatial dimensions may be commensurate with the superlattice period.

The specified obstacles on the way to achieving Bloch generation can be overcome using transient processes that are accompanied by Bloch oscillation excitations, namely, by switching a lateral superlattice, which is in a bi- or multistable state, from the screening regime into the regime of self-induced transparency [10] under the action of an electromagnetic pumping wave. According to [11, 12], lateral superlattice multistability arises during the interaction of a strong electromagnetic field with the superlattice and actually means the appearance of a non-single-valued relation between the strength of the internal self-consistent field in the superlattice and the field incident on the structure from the outside. The physical reason for the multistability of lateral superlattices is dynamic localization of miniband electrons [13, 14] in a self-consistent electromagnetic field [11].

The central idea of this work is to utilize the phenomenon of directly switching a lateral superlattice by a pumping wave for the excitation of Bloch oscillations cyclically repeated during each incident field period. It can be expected that the frequency of Bloch oscillations excited in the transient process, which is determined by the self-consistent field value in the lateral superlattice, will be substantially higher than the incident wave frequency. In essence, this implies the automodulation regime of the interaction of radiation with a superlattice accompanied by a substantial enrichment of the spectrum of radiation as a result of coherent excitation of a system of Bloch oscillators. The purpose of this work was to elucidate the conditions of the arising of the automodulation regime and to determine the spectral characteristics of radiation from a lateral superlattice and the energy effectiveness of the Bloch oscillator.

2. PROBLEM STATEMENT AND MAIN EQUATIONS

Our initial physical model will be a lateral superlattice of thickness h situated on a dielectric substrate with permittivity ε_s (further, we ignore substrate dispersion and absorption; that is, we assume that $\varepsilon_s(\omega)$ = const and $\mathrm{Im}\varepsilon_s = 0$). Let a linearly polarized plane electromagnetic wave be incident on the superlattice in such a way that the electric field vector is oriented along its growth direction (the geometry of the problem is schematically shown in Fig. 1). The transport of electrons in the superlattice will be described by the well-known balance equations obtained in the one-miniband quasi-classical approximation [9, 15],

$$\begin{aligned}\frac{dV}{dt} &= \frac{e}{m(W)}E - \nu_V V,\\ \frac{dW}{dt} &= eEV - \nu_W(W - W_T),\end{aligned} \tag{1}$$

where V and W are the mean (hydrodynamic) electron velocity and energy, respectively; E is the self-consistent field intensity in the superlattice; ν_W is the frequency of inelastic collisions (energy relaxation frequency); $\nu_V = \nu_W + \nu_{el}$ is the frequency of velocity relaxation (ν_{el} is the frequency of elastic collisions); $m(W) = \mu_0/(1 - 2W/\Delta)$; $m_0 = 2\hbar^2/\Delta d^2$ is the effective mass of the electron at the bottom of the miniband; $W_T = \Delta(1 - \mu_0)/2$ is the mean thermal energy of the electron in the absence of an electric field; $m_0 = I_1(\Delta/2k_BT)/I_0(\Delta/2k_BT)$; $I_{0,\,1}(x)$ are the modified Bessel functions; Δ is the miniband width; T is the temperature; and k_B is the Boltzmann constant. The first equation in (1) is the equation of electron gas motion in a miniband with a sinusoidal dispersion law, and the second equation is the law of conservation of energy. The energy dependence of the mass of the electron is described by the dispersion law in the miniband and is physically determined by Bragg reflections from Brillouin zone boundaries. Further, we assume that the concentration of electrons in the miniband is not very low and the polarizability of the lateral superlattice is fully determined by the conduction current, whose density is proportional to the hydrodynamic velocity V,

$$j = en_eV, \tag{2}$$

where n_e is the concentration of electrons recalculated to the whole surface of the structure (see Fig. 1a). It follows that Eqs. (1) can be considered material equations for an electromagnetic field whose polarization coincides with the superlattice axis.

The electromagnetic field in the entire space will be described by the Maxwell equations

$$\nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}, \quad \nabla \times \mathbf{B} = \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi}{c}\mathbf{j} \tag{3}$$

(the medium is taken to be nonmagnetic, $\mathbf{H} = \mathbf{B}$). We also assume that the superlattice thickness is smaller than the wavelength in the medium. The simplest estimates show that such conditions are more than amply satisfied in real structures in the millimeter, submillimeter, and far IR wavelength ranges. Let the z axis of Cartesian coordinates coincide with the incident wave vector direction, and the x axis, with the axis of the superlattice, as is shown in Fig. 1b. Place the $z = 0$ point on the illuminated superlattice surface. The electric current density, $\mathbf{j} = \mathbf{e}_x j$, where $\mathbf{e}_x$ is the unit vector along x axis, takes on the following values along z axis:

$$j = 0, \quad z < 0$$

(vacuum),

$$j = en_eV, \quad 0 \le z \le h$$

(lateral superlattice), and

$$j = \frac{\varepsilon_s - 1}{4\pi}\frac{\partial E}{\partial t}, \quad z > h$$

(polarization current in the substrate).

In order to obtain equations that relate the incident, reflected, and transmitted waves to each other and these waves to the field within the superlattice, taking into account its small thickness and ignoring weak diffraction effects caused by structure discreteness in the y direction, we will treat the superlattice as an equivalent current screen of an infinitesimal thickness. The volume current in the superlattice will be replaced by the equivalent surface current,

$$j(0 \le z \le h) = en_eV \longrightarrow j_s\delta(z), \tag{4}$$

where $j_s = jh$. The $\mathbf{E} = \mathbf{e}_xE$ electric field satisfies the one-dimensional wave equation that directly follows from (3),

$$\frac{\varepsilon(z)}{c^2}\frac{\partial^2 E}{\partial t^2} - \frac{\partial^2 E}{\partial z^2} = -\frac{4\pi}{c^2}\frac{\partial j}{\partial t}, \tag{5}$$

where current density j is given by (4) and the permittivity is

$$\varepsilon(z) = \begin{cases} 1, & z < 0, \\ \varepsilon_s, & z > 0. \end{cases}$$

The boundary conditions for the electric and magnetic fields on the current screen lead to the following equations for E:

$$[E]_{z=0} = 0, \quad \left[\frac{\partial E}{\partial z}\right]_{z=0} = \frac{4\pi}{c^2}\frac{\partial j_s}{\partial t}, \tag{6}$$

according to which the electric field is continuous and the magnetic field experiences a jump on the surface current (square brackets denote value jumps). The incident E_i and reflected E_r electric fields and the electric field transmitted through the superlattice E_t will be described using first-order wave equations related to the well-known D'Alembert solution of the one-dimensional wave equation in a uniform medium without dispersion,

$$\frac{1}{c}\frac{\partial E_i}{\partial t} + \frac{\partial E_i}{\partial z} = 0, \tag{7a}$$

$$\frac{1}{c}\frac{\partial E_r}{\partial t} - \frac{\partial E_r}{\partial z} = 0, \tag{7b}$$

$$\frac{\sqrt{\varepsilon_s}}{c}\frac{\partial E_t}{\partial t} + \frac{\partial E_t}{\partial z} = 0. \tag{7c}$$

The wave fields contained in Eqs. (7) are subject to boundary conditions (6) at $z = 0$. Substituting $E_{i,r,t}$ into (6) and expressing spatial through time derivatives, we obtain

$$E_i + E_r = E_t, \tag{8}$$

$$-\frac{\sqrt{\varepsilon_s}}{c}\frac{\partial E_t}{\partial t} - \frac{1}{c}\frac{\partial E_r}{\partial t} + \frac{1}{c}\frac{\partial E_i}{\partial t} = \frac{4\pi}{c^2}\frac{\partial j_s}{\partial t}. \tag{9}$$

Equation (9) can be integrated in time. Taking into account the zero boundary conditions, we eventually have

$$-\sqrt{\varepsilon_s}E_t - E_r + E_i = \frac{4\pi}{c}j_s. \tag{10}$$

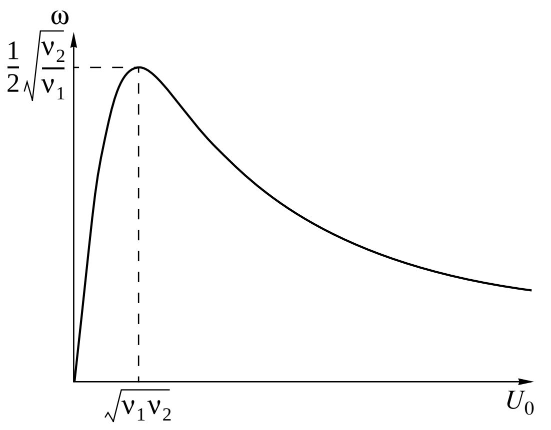


Fig. 2. Esaki–Tsu current–voltage characteristic as a stationary solution to (14).

Note that the electric field strength entering into (1) coincides with the E_t strength of the transmitted wave field. Index "t" will therefore be omitted henceforth. The incident field E_i is determined by an external source and is a set function of time. Excluding E_r from (8) and (9), we obtain a temporally local linear relation between j_s, E_i, and the self-consistent field E in the lateral superlattice,

$$E = \frac{1}{1+\sqrt{\varepsilon_s}}\left[2E_i(t) - \frac{4\pi}{c}j_s\right]. \tag{11}$$

Equation (11), together with electron transport equations (1) and (2), forms a closed nonlinear system of equations with respect to V, W, and E, which completely describes the interaction of a normal incident wave with a thin lateral superlattice. Let the field amplitude of the incident wave vary in time by the harmonic law,

$$E_i(t) = E_m \sin\omega t.$$

In these conditions, we can conveniently introduce the dimensionless variables

$$\tau = \omega t, \quad U = \frac{E}{E_c}, \quad w = \frac{V}{\mu_0 V_0},$$

$$E_c = \frac{\hbar\omega}{ed}, \quad V_0 = \frac{d\Delta}{2\hbar}, \quad \zeta = \frac{2(W-W_T)}{\Delta\mu_0}.$$

The corresponding substitutions in (1) and (11) give

$$\frac{dw}{d\tau} = U(1-\zeta) - \nu_1 w, \tag{12a}$$

$$\frac{d\zeta}{d\tau} = Uw - \nu_2\zeta, \tag{12b}$$

$$U_i(\tau) = U + \theta_0 w, \tag{12c}$$

where

$$\nu_1 = \frac{\nu_V}{\omega}, \quad \nu_2 = \frac{\nu_W}{\omega},$$

$$\theta_0 = \frac{\mu_0}{1+\sqrt{\varepsilon_s}}\frac{\omega_{pe}^2}{\omega^2}\frac{\omega}{c}h, \quad U_i(\tau) = \frac{2}{1+\sqrt{\varepsilon_s}}\frac{E_i(\tau)}{E_c},$$

and $\omega_{pe}^2 = 4\pi e^2 n_e/m_0$ is the electron plasma frequency at the bottom of the miniband. The θ_0 dimensionless parameter is one of the key parameters of the theory determining the dynamics of self-consistent fields in lateral superlattices. It is estimated at a value close to two for $T = 300$ K, $\Delta = 100$ meV, $\omega = 10^{12}$ s^{-1}, $h = 10^{-5}$ cm, $n_s = n_e h = 10^{13}$ cm^{-2}, $m_0 = 0.1m$ (m is the mass of the free electron), and $\varepsilon_s \approx 12.4$ (a GaAs substrate). It will be shown using (12) that electromagnetic radiation incident on a superlattice with a frequency lower than the characteristic collision frequency can cause excitation of Bloch oscillations with terahertz frequencies emitted by the system.

3. BISTABILITY AND TRANSIENT PROCESSES IN LATERAL SUPERLATTICES IN AN EXTERNAL ELECTROMAGNETIC FIELD

First, consider a lateral superlattice in a given stationary electric field. Let us substitute $U = U_0 = \text{const}$ into (12a) and (12b). These equations take into account both quasi-ballistic motion of electrons in a miniband and dissipation caused by scattering [9, 16]. In particular, at $\nu_{1,2} = 0$, the solution to (12a) and (12b) can be written as

$$\omega(\tau) = \sin(U_0\tau) = \sin\left(\frac{eE_0 d}{\hbar}t\right), \tag{13}$$

where U_0 is the dimensionless Bloch frequency. Equation (13) describes spatially localized oscillatory motion of charge carriers with the $\omega_B = eE_0 d/\hbar$ frequency in a constant electric field (Bloch oscillations). Generally, (12a) and (12b) are reduced to the damped harmonic oscillator equation for the velocity of electrons [9],

$$\frac{d^2w}{d\tau^2} + (\nu_1+\nu_2)\frac{dw}{d\tau} + (U_0^2 + \nu_1\nu_2)w = \nu_2 U_0, \tag{14}$$

which describes transient processes in the system. The stationary solution to (14) is the well-known Esaki–Tsu current–voltage characteristic

$$w_0 = \frac{\nu_2 U_0}{U_0^2 + \nu_1 \nu_2} \qquad (15)$$

shown in Fig. 2. The process of velocity relaxation to current–voltage characteristic (15) substantially depends on bias field U_0. It is virtually aperiodic at $U_0 < U_0^{(m)} = \sqrt{\nu_1 \nu_2}$ ($U_0^{(m)}$ corresponds to the current extremum of the current–voltage characteristic) and is accompanied by oscillations of w at the Bloch frequency at $U_0 > U_0^{(m)}$. The relaxation time is here independent of U_0 and determined exclusively by collision frequencies $\tau_{\text{rel}} \approx 2/(\nu_1 + \nu_2)$. It follows that, if we succeed in changing the superlattice state in a time of $\Delta\tau < \tau_{\text{rel}}$, there will be macroscopic current oscillations at the Bloch frequency at a constant field $U_0 \gg U_0^{(m)}$. This will result in an emitting electromagnetic wave train of duration τ_{rel}. In reality, taking such ultrafast actions on a superlattice is an exceedingly difficult task, because the corresponding times do not exceed several hundred femtoseconds. We will show that transient processes accompanied by the emission of electromagnetic waves at terahertz frequencies can be induced by high-power electromagnetic radiation interaction with a superlattice because of the bistability arising in the system.

Let the frequency of the electromagnetic wave incident on the superlattice be not too high, $\nu \geq 1$ (for simplicity, we ignore elastic collisions, $\nu_1 = \nu_2 = \nu$, which does not infringe the physical meaning). This means for typical lateral superlattices [12] that the pumping wave frequency should not exceed several tenths of a terahertz. The velocity of electrons in the quasi-stationary mode is then described by (15), into which we must substitute U (the self-consistent field in the superlattice) for U_0. The relation between U and the incident wave field is given by (12c). As a result, we obtain an approximate equation for the self-consistent field in the lateral superlattice, namely,

$$U_i(\tau) \approx U(\tau) + \frac{\nu\theta_0 U(\tau)}{U^2(\tau) + \nu^2}. \qquad (16)$$

The dependences of U on the current U_i value are shown in Fig. 3. At parameter θ_0 values larger than $\theta_{0c} = 8\nu$, the curves have regions where these dependences are not single-valued, which means the appearance of bistability in the system when the electric field strength of the incident wave falls in the interval

$$U_{c2} < U_i < U_{c1},$$

whose boundaries at $\theta_0 \gg \theta_{0c}$ are approximately determined by the asymptotic equation

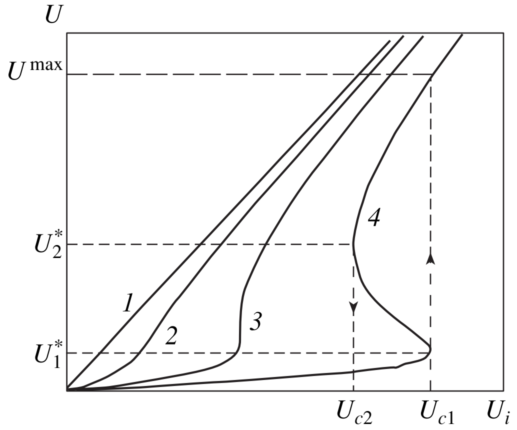


Fig. 3. Dependences of the electric field intensity in a lateral superlattice on the incident wave field intensity at $\nu = 1$ and various θ_0 values: (*1*) $\theta_0 = 0$, (*2*) $\theta_0 = 3 < \theta_{0c}$, (*3*) $\theta_0 = 8 = \theta_{0c}$, and (*4*) $\theta_0 = 20 > \theta_{0c}$; curve *4* characterizes superlattice bistability; $U_1^* \approx \nu$ and $U_2^* \approx \sqrt{\nu\theta_0}$.

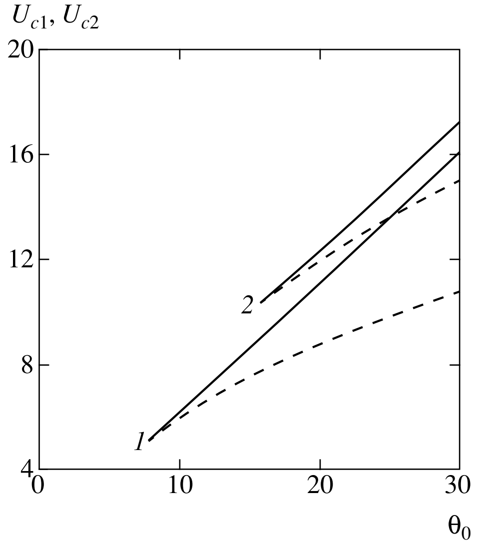


Fig. 4. Dependences of U_{c1} (solid lines) and U_{c2} (dashed lines) on θ_0: (*1*) $\nu = 1$ and (*2*) $\nu = 5$.

$$U_{c1} \approx \theta_0/2, \quad U_{c2} \approx \sqrt{\nu\theta_0}. \qquad (17)$$

The $U_{c1,\,c2}(\theta_0)$ dependences are shown in Fig. 4.

It follows that the dynamics of the interaction of incident radiation with the superlattice at $\theta_0 > \theta_{0c}$ substantially depends on the incident wave amplitude $U_i^{(m)} = E_m/E_c$ and the critical field U_{c1}. If $U_i^{(m)} < U_{c1}$, the field in the superlattice is fully described by the lower branch of curve *4* in Fig. 3. The self-consistent

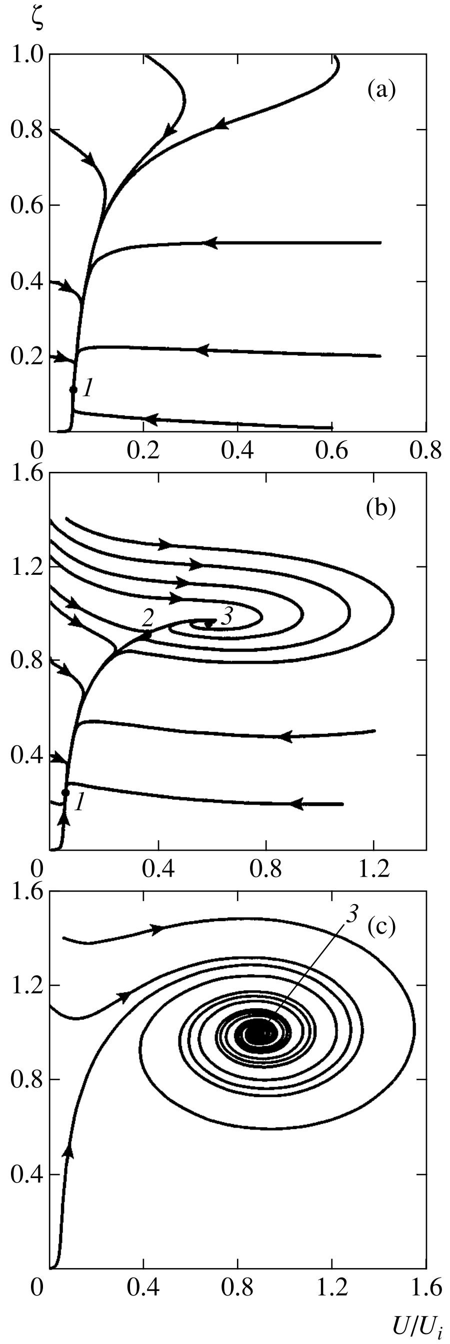


Fig. 5. Phase portraits of system (12) showing the self-consistent field dynamics in a lateral superlattice at $\theta_0 > \theta_{0c}$: (a) $U_i < U_{c2}$, (b) $U_{c2} < U_i < U_{c1}$, (c) $U_i > U_{c1}$; equilibrium states: (*1*) stable node, (*2*) saddle, and (*3*) stable focus.

field U then does not exceed the $U_1^* = \nu$ value, which is, in turn, evidence of the screening state of the superlattice ($U \ll U_i$). If $U_i^{(m)} > U_{c1}$, the system switches to a new state at time instants when $U_i(\tau) > U_{c1}$. This new state is described by the upper branch of curve *4* (Fig. 3). The new state, in which $U \approx U_i$, can be treated as self-induced structure transparency. Indeed, the lateral superlattice in a weak field behaves like an ordinary plasma film with a plasma frequency determined by the concentration of electrons and their effective mass. The electric current in the superlattice is then described by the quasi-linear increasing portion of the current–voltage characteristic (15). The transparency regime in a strong field arises because of the transition of the superlattice to a state with weak electric currents on the decreasing branch of the current–voltage characteristic. At the instants of time when $U_i(\tau)$ becomes equal to U_{c1}, the interaction process ceases to be quasi-stationary [in the sense that the instantaneous local relation between the incident field and the self-consistent field in the lattice described by (16) disappears]. The transition from the screening mode to the self-induced transparency mode shown in Fig. 3 by the arrow upward is qualitatively described by (14) and is accompanied by electric field and electron velocity oscillations at a frequency close to the Bloch frequency. The Bloch frequency is here determined by the strength of the self-consistent field in the superlattice, approximately equal to $\theta_0/2$ in the stationary state. It follows that this transient process is accompanied by the emission of a wave train with a frequency proportional to the concentration of electrons. The duration of such a wave train ("flash") is determined by the condition

$$T_{wt} \sim \min\{\tau_{\rm rel}, \Delta\tau\}, \qquad (18)$$

where $\Delta\tau$ is the time elapsed between the transitions of the system into the self-induced transparency state and back into the screening mode (the downward arrow in Fig. 3). Note that such radiation flashes occur if the negative differential conductivity of the superlattice in the transparent state, which splits the uniform electric field into separate strong field domains [17, 18], does not cause spatial charge instability to arise. This is possible if the corresponding instability increment Γ is smaller than the frequency of Bloch oscillations in the transient process, $\Gamma < \omega_B$.

The complete picture of the self-consistent electric field dynamics in a lateral superlattice can be obtained from Fig. 5, where the phase portraits of system (12) at fixed U_i values are depicted. Fig. 5a corresponds to the condition $U_i < U_{c2}$. The system possesses a single equilibrium state of the type of a stable node. If the pumping field varies fairly slowly, that is, $\tau_{\rm rel} < 1$, the superlattice is always in this equilibrium state (state *1* in Fig. 5a). The system adiabatically follows equilibrium state displacements on the phase plane when $U_i(\tau)$ changes. At the instants of time when $U_{c2} < U_i(\tau) < U_{c1}$, two additional equilibrium states appear on the phase plane, a saddle point (state *2* in Fig. 5b) and a stable focus (state *3* in Fig. 5b). This means the arising of bistability in the system (two of three equilibrium states are stable). Nevertheless, this change in the topology of the phase plane does not qualitatively influence the dynamics of the self-consistent field, because the superlattice is as previously in equilibrium state *1*. The situation dramatically changes when $U_i(\tau)$ attains a value of U_{c1}. The saddle point and the stable node coalesce and disappear, and the stable focus only remains on the phase plane (Fig. 5c). The corresponding phase trajectory coils in toward this equilibrium state and exhibits damped self-consistent field oscillations in the superlattice. The characteristic damping increment of these oscillations, which are associated with the transition of the superlattice into the transparent state, remains close to ν. At the same time, the oscillation frequency in dimensionless variables approximately equals $\Omega \approx \theta_0/2$, as is easy to see from (12). It can be shown that the Ω value obtained exactly corresponds to the Bloch frequency in an electric field whose strength equals the intensity of the self-consistent lateral superlattice field in equilibrium state *3* (Fig. 5c). We also see that the frequency of these oscillations (of course, for $\theta_0 > \theta_{0c}$) is $\theta_0/2$ times higher than the pumping field frequency. For the $\theta_0 = \theta_{0c}$ threshold value, we have the lowest possible Bloch frequency in the system, $\Omega_{\min} \approx 4\nu$, where $\nu \geq 1$. It follows that a high frequency wave train related to Bloch oscillations of electrons in the superlattice will be emitted at each pumping wave half-period. The electromagnetic field frequency in such a flash of Bloch oscillations is determined by the electron concentration in the superlattice (θ_0). The duration of the wave train is given by (18). The switching of the superlattice into the screening mode occurs when the self-consistent field intensity decreases after the focus (state *3* in Fig. 5b) and the saddle (state *2* in Fig. 5b) coalesce and is aperiodic in character because of the node type of the final equilibrium state (state *1* in Figs. 5a, 5b).

4. BLOCH OSCILLATOR EMISSION IN AN EXTERNAL ELECTROMAGNETIC FIELD

The transient process that occurs after the coalescence of equilibrium states *1* and *2* (Fig. 5) and is accompanied by electric field oscillations at the Bloch frequency (Fig. 5c) can be approximated by the equation

$$U(\tau) = U^{\max} - (U^{\max} - U_1^*)e^{-\nu\tau}\cos\left(\frac{\theta_0}{2}\tau\right), \qquad (19)$$

where $U^{\max}$ and U_1^* are shown in Fig. 3. The initial amplitude of oscillations of the self-consistent field in the superlattice, $U_B = U^{\max} - U_1^*$, depends on the position of the representation point on the phase plane after equilibrium states *1* and *2* (Fig. 5b) coalesce at $\tau = 0$. The time dependence of the self-consistent electric field in the superlattice described by (19) is exact in the neighborhood of the stable focus. The approximate equation (19) does not take into account the anharmonic character of the Bloch oscillator, which manifests itself when the representation point is fairly far from the specified equilibrium state. However, this anharmonicity cannot lead to a noticeable error in Bloch oscillator efficiency, because the corresponding nonlinear broadening of the generated spectral line is much smaller than the broadening caused by oscillation damping as a result of electron scattering. The slow $U^{\max}(\tau)$ time dependence in (19) related to the time dependence of the incident field can also be ignored because $U^{\max}(\tau)$ changes comparatively insignificantly during a pulse of excited Bloch oscillations. These assumptions allow us to calculate the energy contained in the excited high-frequency wave train. The maximum duration $\Delta\tau$ of a pulse of Bloch oscillations is determined by the time between mode switchings,

$$\Delta\tau = \pi - \arcsin\frac{U_{c1}}{U_i^{(m)}} - \arcsin\frac{U_{c2}}{U_i^{(m)}}, \qquad (20)$$

where $\arcsin x$ is the lower positive branch of the function. Equation (20) only holds if spatial charge instability in the equilibrium state of the stable focus type (caused by the negative superlattice differential conductivity) does not have sufficient time to develop. The absence of such Gunn instability presupposes that the inequality $\Gamma < \omega_B$, where Γ is the instability increment, is satisfied. The instability increment can roughly be estimated as [19]

$$\Gamma \approx 4\pi e n_e\left(-\frac{dV}{dE}\right)\frac{1}{\varepsilon_L} < \omega_B. \qquad (21)$$

Here, ε_L is the lattice dielectric constant of the superlattice. The superlattice differential mobility dV/dE should be negative near the equilibrium state. Substituting the obtained characteristic field into (21) yields the condition for the concentration of electrons in the superlattice that prevents the development of Gunn instability, namely,

$$\omega_{pe}^4 > 4\frac{\mu_0}{\varepsilon_L}\frac{\nu_w h}{c}\left(\frac{c}{h}\right)^4. \qquad (22)$$

Condition (22) imposes a fairly strict limitation on the concentration of electrons and the frequency of collisions. We assume (22) to be fulfilled. Equation (19) can then be used to calculate the electromagnetic energy emitted by the superlattice into the environment in the form of a pulse of damped Bloch oscillations during each incident pumping wave half-period. Dividing the

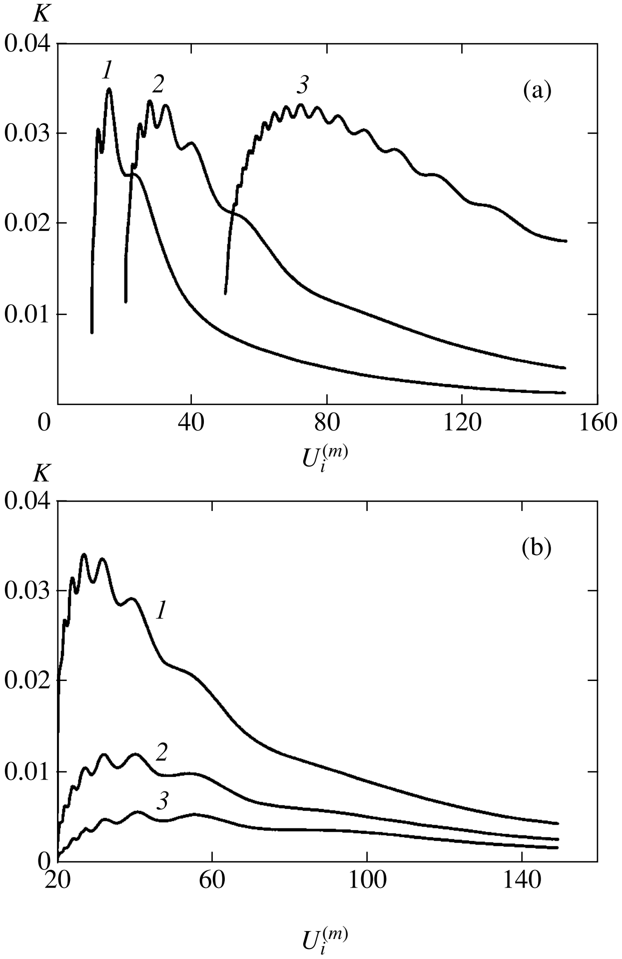


Fig. 6. Energy efficiency K of a Bloch oscillator in a lateral superlattice as a function of $U_i^{(m)}$: (a) $\nu = 1$, $\theta_0 = 20$ (*1*), 40 (*2*), and 100 (*3*) and (b) $\theta_0 = 40$, $\nu = 1$ (*1*), 1.5 (*2*), and 2 (*3*).

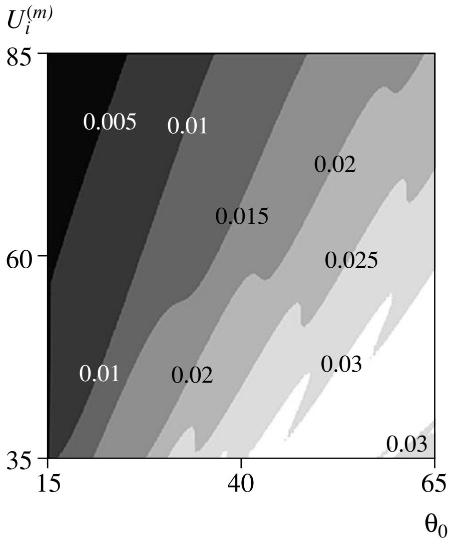


Fig. 7. Lines of constant K efficiencies on the $(U_i^{(m)}, \theta_0)$ parameter plane at $\nu = 1$.

obtained value by the introduced pumping wave energy, we obtain the energy efficiency of the excited Bloch oscillator, which, for $\theta_0 \gg \nu$, can be written in the form

$$K = \frac{1}{\pi U_i^{(m)2}}\left\{\frac{1}{2\nu}\left(e^{-2\nu\tau_1} - e^{-2\nu\tau_2}\right)\right.$$

$$- \frac{2\nu}{\theta_0^2}\left[e^{-2\nu\tau_1}\cos(\theta_0\tau_1) - e^{-2\nu\tau_2}\cos(\theta_0\tau_2)\right] \qquad (23)$$

$$\left. + \frac{1}{\theta_0}\left[e^{-2\nu\tau_1}\sin(\theta_0\tau_1) - e^{-2\nu\tau_2}\sin(\theta_0\tau_2)\right]\right\},$$

$$\tau_1 \approx \arcsin\frac{\theta_0}{2U_i^{(m)}}, \qquad \tau_2 \approx \pi - \arcsin\frac{2\sqrt{\nu\theta_0}}{U_i^{(m)}}.$$

The $K(U_i^{(m)})$ dependences constructed according to (23) at various fixed θ_0 values are shown in Fig. 6a. Note again that the $\theta_0/2$ parameter determines the dimensionless current frequency of emitted Bloch oscillation flashes. We see that the efficiency maximum is almost independent of θ_0 and amounts to several percent at $\nu = 1$. Only the line width grows as θ_0 increases. The $K(U_i^{(m)})$ dependences for various collision frequencies at a fixed $\theta_0 = 40$ value are shown in Fig. 6b. The efficiency maximum of the Bloch oscillator decreases as ν increases. The K = const level lines are shown in Fig. 7 on the $(U_i^{(m)}, \theta_0)$ parameter plane. We see that superlattices with fairly high θ_0 values are most suitable for frequency multiplication purposes. At the same time, a high θ_0 value leads to the necessity of using very strongly doped superlattices. As a consequence, the Bloch generation threshold increases. We should, however, mention the following. In the $\theta_0 \gg \theta_{0c}$ limit, the frequency of generated Bloch oscillations does not depend on the pumping wave frequency. Indeed,

$$\omega_B \approx \omega\frac{\theta_0}{2} = \frac{\mu_0}{2(1+\sqrt{\varepsilon_s})}\omega_{pe}^2\frac{h}{c}. \qquad (24)$$

We must decrease the incident wave frequency to increase θ_0. This, however, decreases efficiency K because of an increase in the relative collision frequency ν. The optimal energy efficiency of Bloch oscillations is achieved at $\nu \approx 1$, when the superlattice exists

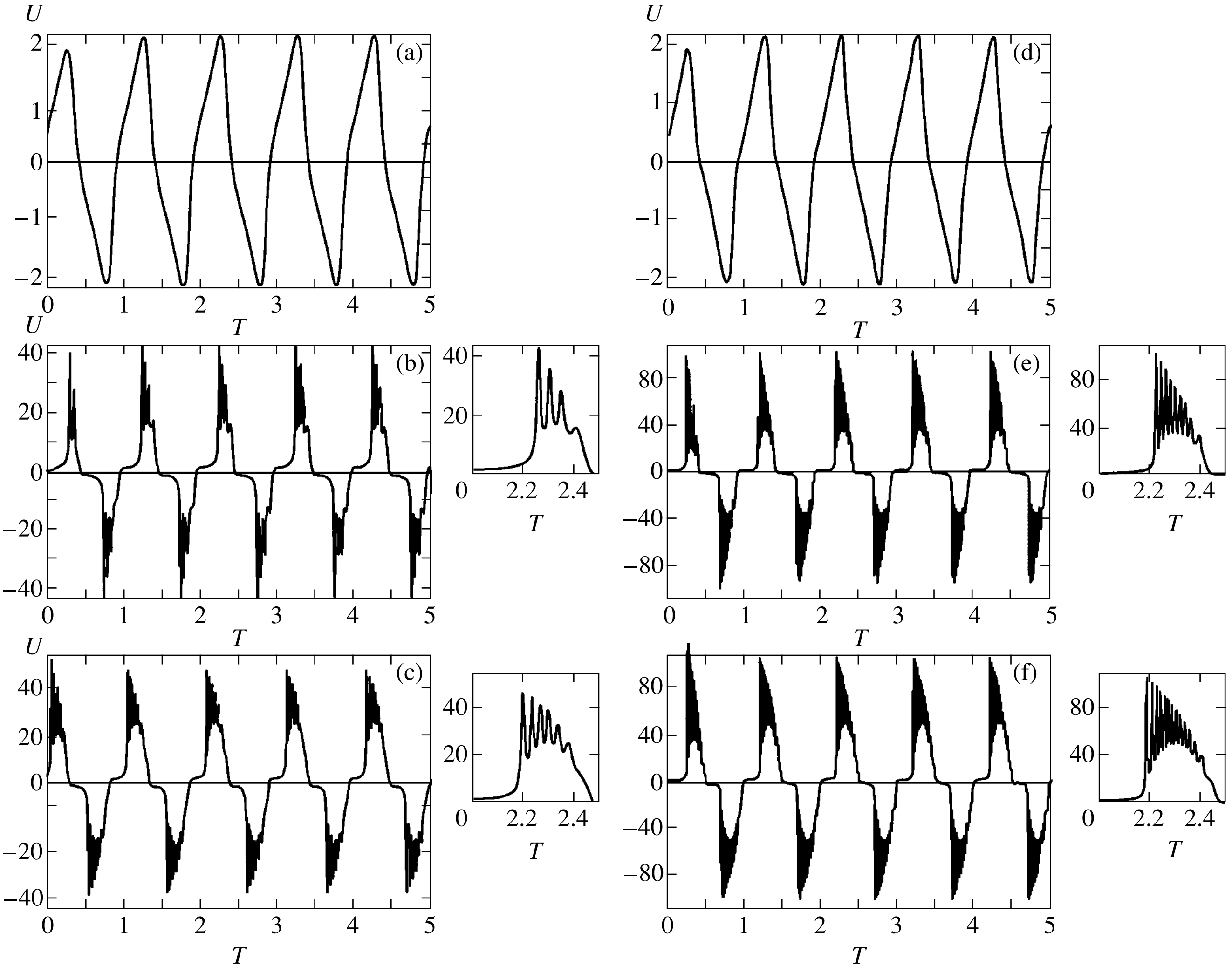


Fig. 8. Oscillograms of the electric fields of waves transmitted through a lateral superlattice at θ_0 = 40 [(a) $U_i^{(m)}$ = 27 ~ U_{c1}, (b) $U_i^{(m)}$ = 30 > U_{c1}, and (c) $U_i^{(m)}$ = 35 > U_{c1}] and θ_0 = 90 [(d) $U_i^{(m)}$ = 58, (e) $U_i^{(m)}$ = 63, and (f) $U_i^{(m)}$ = 70]. Shown by the side of Figs. 7b, 7c, 7e, and 7f are the fine structures of wave train half-periods related to Bloch oscillations in the superlattice.

in the transparent state during one incident wave half-period, $\Delta\tau$, for a time on the order of the Bloch oscillation damping time.

We performed numerical calculations by solving (12). The oscillograms of the electric field strength in a lateral superlattice calculated at various θ_0 and $U_i^{(m)}$ values are shown in Fig. 7. We see that, at $U_i^{(m)}$ below the threshold value (Figs. 7a, 7d), the electric field in the superlattice has a quasi-harmonic time dependence. This corresponds to the screening mode and a weak field in the superlattice. However, if the incident wave amplitude exceeds the threshold value, Bloch oscillations are excited in the transient process when the superlattice switches from the screening to the transparent mode. The spectral composition of these oscillations weakly depends on $U_i^{(m)}$ (Figs. 7b, 7c, 7e, 7f). The value by which the field exceeds the threshold value largely determines the duration of the wave train of Bloch oscillations, because the $\Delta\tau$ time interval increases as $U_i^{(m)}$ grows until $\Delta\tau$ becomes comparable to $1/\nu$. Virtually, the wave train frequency only depends on θ_0. The spectra of radiation transmitted through the lattice are shown in Fig. 9 for two $U_i^{(m)}$ and two θ_0 (see Fig. 7) values to show how the spectral composition of radiation emitted from the superlattice changes depending on $U_i^{(m)}$.

We estimated the effect of radiation of Bloch oscillators excited by an electromagnetic field in a superlattice to obtain the following particular result. At $\nu_W = \nu_V = 6 \times 10^{11}$ s^{-1}, the pumping wave frequency f =

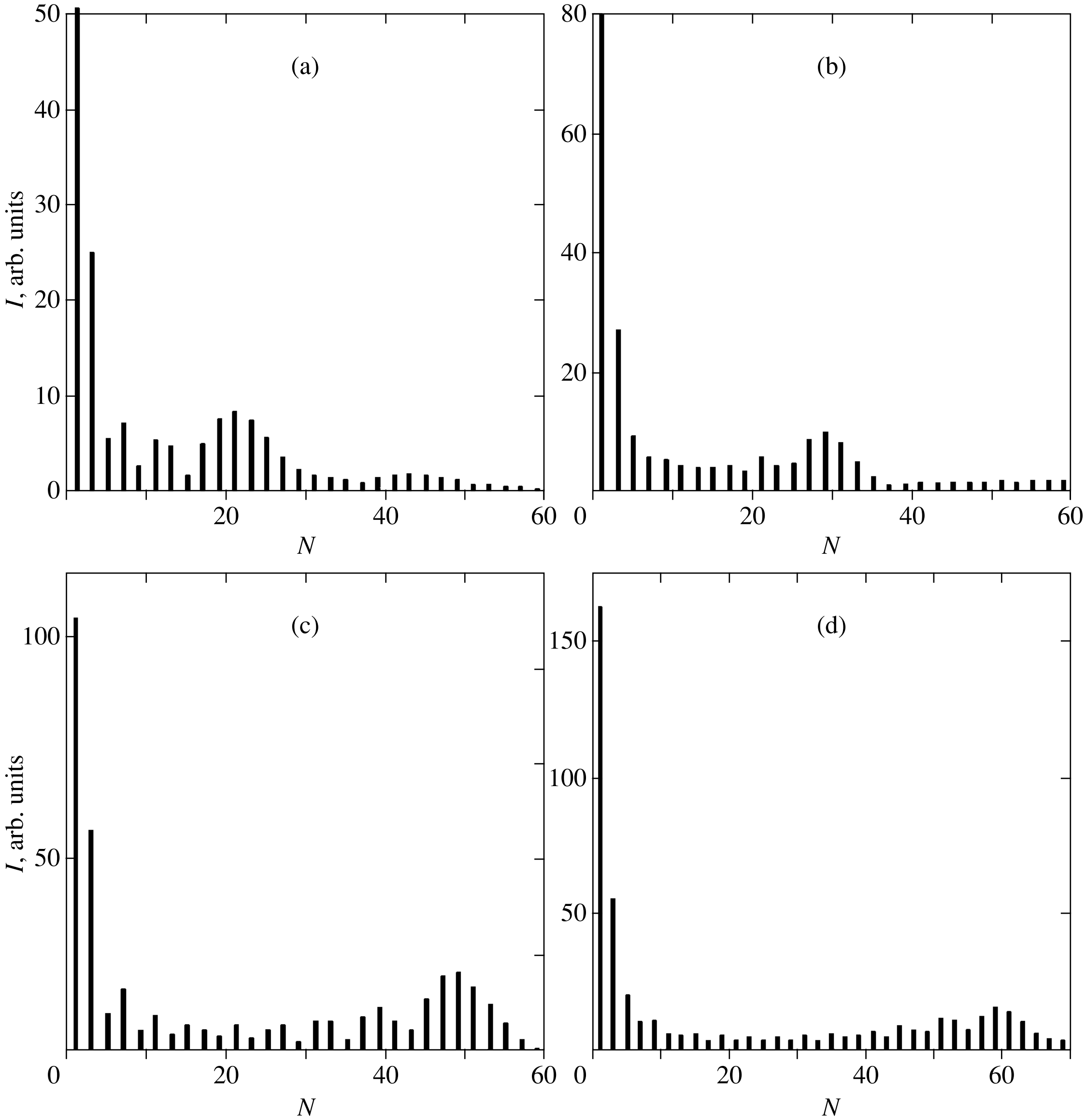


Fig. 9. Spectra of transmitted radiation corresponding to Figs. (a) 7b, (b) 7c, (c) 7e, and (d) 7f (*N* is the harmonic number).

0.1 THz, and the surface electron concentration $n_s = n_e/h = 10^{13}$ cm^{-2}, the critical energy flux density of incident radiation is $S_{in} \approx 5 \times 10^5$ W/cm^2, and the frequency of the emitted wave train of Bloch oscillations is $f_B \approx$ 3 THz.

5. CONCLUSIONS

We theoretically proved the possibility of emission of terahertz Bloch oscillator radiation excited by an electromagnetic field in a lateral semiconductor superlattice. This radiation is a sequence of wave trains resulting from flashes of Bloch oscillations emitted by the superlattice. The Bloch oscillations themselves arise in cyclically repeating transient processes. At a pumping frequency lower than or comparable to the effective electron collision frequency, Bloch oscillations are excited during every half-period of the electromagnetic wave incident on the superlattice. To summarize, a lateral superlattice can be considered a source of terahertz electromagnetic radiation of a new type. The superlattice parameters necessary for transient processes to be excited in it in an external electromagnetic field are quite achievable for modern technologies. In addition, some types of sources of electromagnetic waves in the millimeter and submillimeter wavelength ranges have characteristics suitable for directly observing the emission of Bloch oscillators from semiconductor lateral superlattices. We stress once more that the spatial charge instability is incapable of suppressing Bloch oscillations because its development takes a comparatively long time.

ACKNOWLEDGMENTS

The authors thank I.V. Shadrivov for significant technical assistance. This work was financially supported by the Russian Foundation for Basic Research (project nos. 02-02-16385 and 01-02-16449).

REFERENCES

1. F. Bloch, Z. Phys. **52**, 555 (1928).
2. C. Zener, Proc. R. Soc. London, Ser. A **145**, 523 (1934).
3. A. P. Silin, Usp. Fiz. Nauk **147**, 485 (1985) [Sov. Phys. Usp. **28**, 972 (1985)].
4. A. Ya. Shik, Fiz. Tekh. Poluprovodn. (Leningrad) **8**, 1841 (1974) [Sov. Phys. Semicond. **8**, 1195 (1974)].
5. L. Esaki and L. L. Chang, Thin Solid Films **36**, 285 (1976).
6. H. Sakaki, Jpn. J. Appl. Phys. **28**, 314 (1989).
7. N. Noguchi, J. P. Leburton, and H. Sakaki, Inst. Phys. Conf. Ser., No. 129, 299 (1993).
8. L. Esaki and R. Tsu, IBM J. Res. Dev. **14**, 61 (1970).
9. A. A. Ignatov, E. P. Dodin, and V. I. Shashkin, Mod. Phys. Lett. **5**, 1087 (1991).
10. A. A. Ignatov and Yu. A. Romanov, Phys. Status Solidi B **73**, 327 (1976).
11. E. P. Dodin, A. A. Zharov, and A. A. Ignatov, Zh. Éksp. Teor. Fiz. **114**, 2246 (1998) [JETP **87**, 1226 (1998)].
12. A. A. Zharov, E. P. Dodin, and A. S. Raspopin, Pis'ma Zh. Éksp. Teor. Fiz. **72**, 653 (2000) [JETP Lett. **72**, 453 (2000)].
13. D. H. Dunlap and V. M. Kenkre, Phys. Rev. B **34**, 3625 (1986).
14. A. A. Ignatov, E. Schomburg, J. Grenzer, *et al.*, Z. Phys. B **98**, 187 (1995); A. A. Ignatov, K. Renk, and E. P. Dodin, Phys. Rev. Lett. **70**, 1996 (1993).
15. X. L. Lei, N. J. Horing, and H. L. Cui, Phys. Rev. Lett. **66**, 3277 (1991).
16. A. A. Ignatov, E. P. Dodin, and A. A. Zharov, Phys. Low-Dimens. Semicond. Struct. **7**, 43 (1994).
17. A. A. Gogolin, Zh. Éksp. Teor. Fiz. **88**, 2063 (1985) [Sov. Phys. JETP **61**, 1220 (1985)].
18. B. W. Knight and G. A. Peterson, Phys. Rev. **155**, 393 (1967).
19. M. E. Levinshteĭn, Yu. K. Pozhela, and M. S. Shur, *Gann Effect* (Sovetskoe Radio, Moscow, 1975).

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 138–143.
Translated from Zhurnal Éksperimental’noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 154–160.
Original Russian Text Copyright © 2003 by Magarill, Makhmudian, Éntin.

SOLIDS
Electronic Properties

The Distribution of Equilibrium Magnetization Currents in Systems with Dimensional Quantization in a Finite Magnetic Field

L. I. Magarill*, M. M. Makhmudian, and M. V. Éntin
Institute of Semiconductor Physics, Siberian Division, Russian Academy of Sciences,
pr. Akademika Lavrent’eva 13, Novosibirsk, 630090 Russia
**e-mail: levm@isp.nsc.ru*

Received January 10, 2003

Abstract—The distribution of equilibrium magnetization currents in two-dimensional bounded systems placed in an external magnetic field is studied. A half-plane, a quantum disk, and a wide quantum ring are considered. The passage from classical to quantizing magnetic fields is investigated. The edge currents near the boundary of the half-plane are shown to experience damped (far from the boundaries) spatial oscillations related to the Fermi electron wavelength. The region occupied by currents was found to narrow with increasing field. Apart from these oscillations, the current contains a component that smoothly changes with distance but oscillationally depends on the position of the Fermi level relative to the Landau levels. The suppression of the oscillations by temperature is studied. The spatial distribution of the current in a circular disk and a ring is shown to significantly depend on the position of the Fermi level. *© 2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

The magnetic field that acts on a low-dimensional system is traditionally assumed to coincide with the external magnetic field. Nevertheless, the electron magnetization in such a quantum system as an atom appreciably changes the magnetic fields that act on atomic nuclei and, therefore, is important, for example, in nuclear magnetic resonance (NMR). Previously [1], we studied the magnetic-field distribution in systems with dimensional quantization in the linear (in external magnetic field) approximation. We showed that the equilibrium current density and the current-induced magnetic field experienced oscillations as functions of the distance from the boundaries and the Fermi energy.

This problem is related to the problem of the so-called orbital magnetism—the magnetization of a quantum system in an external magnetic field. The orbital magnetism was intensively studied in various systems with separable and inseparable variables [2–4]. The magnetic susceptibility of large quantum systems at a low temperature T smaller than the energy level spacing ΔE was shown to experience large fluctuations, with the sign of the susceptibility changing when the Fermi level crosses the energy levels of the system. We emphasize that the low-temperature limit may be considered as being opposite to the thermodynamic limit $T \gg \delta E$. Therefore, when $T \longrightarrow 0$, the extensive quantities, in particular, the magnetic susceptibility, are not self-averaged.

Here, our goal is to investigate the spatial distribution of an equilibrium current in various two-dimensional systems with boundaries in a strong magnetic field. We study a semi-infinite medium, a quantum disk, and a quantum ring. Of particular interest to us is the mesoscopic limit where the characteristic size of the system is much larger than the electron wavelength but, at the same time, the quantum properties are still essential.

2. AN EDGE CURRENT IN A TWO-DIMENSIONAL SYSTEM IN A FINITE MAGNETIC FIELD

Let us consider the problem of the equilibrium current density in a semi-infinite sample $x > 0$, $-\infty < y < \infty$, placed in a magnetic field $B_z = B$. This limit describes the current density when the characteristic size of the system is larger than that of the region in which the edge current flows. In particular, this limit is realized when the characteristic size of the system exceeds the cyclotron diameter. The solution is applicable to arbitrarily shaped bounded regions provided that the local curvature of the boundary is smaller than the reciprocal of the cyclotron radius.

We choose the calibration of the vector potential to be $A_y = Bx$. The states in a semi-infinite sample in the presence of a magnetic field can be described by the longitudinal momentum p and the transverse number n: $\psi_{n,p}(x)e^{ipy}$. In what follows, $\hbar = 1$. When the boundary of the system is a rigid wall, the wave functions must

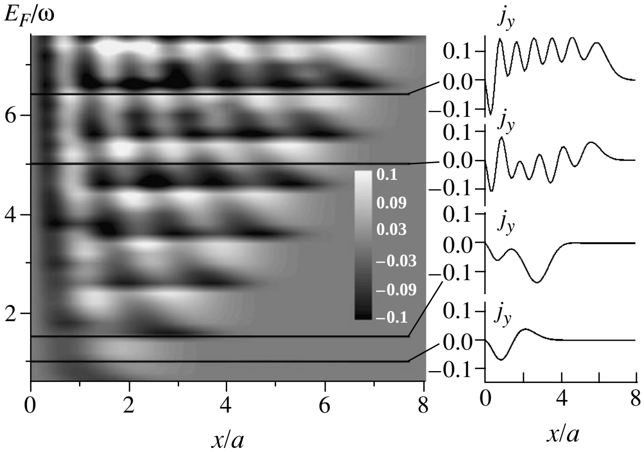


Fig. 1. The left panel shows the edge current density relief, in units of j_0, for a finite magnetic field as a function of the distance to the boundary and the Fermi energy. The gray, white, and black colors correspond to zero, positive, and negative current densities, respectively. The right panel shows a plot of the current density against the distance for selected Fermi energies (marked by the straight lines in the left panel).

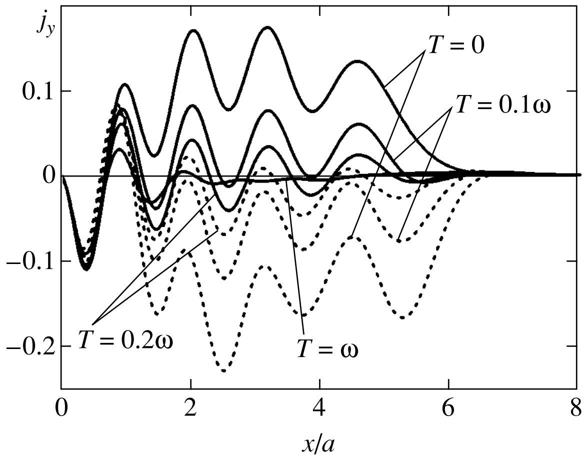


Fig. 2. The edge current density as a function of the temperature. The Fermi level is located near the fifth Landau level. The solid and dotted curves correspond to $E_F = 4.4\omega$ and 4.6ω, respectively; in these cases, the total currents at low temperatures are oppositely directed. At temperatures $T \gg |E_F - 4.5\omega|$, the curves merge together. At $T \gg \omega$, the current density ceases to depend on the magnetic field.

satisfy the zero condition at the $x = 0$ boundary. These wave functions can be expressed in terms of the function of a parabolic cylinder $D_\nu(x)$:

$$\psi_{n,p}(x) = CD_{\nu_n}\left(\frac{\sqrt{2}(x-x_p)}{a}\right),$$
$$C^2\int_{-x_p}^{\infty} D_{\nu_n}\left(\frac{\sqrt{2}x}{a}\right)dx = 1. \tag{1}$$

Here, $a = \sqrt{c/eB}$ is the magnetic length and $x_p = -pa^2$. The boundary condition $D_{\nu_n}(-\sqrt{2}pa) = 0$ defines the energy levels $E_{n,p} = \omega(\nu_n(p) + 1/2)$, where $n = 0, 1, \ldots$; $\omega = eB/m_ec$; and m_e is the effective electron mass.

The current density is

$$j_y(x) = \frac{j_0}{2\pi}\sum_{n=1}^{\infty}\int_{-\infty}^{\infty} dpa \tag{2}$$
$$\times\left[(x_p-x)\psi^2_{n,p}(x)(f_+ + f_-) + \frac{gm_e}{2m_0}(f_- - f_+)\frac{a^2\partial\psi^2_{n,p}(x)}{\partial x}\right],$$

where $j_0 = e\omega/a = e/(m_ea^3)$ is the characteristic current density produced by one electron at the lower Landau level in a magnetic field, $f_\pm = f(E_n(p) \pm g\mu_BB/2)$, $f(E) = (\exp((E-\mu)/T) + 1)^{-1}$ is the Fermi distribution function (μ and T are the chemical potential and the temperature, respectively), and g is the g factor. Expression (2) contains the two contributions that arise when averaging the orbital and spin parts of the current density operator (the first and second lines, respectively). Below, we disregard the spin splitting by assuming that the g factor is small.

The results of our calculations of the edge current using formulas (1) and (2) are presented in Figs. 1–3.

Let us discuss the behavior of the current density at a low temperature. First, the direction of the current is determined by the vector product of the normal and the magnetic field. However, this does not imply that the sign of the surface current density is constant. Indeed, let only one Landau level be occupied. Consider the states located far from the boundary. These states are not perturbed by the wall. The current density given by one state with a specified p is antisymmetric about the point x_p. As the boundary is approached, the level is

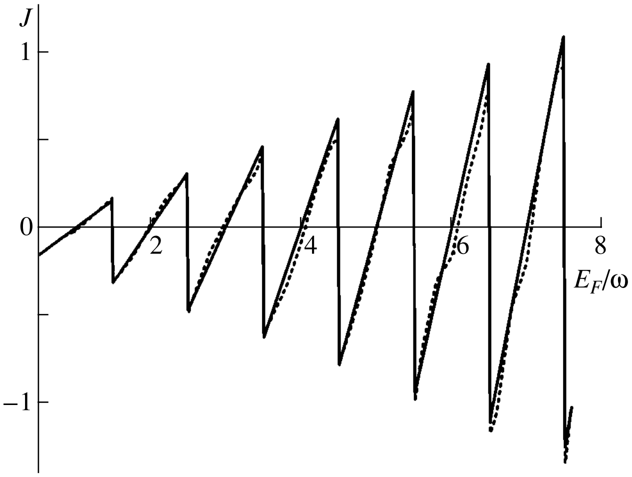


Fig. 3. The total edge current as a function of the magnetic field at $T = 0$.

pushed upward. The occupied levels that underlie the Fermi level have similar wave functions. Momentum averaging compensates for the current density far from the point of intersection of the energy level with the Fermi level, while near this point the uncompensated contribution of constant (in our case, positive) sign remains. As the Fermi level approaches the lower Landau level, the edge state increasingly recedes from the boundary according to the law $\ln^{1/2}(\mu/\omega - 1/2)$ and becomes increasingly ideal. As the Fermi level rises, the wave function begins to be distorted by the boundary, the positive contribution to the average current density is suppressed, but a negative contribution appears—the plot of the current density acquires a negative minimum.

As the next Landau levels are crossed, new contributions from these levels are added to the current density. These contributions are initially located far from the boundary but then approach it and merge together with the contributions from the lower lying states. The higher the state, the wider the region occupied by them and the more oscillations the corresponding contribution to the current density experiences. When the number of occupied levels becomes large, the edge contributions from different Landau levels merge into the Friedel oscillations of the surface current density that we considered previously [1]. However, in contrast to the previously considered limit of a weak magnetic field (a linear response), the number of oscillations proves to be limited—the surface current is distributed in a thickness of the order of the cyclotron diameter $2r_c = 2v_F/\omega$, where v_F is the Fermi velocity. The current density oscillates inside this region and exponentially decays outside. The number of oscillations is determined by the number of occupied Landau levels $N = [\mu/\omega - 1/2]$, where $[\ldots]$ represents the integer part. In the region $x \ll 2r_c$, an expansion can be made in terms of the magnetic field, the magnetic-field effect reduces to a factor, and the surface current density is given by the expression from [1].

A smooth dependence on the coordinate x is superimposed on the spatial oscillations. This smooth contribution determines the total edge current (the coordinate integral) that experiences alternating Shubnikov oscillations as a function of the magnetic field (Fig. 3). In the weak-field limit, the smooth contribution transforms into the contribution studied in [1] that linearly depends on the distance and the oscillations propagate to an infinite distance from the boundary, being damped as a power law.

As the temperature rises (see Fig. 2), the current-density oscillations are suppressed, more strongly at large distances from the surface. This suppression results from the thermal misphasing of the electrons near the Fermi surface at the characteristic length $l_T = k_F/(2\pi m_e T)$, where k_F is the Fermi momentum [1]. Competition between the two lengths that restrict the oscillations arise in a finite magnetic field: the cyclotron diameter and l_T (the smaller length works). The smooth contribution is suppressed by the temperature and scattering in the same way as the surface oscillations at a large distance. This is how the results obtained for weak and strong magnetic fields are joined.

At first glance, the alternating oscillations of the total edge current seem strange if it is considered that the total current at a high Fermi energy must be diamagnetic and be described by the formula

$$J = -\frac{j_0 a}{12\pi}.$$

However, in the limit of a large system, the total edge current J is directly related to the magnetic moment of the system by $M = JS/c$, where S is the area of the system. At the same time, the total moment at $T = 0$ can be determined by using the Ω potential of the system [5]:

$$M = -\frac{\partial\Omega}{\partial B} = \frac{j_0 aS}{\pi c} \times \sum_{n=0}^{\infty}\left[\frac{\mu}{\omega} - (2n+1)\right]\theta\left(\mu - \left(n + \frac{1}{2}\right)\omega\right). \tag{3}$$

This expression oscillates with μ, experiencing jumps when the Fermi level crosses the Landau levels. Another property of this formula is that the moment becomes zero when the Fermi level lies halfway between the Landau levels.

On the other hand, the edge currents under nonequilibrium conditions are related to the quantization of the microcontact resistance in the regime of adiabatic transport in the quantum Hall effect. The total current in a (n, p) state is given by the expression

$$j_{n,p} = -e\frac{\partial E_{n,p}}{\partial p}. \tag{4}$$

Since the states are localized in coordinate x, the edge current is determined by the states with x_p near the boundary. The partial current $j_{n,p}$ exponentially decreases with increasing x_p. Therefore, summing the currents over all states, we obtain the final value. These considerations determine the quantization of the microcontact resistance.

The total edge current in the semi-infinite problem at $T = 0$ is

$$J = -e\sum_{n,p}\frac{\partial E_{n,p}}{\partial p}\theta(\mu - E_n(p)) = -\frac{j_0 a}{\pi} \times \sum_{n=0}^{\infty}\left[\left(n + \frac{1}{2}\right) - \frac{\mu}{\omega}\right]\theta\left(\mu - \left(n + \frac{1}{2}\right)\omega\right). \tag{5}$$

The current through the structure can be expressed in terms of the difference between the edge currents (5) that correspond to the chemical potentials of the edges

μ_1 and μ_2. The difference $\mu_1 - \mu_2$ in the nonequilibrium problem is equal to the potential difference V applied to the microstructure ends. It thus follows that the nonequilibrium current is $J = e^2VN/h$. Under equilibrium conditions, the chemical potentials of the edges are equal, but the edge current from one end must be defined by the same expression as that in the absence of equilibrium. However, formula (5) understood in this way contradicts formula (3) for the moment, as well as numerical calculations.

To elucidate the nature of the error, let us determine the contribution from the Landau states with a momentum between p_1 and p_2 to the current density in the boundaryless problem, with $|x_{p_1} - x_{p_2}| \gg a$:

$$\begin{aligned} j_y(p_1, p_2; x) &= -\frac{j_0}{\pi}\sum_n \int_{p_1}^{p_2} (x - x_p)\varphi_n^2\left(\frac{x - x_p}{a}\right)dp \\ &= -\frac{j_0}{\pi}\sum_n\left(\int_{-\infty}^{p_2} dp\ldots - \int_{-\infty}^{p_1} dp\ldots\right), \end{aligned} \tag{6}$$

where $\varphi_n(\xi)$ are the dimensionless normalized functions of a harmonic oscillator. The contributions from the regions near x_1 and x_2 are independent and separated in space. Thus, although the total current in each Landau state becomes zero, momentum integration yields two edge currents. When we performed the integration over all momenta in formula (5), the correct limit should have been calculated by first limiting the momentum integration below by some point and then letting it tend to $-\infty$. The contribution from the current density near x_1 that corresponds to p_1 to the total current obtained in this case is localized far from the boundary and it should be subtracted from the total surface current.

We determine this part of the current by integrating $j_y(p_1, p_2; x)$ over x near x_p over a region that is large compared to the characteristic size of the wave function:

$$\begin{aligned} &\int_{x_p - \Delta}^{x_p + \Delta} j_y(p_1, p_2; x)dx \\ &= \frac{j_0 a}{\pi}\sum_{n=0}^{\infty}\left(n + \frac{1}{2}\right)\theta\left(\mu - \left(n + \frac{1}{2}\right)\omega\right). \end{aligned} \tag{7}$$

Here, $\Delta \gg a\sqrt{N}$. Subtracting (7) from (5) yields a correct expression for the edge current $J = Mc/S$ with M from formula (3). The numerical calculation (see Fig. 3) agrees with this formula.

3. DIAMAGNETIC CURRENTS IN A QUANTUM DISK AND A QUANTUM RING

Let us consider the distribution of a nondecaying current and magnetic field in systems with a circular geometry or, more specifically, in a plane ring with inner and outer radii of r and R, respectively. A circular disk is the special case of a ring at $r = 0$. We assume that the magnetic field is directed along the z axis of the system and that the vector potential has an azimuthal calibration, $A_\phi = B\rho/2$ (ρ and ϕ are the polar coordinates). The wave function with the moment component m is $\Psi(\rho, \phi) = e^{im\xi}\psi_{nm}(\phi)/\sqrt{2\pi}$, where $\xi = (\rho/a)^2/2$, and the radial function satisfies the equation

$$\begin{aligned} &\xi\psi''_{nm} + \psi'_{nm} \\ &+ \left(-\frac{\xi}{4} + \left(\frac{E_{n,m}}{\omega} - \frac{m}{2}\right) - \frac{m^2}{4\xi}\right)\psi_{nm} = 0, \end{aligned} \tag{8}$$

the zero conditions at the boundary, and the normalization $\int|\psi_{nm}(\xi)|^2 d\xi = 1$. The current has only the azimuthal component j_ϕ:

$$j_\phi(\xi) = -j_0\sqrt{\frac{2}{\xi}}\sum_{n,m}|\psi_{nm}(\xi)|^2(m + \xi)f(E_{n,m}). \tag{9}$$

In the weak-field limit, the current density can be expressed in terms of wave functions in the absence of a magnetic field:

$$\begin{aligned} j_\phi(\rho) &= -j_0 a\rho \\ &\times\sum_{n,m}\left[f(E^0_{n,m}) + \frac{m^2}{m_e\rho^2}\frac{\partial f}{\partial E^0_{n,m}}\right]|\psi^0_{nm}(\rho)|^2, \end{aligned} \tag{10}$$

where $\psi^0_{nm}(\rho)$ and $E^0_{n,m}$ are the unperturbed wave functions and energy levels, respectively. The summation is performed over the transverse quantum numbers n and m.

In the special case of a disk of radius R in the absence of a magnetic field, the wave functions $\psi^0_{nm}(\rho)$ are

$$\psi^0_{nm}(\rho) = \frac{\sqrt{2}J_m(\lambda_{n,m}\rho/R)}{RJ_{m+1}(\lambda_{n,m})}, \tag{11}$$

where $\lambda_{n,m}$ is the nth zero of the Bessel function $J_m(x)$ and the energy levels are $E^0_{n,m} = \lambda^2_{n,m}/(2m_eR^2)$. The first and second terms in (10) are negative and positive, respectively, and correspond to the diamagnetic and paramagnetic contributions to the magnetic susceptibility of the system. At a low temperature, the paramagnetic contribution is nonzero only if the chemical potential coincides with one of the system's energy lev-

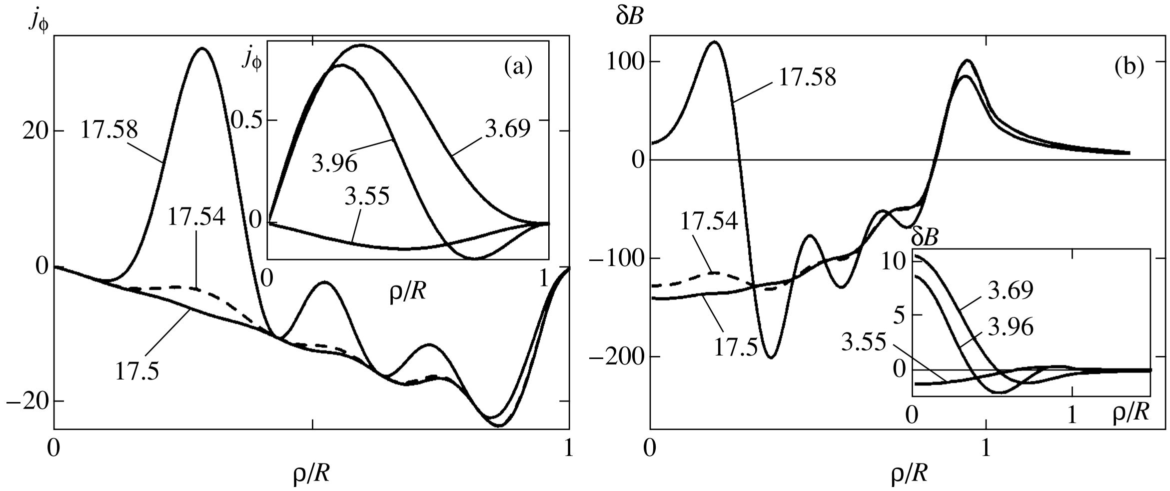


Fig. 4. (a) The evolution of the current density in a disk with the parameter k_FR (indicated on the curves). The external magnetic field is assumed to be weak, $T = 0.5/(2m_eR^2)$. The current density is measured in units of $j_0(a/R)$. The Fermi level is near the transverse quantization level with $E^0_{1,1} = 14.7/(2m_eR^2)$ for $k_FR \approx 3.7$ and near $E^0_{4,4} = 310.3/(2m_eR^2)$ for $k_FR \approx 17.5$. As the Fermi level passes the quantization levels, the diamagnetic and paramagnetic dependences alternate. (b) The distribution of the induced magnetic field, in units of $B(e^2/m_ec^2)/R$.

els, while the diamagnetic contribution is constant if μ lies between the energy levels. The delta-shaped pattern of the paramagnetic contribution stems from the fact that the azimuthally symmetric system is doubly degenerate in the sign of m. This degeneracy is removed by a sufficiently weak magnetic field, which is responsible for the large positive (paramagnetic) peaks in the susceptibility in the low-temperature limit. The diamagnetic contribution originates from the first-order corrections of the term $e^2A^2/2m_ec^2 = (eB\rho/c)^2/8m_e$ in the Hamiltonian. Apart from the factor oscillating with ρ, the current density is proportional to the radius, which corresponds to a quadratic increase of the susceptibility with the system's size.

Figure 4a shows the current-density distribution in the special case of a weak magnetic field that was determined by using formulas (10) and (11). Figure 5 shows the current-density distribution in a quantum ring in a strong magnetic field that was calculated by using the solution of Eq. (8) and formula (9) for the current density.

Note that the current density radially oscillates with a period of π/k_F, while the oscillation phase at the boundary is fixed. In a plate and a strip, the dependence of the current density on the transverse coordinate [1] was antisymmetric. In a ring, the radial dependence becomes antisymmetric relative the circumference of radius $(R - r)/2$ as the ring becomes narrower. For large E_F/ω and $k_F(R - r)$, the radial oscillation amplitude is damped as the distance from the boundaries increases. This contribution represents the edge current. In addition to the radial oscillations, the azimuthal current has a regular, nearly linear dependence. The slope of this dependence and the slope sign are sensitive to the magnetic-field strength. In a strong magnetic field, the current density also oscillates with the magnetic field.

The presence of an edge current causes the magnetic field that acts on a two-dimensional system to slightly change. A nonuniformly distributed current produces a nonuniformly distributed magnetic field. The addition to the z component of the magnetic field in the disk ($z = 0$) plane is described by the expression

$$\delta B(\rho) = \frac{1}{c}\int_0^R d\rho' j_\phi(\rho')\frac{2}{\rho+\rho'}\left[K(k) + \frac{\rho'+\rho}{\rho'-\rho}E(k)\right], \quad k = \frac{4\rho'\rho}{(\rho'+\rho)^2}, \tag{12}$$

where $K(k)$ and $E(k)$ are the elliptic integrals. The magnetic field at the disk center can be expressed in terms of elementary functions:

$$\delta B(0) = \frac{2\pi}{c}\int_0^R \frac{j_\phi(\rho)}{\rho}d\rho. \tag{13}$$

To a first approximation, the electron density for a sufficiently large disk may be assumed to be uniform. In this case, the correction to the external magnetic field at the disk center in the diamagnetic state of the disk is defined by the formula $\delta B = B(e^2/m_ec^2)\pi nR$, where n is the electron surface density. If we increase the disk size while keeping the density constant, then the correction to the magnetic field will increase proportionally with

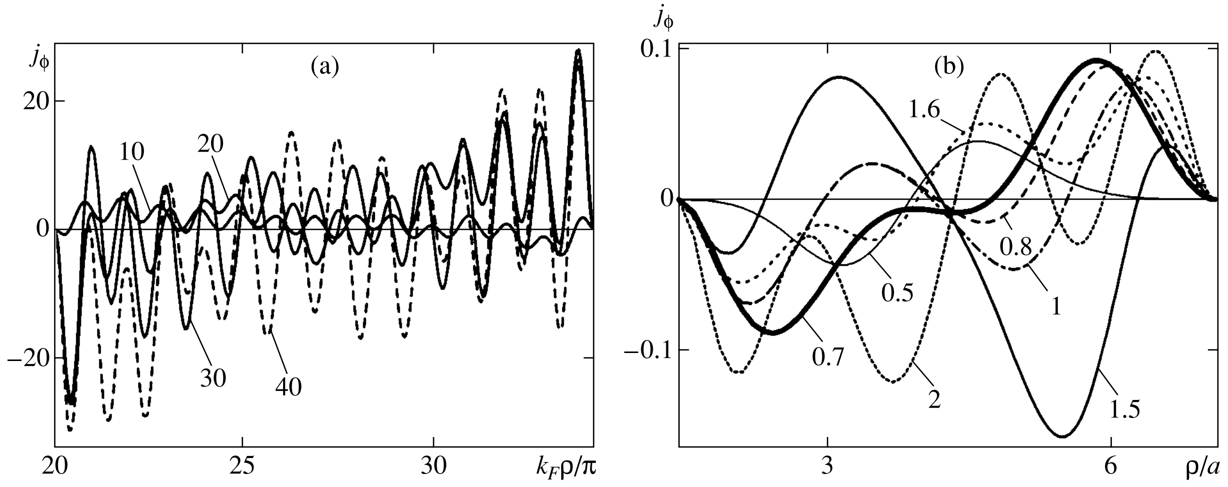


Fig. 5. The current-density distribution in quantum rings in a strong magnetic field. The values of E_F/ω are indicated on the curves. (a) The evolution of the current density (in units of $ek_F^2/m_e = j_0(k_Fa)^3$) with the magnetic field B for fixed E_F. The Fermi wave vector was chosen to be $k_F = 6.4/r$ and the radii are related as $R : r = 1.7$. (b) The evolution of the current density (in units of j_0) with E_F for fixed B. The ratios of the parameters were chosen to be $r : a = 1.4$ and $R : a = 7.1$.

to the disk radius. For a GaAs disk with an electron surface density of 10^{12} cm^{-2} and a radius $R \sim 10^{-5}$ cm, the ratio $\delta B/B \sim 1.2 \times 10^{-4}$.

Figure 4b shows the quantity $\delta B(\rho)$ calculated using formula (12) in the limit of a weak external magnetic field. It should be noted that δB increases with the system's size at $T = 0$. This increase is limited because we disregarded the electron scattering: the system's energy level spacing must be larger than the damping of the electronic states. As the system's size increases, the energy level spacing becomes comparable to the temperature or the damping, and the correction to the external magnetic field ceases to depend on the size. For a disk of radius 10^{-5} cm, the mean energy level spacing is $\Delta E = 1/(m_eR^2) \sim 1$ K and the condition $T < \Delta E$ is easily satisfied.

The spatial nonuniformity of the magnetic field can affect any responses that sharply depend on the magnetic field, in particular, the Shubnikov or geometrical oscillations of the magnetoresistance (the magnetic focusing for small samples). One might expect the field nonuniformity to smear the sharp features in these quantities. Other cases can be based on the sensitivity of the nuclear spins to the local magnetic field. Impurities with nuclear spins located on certain atomic planes can act as a gauge of the local magnetic fields. We can suggest using both a time-constant magnetic field for shifting the position of the NMR line and a variable magnetic field for the excitation of transitions. A time-varying and nonuniform magnetic field can be produced by varying the wave functions using, for example, a field electrode. A periodic variation of the electron gas temperature through a variable pulling voltage (a rise in temperature suppresses the Friedel oscillations by changing δB) can be suggested as another way of producing a variable magnetic field.

ACKNOWLEDGMENTS

This study was supported by the Russian Foundation for Basic Research (project nos. 02.02-16377 and 00-02-17658).

REFERENCES

1. L. I. Magarill, M. M. Makhmudian, and M. V. Éntin, Pis'ma Zh. Éksp. Teor. Fiz. **75**, 560 (2002) [JETP Lett. **75**, 470 (2002)].
2. M. Ya. Azbel, Phys. Rev. B **48**, 4592 (1993).
3. K. Richter, D. Ullmo, and R. A. Jalabert, Phys. Rep. **276**, 1 (1996).
4. E. Gurevich and B. Shapiro, J. Phys. (Paris) **7**, 807 (1997).
5. A. A. Abrikosov, *Fundamentals of the Theory of Metals* (Nauka, Moscow, 1987; North-Holland, Amsterdam, 1988).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 144–153.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 161–171.
Original Russian Text Copyright © 2003 by Voloshin, Kalinov, Fisher, Levchenko, Yampol'skiĭ.

SOLIDS
Electronic Properties

Collapse of Static Magnetization of Type II Anisotropic Superconductors

I. F. Voloshin[a], **A. V. Kalinov**[a], **L. M. Fisher**[a],*,
A. A. Levchenko[b], **and V. A. Yampol'skiĭ**[b]

[a]State Research Center All-Russia Institute of Electrical Engineering, Moscow, 111250 Russia
[b]Institute of Radiophysics and Electronics, National Academy of Sciences of Ukraine, Kharkov, 310085 Ukraine
**e-mail: lfisher@online.ru*

Received January 13, 2003

Abstract—The effect of suppression of static magnetization (magnetization collapse) of an anisotropic hard superconductor by an alternating magnetic field is studied experimentally and theoretically. The measurements are performed in constant and alternating magnetic fields **H** and **h**(t) which are parallel to the sample surface and are mutually orthogonal. Superconducting samples of the Y-123 system are cut from a grain-oriented piece in such a way that the c axis lies in the plane of the plate. It has been established that the amplitude of the alternating field for which the complete suppression of the magnetization occurs significantly depends on the field orientation relative to the crystallographic axes of the sample. The dynamics of the magnetization collapse is studied, and it is established that, in some cases, the magnetic moment varies nonmonotonically with the growth in amplitude h_0 of the alternating field. A peculiar effect of transition (induced by field $h(t)$) of the superconductor from the paramagnetic into the diamagnetic state is discovered. The results of observations are explained in the framework of the model of critical state generalized to the anisotropic case. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

For many years, the physics of turbulent matter of hard superconductors has attracted the attention of researchers. Recently, a group of new interesting effects in a turbulent liquid was discovered. Among these effects, it is necessary to distinguish the phenomenon of macroturbulence observable in a system containing vortices of opposite directions [1–3]. The physical nature of macroturbulence is connected with the magnetic flux anisotropy in superconductors of the 1–2–3 system [4] due to the effect caused by twin boundaries [5, 6]. Another peculiar instability effect, namely, the collapse of the static magnetization of type II superconductors for a small variation of the direction of an external magnetic field, has been observed in studies [7–13]. The essence of this effect is the following. Let a plane superconducting sample cooled in zero magnetic field be placed into an external magnetic field $H > H_{c1}$ (here, H_{c1} is the lower critical magnetic field) which is parallel to the superconductor surface. Pinning leads to the emergence of a nonuniform distribution of the magnetic induction and, accordingly, a static magnetization, in the sample. Although such a distribution of magnetic induction is metastable, for sufficiently low temperatures, it can be quite stable and can hold for many years. If an alternating magnetic field $\mathbf{h}(t) = \mathbf{h}_0\cos(\omega t)$ is applied to the magnetized sample in a direction parallel to the sample surface and perpendicular to a constant field **H**, then magnetization M of the sample decreases. Direct measurements [14] show that, everywhere in the sample where the alternating magnetic field penetrates, the flow of nondissipative currents becomes impossible. As a result, in the surface region of the sample, the currents that previously screened field **H** and contributed to the magnetization of the superconductor disappear. If the amplitude is sufficiently large, $h_0 \sim H_p = 2\pi J_c(H)/cd$, (the alternating field practically pierces the entire sample), magnetization M is suppressed; i.e., the collapse of the sample magnetization occurs. Here, J_c is the critical current density, d is the sample thickness, and c is the velocity of light. For sufficiently large values of H, when the critical current density noticeably decreases, the collapse occurs for relatively small values h_0, for which the deviation of the resultant magnetic field from the initial direction of **H** (from the z axis) is small. Physically, this means that the nonuniform distribution (stable for $h_0 = 0$) of magnetic induction B_z over the sample thickness can be easily violated due to a small deviation of the magnetic field from the z axis. It is shown in [13] that the magnetic moment decreases to less than half even after a simple small deviation of the constant magnetic field from the z axis; i.e., the effect can occur in the absence of the alternating field.

The nature of the magnetization collapse can be easily understood in the framework of the simplest isotropic model of critical state proposed by Bean [15]. In this model, the distribution of magnetic induction is described by the equation

$$\mathrm{curl}\mathbf{B} = \frac{4\pi}{c} J_c \frac{\mathbf{E}}{E}, \tag{1}$$

where **E** is the electric field strength. According to this equation, a current directed along field **E** flows in all regions of the superconductor where the electric field is present, the current density being equal to the critical density. It is important to emphasize that, if an electric field arises in a region of the superconductor, then the memory about the currents flowing here before is erased and current J_c is induced with the new direction along vector **E**. Therefore, at all places where alternative field $h(t)$ penetrates, critical currents screening $h(t)$ arise and the nondissipative magnetization currents flowing before disappear.

It is clear that the model of critical state is a rather rough approximation that does not take into account the complex phenomena connected with the intersection and reconnection of the vortex lines in the turbulent system of the superconductor. Therefore, to explain the collapse process in greater detail, more adequate models are needed [11, 13, 16, 17]. However, the main reason for the collapse lies in the mutual influence of different components of the critical current density, which exists at a point of the superconductor as it is explicitly reflected in Eq. (1). Indeed, the excitation of current J_c in the direction of the z axis is necessarily accompanied by the suppression of the other current components at this point of the sample. Thus, the critical state model is a convenient and quite satisfactory tool for the qualitative description of the collapse.

Although a large number of papers are devoted to the study of magnetization collapse, up to now, only the isotropic situation has been considered, where the value of the critical current density is independent of its direction. At the same time, it is well known that HTSC systems are characterized by a clearly manifested anisotropy of practically all physical properties. The anisotropy of current-carrying capacity of superconductors results in a qualitative change in the character of the electromagnetic field penetration in the sample. It is shown in [18–20] that an alternating magnetic field in anisotropic superconductors is represented by two components which are screened by currents with quite different values of density. Therefore, these components die down at different depths. Due to this peculiar character of the field penetration, a number of well-known electrodynamic effects manifest themselves in peculiar forms. In particular, the dependence of the relative electromagnetic losses on amplitude h_0 in anisotropic superconductors is characterized by the presence of two dimensional maximums, each of which is connected with the penetration of one of the two field components to the center of the sample [18, 20]. In the static case, the anisotropy of the current-carrying capacity results in the arising of additional extremums on the magnetization curves, i.e., a peculiar new type of the peak effect [19]. For these reasons, it seems important to investigate the phenomenon of the magnetization collapse in anisotropic hard superconductors. The present paper is devoted to solving this very problem.

2. MEASURING TECHNIQUE AND RESULTS

2.1. Samples and Measurements

Our main task in this work is to study the influence of the anisotropy in the critical current density of a superconductor on the collapse of its static magnetization. Therefore, the experimental investigations were performed on samples with clearly manifested anisotropic properties. The main measurements were made on a plate of Y1 yttrium superconductor with the c axis in the plane of this plate. A sample $3.4 \times 2.2 \times 0.46$ mm^3 in size was cut from a piece of fused grain-oriented ceramic of the Y-123 system, which was synthesized with the help of a seed placed at the top at a high temperature. To choose a perfect monodomain part of the piece, we examined it with the help of an optical microscope in polarized light and also used chart-making of the frozen magnetic flux with the help of a Hall probe. The sample was cut out from the chosen monodomain by a diamond milling cutter in such a way that the c axis was in the sample plane. Before the measurements, the sample was mechanically polished. The superconducting transition temperature T_c of the sample was 89 K, and the transition width was 0.5 K. Preliminary measurements of dynamic magnetic susceptibility allowed us to determine the critical current densities of the sample along the principal anisotropy directions and their dependence on the magnitude and direction of field **H** using the technique described in [18, 20]. Representative values of the critical current density $J_c^{(c)}$ along the c axis and $J_c^{(ab)}$ in the ab plane in field $H = 5$ kOe perpendicular to the current were equal to 7×10^3 and 6×10^4 A/cm^2, respectively.

Projection M of the magnetization vector of the superconductor onto the direction of the external constant magnetic field **H** was measured by a vibrating-coil magnetometer with a vibrating sample. The sample was mounted in such a way that the constant magnetic field created by the electromagnet was parallel to the sample surface. The direction of the field could be varied in the sample plane (the yz plane) by turning the electromagnet.

In the experiment, we studied the behavior of the magnetic moment of the superconductor in an external alternating magnetic field $\mathbf{h}(t) = \mathbf{h}_0\cos(\omega t)$ parallel to the sample surface and perpendicular to a constant magnetic field **H** (see Fig. 1). The measurements were performed at a frequency of $\omega = 2\pi f = 2\pi \times 343$ s^{-1}. The

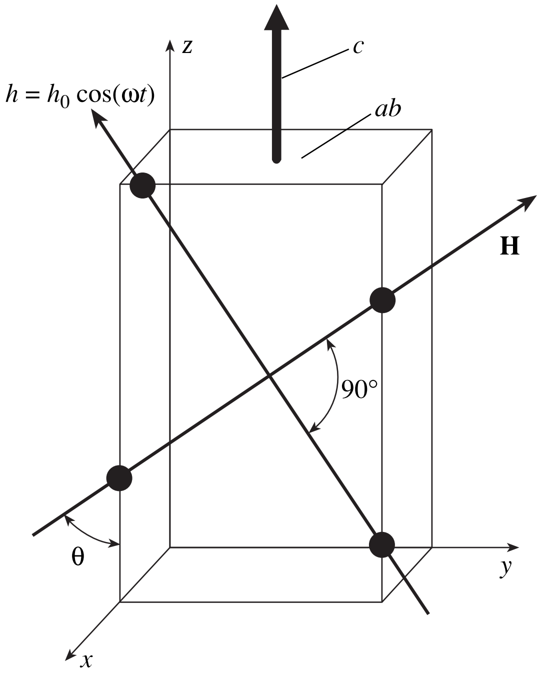


Fig. 1. Measurement geometry.

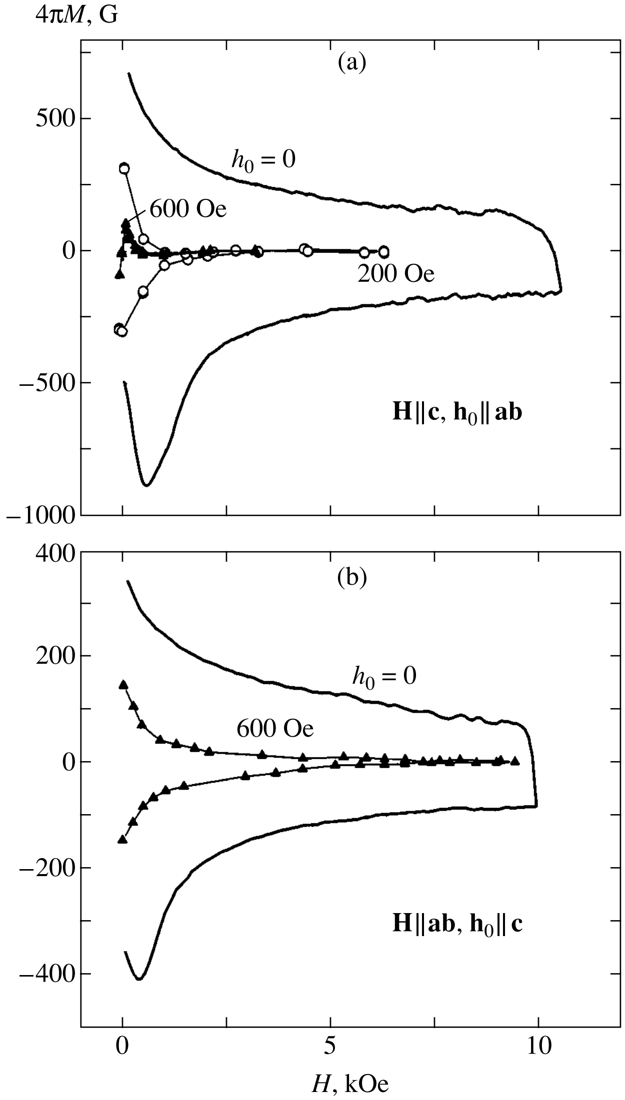


Fig. 2. Magnetization loops of an anisotropic superconductor for different amplitudes of the alternating magnetic field and directions of vector **H** along the principal directions of the anisotropy: θ = 0 (a) and θ = 90° (b). Temperature $T = 77$ K.

possibility of a continuous rotation of field $\mathbf{h}(t)$ was achieved with the help of two field-exciting coils with mutually orthogonal axes. All measurements were performed at a temperature of $T = 77$ K in the cooling regime of the sample in zero magnetic field.

In the experiment, it was possible in principle to measure the dependence of static magnetization on the alternating magnetic field amplitude in the continuous regime. However, in this case, the amplitude of the noise of the signal being measured increased. Therefore, point by point measurements of $M(h_0)$ were made. In this case, after the action of the alternating magnetic field of a specified amplitude on the magnetization of the sample during several seconds, the field was switched off and the magnetization was measured. To escape the magnetic flux trapping connected to the alternating magnetic field, the switching off was performed by a slow monotonic decrease in amplitude h_0.

2.2. Measurement Results

The results of measurements of magnetization loops for the Y1 sample under orientations of external magnetic field **H** in the plane of the plate along the *c* axis (angle θ between **H** and the *c* axis is equal to zero) and perpendicularly to this axis (θ = π/2) are presented in Fig. 2. The curves mainly differ in the loop width. This difference is connected to the anisotropic current-carrying capacity of the superconductor. Indeed, the sample magnetization for $\mathbf{H} \parallel \mathbf{c}$ is determined by the screening currents of high density $J_c^{(ab)}$ flowing in the *ab* plane, whereas, in conditions $\mathbf{H} \perp \mathbf{c}$, the screening is mainly ensured by weak currents $J_c^{(c)}$ flowing along the *c* axis.

The influence of an alternating magnetic field $\mathbf{h}(t) = \mathbf{h}_0 \cos(\omega t)$ orthogonal to **H** on the magnetization is illustrated by the curves in Fig. 2b. The switching on of the alternating magnetic field results in a significant decrease in the magnetization and the formation of reversible segments on the curves for both geometries. Note that the complete suppression of magnetization *M* of the sample located in an external field parallel to the *c* axis (Fig. 2a) is observed for significantly smaller amplitudes of the alternating field than in conditions $\mathbf{H} \perp \mathbf{c}$ (Fig. 2b). It is interesting to note that a moment which is significantly greater in magnitude is suppressed at much smaller amplitudes. This fact, quite paradoxical at first glance, can be rather naturally explained within the framework of the collapse theory [9, 13]. Indeed, the alternating magnetic field in conditions $\mathbf{h}(t) \perp \mathbf{c}$ screened by weak currents with density $J_c^{(c)}$ penetrates the sample more deeply and efficiently suppresses static magnetization.

A similar pattern of suppressing the magnetic moment of the superconductor is also observed for other orientations of mutually orthogonal alternating and constant magnetic fields. In Fig. 3, a set of curves $M(H)$ obtained for $\theta = 45°$ is presented. As should be expected, the efficiency of the influence of the alternating field on the magnetization for intermediate angles $0 < \theta < \pi/2$ is less than that for $\theta = 0$, but greater than for $\theta = \pi/2$.

To study in more detail the role of the anisotropy in the development of the magnetization collapse, we measured the dependence of the magnetic moment M on the amplitude of the alternating magnetic field for the same values of external constant magnetic field $\mathbf{H}$, but for different field orientations relative to the principal anisotropy axes. The results of these measurements are presented in Fig. 4. Different curves in this figure correspond to different values of angle θ. One can see that the magnetic moment is monotonically suppressed with the increase in h_0. At a certain amplitude $h_0^{(\mathrm{col})}$, total collapse of magnetization is achieved. It should be noted that the fact itself of the magnetization vanishing is very difficult to determine; hence, the error in the determination of the value $h_0^{(\mathrm{col})}$ is rather large. Nevertheless, the measurement results imply that $h_0^{(\mathrm{col})}$ slightly increases for small θ and that it has a relatively steep growth for $\theta > 70°$. The dependence of the suppression amplitude $h_0^{(\mathrm{col})}$ on angle θ is depicted by the squares in Fig. 5. The presented data are normalized to the value $H_p \approx 480$ Oe of the penetration field that corresponds to the value H of the external field H directed along the c axis and penetrating to the center of the sample. With the angle θ increasing from zero to 90°, the value $h_0^{(\mathrm{col})}$ increases by approximately a factor of 7–8. This significant variation of field $h_0^{(\mathrm{col})}$ correlates well with the value of anisotropy parameter $\delta = J_c^{(c)}/J_c^{(ab)} \sim 0.13$. The behavior of function $h_0^{(\mathrm{col})}(\theta)$ slightly varies with the growth of field H, in complete agreement with the dependence of the anisotropy parameter on the magnetic field.

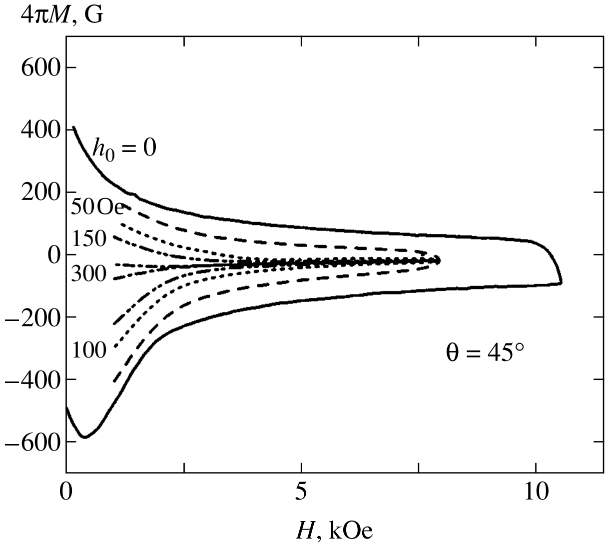


Fig. 3. Magnetization loops of an anisotropic superconductor for different amplitudes of the alternating magnetic field and the direction of vector $\mathbf{H}$ at an angle of 45° to the c axis. Temperature $T = 77$ K.

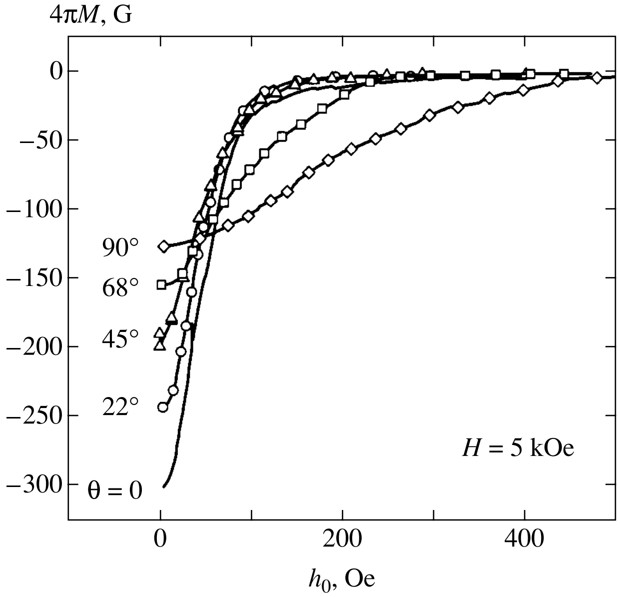


Fig. 4. Suppression of the sample magnetization for different orientations of field $\mathbf{H}$ relative to the c axis. The magnetic prehistory of the sample consisted in a monotonic increase of the field H from zero to 5 kOe. Temperature $T = 77$ K.

Studying the collapse process in more detail, we established that the suppression of moment $M(h_0)$ is determined not only by the magnitude and direction of vector $\mathbf{H}$, but also by the magnetic prehistory of the sample. The curves in Fig. 4 were obtained after a monotonic increase in an external magnetic field H from zero to the value at which the $M(h_0)$ measurements were performed. This corresponds to starting point *1* in the inset in Fig. 6. With the growth of amplitude h_0, the dynamics of the magnetic moment suppression becomes qualitatively different if the external field varied preliminarily in a nonmonotonic way. Starting points *2–4* in Fig. 6 correspond to growth in the magnetic field up to its maximal value (up to the sweep

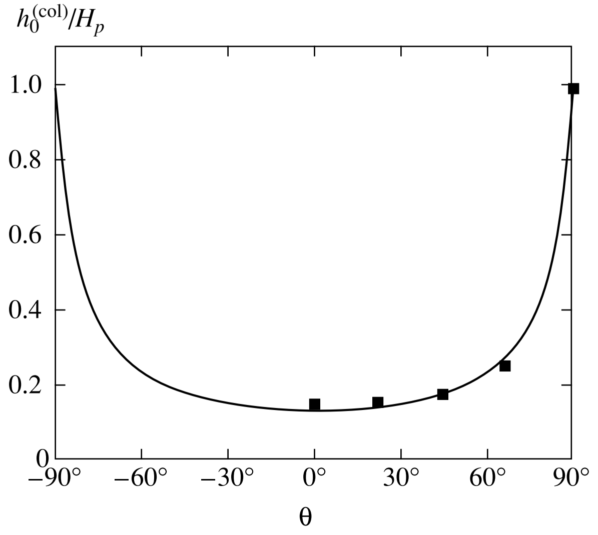


Fig. 5. Dependence of the dimensionless amplitude $h_0^{(col)}/H_p$ of magnetization suppression on the angle θ between the c axis and field **H**. The squares on the graph are the measurement results; the solid curve is the calculation results for the value of the anisotropy parameter $\delta = 0.13$.

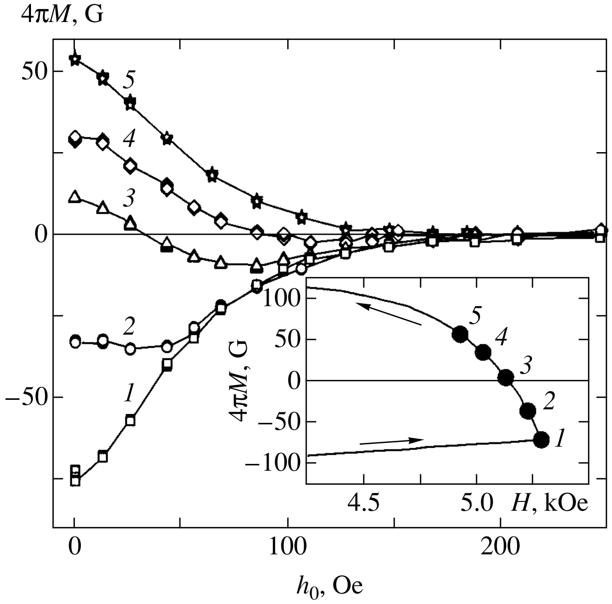


Fig. 6. Measured dependence of the sample magnetization on the amplitude of the alternating magnetic field for different magnetic prehistories of the sample. The numbers on the curves correspond to the enumeration of starting points *1–5* shown in the inset, where a fragment of the curve of sample magnetization for θ = 45° and $h_0 = 0$ is presented.

field H_m, point *1*) followed by its decrease. In Fig. 6, a series of measurements of the dependence $M(h_0)$ for starting points *1–5* is depicted. The curves in this figure show the nontrivial character of the magnetization suppression. The monotonic suppression of the diamagnetic moment corresponding to staring point *1* changes to the nonmonotonic behavior of $M(h_0)$ (starting point *2*). The magnetic moment remains diamagnetic on the entire interval $0 < h_0 < h_0^{(col)}$ of the amplitude variation. For starting points *3* and *4* corresponding to the initial paramagnetic state of the sample, the moment first decreases upon an increase in h_0, changes its sign, becomes diamagnetic, attains a minimal value, and only then begins to tend monotonically to zero. Finally, for a sufficiently large decrease in H (starting point *5*), the paramagnetic moment decreases monotonically with increasing h_0.

To interpret the results of measuring the dynamics of the magnetic moment collapse in an anisotropic superconductor, an adequate theoretical analysis is needed where not only the nature of the collapse phenomenon, but also its peculiarities connected to the anisotropy of the current-carrying capacity of the hard superconductor, should be taken into account. This will be done in the next section.

3. THEORETICAL STUDY OF THE MAGNETIZATION COLLAPSE IN ANISOTROPIC SUPERCONDUCTORS

3.1. Theoretical Model

The theory of suppression of the static magnetization by an orthogonal alternating magnetic field elaborated in [9] relates to the isotropic situation, where the current-carrying capacity of a sample is described by a single phenomenological parameter, namely, critical current density J_c. The subsequent development of collapse theory [11–13, 16] also relates to isotropic superconductors. Unfortunately, at present, there exists no conventional model describing the penetration of the magnetic flux into an anisotropic superconductor. A phenomenological description of this process based on a nonlinear tensor relation between the current density and the electrical field was given in [18–20]. Recently [21], a new approach to solving electrodynamic problems on the critical state of hard superconductors based on the principle of the minimum of a certain energy functional was proposed. The authors of [21] presume that their approach has a fundamental character and can be applied to solving any problems of a critical state, in particular, to the description of anisotropic superconductors. This conclusion is justified by the fact that the minimum principle used in [21] is completely equivalent to the Prigogine principle of the minimum entropy production rate. However, it should be noted that the applicability of the Prigogine principle to a nonlinear strongly nonequilibrium system (a hard superconductor in the critical state represents such a system) should be specially proved in every case, because it is not valid in the general case. Therefore, the method developed in [21], which is convenient for numerical calculations, cannot be considered as fundamentally justified. Due to this circumstance, we will remain within the framework of the critical state model [18–20] generalized to the anisotropic case. The application of this model, in contrast with model [21], allows us to obtain an analytic solution to the problem and to trace the physical nature of the investigated phenomenon. In addition, it occurs that model [18–20] in a number of cases agrees well (both qualitatively and quantitatively) with experiment [20].

Consider an infinite plane-parallel superconducting plate of thickness d located in external constant magnetic field $\mathbf{H}$ and alternating magnetic field $\mathbf{h}(t) = \mathbf{h}_0\cos(\omega t)$, which are mutually perpendicular and parallel to the plate surface. We assume that all fields and currents depend on only one spatial coordinate x directed along the normal to the plate. The origin $x = 0$ is located at the center of the sample. We will study the simplest case of anisotropy, when it is characterized by two principal directions y and z in the plane of the plate. These directions coincide either with the crystallographic axes of the sample or with the representative directions of the defect structure. In this case, the equations of the generalized model of the critical state for magnetic induction $\mathbf{B}$ can be written in the form

$$\begin{aligned} \frac{\partial B_z}{\partial x} &= -\frac{4\pi}{c}J_{cy}(B_y, B_z)\cos(\phi(x)), \\ \frac{\partial B_y}{\partial x} &= -\frac{4\pi}{c}J_{cz}(B_y, B_z)\sin(\phi(x)), \end{aligned} \tag{2}$$

where $\phi(x)$ is the angle between electric field $\mathbf{E}$ and the y axis. The electric field is connected to the magnetic induction by the Faraday law:

$$\frac{\partial E_y}{\partial x} = -\frac{1}{c}\frac{\partial B_z}{\partial t}, \quad \frac{\partial E_z}{\partial x} = \frac{1}{c}\frac{\partial B_y}{\partial t}. \tag{3}$$

The boundary conditions to the Maxwell equations have the form

$$\begin{aligned} B_z(-d/2) &= B_z(d/2) \\ &= H\cos\theta - h_0\sin\theta\cos(\omega t), \\ B_y(-d/2) &= B_y(d/2) \\ &= H\sin\theta + h_0\cos\theta\cos(\omega t), \end{aligned} \tag{4}$$

where θ is the angle between vector $\mathbf{H}$ and the z axis.

Below, the solution to the Maxwell equations with boundary conditions will be used for determining the constant component of the projection of the magnetic moment of the sample on the direction of external constant magnetic field $\mathbf{H}$,

$$M = \frac{1}{4\pi}\left[\frac{\omega}{2\pi d}\int\limits_0^{2\pi/\omega} dt \times \int\limits_{-d/2}^{d/2} (B_y(x)\sin\theta + B_z(x)\cos\theta dx - H)\right]. \tag{5}$$

By virtue of the symmetry of the problem, it is sufficient to solve the system of equations on the space interval $0 \le x \le d/2$ and take into account in formula (5) the fact that function $\mathbf{B}(x)$ is even. For the sake of simplifying of the calculations, we neglect the dependence of the components $J_{ci}(\mathbf{B}(x))$ of the critical current density on the x coordinate caused by the nonuniformity of the magnetic induction distribution and assume that $J_{ci}(\mathbf{B}(x)) = J_{ci}(\mathbf{H})$. This is justified for small amplitudes $h_0 \ll H$ if the induction in the sample slightly differs from the magnitude of the external constant magnetic field. In addition, we replace the x-coordinate-dependent angle ϕ between the electric field and the y axis by angle $\pi/2 + \theta$ between the external alternating magnetic field and the z axis (as in the anisotropic situation). It was shown in [20] that this simplification has no qualitative influence on the results obtained.

3.2. Calculation of the Magnetization Collapse and Discussion

The effect of suppression of the sample magnetization by an alternating magnetic field occurs due to the field penetration into the superconductor volume and the field interaction with static screening currents. Therefore, first of all, it is necessary to determine the depth of the penetration of field $h(t)$. It is convenient to perform the calculations in the following dimensionless variables:

$$\begin{aligned} &\xi = 2x/d, \quad \mathcal{H} = H/H_p, \quad g_0 = h_0/H_p, \\ &b_i = B_i/H_p, \quad H_p = 2\pi J_{cy}d/c, \quad \delta = J_{cz}/J_{cy} < 1. \end{aligned} \tag{6}$$

Here and below, we measure the values of all magnetic fields in units of penetration field H_p. Field H_p is the value of field H directed along the z axis, which, according to the critical state equation (2), penetrates to the middle of the sample. It should be borne in mind that quantity H_p has different values at different points of the magnetization loop, because, at each point, J_{cy} takes a certain value corresponding to the current magnitude of the external constant magnetic field $\mathbf{H}$. For definiteness, we assume that the critical current density along the z axis is less than the current density along the y axis; i.e., the anisotropy parameter δ is less than unity.

One can easily show that, at all places where the alternating field $h(t)$ penetrates, the magnetic induction can be represented as a sum of two terms. One of these terms is a constant homogeneous quantity coinciding with vector $\mathcal{H}$. The second term $\tilde{b}_i$ varies in time and is determined by the following equations:

$$\frac{\partial \tilde{b}_z}{\partial \xi} = \sin\theta, \quad \frac{\partial \tilde{b}_y}{\partial \xi} = \delta\cos\theta. \tag{7}$$

These equations hold in the superconductor region, where both induction components $\tilde{b}_y$ and $\tilde{b}_z$ are present. In the region where component $\tilde{b}_z$ vanishes and there is the only induction component $\tilde{b}_y$, the distribution of this component is described by the equation

$$\frac{\partial \tilde{b}_y}{\partial \xi} = \delta. \tag{8}$$

The boundary conditions to Eqs. (7) and (8) are the equalities

$$\begin{aligned} \tilde{b}_y(\xi = 1) &= g_0\cos(\omega t)\cos\theta, \\ \tilde{b}_z(\xi = 1) &= g_0\cos(\omega t)\sin\theta. \end{aligned} \tag{9}$$

At the instants when the alternating field most deeply penetrates into the sample (for $\cos(\omega t) = 1$), the system of equations (7) and (8) with boundary conditions (9) gives the following distribution of inductions $\tilde{b}_y(\xi)$ and $\tilde{b}_z(\xi)$:

$$\tilde{b}_y(\xi) = \begin{cases} \cos\theta(\delta(\xi - 1) + g_0), & \xi_z < \xi < 1, \\ \delta(\xi - 1) + g_0(\cos\theta(1 - \delta) + \delta), & \xi_y < \xi < \xi_z, \\ 0, \quad 0 < \xi < \xi_y, \end{cases} \tag{10}$$

$$\tilde{b}_z(\xi) = \begin{cases} \sin\theta(\xi - 1 + g_0), & \xi_z < \xi < 1, \\ 0, \quad 0 < \xi < \xi_z. \end{cases} \tag{11}$$

Here, ξ_y and ξ_z are space coordinates determining the penetration boundaries for the y and z components of the magnetic induction:

$$\xi_y = 1 - g_0\left(1 + \left(\frac{1}{\delta} - 1\right)\cos\theta\right), \quad \xi_z = 1 - g_0. \tag{12}$$

One can see from the latter formulas that, as can be expected, the y component of the magnetic induction, which is screened by a small component of the critical current density, penetrates more deeply into the sample as compared with the z component. Therefore, it is at the instant when the y component of the alternating magnetic field attains the middle of the sample that the distribution of the constant component of the induction becomes spatially uniform all over the superconductor; i.e., static magnetization collapse occurs. We determine amplitude $g_0^{(\mathrm{col})} = h_0^{(\mathrm{col})}/H_p$ by equating ξ_y to zero:

$$g_0^{(\mathrm{col})} = \frac{\delta}{\delta + (1 - \delta)\cos\theta}. \tag{13}$$

Function $g_0^{(\mathrm{col})}(\theta)$ has a universal character; it is defined via the unique parameter δ of the theory. However, it should be taken into account that, in principle, the anisotropy parameter may depend on the orientation of the external constant magnetic field, $\delta = \delta(\theta)$. According to formula (13), the amplitude of the alternating field at which the total magnetization collapse occurs monotonically increases by a factor of δ^{-1} with an increase in angle θ from zero to $\pi/2$. This conclusion agrees well with the measurement results (see Fig. 5). In this figure, the theoretical curve is calculated disregarding a possible dependence $\delta(\theta)$.

Let us now consider the dynamics of suppression of the static magnetic moment of the sample by an orthogonal alternating magnetic field. As has been noted, the suppression effect essentially depends on the magnetic prehistory of the sample, i.e., on the position of the starting point on the magnetization loop. We begin with the simplest case of the magnetic prehistory, when the external magnetic field monotonically increased up to a certain maximal value $H_m \gg H_p$ (in dimensionless units, up to $\mathcal{H}_m = H_m/H_p \gg 1$). As has been shown, the switching on of the alternating field results in establishing a uniform distribution of static magnetic induction in the entire region $\xi_y < \xi < 1$, where field $h(t)$ penetrates. In the bulk of the sample, the critical induction profile created by the screening currents J_{cy} and J_{cz} (in dimensionless units, 1 and δ) is preserved. Thus, the magnetic induction distribution is described by the following formulas:

$$b_y(\xi) = \begin{cases} \mathcal{H}\sin\theta, & \xi_y < \xi < 1, \\ \mathcal{H}\sin\theta + \delta(\xi - \xi_y)\sin\theta, & 0 < \xi < \xi_y, \end{cases} \tag{14}$$

$$b_z(\xi) = \begin{cases} \mathcal{H}\cos\theta, & \xi_y < \xi < 1, \\ \mathcal{H}\cos\theta + (\xi - \xi_y)\cos\theta, & 0 < \xi < \xi_y. \end{cases} \tag{15}$$

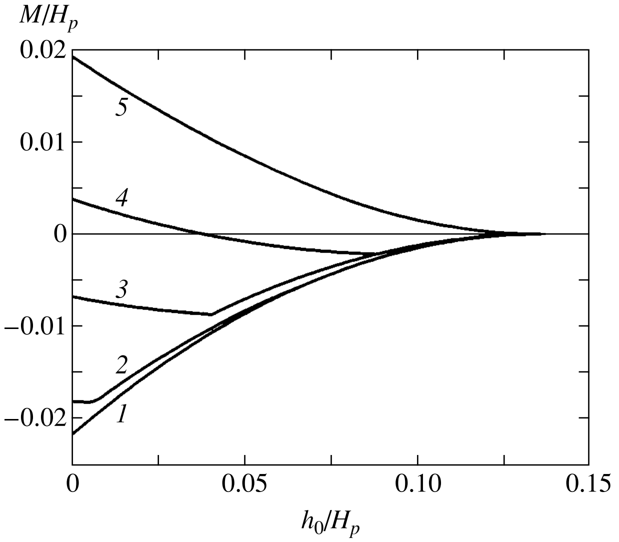


Fig. 7. Calculated dynamics of the suppression of the magnetic moment *m*. Curves *1–5* correspond to the values of the dimensionless constant magnetic field $\mathcal{H} = \mathcal{H}_m = 5$, 4.85, 4.7, 4.35, and 3.5. Anisotropy parameter $\delta = 0.1$.

Dimensionless magnetization $m = M/H_p$ corresponding to this distribution is defined by the expression

$$m = -\frac{\cos^2\theta + \delta\sin^2\theta}{8\pi}\left(1 - \frac{g_0}{g_0^{(\mathrm{col})}(\theta)}\right)^2. \tag{16}$$

The results of the magnetization calculation by this formula are presented in Fig. 7 (curve *1*). One can see that the diamagnetic moment *m* is monotonically suppressed with increase in amplitude g_0 and vanishes at $g_0 = g_0^{(\mathrm{col})}$. This behavior of the magnetization agrees with the measurement results (see curve *1* in Fig. 6).

For a more complex magnetic prehistory of the sample, the picture of the magnetization collapse becomes nontrivial. Unfortunately, due to an infinite number of ways for establishing the initial distribution of the magnetic induction, it is impossible to exhaustively analyze all kinds of dependence $m(g_0)$. Therefore, we demonstrate here a method for calculating $m(g_0)$ for one of the simplest magnetic prehistories tracing different stages in the development of the collapse effect. For this purpose, we consider the case where a monotonic increase in field $\mathcal{H}$ changes to a decrease. Even in this simple situation, the character of the initial distribution of the magnetic induction and, hence, the development of collapse are qualitatively different in different variation domains of field $\mathcal{H}$. Therefore, as an example, we consider one of such intervals,

$$\mathcal{H}_m - 1 < \mathcal{H} < \mathcal{H}_m - 2\delta, \quad \mathcal{H}_m - 2 + \sqrt{2(1+\delta)}, \tag{17}$$

when the initial distribution profile of the magnetic induction has the form schematically depicted in Fig. 8a. For the sake of simplicity, we take $\theta = \pi/4$ and $\delta < 1/2$. Under conditions (17) and in the absence of the alternating magnetic field, the sample is in the paramagnetic state. As the amplitude of the alternating magnetic field increases in the interval

$$0 < g_0 < g_0^{(\mathrm{col})}(\mathcal{H}_m - \mathcal{H}), \tag{18}$$

the profile of the magnetic induction distribution is deformed as is shown in Fig. 8b. Near the sample surface, a region arises in which the induction is uniform. This deformation of the profile results in a decrease in the magnetic moment, which becomes negative at a cer-

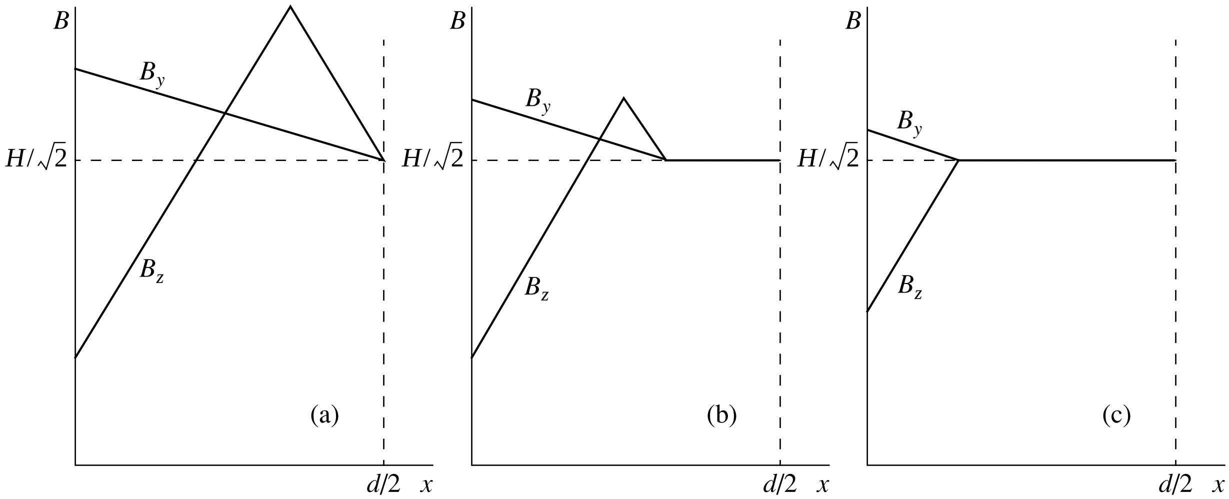


Fig. 8. Schematic representation of the profiles of the magnetic induction distribution in a sample at different stages of the development of the magnetization collapse.

tain value of the alternating field amplitude; i.e., the sample passes to the diamagnetic state. In the entire interval (18), the magnetic moment monotonically decreases with the growth of g_0 according to the parabolic law

$$8\pi m = 4(\mathcal{H}_m - \mathcal{H})\left(1 - \frac{g_0}{2g_0^{(\mathrm{col})}}\right) + 2\delta\left(1 - \frac{g_0}{2g_0^{(\mathrm{col})}}\right)^2 - (\mathcal{H}_m - \mathcal{H})^2 - 2 + \left(\frac{g_0}{g_0^{(\mathrm{col})}}\right)^2. \tag{19}$$

In the interval

$$g_0^{(\mathrm{col})}(\mathcal{H}_m - \mathcal{H}) < g_0 < g_0^{(\mathrm{col})}, \tag{20}$$

due to the extension of the region where the induction is uniform, the typical paramagnetic “tooth” on the distribution profile $b_z(\xi)$ disappears (see Fig. 8c). In this interval of amplitudes, as g_0 grows, the magnetization is monotonically suppressed down to its total vanishing. Here, the magnetic moment is described by formula (16), in which the sign of δ must change. The results of the magnetization calculation are presented in Fig. 7 (curve *4*). The typical break of the curve at the border of intervals (18) and (20) is worth noting.

The analysis shows that, in the interval of the magnetic fields $\mathcal{H}_m - 1 < \mathcal{H} < \mathcal{H}_m$, moment m in the amplitude function g_0 varies nonmonotonically (see curves *2–4* in Fig. 7). For $\mathcal{H} < \mathcal{H}_m - 1$, the paramagnetic moment monotonically decreases with the growth of the alternating field amplitude (curve *5* in Fig. 7). The comparison of the curves in Figs. 6 and 7 shows a good qualitative agreement of the theory and experiment. The absence of breaks on the experimental $M(h_0)$ curves can be connected with the finite size of the sample and its possible inhomogeneity.

4. CONCLUSIONS

Our investigation shows that the alternating magnetic field exerts a significant influence on the static magnetic properties of anisotropic hard superconductors. Switching on a sufficiently strong field $\mathbf{h}(t)$ orthogonal to the static magnetizing field results in the complete suppression of the magnetic moment of the sample. The reason for the magnetization collapse is that, at all places where the alternating field penetrates, leveling of the distribution profile of the static magnetic induction is observed. In other words, in the same spatial region of the sample, the constant and alternating screening currents cannot coexist. In the conditions when the alternating field penetrates into the entire volume of the sample, complete suppression of the static magnetization takes place. The nature of the collapse, which consists in a local effect of the mutual influence of different components of the critical current density vector, is manifested in the anisotropic situation in a rather peculiar way. Different components of the magnetic field penetrate at different depths, since they are screened by critical current density components of quite different magnitudes. This is the reason why the magnetization collapse is primarily caused by the alternating field component deeply penetrating into the sample. As a result, the anisotropy induces a quite unexpected effect: to suppress a large magnetic moment arising under the conditions of screening by a strong current, a small amplitude of the alternating signal is sufficient. This means that the instability of the critical profile of the vortex density distribution is most clearly manifested precisely in the anisotropic case, In the paper, the dynamics of the magnetization collapse with the increase in the amplitude of the alternating field h_0 is also studied. It was found that, in some cases, the dependence of the moment on h_0 is nonmonotonic. In addition, during the collapse, a quite peculiar transition of the sample from the paramagnetic state into the diamagnetic state sometimes occurs. This transition is stimulated by an external orthogonal alternating magnetic field. All discovered peculiarities of the effect of magnetization collapse can be interpreted within the framework of a simple model of critical state generalized to the anisotropic case.

ACKNOWLEDGMENTS

This study was carried out under the Russian Federal Goal Program on Superconductivity (contract no. 40.012.1.1.11.46) and was supported by the International Association for the Promotion of Cooperation with Scientists from the New Independent States of the Former Soviet Union (INTAS) (project no. 01-2282) and the Russian Foundation for Basic Research (project no. 00-02-17145).

REFERENCES

1. V. K. Vlasko-Vlasov, V. I. Nikitenko, A. A. Polyanskii, *et al.*, Physica C (Amsterdam) **222**, 361 (1994).
2. M. V. Indenbom, Th. Schuster, M. R. Koblischka, *et al.*, Physica C (Amsterdam) **209**, 259 (1993).
3. T. Frello, M. Baziljevich, T. H. Johansen, *et al.*, Phys. Rev. B **59**, R6639 (1999).
4. L. M. Fisher, P. E. Goa, M. Baziljevich, *et al.*, Phys. Rev. Lett. **87**, 247005-1 (2001).
5. A. K. Niessen and C. H. Weijsenfeld, J. Appl. Phys. **40**, 384 (1969).
6. H. Pastoriza, S. Candia, and G. Nieva, Phys. Rev. Lett. **83**, 1026 (1999).
7. I. V. Baltaga, V. V. Borisovskii, L. M. Fisher, *et al.*, in *AIP Conference Proceedings on Superconductivity and Its Applications*, Ed. by Yi-Han-Kao, Philip Coppers, and Hoi-Sing Kwok (AIP, New York, 1991), p. 153, AIP Conf. Proc., No. 219.

8. I. V. Baltaga, L. M. Fisher, N. M. Makarov, *et al.*, Fiz. Nizk. Temp. **21**, 411 (1995) [Low Temp. Phys. **21**, 320 (1995)].
9. L. M. Fisher, A. V. Kalinov, I. F. Voloshin, *et al.*, Solid State Commun. **97**, 833 (1996).
10. S. J. Park, S. J. Kouvel, H. B. Radousky, and J. Z. Liu, Phys. Rev. B **48**, 13998 (1993).
11. I. F. Voloshin, A. V. Kalinov, S. E. Savel'ev, *et al.*, Zh. Éksp. Teor. Fiz. **111**, 1071 (1997) [JETP **84**, 592 (1997)].
12. L. M. Fisher, A. V. Kalinov, S. E. Savel'ev, *et al.*, Physica C (Amsterdam) **278**, 169 (1997).
13. L. M. Fisher, K. V. Il'enko, A. V. Kalinov, *et al.*, Phys. Rev. B **61**, 15382 (2000).
14. I. F. Voloshin, N. V. Il'in, N. M. Makarov, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **53**, 109 (1991) [JETP Lett. **53**, 115 (1991)].
15. C. P. Bean, Phys. Rev. Lett. **8**, 250 (1962).
16. S. E. Savel'ev, L. M. Fisher, and V. A. Yampol'skiĭ, Zh. Éksp. Teor. Fiz. **112**, 936 (1997) [JETP **85**, 507 (1997)].
17. L. M. Fisher, S. E. Savel'ev, and V. A. Yampol'skii, Physica B (Amsterdam) **284–288**, 735 (2000).
18. L. M. Fisher, A. V. Kalinov, S. E. Savel'ev, *et al.*, Physica C (Amsterdam) **350**, 152 (2001).
19. I. F. Voloshin, A. V. Kalinov, L. M. Fisher, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 324 (2001) [JETP Lett. **73**, 285 (2001)].
20. I. F. Voloshin, A. V. Kalinov, L. M. Fisher, *et al.*, Zh. Éksp. Teor. Fiz. **120**, 1273 (2001) [JETP **93**, 1105 (2001)].
21. A. Badia and C. Lopez, Phys. Rev. Lett. **87**, 127004 (2001).

Translated by E. Pankratiev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 154–161.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 172–179.
Original Russian Text Copyright © 2003 by Marchenko, Podolyak.

SOLIDS
Electronic Properties

Phase Diagram of Surface Superconductivity

V. I. Marchenko* and E. R. Podolyak

Kapitza Institute of Physical Problems, Russian Academy of Sciences,
Moscow, 119334 Russia

**e-mail: mar@kapitza.ras.ru*

Received January 31, 2003

Abstract—The Ginzburg–Landau theory is used to numerically analyze the stability of surface superconductivity. The singularities in the behavior of the solution on approaching the stability limit are described within the Landau theory of second-order phase transitions applied to a metastable state. It is found that the wetting must be observed upon transition from type I to type II superconductors: the thickness of surface superconductivity goes to infinity when the magnetic field tends to critical. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Saint-James and de Gennes [1] demonstrated the instability of the normal phase of a superconductor relative to the emergence of a layer of superconducting state in the vicinity of the sample boundary in fields below $H_{c3} = 2.3965\kappa H_c$. As was found by Feder [2], the critical value of the Ginzburg–Landau parameter $\kappa_F \approx 0.41$ exists, which separates superconductors by their behavior in fields below H_{c3}. For materials with $\kappa < \kappa_F$, no equilibrium solution of equations exists for the surface layer; i.e., the field H_{c3} is a field of supercooling of the normal state. In the case of materials with $\kappa > \kappa_F$, the surface superconductivity with a finite thickness on the order of the Ginzburg–Landau coherence length must be observed in a field below H_{c3}. The critical point found by Feder (we will refer to it as the Feder point) is in the region of metastability of the normal state $H < H_c$.

The supercooling of surface superconductivity was experimentally found as a result of measurements of the hysteresis of magnetization curves in tantalum and lead by McEvoy, Jones, and Park [3]. Khlyustikov and Berezin [4] have recently observed the effect of supercooling of surface superconductivity in lead when measuring the surface impedance.

Therefore, some region of metastability exists in the (κ, H) phase diagram of surface superconductivity. Park [5] determined the boundary $H_s(\kappa)$ of this region for the values of the Ginzburg–Landau parameter from the Feder point to $\kappa = 0.575$ where the limit approached the $H = H_c$ line. In this paper, we will clarify the further behavior of this boundary (see Fig. 1).

In order to determine the range of existence of surface superconductivity, we will perform a complete investigation of the stability (to small perturbations) of surface superconductivity. We will find that, for type I superconductors, the boundary is defined by the loss of stability of the surface solution relative to uniform (in the boundary plane) perturbations. It turns out that, at $\kappa > 0.575$, the field of supercooling continues to be pressed against the critical field H_c and becomes equal to H_c only at the point $\kappa = 1/\sqrt{2}$. For type II superconductors, the $H_{c2}(\kappa)$ serves as the stability boundary.

We will demonstrate that the singularities in the behavior of the solution on approaching the line of the loss of stability in type I superconductors may be understood from very general considerations within the Landau theory of second-order phase transitions applied to metastable states. In the neighborhood of the point ($1/\sqrt{2}$, H_c), the thickness of the superconducting layer diverges; i.e., a wetting transition occurs. In this case, we suggest a macroscopic approach enabling one to explain the very unusual nonanalytical behavior of the surface solution.

2. THE STABILITY REGION OF SURFACE SUPERCONDUCTIVITY

In the units of $H_c = \delta = 1$, which are natural for the Ginzburg–Landau theory (see, for example, Paras. 45 and 46 in [6]), the free energy of a superconductor is

$$F = \frac{1}{8\pi} \times \int\left\{-2|\psi|^2 + |\psi|^4 + \left|\left(\frac{\sqrt{2}}{\kappa}\nabla - i\mathbf{A}\right)\psi\right|^2 + \mathbf{B}^2\right\}dV. \tag{1}$$

In the problem being treated, the external field **H** is directed along the sample boundary (z axis). For an equilibrium solution localized in the vicinity of the sur-

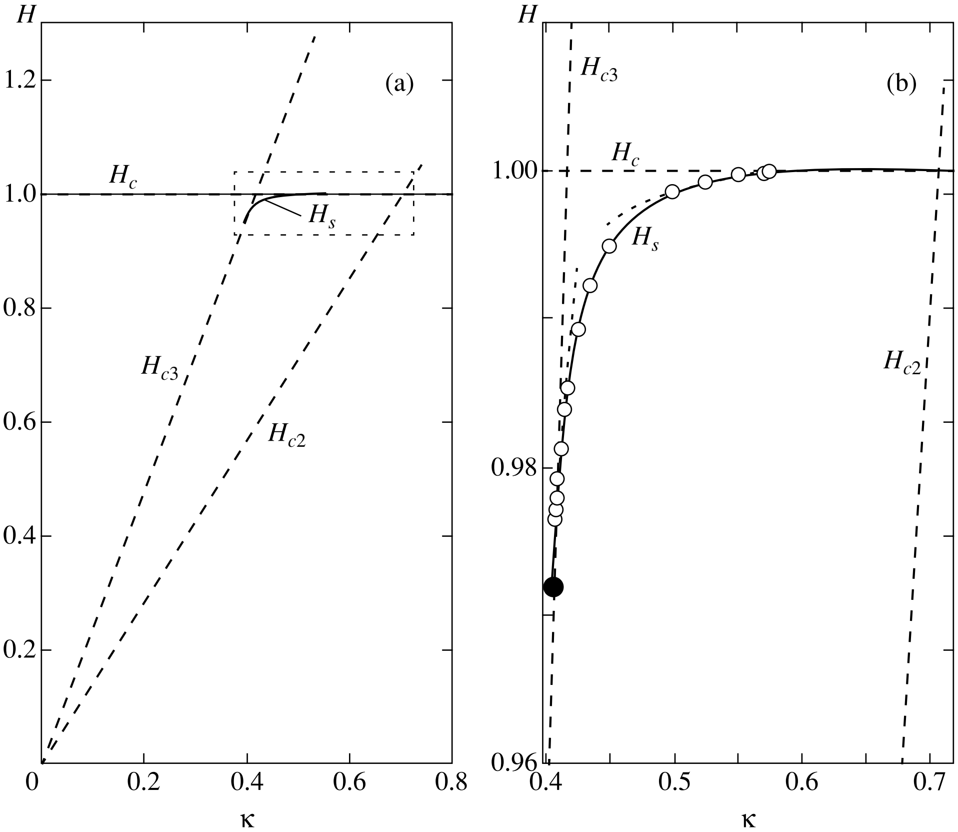


Fig. 1. (a) Phase diagram of surface superconductivity and (b) a part of the H_s curve indicated by dashed contour. The dark circle indicates the Feder point $\kappa_F = 0.4054$. The light circles indicate Park's results [5]. The dashed lines indicate the asymptotes in the vicinity of the Feder point: $H_s = 0.9623 + 0.211(\kappa - 0.4035)^{1/2}$ (see Section 2), and in the vicinity of the point $(1/\sqrt{2}, H_c)$: $H_s = 1 - 0.304\varepsilon e^{-0.793/\varepsilon}$ (see Section 4).

face, we select the gauge for the vector potential **A** such that the order parameter ψ should be a real function of distance x from the boundary,

$$A_y(\mathbf{r}) = A(x), \quad A_x(\mathbf{r}) = A_z(\mathbf{r}) = 0.$$

We will investigate the stability (to small perturbations) of the equilibrium solution $\{\psi(x), A(x)\}$ in the general case of nonuniform perturbations. For this purpose, it is necessary to check the positiveness of the variation of free energy in the case of arbitrary small fluctuations of the order parameter $\delta\psi(\mathbf{r})$ and of the vector potential $\delta\mathbf{A}(\mathbf{r})$. It is natural to expect that the most dangerous fluctuations will be those in which the magnetic field variation $\delta\mathbf{B}(\mathbf{r})$ will remain parallel to the external field **H** (i.e., the magnetic field lines will not bend). We use the gauge invariance to select $\delta\mathbf{A}(x, y) = (0, \delta A(x, y), 0)$. We expand the perturbation $\{\delta\psi(x, y), \delta A(x, y)\}$ into a Fourier series in $\exp(iqy)$ to represent the second variation of free energy as

$$\delta^2 F = \frac{1}{4\pi}\sum_q \int_0^\infty dx \left\{ \frac{1}{\kappa^2}|\partial_x \delta\psi_q|^1 + |\delta\psi_q|^2\left(\frac{A}{\sqrt{2}} - \frac{q}{\kappa}\right)^2 \right.$$

$$+ \psi\left(A - \frac{q}{\sqrt{2}\kappa}\right)(\delta A_q \delta\psi_q^* + \delta A_q^* \psi_q) + |\delta\psi_q|^2(\psi^2 - 1) \tag{2}$$

$$\left. + \psi^2|\delta\psi_q + \delta\psi_{-q}^*|^2 + |\partial_x A_q|^2 + \psi^2|A_q|^2 \right\}.$$

The quantity $\delta A(x, y)$ is a real function; therefore, $\delta A_{-q}(x) = \delta A_q^*(x)$. No such correlation exists for the components of $\delta\psi_q$, and expression (2) contains cross terms $\delta\psi_q$ and $\delta\psi_{-q}^*$. We introduce the Fourier components of the real, $\delta\psi'$, and imaginary, $\delta\psi''$, parts of $\delta\psi$,

$$\delta\psi'_q = \frac{1}{2}(\delta\psi_q + \delta\psi_{-q}^*), \quad \delta\psi''_q = \frac{1}{2}(\delta\psi_q + \delta\psi_{-q}^*),$$

and proceed to perform the summation over positive values of q to reduce the second variation (2) to the sum of independent contributions $\delta^2 F_q\{\delta\psi'_q, \delta\psi''_q, \delta A_q\}$. In order to investigate the stability, it is sufficient to find

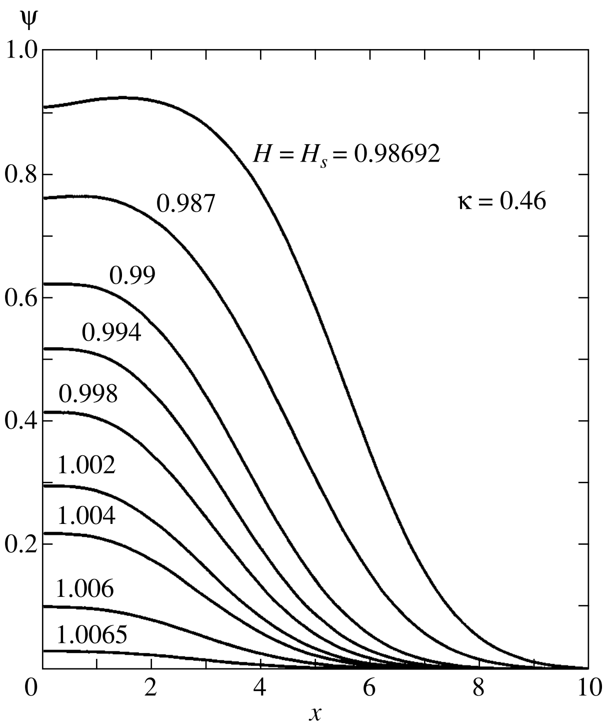


Fig. 2. The evolution of the Ginzburg–Landau ψ-function for surface superconductivity upon variation of the field from H_{c3} to H_s at $\kappa = 0.46$.

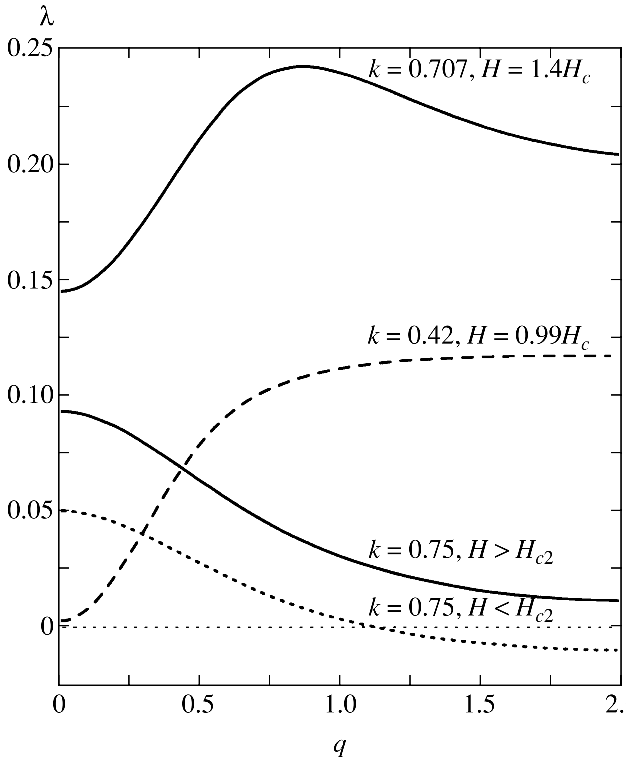


Fig. 3. Examples of spectra of eigenvalues $\lambda(q)$.

the minimum of quadratic functional $\delta^2 F_q$ with a fixed value of

$$\int_0^\infty |\delta\psi_q'|^2 dx. \qquad (3)$$

This formulation of the problem turns out to be convenient from the standpoint of numerical computation, where the asymptotic behavior of eigenfunctions at $x \longrightarrow \infty$ is of importance. Because of their awkwardness, we will not give here the appropriate Euler equations with natural (i.e., obtained as a result of variation) boundary conditions. They were numerically solved using the method of "shooting" from the point $x \longrightarrow \infty$ to the point $x = 0$. Three parameters were preassigned on the right-hand boundary, namely, the Lagrange multiplier λ_q corresponding to the constancy of integral (3) and the amplitudes of the asymptotics $\delta\psi_{-q}^*$ and δA_q. By virtue of the linearity of the problem, the amplitude of $\delta\psi_q$ may be arbitrary. On the left-hand boundary with $x = 0$, three conditions were to be met, namely, $\partial_x \delta\psi_q = \partial_x \delta\psi_{-q}^* = \partial_x \delta A_q = 0$. The latter of these conditions follows from the trivial solution $\delta\mathbf{B} = 0$ in the treated case of Maxwell equations ($\partial_x B_z = \partial_y B_z = 0$) outside the superconductor ($x < 0$).

The problem involves an infinite set of discrete levels λ (in the case of $\psi(x) = 0$, these are simply electron levels in a uniform magnetic field) for each value of the wave vector q. In order to analyze the stability, it is sufficient to find the ground level for which the eigenfunctions have no zeros. Figure 2 demonstrates the typical evolution of the order parameter of surface superconductivity upon variation of the field. Figure 3 gives the results of numerical counting for the spectrum of λ_q for different values of κ and H.

In type I superconductors in some field of supercooling H_s such that $H_{c2} < H_s < H_c$, the surface state loses stability at $q = 0$ and the unstable mode turns out to be localized in the region of surface superconductivity. This instability leads to the propagation of superconductivity from the surface into the bulk of the sample.

In type II superconductors at $q \longrightarrow \infty$, the function $\delta\psi_q(x)$ is localized at distances far away from the surface in the neighborhood of the point $x = x_0$, where $A(x_0) = \sqrt{2}\,q/\kappa$. Here, the function $\psi(x_0)$ is small, and the field $B(x_0)$ is almost equal to the external field. Therefore, the stability of the surface state deteriorates because of nucleation in the normal-phase volume; i.e., the stability limit of surface superconductivity for type II superconductors coincides with the stability limit of the normal state in the volume $H = H_{c2}$.

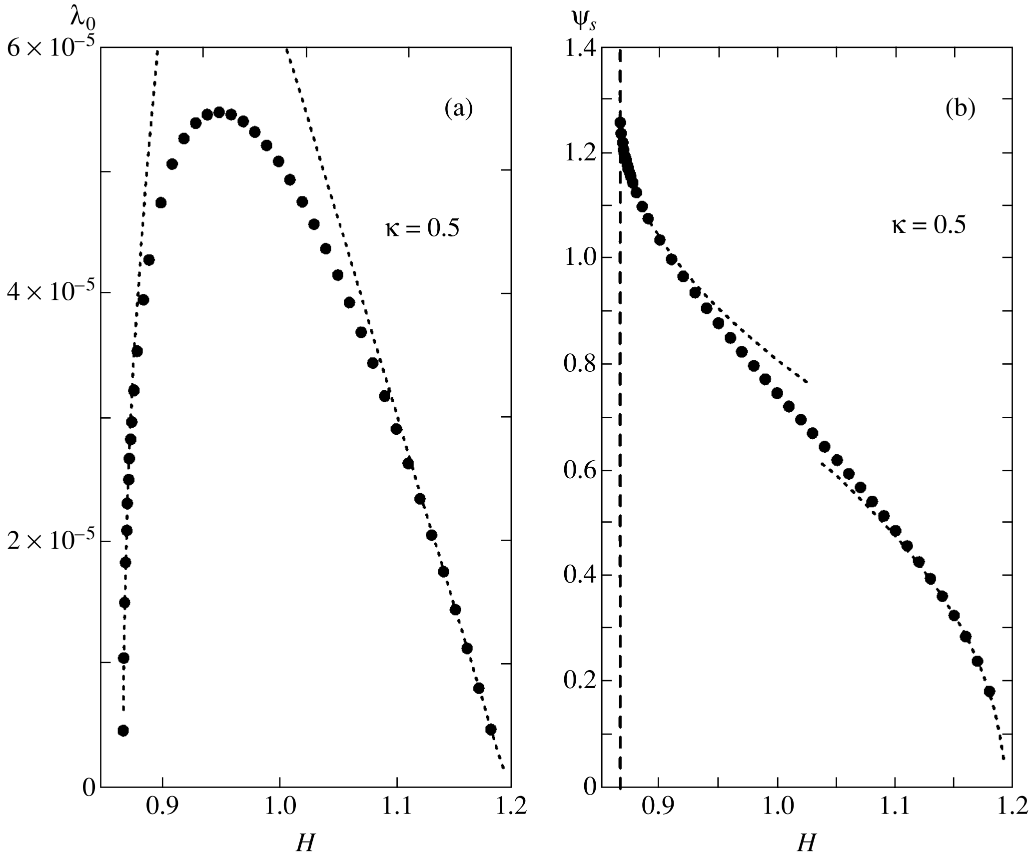


Fig. 4. The evolution of the "gap" in the excitation spectrum and of the integral order parameter ψ_s upon variation of the field from H_{c3} to H_s at $\kappa = 0.5$. The dotted line indicates the asymptotics in the vicinity of $H_{c3} = 1.19$: $\lambda_0 = 3.19 \times 10^{-4}(H_{c3} - H)$, $\psi_s = 1.56(H_{c3} - H)^{1/2}$ (see Section 2); and in the vicinity of $H_s = 0.866$: $\lambda_0 = 3.35 \times 10^{-4}(H - H_s)^{1/2}$, $\psi_s = 1.276 - 1.28(H - H_s)^{1/2}$ (see Section 3).

Figures 4–6 give the results of computation of the evolution of the quantity λ_0 ($q = 0$) and of the integral order parameter,

$$\psi_s = \int_0^\infty \psi dx, \tag{4}$$

during motion over the field from H_{c3} to H_s at $\kappa = 0.5$ and $\kappa = 1/\sqrt{2}$, as well as during motion with respect to κ with the critical field H_c.

3. THE FEDER POINT IN THE LANDAU THEORY

The emergence of surface superconductivity in the vicinity of the field H_{c3} is an example of second-order phase transition in a two-dimensional system. The specific features in the behavior of the stability limit of metastable state in the vicinity of the Feder point may be clarified in the Landau theory of second-order phase transitions. The fact that the curve of the loss of stability of surface superconductivity turns out to be tangent to the straight line $H_{c3}(\kappa)$ at the Feder point is a property that is common to critical points of this type.

The Landau expansion for free energy of a metal surface has the standard form

$$A(\kappa, H)|\psi_s|^2 + B(\kappa, H)|\psi_s|^4 + C(\kappa, H)|\psi_s|^6 + \dots. \tag{5}$$

Note that the vector potential in the case of such (two-dimensional) description is of no significance if the field is parallel to the surface. For comparison with the results of numerical solution of Ginzburg–Landau equations, we will take the integral characteristic given by Eq. (4) as the order parameter of two-dimensional theory ψ_s. The range of validity of the two-dimensional approach away from the Feder point, when the effect of the sixth-order contribution in expansion (5) may be ignored, is seen in Figs. 4–6. Asymptotic curves are given in the vicinity of the field H_{c3}; these curves follow from expression (5) on the assumption (made in the Landau theory) that the coefficient A linearly goes to zero. It is clear that, in the vicinity of the field H_{c3}, the second derivative of free energy (5) with respect to the

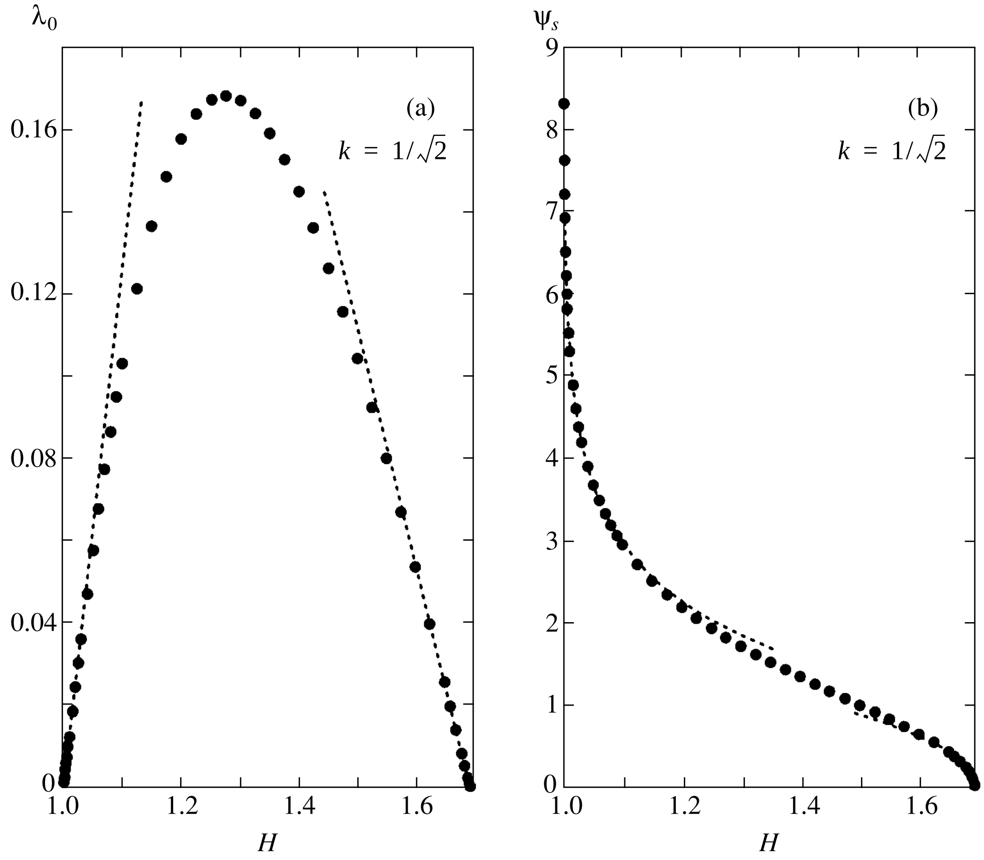


Fig. 5. The evolution of the "gap" in the excitation spectrum and of the integral order parameter ψ_s upon variation of the field from H_{c3} to H_s at $\kappa = 1/\sqrt{2}$. The dotted line indicates the asymptotics in the vicinity of $H_{c3} = 1.70$: $\lambda_0 = 0.576(H_{c3} - H)$, $\psi_s = 2.03(H_{c3} - H)^{1/2}$ (see Section 2); and in the vicinity of $H_c = 1$: $\lambda_0 = 1.27(H - 1)$, $\psi_s = -2.33\ln[0.52(H - 1)]$ (see Section 4).

two-dimensional order parameter ψ_s, which defines the stability of the solution, is linearly related to λ_0.

At the Feder point, the coefficient B obviously goes to zero (again by the linear law). In order to describe the behavior in the vicinity of this point, the sixth-order term must be taken into account. Here, however, in contrast to the critical point treated by Landau [7] (see also Para. 150 in [8]), the coefficient C must be taken to be negative, which corresponds to the metastability of surface superconductivity being investigated. With regard for the sixth-order term, we derive, in addition to the trivial $\psi_s = 0$, two possible extrema of functional (5):

$$|\psi_{s1}|^2 = \frac{-B - \sqrt{B^2 - 3AC}}{3C},$$
$$|\psi_{s2}|^2 = \frac{-B + \sqrt{B^2 - 3AC}}{3C}. \qquad (6)$$

One solution ψ_{s1} exists at $A > 0$ and $B > 0$. This solution corresponds to a local maximum ensuring the metastability of normal state. With $B > 0$ and a change of sign for A, a new solution ψ_{s2} arises, which corresponds to a local minimum, i.e., to surface superconductivity. With further variation of A, when the quantity $B^2 - 3AC$ changes sign, both solutions coincide and the nontrivial extrema disappear. The equation

$$B^2 = 3AC \qquad (7)$$

apparently corresponds to the critical curve of $H_s(\kappa)$. One can readily see that it is the quadratic dependence given by Eq. (7) between the small coefficients A and B that leads to the characteristic root approximation of curve $H_s(\kappa)$ to line $H_{c3}(\kappa)$ at the Feder point (see Fig. 1). One solution ψ_{s2}—a local maximum—exists in the region of $A > 0$ and $B > 0$. On the $A = 0$ line, this solution goes to zero and only one trivial extremum—a maximum—remains, $\psi_s = 0$.

4. SPECIAL FEATURES IN THE LOSS OF STABILITY IN METASTABLE STATE

A loss of stability of the surface solution relative to uniform perturbation ($q = 0$) occurs on the $H_s(\kappa)$ line. This means that two solutions merge on this line,

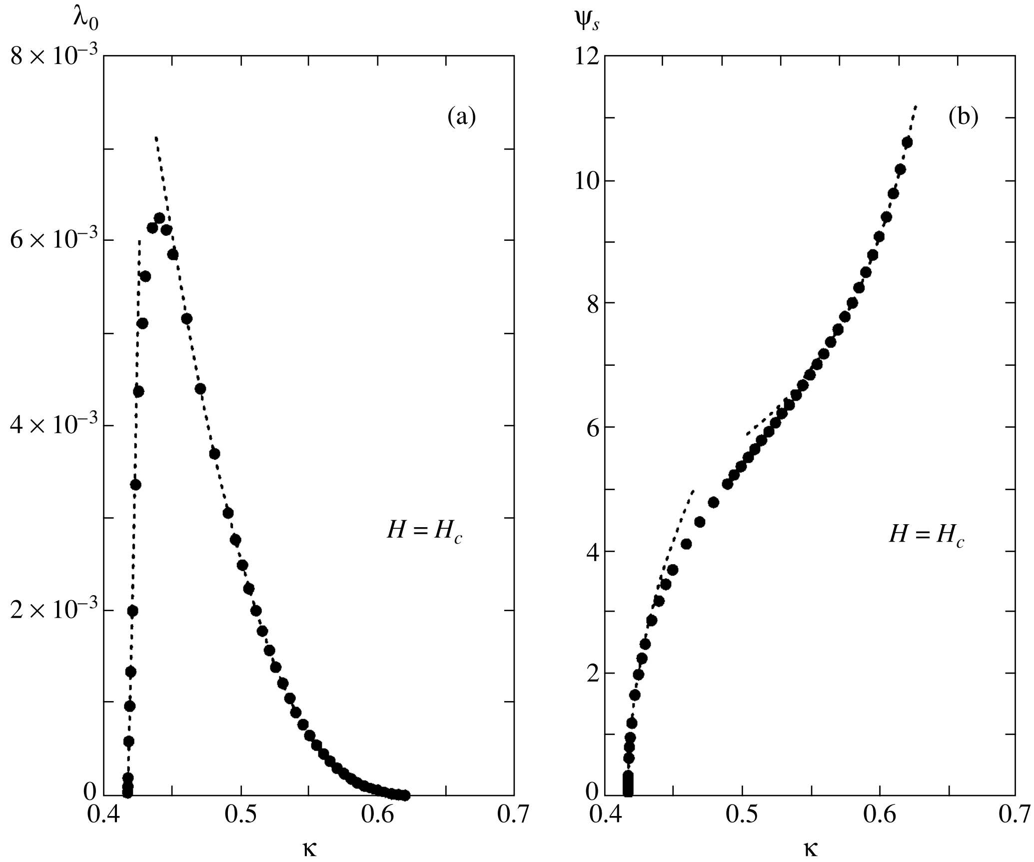


Fig. 6. The evolution of the "gap" in the excitation spectrum and of the integral order parameter ψ_s upon variation of κ from 0.417 to $1/\sqrt{2}$ in a field equal to the critical field. The dotted line indicates the asymptote in the vicinity of $\kappa = 0.417$: $\lambda_0 = 0.83(\kappa - 0.417)$, $\psi_s = 22.8(\kappa - 0.417)^{1/2}$ (see Section 2); and in the vicinity of $\kappa = 1/\sqrt{2}$: $\lambda_0 = 0.356\varepsilon e^{-0.701/\varepsilon})$, $\psi_s = 0.721\varepsilon^{-1} + 2.34$ (see Section 4).

namely, the local minimum (the surface solution being investigated) and maximum (the barrier providing for the metastability of the solution). It was such a merger of two solutions that was found by Park [5].

Let ψ_{s0} be the limiting solution on the line of the loss of stability. The Landau expansion with respect to the small parameter $\eta = \psi_s - \psi_{s0}$ in the vicinity of this line has the form

$$A(\kappa, H)\eta + B(\kappa, H)\eta^2 + C(\kappa, H)\eta^3 + \dots . \quad (8)$$

Because we treat the states with disturbed gage invariance ($\psi_{s0} \neq 0$), no symmetry exclusions exist in this case and odd-power terms are present in the expansion. The coefficient A goes to zero on the line being discussed simply by definition of the parameter η (equilibrium equations must have the solution of $\eta = 0$). The vanishing of the coefficient B corresponds to the merger of solutions. On approaching the critical line, we have two extrema of functional (8),

$$\eta_1 = \frac{-B + \sqrt{B^2 - 3AC}}{3C}, \quad \eta_2 = \frac{-B - \sqrt{B^2 - 3AC}}{3C}. \quad (9)$$

The stability of solutions will be defined by the positiveness of the second derivative,

$$F'' = 2B + 6C\eta > 0. \quad (10)$$

In the surface superconductivity problem under discussion, where the parameter ψ_s increases in the vicinity of the critical line, we must have $\eta < 0$ for a metastable solution. One can readily see that, for this purpose, it is necessary for the inequalities $A < 0$, $B > 0$, and $C < 0$ to be valid above the line of the loss of stability.

Given the linear vanishing of the coefficients A and B (which is standard for the Landau theory), on approaching the critical line ($\varepsilon \ll 1$), we obtain the following special features in the order parameter and in the "energy gap" (10):

$$\eta \propto -\sqrt{|\varepsilon|}, \quad F'' \propto \sqrt{|\varepsilon|}. \quad (11)$$

The resultant asymptotics in the vicinity of H_s are shown in Fig. 4.

5. WETTING TRANSITION

According to our numerical results, the integral order parameter (4) significantly increases on approaching the point $\kappa = 1/\sqrt{2}$. In this case, the Ginzburg–Landau ψ-function tends to unity almost everywhere within the layer; therefore, the parameter given by Eq. (4) defines the surface layer thickness in the limit being treated to an accuracy within the order of the width of the *ns*-boundary. The anomalous growth of this thickness makes possible a macroscopic description along the lines of the wetting theory (see Para. 160 in [8]). The specific features of superconductivity bring about the peculiarity of the thickness dependence of the free energy of the wetting layer.

In a homogeneous superconducting state, small perturbations of the order parameter and vector potential at distances far away from the plane source of perturbations decrease as $e^{-x/\xi}$ and $e^{-x/\delta}$, respectively, where $\xi = \delta/\sqrt{2}\,\kappa$ is the coherence length. Two sources of perturbations exist in the superconducting layer at the surface, namely, the sample surface and the boundary between the superconducting and normal phases. By complete analogy with, for example, the derivation of the law of interaction of electric charges in terms of the field of electric potential or elastic interaction of defects in crystals, one can state that the expression for the energy of interaction between the boundaries of the superconducting layer with its macroscopic thickness $L \gg \xi, \delta$ in the general case reduces to the sum

$$F_{int} = Ae^{-L/\xi} + Be^{-L/\delta}, \qquad (12)$$

where A and B are some functions of parameter κ (the dependence on the magnetic field may be ignored, because the superconducting layer thickness may be significant only in the immediate neighborhood of $H = H_c$). Corresponding to the numerically established behavior of surface superconductivity in the neighborhood of the point $\kappa = 1/\sqrt{2}$ is the case of $A < 0$ (attraction), $B > 0$ (repulsion), with $B > |A|$. The free energy of superconducting layer, reckoned from the normal state, includes the energy of the superconductor boundary, the energy of the interface between superconducting and normal phases, the interaction energy given by Eq. (12), and the bulk contribution

$$\frac{1}{8\pi}(H^2 - H_c^2)L. \qquad (13)$$

The equilibrium thickness of the superconducting layer is defined by the condition of the minimum of the sum of contributions (12) and (13),

$$F' = -\frac{A}{\xi}e^{-L/\xi} - \frac{B}{\delta}e^{-L/\delta} + \frac{1}{8\pi}(H^2 - H_c^2) = 0. \qquad (14)$$

The solution will be stable in the case of the positive second derivative of energy with respect to L,

$$F'' = \frac{A}{\xi^2}e^{-L/\xi} + \frac{B}{\delta^2}e^{-L/\delta}. \qquad (15)$$

On the $H_s(\kappa)$ curve, the stability of the solution is lost. From the condition $F'' = 0$, we find

$$L = \frac{\delta\xi}{\xi - \delta}\ln\left(\frac{B\xi^2}{|A|\delta^2}\right) \qquad (16)$$

and substitute this value of L into Eq. (14) to derive

$$H_s^2 - H_c^2 = -8\pi\frac{|A|}{\xi^2}(\xi - \delta)\exp\left\{-\frac{\delta}{\xi - \delta}\ln\frac{B\xi^2}{|A|\delta^2}\right\}. \qquad (17)$$

In view of the proximity to the critical point at which $\varepsilon = (\xi - \delta)/\delta \ll 1$, results (16) and (17) may be represented in the form of $L \propto \varepsilon^{-1}$ and

$$H_s - H_c \propto -\varepsilon e^{-a/\varepsilon}, \qquad (18)$$

where $a = \ln(B/|A|)$ is a number on the order of unity. As we see in Fig. 1, the asymptotic form given by Eq. (18) is valid even at a considerable distance from the critical point being treated.

Following are the extreme expressions on approaching the critical point $(1/\sqrt{2}, H_c)$ along two preferred trajectories. For a superconductor with $\kappa = 1/\sqrt{2}$, the superconducting layer thickness logarithmically goes to infinity: $L \propto \ln\{1/(H - H_c)\}$, with $F'' \propto H - H_c$. In the case of motion along a straight line ($H = H_c$), the thickness increases by the law $L \propto \varepsilon^{-1}$ (as in the case of motion along a curve, $H = H_s$), and the "gap" in the excitation spectrum decreases by the law of $F'' \propto \varepsilon e^{-a/\varepsilon}$. The asymptotic laws obtained are given in Figs. 5 and 6. We used the above-identified relation between the integral parameter (4) and the superconducting layer thickness L, as well as the obvious linear dependence between the second derivative of energy F'' (15) and the quantity λ_0 in the vicinity of the critical lines.

ACKNOWLEDGMENTS

We are grateful to I.N. Khlyustikov for attention to our work and discussions and to I.A. Fomin for discussion of our work.

This study was supported by the Russian Foundation for Basic Research (project nos. 00-02-16250 and 03-02-16958).

REFERENCES

1. D. Saint-James and P. G. de Gennes, Phys. Lett. **7**, 306 (1963).
2. J. Feder, Solid State Commun. **5**, 299 (1967).
3. J. P. McEvoy, D. P. Jones, and J. G. Park, Solid State Commun. **5**, 641 (1967).
4. I. N. Khlyustikov and V. A. Berezin, Zh. Éksp. Teor. Fiz. **121**, 1370 (2002) [JETP **94**, 1176 (2002)].
5. J. G. Park, Solid State Commun. **5**, 645 (1967).
6. E. M. Lifshitz and L. P. Pitaevskiĭ, *Course of Theoretical Physics*, Vol. 5: *Statistical Physics* (Nauka, Moscow, 1978; Pergamon, New York, 1980), Part 2.
7. L. D. Landau, Phys. Z. Sowjetunion **8**, 113 (1935); *Collected Works* (Nauka, Moscow, 1969), p. 123.
8. L. D. Landau and E. M. Lifshitz, *Statistical Physics*, 4th ed. (Nauka, Moscow, 1995; Pergamon Press, Oxford, 1980), Part 1.

Translated by H. Bronstein

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 162–173.
Translated from Zhurnal Éksperimental’noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 180–193.
Original Russian Text Copyright © 2003 by Vinoslavskii, Kravchenko.

SOLIDS
Electronic Properties

High-Field Autosolitons in an Electron–Hole Plasma in *p*-Ge

M. N. Vinoslavskii* and A. V. Kravchenko
Institute of Physics, National Academy of Sciences of Ukraine, Kiev, 03028 Ukraine
**e-mail: mvinos@iop.kiev.ua*

Received August 28, 2002

Abstract—The formation of high-field thermodiffusion autosolitons was studied experimentally in a photogenerated electron–hole plasma heated up by an electric field in *p*-Ge samples oriented along the ⟨111⟩ axis at $T = 77$ K. Measurements of the current–voltage characteristics, electric field distributions along the samples, and IR emission in the wavelength range $\lambda = 1.65–10$ μm showed that the arising of an autosoliton was accompanied by the appearance of N-shaped current–voltage characteristic regions. Autosolitons were formed at electron–hole plasma concentrations $n \geq 1 \times 10^{14}$ cm^{-3} and field strengths $E \geq 500$ V/cm. They manifested themselves as static, moving, and pulsating strata with field strengths $E_{as} = 1000–20000$ V/cm and carrier temperatures $T_e \geq 1000$ K. We also observed a turbulent electron–hole plasma state when autosolitons chaotically appeared and disappeared in the samples. The multivalley band structure of germanium influenced the dynamics of autosoliton formation; intervalley transfer of electrons in the strong field of autosolitons caused a three-step autosoliton field growth. © *2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

The arising of current instabilities accompanied by the formation of spatially nonuniform dissipative structures in the form of current filaments or electric field domains in a monopolar system of charge carriers requires the presence of negative differential *S*- or *N*-type conductivity in a semiconductor crystal (e.g., see [1, 2]). Similar dissipative structures in a bipolar semiconductor plasma can also arise at a positive differential conductivity of samples and in much lower electric fields (e.g., see reviews in [3]). Among the dissipative structures in an electron–hole plasma of a semiconductor are autosolitons; a theory of autosolitons is described in [3]. Several authors relate the observation of luminous spots in GaAs films [4], current filaments in reverse biased *p*–*n* junctions in α-SiC [5] and silicon p^+–n^+–p–n^- structures [6], and current instabilities caused by heating InSb structures [7] to the formation of autosolitons. The possibility of observing autosolitons in a photogenerated electron–hole plasma in germanium and silicon crystals was predicted in [8].

Earlier, we studied high-field thermodiffusive autosolitons in electron–hole plasmas in *n*-Ge [9–13] and *p*-Si [14, 15]. These autosolitons manifested themselves as narrow, strong, electric field strata (E_{as} = 1000–5000 V/cm) transverse with respect to the current direction. The mechanism of formation of such autosolitons was related to the appearance of a positive feedback for temperature growth of carriers heated up by an electric field to the Debye temperature. The multivalley energy band structure in samples with the electric field direction along the axis corresponding to a large effective mass of electrons (field *E* is parallel to the ⟨111⟩ axis in *n*-Ge and the ⟨100⟩ axis in *p*-Si) influenced the behavior of the autosolitons, which moved in the direction of minority carriers (holes) in *n*-Ge and formed a series of static autosolitons in *p*-Si.

The purpose of this work was to determine the conditions for the arising of high-field autosolitons in *p*-Ge crystals, verify the mechanism of formation of high-field autosolitons suggested earlier for *n*-Ge, and study the influence of the multivalley energy band structure and the type of junctions on the behavior of autosolitons. Next, we will consider the phenomena of spatial ordering in a photogenerated electron–hole plasma. They arise with increased voltage applied to *p*-Ge samples with junctions of two types, p^+–p and n–p.

2. EXPERIMENTAL

The samples were cut from a *p*-Ge crystal ($\rho \approx$ 40 Ω cm) as rectangular $0.05 \times 0.1 \times 0.8$ cm^3 plates. The longer plate axes were aligned with the ⟨111⟩ crystallographic axis. The broad faces of the samples were polished and etched in an SR-4 polishing etcher. Indium p^+–p contacts were burned into the opposite smaller ends of the samples of the first type, and n–p In + 3% As contacts, into the opposite ends of the samples of the second type at $T \approx 450$ K. Measurements were taken at $T_0 = 77$ K.

In the first unit described in [11, 12], an electron–hole plasma of a high density ($n \approx p = 5 \times 10^{13}$–$1 \times 10^{17}$ cm^{-3}) was generated uniformly along the broad

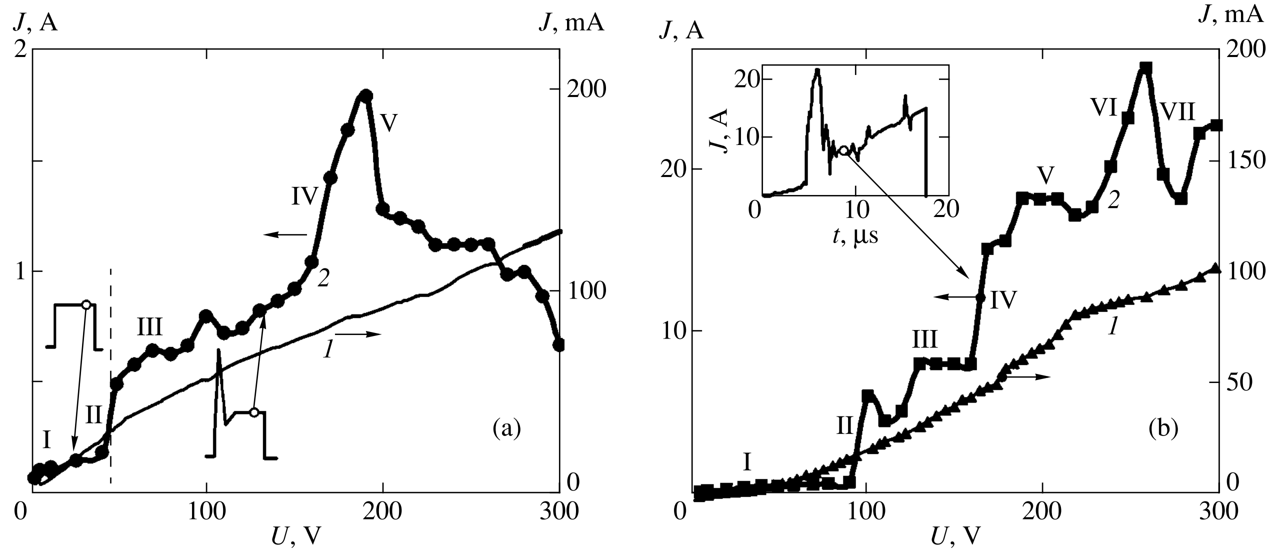


Fig. 1. Current–voltage characteristics of samples with two (a) p^+–p and (b) n–p junctions obtained using rectangular voltage pulses of width $\tau_U \approx 20$ μs (*1*) without illumination and (*2*) under illumination (CVC-1) at the instant corresponding to the stationary current J value.

sample face by a bell-shaped light pulse of intensity I_p from an IFP-800 pulsed lamp. The pulse width was τ_p = 0.5 ms. We used single rectangular or saw-tooth voltage U pulses of width τ_U = 5–300 μs and amplitude up to 300 V. The voltage pulses were applied to sample contacts simultaneously with light pulses. Under these conditions, the temperature rise, the mean over the sample during a current pulse, did not exceed 20 K. The temperature rise was estimated from the heat balance equation without taking into account heat transfer to the holder. Measurements were performed in three regimes: (1) the current–voltage characteristic (CVC-1) of a sample was measured at a light pulse maximum for rectangular voltage pulses, and the time evolution of high-field autosolitons was studied at constant illumination intensity and voltage values; (2) the dynamic current–voltage characteristic (CVC-2) was measured during growth of the leading voltage pulse edge at a light pulse maximum, and the dynamics of high-field autosoliton formation and evolution was studied; and (3) the dynamics of high-field autosoliton development was studied depending on the concentration of the electron–hole plasma while the intensity of light increased during a rectangular voltage pulse.

The $E(x)$ distribution of electric field strength along the sample was measured using a microcircuit of 26 potential probes made of phosphor bronze. The distance between neighboring probes was within 0.15–0.4 mm. The voltage measured by each of the two neighboring probes was recorded using two differential inputs of oscilloscopes with memory. The oscillograms of local fields in all interprobe intervals $E_{i-j}(t) = \Delta U_{i-j}(t)/l_{i-j}$, where l_{i-j} is the distance between probes, were used to construct the $E(x)$ electric field distributions along the samples at the selected instants of time during a voltage pulse and to determine high-field autosoliton parameters such as the time of formation, the velocity of motion, and the shape and size. The dynamics of carrier and crystal lattice temperature variations during the formation and decay of high-field autosolitons was studied with the second unit described in [13], which measured IR emission from the samples with a Ge/Au photodetector at T = 77 K. The measurements were taken in the wavelength range λ = 1.65–10 μm (I_{IR}) and, using a glass filter, at λ = 1.65–2.5 μm (I_g). Simultaneously, the field distribution along the sample was controlled by an 18-probe head.

3. SAMPLE CURRENT–VOLTAGE CHARACTERISTICS

The current–voltage characteristics of unilluminated p-Ge samples with p^+–p junctions are approximately linear and have a slightly decreasing slope at high voltages (Fig. 1a, curve *1*) because of a slightly non-Ohmic contact behavior. The CVC-1 curves of the same samples under illumination, obtained for steady currents, contain three to five distinct regions (Fig. 1a, curve *2*). In region I (U < 50 V), current J grows weakly as voltage increases. Regions II and IV are characterized by a steep increase in the current at $U_1 \approx 50$ V and $U_2 \approx 170$ V. In regions III and V, the current reaches saturation or has an N-shaped dependence. The dynamic CVC-2 of illuminated samples with p^+–p junctions shows a weak current increase during the onset of voltage pulse growth (see below, Fig. 2a, region I). The current then sharply increases (region II). After this, we observe either current saturation or an N-shaped dependence of the current (region III) often followed by a monotonic current increase in time, even under voltage saturation.

The initial region of the current–voltage characteristic of unilluminated *p*-Ge samples with *n*–*p* junctions is a quadratic dependence caused by the injection of charge carriers from the forward biased negative *n*–*p* junction (Fig. 1b, curve *1*). At large voltages, the slope of the current–voltage characteristic decreases because of blocking of the reverse biased positive *n*–*p* junction. The CVC-1 curves of illuminated samples with *n*–*p* junctions obtained for the lowest steady current value contain region I of a weakly growing current (Fig. 1b, curve *2*) followed by regions II, IV, and VI of a steeply increasing current and regions III, V, and VII of current saturation or an N-shaped dependence. The dynamic CVC-2 curves of illuminated samples with *n*–*p* junctions also contain region I of a weak current during a *U* pulse leading edge growth (see below, Fig. 3a). Next, we observe region II of a sharp increase in the current and N-shaped region III.

4. CONTACT EXCLUSION OF AN ELECTRON–HOLE PLASMA IN SAMPLES WITH p^+–p JUNCTIONS

We explain a weak increase in the current in regions I of CVC-1 (Fig. 1a) and CVC-2 (Fig. 2a) in an electron–hole plasma photogenerated in samples with p^+–p junctions by contact exclusion. The electron–hole plasma is then carried by the field from the negative to positive junction in the drift direction of minority carriers (electrons). This causes the formation of an exclusion region with a decreased electron–hole plasma concentration and a high electric field strength at the negative junction. For a growing voltage pulse, this is shown in Fig. 2d (curve t_1, the $E(x, t_1)$ field distribution obtained from the oscillograms of local fields $E_{i-j}(t) = \Delta U_{i-j}/l_{i-j}$, $i = 0$–8, $j = 1$–9 at time t_1). In the remaining sample regions, a high plasma concentration and a weak field are established [Fig. 2c, curves $E_{2-3}(t_1)$ and $E_{3-4}(t_1)$], and the $E(x)$ field distribution profile in the exclusion region of the transition from a strong to a weak field acquires the character of a steep wall

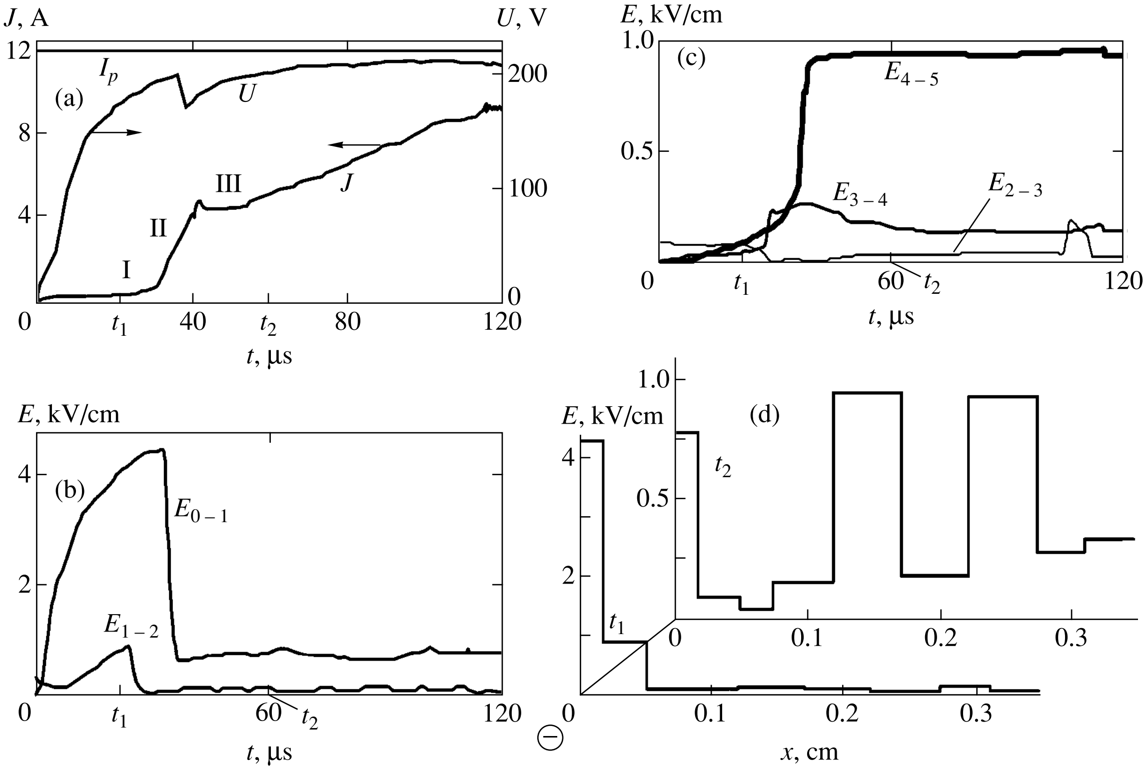


Fig. 2. (a–c) One-step local field growth in the formation of two static high-field autosolitons with small amplitudes; $E_{as1} = E_{4-5}(t_2) = \Delta U_{4-5}(t_2)/l_{4-5}$ and $E_{as2} = E_{6-7}(t_2)$ near the exclusion region at the negative junction in a sample with p^+–p junctions during voltage pulse *U* growth at a constant illumination intensity, at a light pulse maximum of I_p; *J* is the current. The oscillogram of field E_{6-7} is similar to that of E_{4-5} and is therefore not shown. (d) $E(x, t_1)$ and $E(x, t_2)$ are the field strength distributions close to the negative junction before and after the formation of autosolitons obtained from local field E_{i-j} oscillograms at times t_1 and t_2, respectively; $l_{0-1} = 0.017$ cm (the distance between the negative junction and the first probe), $l_{1-2} = 0.032$ cm, $l_{2-3} = 0.025$ cm, $l_{3-4} = 0.045$ cm, $l_{4-5} = 0.05$ cm, $l_{5-6} = 0.05$ cm, and $l_{6-7} = 0.05$ cm; and $l = 0.8$ cm is the sample length. In Figs. 2–4, minus sign in the circle indicates negative polarity of the voltage applied to this end of a sample.

(Fig. 2d, curve t_1) at a sufficiently high voltage, in agreement with the results obtained in [16].

5. CURRENT–VOLTAGE CHARACTERISTIC REGIONS OF A STEEP INCREASE IN THE CURRENT

A sharp increase in the current in regions II of CVC-1 (Fig. 1a) and CVC-2 (Fig. 2a) in samples with p^+–p junctions is accompanied by a decrease in the field strength in the exclusion region (Fig. 2b, 2c, curves E_{0-1}–E_{2-3}, and Fig. 2d, curve t_2) and an increase in the field in the adjacent sample regions. Such an increase in the current is accompanied by a small decrease in the voltage on the sample because of a strong decrease in sample resistance, which becomes comparable to the output resistance of the generator of voltage pulses. This causes the formation of S-shaped region II in dynamic CVC-2 (Fig. 2a). These data are evidence that an electron–hole plasma flow comes to the near-junction high-field region from the neighboring sample region enriched in carriers. We relate this to the reversal of the bipolar drift direction of the electron–hole plasma heated up by the electric field in the exclusion region. The reason for the bipolar drift reversal is the different dependences of electron and hole mobilities on their temperatures in a heating electric field [17–20]. We observed a similar phenomenon for *n*-Ge samples at much lower local fields in the exclusion region [15].

Initial regions I of CVC-1 (Fig. 1b) and dynamic CVC-2 (Fig. 3a) characterized by a weak increase in the current in samples with *n*–*p* junctions is caused by blocking of the reverse biased positive *n*–*p* junction. This is evidenced by an initial increase in the field strength to a high value at this junction, which accompanies voltage pulse growth (Fig. 3c). We explain the appearance in these samples of region II with a sharp

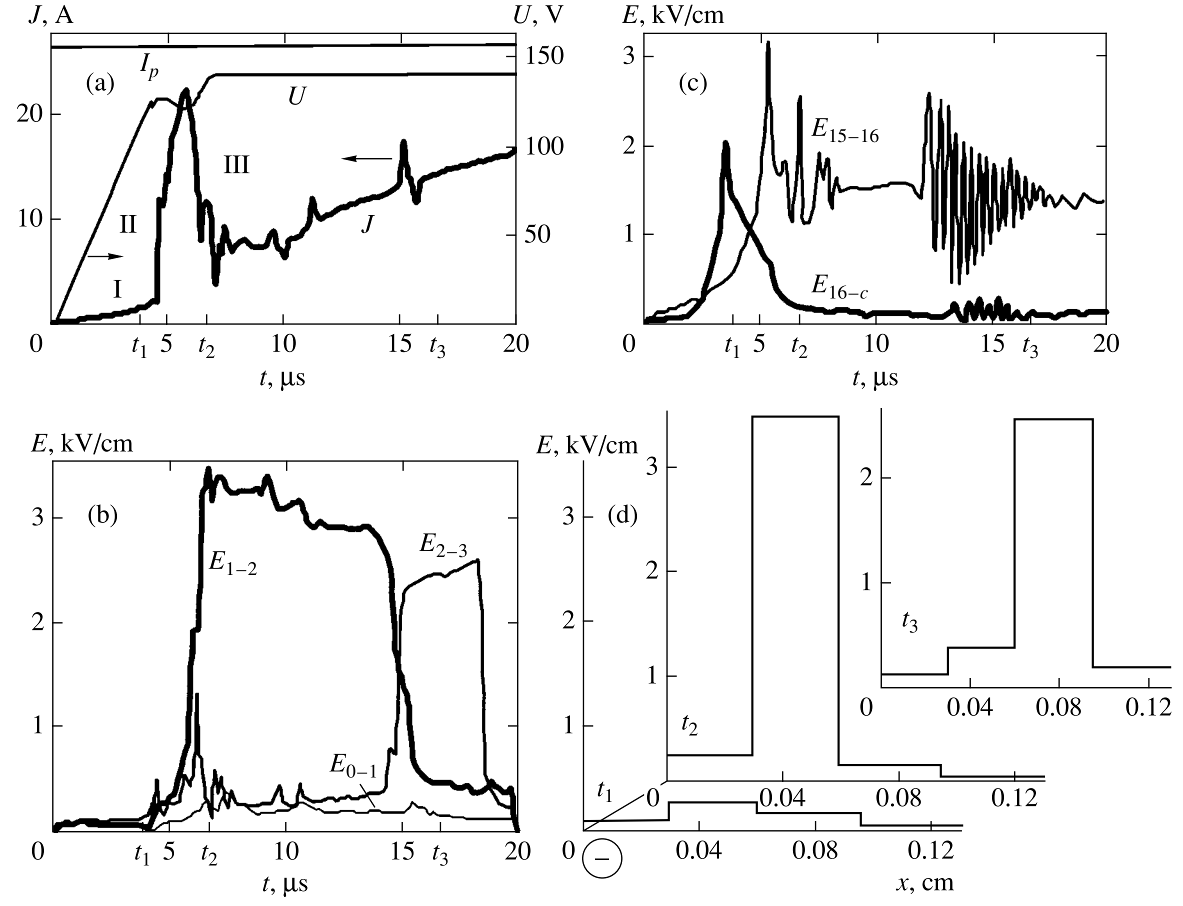


Fig. 3. (a–c) Three-step field strength growth in the formation of a moving high-field autosoliton with a large amplitude ($E_{as1} = E_{1-2}$) close to the injecting negative junction and the formation of a low-amplitude static soliton ($E_{as2} = E_{15-16}$) with an intermediate pulsating stage close to the positive blocking junction in a sample with *n*–*p* junctions during voltage pulse U growth at a constant illumination intensity I_p, at a light pulse maximum; J is the current. (d) $E(x, t_1)$, $E(x, t_2)$, and $E(x, t_3)$ are the field strength distributions near the negative junction before and after the formation of a moving autosoliton determined from local field E_{i-j} oscillograms at times t_1, t_2, and t_3, respectively; $l_{0-1} = 0.03$ cm, $l_{1-2} = 0.03$ cm, $l_{2-3} = 0.035$ cm, $l_{3-4} = 0.035$ cm, $l_{15-16} = 0.025$ cm, and $l_{16-c} = 0.015$ cm (the distance between the last probe and the positive junction); and $l = 0.8$ cm is the sample length.

increase in the current in the stationary (Fig. 1b) and dynamic (Fig. 3a) current–voltage characteristics by the blocking *n–p* junction being poured over with the photogenerated electron–hole plasma as a result of its bipolar drift. The E_{16-c} field at the positive junction then sharply decreases. A small decrease in the voltage at the instant of a sharp current rise creates an S-shaped region of the dynamic current–voltage characteristics (CVC-2) (Fig. 3a).

6. THE ARISING OF HIGH-FIELD AUTOSOLITONS

The appearance of branches III and V in stationary current–voltage characteristics (CVC-1) (Fig. 1) and of branch III in dynamic current–voltage characteristics (CVC-2) (Figs. 2, 3) with either current saturation or N-shaped current dependences is accompanied by the arising and development of high-field domains in samples (Figs. 2, 3d). We treat these domains as high-field thermodiffusion autosolitons in conformity with the theory described in [3] and the results obtained in [9]. These autosolitons are formed at electron–hole plasma concentrations of $n \geq 1 \times 10^{14}$ cm^{-3} and mean field strengths in the sample of $E \geq 100$ V/cm. They manifest themselves as solitary strata with a high field strength (E_{as} = 1000–20000 V/cm), carrier temperatures exceeding 1000 K, and a decreased concentration of carriers. Depending on the light intensity and applied voltage, static, moving, and pulsating high-field autosolitons arise in the sample. In addition, we observed a turbulent electron–hole plasma state, when autosolitons chaotically formed and disappeared in various sample regions.

6.1. The Static High-Field Autosoliton

An increase in the voltage over time at a constant illumination intensity stretches the exclusion region in samples with p^+–p junctions. When the threshold voltage is attained, the field strength in the extended part of the exclusion region decreases (Fig. 2b, curves E_{0-1}–E_{2-3}) and the current increases because of the bipolar drift reversal. After this, the current slightly decreases and either one high-field autosoliton with a large amplitude or simultaneously two static autosolitons with smaller amplitudes are formed at some distance (l = 1–2 mm) from the negative junction (Fig. 2, $E_{as1} = E_{4-5} \approx E_{as2} = E_{6-7} \approx 1000$ V/cm), in the region characterized by the local carrier concentration $n \approx p \geq 1 \times 10^{14}$ cm^{-3} and field strength $E_{i-j} \geq 500$ V/cm necessary for the formation of autosolitons. An increase in the field of a small-amplitude autosoliton has a one-step dynamics. The high field at the junction is largely determined by exclusion regions depleted of carriers, and the high field of the autosoliton is largely determined by the high temperature of carriers. The formation of autosolitons in samples with p^+–p junctions insignificantly (by 5–15%) decreases the current, because the current value is already substantially limited by the high resistance of the exclusion region.

The formation of a large-amplitude static soliton ($E_{as} = E_{1-2} \approx 5200$ V/cm) in the exclusion region in samples with p^+–p junctions at high light intensities and applied voltage values ($U \geq 170$ V, regime 3) is accompanied by a strong Joule heating of the autosoliton region. This causes substantial thermal generation of the electron–hole plasma (comparable to its photogeneration) and a decrease in the autosoliton field, which in turn results either in a monotonic increase in the current and autosoliton transfer to neighboring sample regions or in autosoliton destruction and a jump increase in the current. A large-amplitude static soliton ($E_{as} \approx 3200$ V/cm) formed under a weak illumination intensity and at a low voltage ($U \approx 100$ V) and a low steady current ($J \leq 600$ mA) can exist for a long time (up to 200 μs) until it decays because of Joule lattice heating.

Note that the reversal of the bipolar drift direction of the electron–hole plasma in these samples occurs at voltage values smaller than those necessary for high-field autosoliton formation, and both these effects are independent. The exclusion region often plays the role of the main seed for preliminarily heating carriers and the formation of such autosolitons. At the same time, the electron–hole plasma drift reversal lowers the field in the exclusion region, increases the field in the other sample regions, and enhances the role played by intrinsic sample inhomogeneities far from junctions in the formation of autosolitons.

To eliminate the initial influence of the exclusion region on the formation of high-field autosolitons, we performed measurements for samples with two *n–p* junctions subjected to the action of rectangular voltage pulses at a constant illumination intensity. Field strength E_{0-1} at the negative injecting junction was then low; that is, the exclusion region was absent (Fig. 3d, curve t_1). At the same time, the high field intensity E_{16-c} during the increase in voltage pulse at the positive blocking junction then decreases because of the blocking junction being poured over by the photogenerated electron–hole plasma simultaneously with an increase in the current, which increases the field strength in the other sample regions (E_{1-2}, E_{2-3}, and E_{15-16}). As a result, at a voltage exceeding the threshold of autosoliton formation ($U \geq 120$ V), one large-amplitude static autosoliton is formed near the negative junction, $E_{as1} = E_{1-2} \approx 3300$ V/cm. The second autosoliton with a smaller amplitude is formed near the positive junction, $E_{as2} = E_{15-16} \approx 1600$ V/cm (Fig. 3). Because of the absence of the exclusion region in these experiments, autosoliton formation causes a substantial (by a factor of 1.5–2) decrease in the current. At the same time, the currents in these samples were much stronger than in the samples with p^+–p junctions (cf. Figs. 1a and 1b or 2 and 3). While the leading U pulse edge increases, the

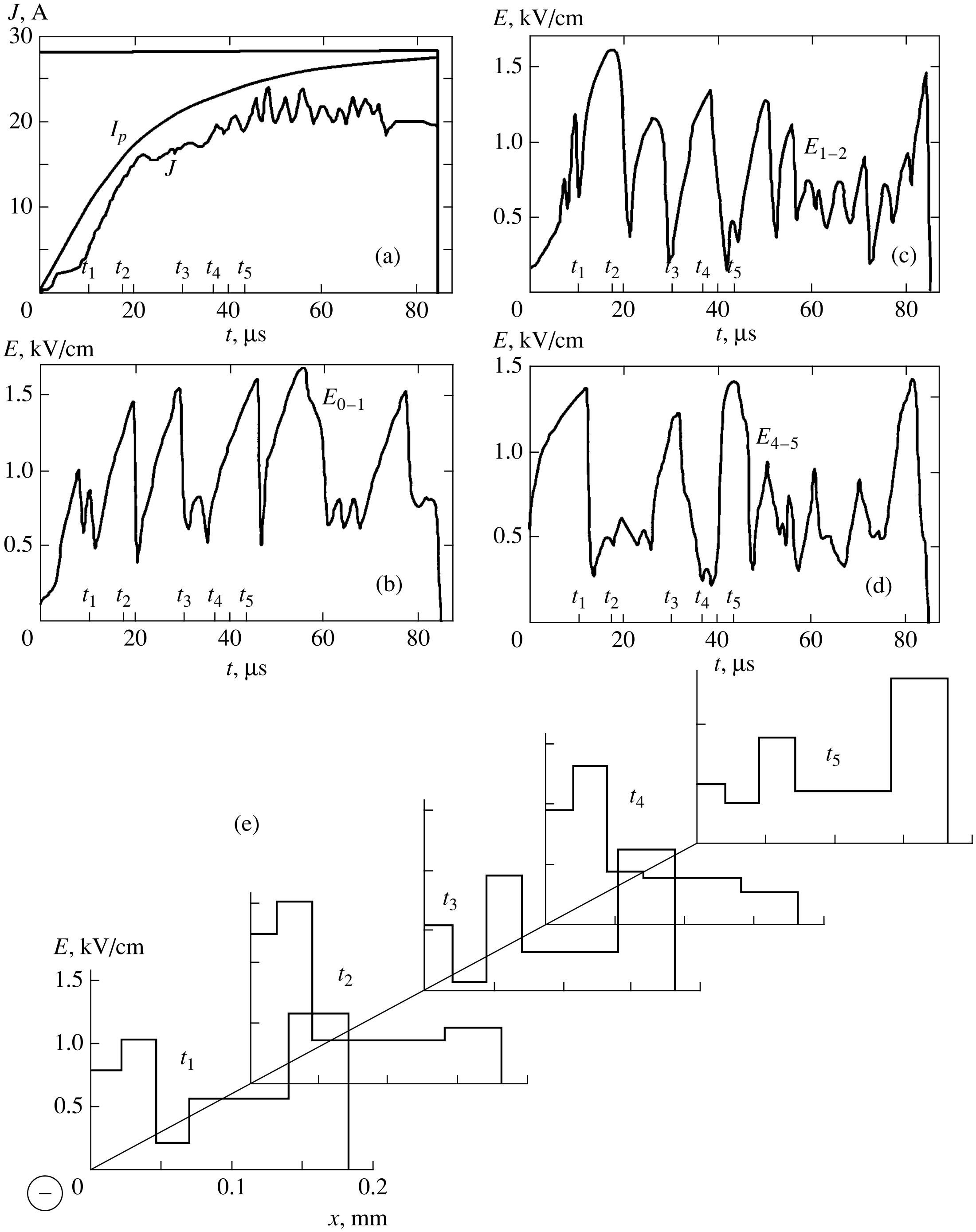


Fig. 4. Turbulent electron–hole plasma state with chaotic formation and disappearance of autosolitons in various regions of a sample with p^+–p junctions during illumination intensity I_p growth at a constant high voltage of $U = 300$ V: (a) current, (b)–(d) local field oscillograms, and (e) field strength distributions near the negative junction at various U pulse instants; $l_{0-1} = 0.03$ cm, $l_{1-2} = 0.03$ cm, $l_{2-3} = 0.03$ cm, $l_{3-4} = 0.03$ cm, and $l_{15-16} = 0.03$ cm; $l = 0.8$ cm is the sample length.

dynamics of field growth in the first autosoliton close to the negative *n*–*p* junction, has a three-step character, and this autosoliton then moves toward the positive junction. Field growth in the second autosoliton close to the positive junction has a one-step character, and its static state transforms into the pulsating state and back.

At a large amplitude of rectangular voltage pulses and a growing light intensity (regime 3), a temporary field decrease was observed in some samples in the region of the large-amplitude autosoliton, $E_{as1} = E_{1-2} \geq 5000$ V/cm, at a maximum light intensity. This was accompanied by a strong current increase followed by field growth to its initial value. We explain the field decrease by an interband breakdown in the autosoliton region, because the autosoliton width decreases as the light intensity (the electron–hole plasma concentration) increases as a result of a decrease in the velocity of electron–hole plasma bipolar drift [11, 12] to a value smaller than the distance between neighboring probes, and the actual autosoliton field value substantially increases to the breakdown field value.

Note that the accuracy of determining the maximum field strength value in a static autosoliton largely depends on the ratio between the autosoliton width L_{as} and the l_{i-j} distance between neighboring probes in the autosoliton region. If $L_{as} \approx l_{i-j}$, this accuracy can amount to 60–80% of the real maximum autosoliton field value. If $L_{as} \ll l_{i-j}$, the real autosoliton field is determined either indirectly, as with the interband breakdown in the soliton, or more accurately for the example of a moving autosoliton, as is shown in the next paragraph.

6.2. A Moving High-Field Soliton

Moving autosolitons were largely observed in samples with *n–p* junctions at a large amplitude of a rectangular voltage pulse ($U \geq 140$ V, regime 1). The large-amplitude autosoliton ($E_{as} = E_{1-2} \approx 3700$ V/cm, Fig. 3) that formed close to the injecting negative junction moved toward the positive junction [in the minority carrier (electron) drift direction] at a velocity of $v_{as} \approx 9 \times 10^3$ cm/s. The velocity of an autosoliton was determined from the time of its passage ($t_{2-3} \approx 4$ μs) through the interprobe gap, $l_{2-3} = 0.035$ cm. The time at which an autosoliton enters the interprobe gap l_{2-3} can be determined from field oscillograms, $E_{2-3}(t_2)$ ($\Delta t_{2-3} \approx 0.5 \times 10^{-6}$ s), which allows us to estimate the autosoliton width at its base, $L_{as} \approx 4.5 \times 10^{-3}$ cm. The mean autosoliton field strength can then be estimated as $E_{as} = \Delta U_{2-3}/L_{as} = 90$ V/4.5×10^{-3} cm $\approx 2 \times 10^4$ V/cm. It may well be still higher at the maximum, taking into account an approximately Gaussian autosoliton field strength distribution. The rapid entrance of a narrow autosoliton into a wider interprobe gap ($L_{as} \ll l_{i-j}$), its motion in this gap, and the rapid exit from the gap form a voltage drop pulse $\Delta U_{as}(t)$ of a shape close to rectangular.

6.3. Pulsating High-Field Autosolitons and Turbulent State of Electron–Hole Plasma

At low (near-threshold) light intensity and applied voltage values, an autosoliton with a steadily pulsating amplitude could form in a sample. For instance, switching on a low-amplitude ($U \approx 120$ V) voltage pulse simultaneously with the onset of light pulse growth (regime 3) caused the formation of a steadily pulsating autosoliton with a field of $E_{as1} = E_{1-2} \approx 2200$ V/cm in the exclusion region in a sample with p^+–p junctions. The amplitude of autosoliton field oscillations was $\Delta E_{as} \approx 0.1 E_{as}$, and the frequency of oscillations was $f \approx 100$ kHz at a current of $J \approx 0.8$ A. Simultaneously, the second, static autosoliton with a field of $E_{as2} = E_{7-8} \approx 1200$ V/cm appeared in the low-field sample region. The autosolitons of the two types coexisted for a fairly long time, $t_{as} \approx 250$ μs. After the decay of the pulsating soliton (because of Joule heating), the current sharply increased.

A different type of pulsating autosoliton characterized by a large and constant amplitude ($E_{1-2} \approx 4800$ V/cm) was observed in some samples at high applied voltage values of $U \geq 180$ V. This autosoliton experienced oscillatory displacements to neighboring interprobe gaps, where it caused E_{0-1} and E_{2-3} field oscillations with amplitudes of up to 20% of $E_{\max}$. We explain such oscillatory autosoliton shifts by a higher rate of sample surface recombination determined by the special features of its etching. The conditions of formation of pulsating autosolitons of both types coincide with theoretically predicted conditions [3].

At high voltages of $U = 200$–300 V and intense illumination, the electron–hole plasma exhibits turbulent behavior in samples with p^+–p junctions. This is accompanied by the chaotic appearance and disappearance of autosolitons in different sample regions. The amplitude of field strength oscillations then amounts to $\Delta E_{i-j} = (0.7$–$0.9)E_{i-j}$, the frequency of oscillations is $f \approx 200$ kHz, and the autosoliton field is as high as $E_{1-2} \approx E_{4-5} \approx 1500$ V/cm (Fig. 4). At the same time, only small current oscillations whose amplitude is about 15% of $J_{\max}$ go to the external circuit. Note that the turbulent autosoliton state was not reached in similar experiments with *n*-Ge [9–13].

7. IR EMISSION DURING FORMATION AND DECAY OF AUTOSOLITONS

A weak IR radiation signal I_{IR} from the exclusion region appeared and disappeared in the course of exclusion during a rectangular voltage pulse whose amplitude was below the threshold of autosoliton formation. The signal approximately reproduced the shape of the light pulse, as in [13]. The spectrum of IR radiation was observed in the wavelength range $\lambda = 6$–10 μm and corresponded to carrier temperatures of $T_e = 150$–300 K.

The dynamics of IR emission from the samples with p^+–p junctions during the formation and decay of high-field autosolitons was studied by switching on a rectangular voltage pulse ($U \geq 80$ V) simultaneously with the onset of an increase in light pulse (Fig. 5). Under these conditions, the current and field E_{0-1} first sharply decrease, whereas the E_{1-2} field increases, because of the development of electron–hole plasma exclusion at the negative junction. Next, field E_{0-1} again sharply decreases, and the current sharply increases because of an increase in the concentration of the electron–hole

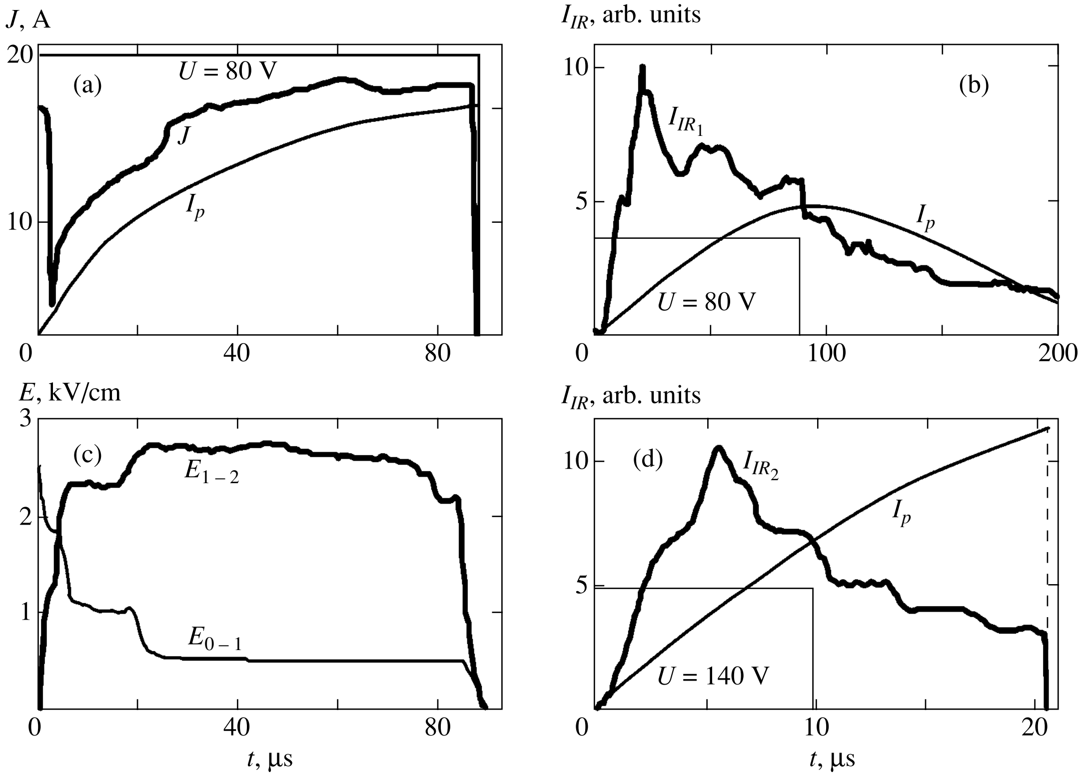


Fig. 5. (a, b) Dynamics of IR emission ($\lambda = 1.65$–$10\ \mu$m) from the near-junction region of a sample with p^+–p contacts in the course of the formation and decay of an autosoliton under the conditions of growing illumination intensity I_p during a long voltage pulse of small amplitude; (d) multistep decay of IR radiation from the autosoliton region after termination of a short voltage pulse of large amplitude; (c) oscillograms of local fields.

plasma as a result of photogeneration and bipolar drift reversal. Simultaneously, field E_{1-2} sharply increases to produce a high-field static autosoliton. At this instant, the I_{IR_1} IR radiation signal from hot carriers in the autosoliton region rapidly increases.

As illumination and current increase further, the second I_{IR_1} signal rapidly grows to its maximum (Fig. 5). We explain this by thermal radiation from the high-field autosoliton region caused by its strong Joule heating, which reaches hundreds of Kelvin. As the light intensity further increases, the first sharp decrease in the I_{IR_1} signal after its peak value is observed while field E_{1-2} increases. This decrease in the I_{IR_1} signal can be related to a decrease in the temperature of carriers as a result of a decrease in power supplied to the autosoliton region. The decrease in power is explained by a decrease in the mobility of electrons caused by their transition from hot valleys to the cold valley in the strong autosoliton field. According to our calculations, at T = 80 K, the concentration of the electron–hole plasma $n = 1 \times 10^{15}$ cm^{-3}; field E = 2000 V/cm directed along the $\langle 111\rangle$ axis; and the mobility of electrons decreases by 25% compared with the case when the field is directed along the $\langle 100\rangle$ axis.

After the initial I_{IR_1} signal peak, we often observe substantial signal oscillations during the existence of a stationary high-field autosoliton when the E_{as} field is constant and the light intensity and current increase. This is evidence of the complex character of the development of carrier and lattice heating in the autosoliton region and the transfer of this heat into the surrounding sample volume. The temperature of carriers in such an autosoliton was estimated from the experimental I_{IR_1} and I_g signals and the spectrum of photodetector sensitivity. The temperature of electrons that made the major contribution to hot carrier radiation signals was found to exceed 1000 K.

A short-term decrease in the autosoliton field and an increase in the current caused by interband breakdown in the autosoliton region were accompanied by a sharp increase in short-wave IR radiation I_g.

Abnormally strong total IR radiation signals from the sample, I_{IR} = 70–120 mV, and IR radiation through the glass filter, I_g = 15–60 mV (the usual signal and noise levels being I_{IR} = 10–15 mV and 1 mV, respectively), were observed when a narrow (t_U = 10–20 µs) rectangular voltage pulse of $U \geq 140$ V was applied to the sample through which a strong current ($J \sim 20$–40 A) flowed at the instant of maximum light pulse intensity.

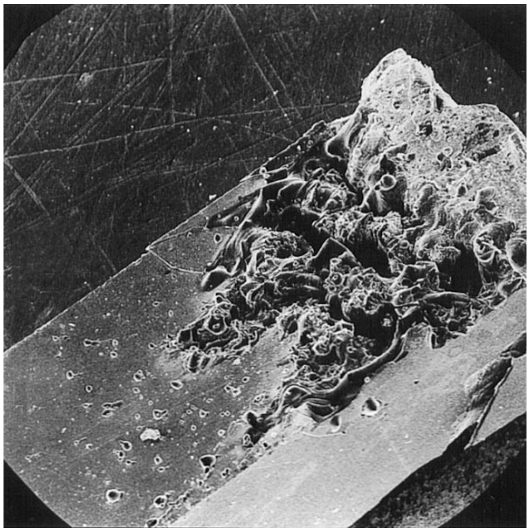


Fig. 6. Traces of sample material boiling up in the region of a high-field autosoliton in the exclusion region near the negative junction after applying a short $U \approx 300$ V pulse to a sample with two p^+–p contacts at a light pulse maximum.

The high I_{IR} and I_g signal values were caused, first, by the passage of recombination electron–hole plasma radiation from the autosoliton region through a Ge filter to the photodetector. This occurred because of a strong autosoliton region heating (to $T \gg 300$ K) and a narrowing of the forbidden band in the crystal in this region to a value smaller than the forbidden bandwidth of the Ge filter, which was at $T = 300$ K and served to cut off flash lamp light reflected from the sample. Secondly, apart from hot carriers, the heated lattice in the high-field autosoliton made a large contribution to the strong IR radiation signal.

Autosoliton region heating has a strong influence on the IR radiation signal from the sample after the destruction of the autosoliton. For instance, after voltage pulse decay, we first observe a rapid decrease in the I_{IR} signal by 20–30% caused by autosoliton destruction. The I_{IR_2} signal then slowly decreases in two to three steps 4–10 μs in duration (Fig. 5d). We explain these steps by photogenerated carrier radiation from the "hot spot" formed in the autosoliton region heated by hundreds of Kelvin. The width of these steps in the tail of the IR signal increases to 50 μs as the voltage pulse width grows to $t_U \approx 100$ μs. At still wider voltage pulses ($t_U \approx 200$ μs), the I_{IR} signal decreases monotonically after its decay.

Sample surface boiling close to the negative junction was observed in some samples at a maximum applied power (Fig. 6). This occurred as a result of rapid Joule heating of the crystal in the high-field autosoliton region, which occupied a thin ($d < 10$ μm) subsurface layer with bipolar conductivity in the exclusion region. As a BaF_2 plate was pressed to the sample from above, the boiling caused the deposition of the sample material on its surface. This result shows that a hot spot is formed in the autosoliton region. The temperature in this spot can reach $T > 1300$ K. In addition, this result substantiates estimates of the depth of electron–hole plasma penetration into the sample in a longitudinal electric field. This depth is determined by the depth of light absorption and the depth of minority carrier (electron) diffusion during their drift along the sample. It varies from several μm in the exclusion region to dozens of μm at the rear junction.

8. RESULTS AND DISCUSSION

Earlier, the phenomenon of the reversal of the bipolar electron–hole plasma drift direction was observed for an electron–hole plasma packet light-generated in germanium [7]. This phenomenon was also studied for the contact injection of an electron–hole plasma [18–20]. We were the first to observe this phenomenon under contact exclusion conditions in n-Ge samples [13]. It substantially influenced current–voltage characteristics and field strength distributions.

To analyze the state of a nonequilibrium electron–hole plasma, we used a computer program to solve the energy balance equation

$$\left(\frac{d\varepsilon}{dt}\right)_{\text{field}} = \left(\frac{d\varepsilon}{dt}\right)_{ac} + \left(\frac{d\varepsilon}{dt}\right)_{\text{opt}} + \left(\frac{d\varepsilon}{dt}\right)_{ee} + \left(\frac{d\varepsilon}{dt}\right)_{eh}, \quad (1)$$

where ε is the energy of particles, and calculated electron and hole characteristics depending on the electric field strength and its direction with respect to the crystallographic axes of the sample at crystal temperatures from liquid nitrogen temperature and higher. For this purpose, the symmetrical distribution function part was described by the Maxwell function with the electronic temperature determined from (1). Equation (1) includes charge carrier scattering by acoustic $(d\varepsilon/dt)_{ac}$ and optical $(d\varepsilon/dt)_{\text{opt}}$ phonons of different branches [21] and electron–electron $(d\varepsilon/dt)_{ee}$ [22] and electron–hole $(d\varepsilon/dt)_{eh}$ [22, 23] interactions. The asymmetric distribution function part was determined from the Boltzmann equation in the τ approximation taking into account scattering of carriers by acoustic and optical phonons [21] and impurities according to the Brooks–Herring equation [21, 24]. The field term has the form

$$(d\varepsilon/dt)_{\text{field}} = e\mu_{e,h}E^2,$$

where $\mu_{e,h}$ represents the mobilities of carriers. The concentration of electrons in different valleys in an electric field directed asymmetrically with respect to the valleys was determined from the balance equation for the number of particles, taking into account interactions with intervalley phonons [21, 24].

According to the calculated dependence of the ratio between the mobilities of electrons and holes on the electric field,

$$b(E) = \mu_e(E)/\mu_p(E)$$

(Fig. 7) at the crystal temperature $T = 80$ K and the field directed along the $\langle 111 \rangle$ axis, the mobility of holes becomes higher than that of electrons in limited electron–hole plasma concentration and electric field ranges. Under these conditions, the reversal of the electron–hole plasma bipolar drift direction can occur. The electron–hole plasma density then moves in the drift direction of more mobile hole carriers rather than in the drift direction of minority (electron) carriers. According to the estimates obtained by (1), the necessary electron–hole plasma concentrations are attained close to the illuminated surface in the l_{1-2} sample region ($E_{1-2} \approx 1000$ V/cm) at time t_1 (Fig. 2), preceding a sharp increase in the current. Local sample heating during a pulse and during voltage increase shifts not only the specified parameter ranges but also the place where the reversal of the electron–hole plasma drift direction begins.

The electron and hole parameters obtained also enter into the continuity equations for particle flows and energy fluxes in the theory described in [3]. In our experiments, they are characteristic of N-shaped and current-saturation regions of the current–voltage characteristics. These parameters satisfy the requirements that determine the conditions of formation of high-field transverse thermodiffusion autosolitons [3], namely:

(1) the strong inequality $\tau_{\mathcal{E}}/\tau \ll 1$ or $l_{\mathcal{E}}/L \ll 1$ is satisfied ($\tau_{\mathcal{E}} \approx 1 \times 10^{-11}$ s and $l_{\mathcal{E}} \approx 1 \times 10^{-5}$ cm are the particle energy relaxation time and length, $\tau \approx (1–5) \times 10^{-6}$ s is the lifetime of the electron–hole plasma, and $L \approx 1 \times 10^{-2}$ cm is the bipolar carrier diffusion length);

(2) the nonlinear condition $\alpha + s > –1$ is met in fields $E \geq 800$ B/cm (Fig. 8, $\tau_p \propto (T_e)^\alpha$, $\alpha = d\ln\tau_p/d\ln T_e$, $\tau_{\mathcal{E}} \propto (T_e)^s$, $s = d\ln\tau_{\mathcal{E}}/d\ln T_e$, and τ_p is the particle momentum relaxation time), which enables the mechanism of positive feedback for carrier temperature T_e growth to operate;

(3) the inequalities

$$\tau_p(\approx 1 \times 10^{-13}\ \text{s}) \ll \tau_{ee}(\approx 1 \times 10^{-12}\ \text{s})$$
$$\ll \tau_{\mathcal{E}}(\approx 1 \times 10^{-11}\ \text{s})$$

ensure the regime of energy control for current carriers and the formation of an autosoliton in the form of a high-field domain (at $n \leq 2 \times 10^{16}$ cm^{-3}).

The $\tau_{\mathcal{E}}$ value was calculated by the equation

$$\int d^3k f \frac{d\varepsilon}{dt} = \frac{3nk(T_e - T)}{2\tau_{\mathcal{E}}} \qquad (2)$$

using (1). Here, f is the Maxwell distribution function and k is the Boltzmann constant. The simultaneous fulfillment of the inequalities $\tau_{\mathcal{E}}/\tau \ll 1$ and $l_{\mathcal{E}}/L \ll 1$ for the electron–hole plasma allows this plasma to be classified with $K\Omega$ systems [3], in which autosolitons of all types observed in this work can form depending on the concentration and heating level of the electron–hole plasma.

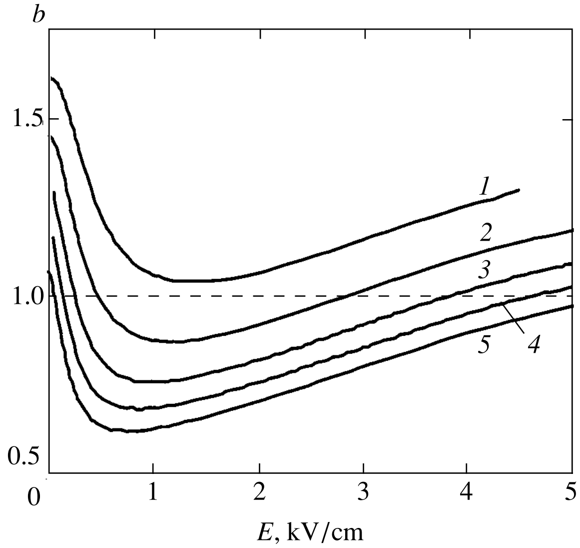


Fig. 7. Electric field strength dependences of the ratio between electron and hole mobilities $b = \mu_e/\mu_h$ at various carrier concentrations obtained from the energy balance equation: $n = 4 \times 10^{15}$ (*1*), 2×10^{15} (*2*), 1×10^{15} (*3*), 5×10^{14} (*4*), and 2×10^{14} cm^{-3} (*5*); $b = \mu_e/\mu_h$, where μ_e and μ_h are the electron and hole mobilities, respectively.

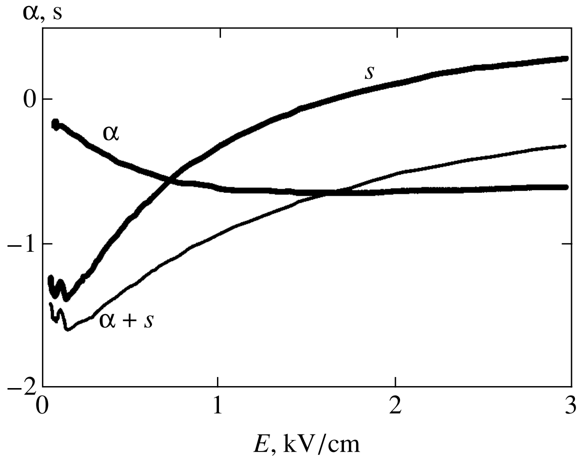


Fig. 8. Electric field strength dependences of the temperature exponents for relaxation times with respect to momentum, $\alpha(E) = d\ln\tau_p/d\ln T_e$, and energy, $s(E) = d\ln\tau_{\mathcal{E}}/d\ln T_e$, and their sums $\alpha(E) + s(E)$ obtained from energy balance equation (1) and from (2). The $\alpha + s > –1$ criterion of the formation of high-field autosolitons is satisfied in fields $E > 750$ V/cm at a $T = 80$ K temperature of the crystal, an $n = 2 \times 10^{15}$ cm^{-3} concentration of the electron–hole plasma, and the field direction along the $\langle 111 \rangle$ axis.

The first stage of high-field autosoliton formation in a sample under the action of an increasing leading voltage pulse edge is characterized by the first jump of the local field upward to a high value ($E_{as} \geq 1000$ V/cm) and a small decrease in the current (Figs. 2, 3). We relate this stage of autosoliton formation [9–12] to a change in the form of the dependence of the electron energy relaxation time $\tau_{\mathscr{E}}$ on the electron temperature (electric field) from decreasing to increasing near the Debye temperature ($\Theta_D = 430$ K for germanium) [25]. According to our calculations, this temperature is reached in field $E \approx 1300$ V/cm. In reality, the $\alpha + s > -1$ criterion is met even earlier, when electron temperature T_e only approaches Θ_D and the inequality $s > -1/2$ at $\alpha \approx -1/2$ is satisfied. This triggers positive feedback for electron temperature growth under the conditions of strong (although short-term) heating of electrons by an electric field [3]. Energy transfer from the electrons, which are powered by the electric field, to the lattice then slows down, the temperature of electrons grows, and their mobility decreases. This causes an increase in the local field and, at a small change in the current value, a still greater heating of electrons and the formation of high-field autosolitons. A stationary autosoliton can exist because the thermodiffusion flow of hot carriers directed from the center of the autosoliton outward is balanced at its boundary by the diffusion counterflow of cold carriers.

The second field intensity jump in the formation of a large-amplitude autosoliton is caused by electron transfer between equivalent valleys. The electrons are transferred from the valleys with a low effective mass to the valley with a high effective mass. This decreases the mobility of the electrons and increases the field strength in the autosoliton region. The third jump of autosoliton field growth can be related to throwing electrons to a higher X valley in a strong autosoliton field ($E_{as} > 5000$ V/cm). In [26], the appearance of current instabilities was related to such an intervalley electron transfer in a strong electric field in n-Ge samples at $T \approx 80$ K. Enhanced field regions situated near the exclusion area at the negative p^+–p junction, close to the blocking positive n–p junction, and on intrinsic sample inhomogeneities can play the role of nuclei for autosoliton formation.

The static state of autosolitons in the p-Ge samples under study can be explained as follows. First, at a low concentration of impurities ($N_a \approx 1 \times 10^{13}$ cm^{-3}) and a high concentration of photogenerated carriers ($n = 10^{14}$–10^{16} cm^{-3}), the electron–hole plasma becomes almost intrinsic, and the bipolar drift velocity in it tends to zero. Secondly, the stationary state is favored by stronger heating of the electrons, which are more mobile than the holes, and the transfer of the electrons from "light" valleys to the "heavier" one in field E directed along the $\langle 111 \rangle$ axis; this decreases the mobility of the electrons and makes the plasma more symmetric. A similar situation arose in our experiments with n-Ge samples when the field was oriented along the $\langle 110 \rangle$ axis [10] and p-Si samples oriented along the $\langle 100 \rangle$ axis [14, 15], in which the formation of static solitons was observed.

According to field strength measurements in a high-field autosoliton and our calculations based on energy balance equation (1) for $T = 77$ K, the temperature of the electrons in the region of such an autosoliton amounts to $T_e = 1000$–3000 K and higher, in agreement with the results of measuring IR radiation from the sample. Because of thermal diffusion, the concentration of the carriers in these autosolitons is lower than in the neighboring sample regions. In addition, the energy of the carriers in the region of such an autosoliton can reach very high values and cause interband breakdowns because of the smallness of the $\tau_{\mathscr{E}}/\tau$ ratio [3]. According to our calculations, the mean electron energy becomes $\mathscr{E}_e \geq 1$ eV at $n \approx 1 \times 10^{15}$ cm^{-3} and $E_{as} \approx 2 \times 10^4$ V/cm.

9. CONCLUSIONS

To summarize, an electron–hole plasma photogenerated uniformly along p-Ge samples and heated by an electric field at $T = 77$ K becomes unstable and stratifies along the current direction into transverse high-field thermodiffusion autosolitons. They are formed in a sample irrespective of the type of junctions (antiblocking, injecting, or blocking) both near junctions and far from them.

The reversal of the bipolar drift direction of an electron–hole plasma in samples with p^+–p junctions, the injection of an electron–hole plasma, and the pouring over of a blocking junction by photogenerated carriers in samples with n–p junctions create the first regions of a sharp increase in the current in current–voltage characteristics and influence the site of autosoliton formation. Repeated current–voltage characteristic regions of a sharp current increase and saturation are related to the rearrangement of autosolitons and require further inquiries.

The multivalley structure of the energy bands of germanium manifests itself in the dynamics of autosoliton formation in the form of three-step autosoliton field growth.

Autosolitons of various types are formed in p-Ge samples depending on the intensity of photogeneration and the applied voltage, namely, static, moving, and pulsating solitons and the turbulent state of the electron–hole plasma.

Note for comparison that we only observed autosolitons of one type in n-Ge samples oriented along the $\langle 111 \rangle$ axis [11]. These were autosolitons that moved in the hole drift direction under bipolar drift conditions; the holes were minority carriers, and they were more mobile because of the transfer of electrons to the valley with a heavy effective mass. In these samples, we did not observe the third field jump during autosoliton formation and the accompanying interband breakdown. At the same time, we only observed static solitons in *p*-Si samples oriented along the ⟨100⟩ axis, also under the conditions of electron transfer to the valley with a large effective mass [14].

ACKNOWLEDGMENTS

This work was financially supported by the State Foundation for Basic Research (project no. 2.4/816), International Science Foundation CRDF (grant no. UP1-368), and Ukrainian–Israeli Scientific Project (grant no. 2M/1807-97). The authors are deeply indebted to O.G. Sarbei (Institute of Physics, National Academy of Sciences of the Ukraine) for many fruitful discussions and help with the calculations.

REFERENCES

1. V. L. Bonch-Bruevich, I. P. Zvyagin, and A. G. Mironov, *Domain Electrical Instabilities in Semiconductors* (Nauka, Moscow, 1972; Consultants Bureau, New York, 1975).
2. E. Scholl, *Nonequilibrium Nuclear Fusion in Semiconductors* (Springer, Berlin, 1987).
3. B. S. Kerner and V. V. Osipov, Usp. Fiz. Nauk **157** (2), 201 (1989) [Sov. Phys. Usp. **32**, 101 (1989)]; **160** (9), 1 (1990); *Autosolitons: a New Approach to Problems of Self-Organization and Turbulence* (Nauka, Moscow, 1991; Kluwer Academic, Dordrecht, 1994).
4. I. A. Vashchenko, B. S. Kerner, V. V. Osipov, and V. F. Sinkevich, Fiz. Tekh. Poluprovodn. (Leningrad) **23**, 1378 (1989) [Sov. Phys. Semicond. **23**, 857 (1989)].
5. B. S. Kerner, D. P. Litvin, and V. I. Sankin, Pis'ma Zh. Tekh. Fiz. **13**, 819 (1987) [Sov. Tech. Phys. Lett. **13**, 342 (1987)].
6. F.-J. Niedernostheide, B. S. Kerner, and H.-G. Purvins, Phys. Rev. B **46**, 7559 (1992).
7. I. K. Kamilov, A. A. Stepurenko, and A. S. Kovalev, Fiz. Tekh. Poluprovodn. (St. Petersburg) **34**, 433 (2000) [Semiconductors **34**, 418 (2000)].
8. B. S. Kerner and V. V. Osipov, Zh. Éksp. Teor. Fiz. **89**, 589 (1985) [Sov. Phys. JETP **62**, 337 (1985)].
9. M. N. Vinoslavskiy, B. S. Kerner, V. V. Osipov, and O. G. Sarbei, J. Phys.: Condens. Matter **2**, 2863 (1990).
10. O. G. Sarbey, M. N. Vinoslavskiy, and D. A. Khizhnyak, in *Proceedings of 9th International Conference on Hot Carriers in Semiconductors*, Ed. by K. Hess *et al.* (Plenum, New York, 1995), p. 461.
11. M. N. Vinoslavskiy and O. G. Sarbey, in *Proceedings of 23rd International Conference on the Physics of Semiconductors*, Ed. by M. Scheffler and R. Zimmermann, World Sci. Publ. Comp. (1996), Berlin (1996), p. 117.
12. O. G. Sarbeĭ, M. N. Vinoslavskii, and A. V. Kravchenko, Ukr. Fiz. Zh. **44**, 190 (1999).
13. M. N. Vinoslavskii and A. V. Kravchenko, Fiz. Tekh. Poluprovodn. (St. Petersburg) **35**, 390 (2001) [Semiconductors **35**, 377 (2001)].
14. M. Vinoslavskiy, A. Kravchenko, and V. Annin, J. Phys.: Condens. Matter **13**, 11623 (2001).
15. M. N. Vinoslavskii and A. V. Kravchenko, Izv. Akad. Nauk, Ser. Fiz. **66**, 226 (2002).
16. A. A. Akopyan, S. A. Vitusevich, and V. K. Malyutenko, J. Phys.: Condens. Matter **4**, 4267 (1992).
17. A. C. Prior, Proc. Phys. Soc. London **76**, 465 (1960).
18. S. Tosima and K. Ando, J. Phys. Soc. Jpn. **23**, 812 (1967).
19. A. A. Akopyan and Z. S. Gribnikov, Fiz. Tekh. Poluprovodn. (Leningrad) **9**, 1485 (1975) [Sov. Phys. Semicond. **9**, 981 (1975)].
20. A. P. Dmitriev, A. E. Stefanovich, and L. D. Tsendin, Phys. Status Solidi A **46**, 45 (1978).
21. M. Ashe and O. G. Sarbey, Phys. Status Solidi A **33**, 9 (1969).
22. Ya. M. Dykman and P. M. Tomchuk, *Transport Phenomena and Fluctuations in Semiconductors* (Naukova Dumka, Kiev, 1984).
23. M. I. D'yakonov, V. I. Perel', and I. N. Yasievich, Fiz. Tekh. Poluprovodn. (Leningrad) **11**, 1364 (1977) [Sov. Phys. Semicond. **11**, 801 (1977)].
24. E. M. Conwell, *High Field Transport in Semiconductors* (Academic, New York, 1967; Mir, Moscow, 1970).
25. K. Seeger, *Semiconductor Physics* (Springer, Berlin, 1974; Mir, Moscow, 1977).
26. J. C. McGroddy and M. I. Nathan, IBM J. Res. Dev. **11**, 337 (1967).

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 174–179.
Translated from Zhurnal Éksperimental’noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 194–199.
Original Russian Text Copyright © 2003 by Éntin, Braginsky.

SOLIDS
Electronic Properties

The Surface Energy of an Electron Gas in Model Crystals

M. V. Éntin* and L. S. Braginsky

Institute of Semiconductor Physics, Siberian Division, Russian Academy of Sciences,
Novosibirsk, 630090 Russia
**e-mail: entin@isp.nsc.ru*

Received January 27, 2003

Abstract—The surface energy of an electron gas in a crystal is considered. The results obtained for a quadratic spectrum are generalized to an arbitrary energy spectrum in certain crystal models. The surface energy of an electron gas with a quadratic spectrum is found for a sample with a rough boundary when the height of irregularities is small compared with the electron wavelength. *© 2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

The quantization of electron states affects the thermodynamic and kinetic properties of a bounded crystal even in systems that are large compared with the characteristic electron wavelengths λ. Thermodynamic quantities are expanded in powers of the ratio of λ to the system size d. The main correction to an extensive quantity can be represented as a surface contribution proportional to the surface area of a sample (see, for example, review [1]). When the size of a large system decreases, this quantization effect precedes the manifestation of properties associated with the fluctuations of electron levels. Indeed, when the correlation of levels is neglected, the fluctuations of the number of levels are inversely proportional to the square root of this number, i.e., to the square root of the volume (the inclusion of the correlation gives rise to additional logarithmic factors depending on the statistics of levels). In facetted monocrystals, there exist edge and vertex contributions in addition to the surface contribution [2].

The dependence of the chemical potential of an electron gas on the surface-to-volume ratio gives rise to a contact potential between grains of the same material but of different sizes and shapes, which leads to a spontaneous recharging of the grains [1, 3]. As a result, a granulated metal may undergo transition to the state of the Hubbard gapless dielectric [3].

Since the surface contributions are formed at a distance of about the electron wavelength from the surface, they are sensitive to the surface perfectness, the crystalline structure, and the surface potential. These factors are especially important in the case of metal particles in which the wavelength is comparable to the lattice constant. In the present paper, we consider surface corrections to thermodynamic quantities in a crystal with regard to the crystalline structure.

2. CONTINUUM MODEL

Consider an Ω potential of electrons with the energy spectrum $\epsilon(\mathbf{k})$ in a plate $0 < z < d$. Suppose that the electron wave function can be described within the model of envelopes $\chi(\mathbf{r})$ that are subject to the identical homogeneous boundary conditions

$$\chi = \pm\zeta\partial\chi/\partial\mathbf{n}|_{z = 0;\, d}. \tag{1}$$

The quantity ζ has a dimension of length and can be regarded as a characteristic length of reflection. In a continuum model with the energy barrier $U(z) = U_0\theta(z)$ on the boundary, parameter $\zeta = 1/\sqrt{2mU_0 - k_z^2}$ (here, m is the effective mass of an electron) represents the penetration depth of electrons into the barrier. In the general case, ζ also depends on the component of the momentum of an electron along the boundary. Boundary condition (1) is satisfied by the wave functions $\chi = \sin(k_z z + \varphi)$, where $\tan\varphi = -k_z\zeta$. The transverse wave vector takes quantized values of $k_z = (\pi n - 2\varphi)/d$, where n is an integer. Potential Ω at low temperature is expressed as

$$\Omega = -2S\int\frac{d^2k}{(2\pi)^2}\sum_{k_z l}(\mu - \epsilon(\mathbf{k}))\theta(\mu - \epsilon(\mathbf{k})).$$

Here and below, S is the lateral surface and $V = Sd$ is the sample volume. To determine the surface contribution to the Ω potential, we replace the summation by an integral using the summation formula

$$\sum_{n=1}^{N} F(n) = \int_0^N dnF(n) - \frac{1}{2}[F(0) + F(N)] + \dots .$$

Taking into account that the function to be summed vanishes at the upper limit, we obtain a decomposition of the Ω potential into the volume, Ω_V, and surface, Ω_S, contributions:

$$\Omega = \Omega_V + 2\Omega_S + \ldots, \quad \Omega_S = \Omega_{S1} + \Omega_{S2},$$

$$\begin{gathered} \Omega_V = -2V\int\frac{d^3k}{(2\pi)^3}(\mu-\epsilon(\mathbf{k}))\theta(\mu-\epsilon(\mathbf{k})), \\ \Omega_{S1} = \frac{S}{2}\int(\mu-\epsilon(\mathbf{k}))\theta(\mu-\epsilon(\mathbf{k}))|_{k_z=0}\frac{d^2k}{(2\pi)^2}, \\ \Omega_{S2} = 2S\int\frac{d^3k}{(2\pi)^3}\frac{\partial\varphi}{\partial k_z}(\mu-\epsilon(\mathbf{k}))\theta(\mu-\epsilon(\mathbf{k})). \end{gathered} \tag{2}$$

According to (2), potential Ω_S contains two contributions. The first one, Ω_{S1}, is determined by electrons with $k_z = 0$. This means that the electron spectrum, which determines this contribution, becomes two-dimensional. At the same time, Ω_{S1} is independent of the boundary condition imposed on the wave function. The contribution Ω_{S2} is determined by all electrons below the Fermi surface and depends on the phase of reflection from the boundary.

In the particular case of a boundary between semiconductor and vacuum, the Fermi energy is much less than the work function U_0. In this case, $\zeta = \text{const}$, $k_z\zeta \ll 1$, and $\varphi = -k_z\zeta$, and we have the following expression for Ω_{S2}:

$$\Omega_{S2} = -2S\zeta\int\frac{d^3k}{(2\pi)^3}(\mu-\epsilon(\mathbf{k}))\theta(\mu-\epsilon(\mathbf{k})). \tag{3}$$

The comparison of Ω_{S1} and Ω_{S2} shows that the latter is $k\zeta$ times smaller than the former. The contribution of Ω_{S2} is proportional to the volume Ω potential and implies that the sample thickness is effectively increased by ζ at each face. The signs of Ω_{S1} and Ω_{S2} are opposite. Thus, the reflection length determines the increase of the effective thickness.

For a given number of electrons in the sample, the presence of the boundary results in the variation of the chemical potential

$$\delta\mu = \frac{1}{V}\frac{\partial\Omega_S}{\partial\mu}/\frac{\partial n(\mu)}{\partial\mu}.$$

Here, $n(\mu)$ is the density of a three-dimensional electron gas; for the quadratic spectrum, $\partial n(\mu)/\partial\mu = \pi^{-2}(2m^3\mu)^{1/2}$.

In a real crystal, the value of ζ is determined by the crystalline structure, the orientation, the perfectness of the boundary, the presence or absence of an insulating layer on the boundary, etc. To analyze possible dependence of ζ on electron momentum, we introduce a boundary condition for the wave function in certain models.

3. SURFACE TENSION OF AN ELECTRON GAS IN A DIELECTRIC

Consider a strong-coupling model with the Hamiltonian

$$H = \varepsilon_{n\nu}\delta_{n\nu,n'\nu'} + t_{n\nu,n'\nu'} + \text{H.c.}$$

Here, $\varepsilon_{n\nu}$ is the energy level of state ν of an atom n and $t_{n\nu,n'\nu'}$ is the transition amplitude between the levels of atoms that constitute a lattice. Suppose that the crystal is bounded and all the states in the valence band are filled. Using the well-known relation

$$\sum_n E_n^k = \operatorname{Sp} H^k,$$

we find that

$$\sum E_n = \sum \varepsilon_{n\nu};$$

i.e., the sum of energies of levels coincides with the sum of energies of levels that are not perturbed by the overlapping of states.

Suppose that atomic energy levels are independent of the distance to the surface. Consider the Ω potential of an electron gas in a dielectric at zero temperature disregarding the interaction of electrons and the overlapping of different nondegenerate levels belonging to different atoms:

$$\Omega = \sum(\mu-E_n)\theta(\mu-E_n).$$

Hence, atomic energy levels that are not perturbed by overlapping determine the total energy of the electron gas in a dielectric, as well as any thermodynamic function of the gas of noninteracting electrons. The electron energy of a crystal is merely the sum of energies of individual atoms; i.e., the surface contribution to the electron energy is zero. At the same time, in a metal, the surface makes a contribution to the crystal energy. Thus, as a band is being filled, the positive contribution of the surface to the energy that arises at low filling should be compensated for by the negative contribution of the hole states in the almost filled band.

4. THE CASE OF A SIMPLE CUBIC LATTICE

Consider an interface between two simple cubic lattices with the boundary plane (001) on the basis of the strong-coupling model. Suppose that there are transitions only between nearest neighbors, described by the transition amplitudes t_1 and t_2 on the left and right of the interface $j = 0$, respectively (Fig. 1), and the energy lev-

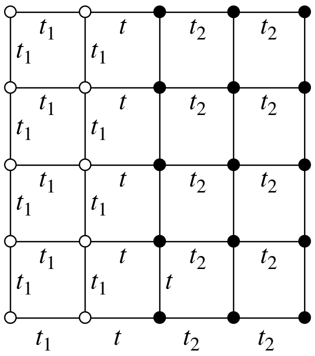


Fig. 1. Model of a cubic lattice.

els of these atoms are ε_1 and ε_2. The transition amplitudes between interface atoms are equal to t. Medium 2 is assumed to be impenetrable for electrons (the energy of electrons localized in the region $z < 0$ proves to be lower than the bottom of the band in the region $z > 0$). The electron spectra in unbounded crystals 1 and 2 are given by the equations

$$\varepsilon - \tilde{\varepsilon}_i - 2t_i \cos k_z l = 0, \quad i = 1, 2,$$

$$\tilde{\varepsilon}_i = \varepsilon_i + 2t_i(\cos k_x l + \cos k_y l).$$

Let us introduce the Fourier images of the amplitudes of electron wave functions c_{mnj} in the plane:

$$c_j = \sum_{m,n} e^{ik_x lm + ik_y ln} c_{mnj}.$$

The strong-coupling equations for boundary atoms allow us to relate the quantities c_j on the left and right of the boundary:

$$(\varepsilon - \tilde{\varepsilon}_1)c_0 - t_1 c_{-1} - tc_1 = 0,$$

$$(\varepsilon - \tilde{\varepsilon}_2)c_1 - tc_0 - t_2 c_2 = 0.$$

Assuming that $c_n = A\sin(-nk_z l + \varphi)$ for $n \le 0$ and $c_n = \exp(-(n-1)\kappa l)$ for $n > 0$, we obtain the following equation for φ:

$$\cot\varphi = -\cot k_z l + \frac{t^2}{t_1 t_2}\frac{e^{-\kappa l}}{\sin k_z l}. \quad (4)$$

A particular case of a truncated lattice of a semi-infinite crystal is equivalent to the vanishing of t. In this case, the phase satisfies the condition $\varphi = -k_z l$. This means that $\varphi = 0$ on an atom next to the boundary atom. For small k_z, the phase tends to zero irrespective of t because

$$\varphi \propto -k_z l t_1 t_2 (t_1 t_2 - t^2 e^{-\kappa l})^{-1}.$$

Thus, the vanishing of a wave function on the boundary is justified if the typical wave vectors of electrons are small, in particular, in the case of a semiconductor. A passage to the limit of a continuum from (4) occurs only for $t^2 = t_1 t_2$ and $\kappa \longrightarrow 0$. The first condition guarantees that the reflection of electrons is small because of the inhomogeneity of t (see [4]), while the second condition guarantees that the below-barrier wavelength is large compared with the lattice constant. In the general case, when the continuum approximation is inapplicable, the quantity ζ in (1) and (3) is replaced by a more complicated expression $(1 - t^2 e^{-\kappa l}/t_1 t_2)|_{k=0}/l$.

Let us dwell on the case of a truncated lattice. Truncating the system at the atoms $n = 1$ and $n = N$ is equivalent to imposing the boundary conditions $c_0 = c_{N+1} = 0$. The solution $c_n = \sin k_z nl$ satisfies the condition at the zero atom and, for $k_z l = \pi K/l(N+1)$, the condition at the $(N+1)$th atom. Thus, in a strong-coupling model, one may assume that the condition $\varphi = 0$ holds on the boundary situated at a distance of the lattice constant from the boundary atoms. Measuring the crystal size by the number of unit cells, we should increase this number by one, i.e., we should replace $V/S = d$ in (2) by $d + l$. This means that $\zeta = l$ and the term

$$-Sl\int \frac{d^3k}{(2\pi)^3}(\mu - \epsilon(\mathbf{k}))\theta(\mu - \epsilon(\mathbf{k})) \quad (5)$$

is added to the value of Ω given by (2).

The potential Ω is determined by the density of states: $\Omega_S = -S\int(\mu - E)\theta(\mu - E)\nu(E)dE$. Let us calculate the density of two-dimensional states in this model, that corresponds to contribution (5):

$$\nu = 2\int \frac{d^2\mathbf{k}}{(2\pi)^2}\delta(E - 2t(\cos k_x l + \cos k_y l))$$

$$= \frac{1}{\pi^2 l^2 t}K\left(\sqrt{1 - \left(\frac{E - \epsilon_0}{4t}\right)^2}\right).$$

Here, $K(k)$ is a complete elliptic integral of the first kind. The density of states has a singularity at $E = \epsilon_0$, which corresponds to nesting.

Bevel cut. Next, consider a simple cubic lattice truncated at the boundaries (011) and (111). In these cases, the boundary also represents a symmetry plane, which facilitates the solution of the problem. Truncating a lattice is equivalent to finding a solution to the equations for amplitudes in an unbounded crystal, such that the amplitudes vanish on an imaginary atomic plane next to the last row. Such a solution is given by

$$c_{nmj} \sim \sin(k_3 l_3 (n_3 + 1)),$$

where axis 3 is directed along the normal to the boundary and l_3 is the interplanar spacing along this axis.

For the boundary (011), we have $\varphi = k_3\zeta = k_3 l/\sqrt{2}$, and, for the boundary (111), we have $\varphi = k_3\zeta = k_3 l/\sqrt{3}$. Since the interplanar spacing decreases with the slope of a surface, it is as if the crystal thickness decreases compared with the case of the boundary (001).

Two atoms in a unit cell. Consider a one-dimensional lattice with a unit cell consisting of two atoms with the energies of states ϵ_1 and ϵ_2. The transition amplitudes between the atoms of a unit cell are t_1, and those between the nearest atoms of adjacent cells are t_2. A system of strong-coupling equations is expressed as

$$(\epsilon_1 - E)c_{1,n} + t_2 c_{2,n-1} + t_1 c_{2,n} = 0,$$

$$(\epsilon_2 - E)c_{2,n} + t_1 c_{1,n} + t_2 c_{1,n+1} = 0.$$

The energy spectrum of the unbounded chain is given by

$$(\epsilon_1 - E)(\epsilon_2 - E) - (t_1^2 + t_2^2 + 2t_1 t_2 \cos k_z l) = 0.$$

The boundary conditions depend on the type of the atom at which the lattice is truncated. If the lattice is truncated at an atom of type 1, then $c_{2,0} = 0$; if the truncation occurs at an atom of type 2, we have $c_{1,1} = 0$. In the first case, the boundary conditions yield the equation

$$e^{2i\varphi} = \frac{t_1 + t_2 e^{ik_x l}}{t_1 + t_2 e^{-ik_x l}}, \tag{6}$$

while, in the second case, we have to change $t_1 \longleftrightarrow t_2$ in Eq. (6).

The phase φ depends on k_z and the ratio t_1/t_2. However, when $k_z \longrightarrow 0$, we have $\varphi \longrightarrow 0$.

5. THE EFFECT OF THE BOUNDARY ROUGHNESS ON THE SURFACE TENSION OF AN ELECTRON GAS

Above, we obtained the reflection length on a smooth crystal surface. A surface inclined to the symmetry plane inevitably has a steplike character. Irregularities whose height does not exceed the Fermi wavelength form an effective smooth wall for electrons. The position of this wall, which determines the reflection length, can be specified to within the height of irregularities. In this section, we determine the reflection length from a steplike rough surface with the height of steps smaller than the wavelength. We will apply a continuum model corresponding to the case of a wavelength large compared to the size of an atom, which is important for semiconductors and semimetals.

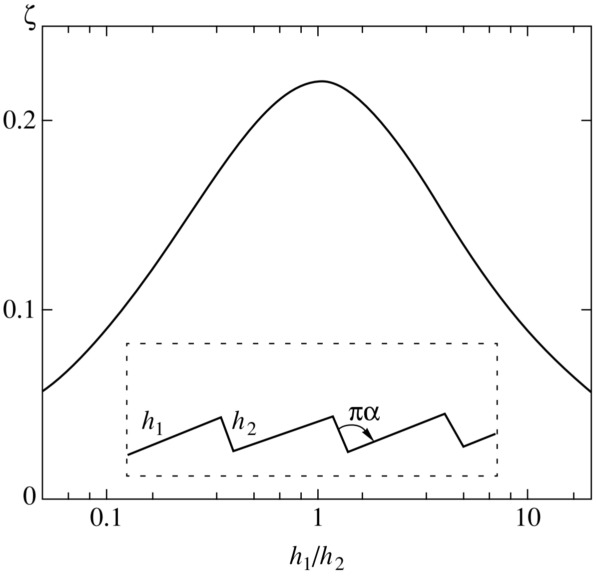


Fig. 2. Reflection length ζ as a function of the ratio of sides of right-angle steps. The inset depicts the shape of the surface.

Formally, the problem is stated as follows: the wave function vanishes on the surface $y = u(x, z) = 0$. The function $u(x, z)$ is bounded, does not decrease at infinity, and equals zero on the average. The characteristic scales of irregularities—the height h and the length L—are less than the wavelength. At a distance of $\lambda \gg y \gg h$ from the surface, we can neglect energy in the Schrödinger equation, thus reducing it to the Laplace equation; a solution to the latter equation behaves as $\psi = -F(y - \zeta)$ as $y \longrightarrow \infty$. We have to relate ζ to $u(x, z)$. The above asymptotics implies that one can impose the effective boundary condition $\psi = 0|_{y=\zeta}$ on the solution of the Schrödinger equation.

Physical analysis. Suppose that the irregularities are smooth, $h/L \ll 1$. In this case, different regions along the surface weakly influence each other, and the effective boundary condition must be imposed near the averaged line of the surface shape $\zeta = \langle u(x, z)\rangle = 0$. Now, consider the case of $h/L \gg 1$. In this limit, the wave function does not penetrate into narrow wells; therefore, in the first approximation, we have to impose the zero boundary condition on the peaks of the surface, and $\zeta = \max(u(x, z)) - \gamma L$. The correction term γL with an arbitrary coefficient γ is attributed to the penetration of the wave function into the gaps to a distance on the order of the gap width. Finally, suppose that the surface consists of right-angle steps of the same height (see the inset in Fig. 2). If $L \sim h$, then $\zeta \sim h$. If $L \gg h$, then the steps represent a rarefied “gas” on the surface and ζ is proportional to its concentration $1/L$, i.e., $\zeta \sim h^2/L$. According to the analysis carried out, the effective boundary ζ can be represented as $\zeta = hf(h/L)$, where $f \longrightarrow 0$ if $h/L \longrightarrow 0$ and $f \longrightarrow 1$ if $h/L \longrightarrow \infty$.

Exactly solvable model. In the two-dimensional case, the Laplace equation with the boundary condition $\psi = 0|_{y = u(x)}$ is solved by the conformal mapping technique. This fact stimulates the consideration of one-dimensional irregularities of the surface $y = u(x)$. Among such irregularities are terraces that are frequently occur during the crystal growth. In addition, such a statement may be adequate to two-dimensional quantum dots.

Consider a complex plane of variables $\eta = \eta_1 + i\eta_2$, bounded from below by a sawtooth boundary of the form

$$u(\eta_1) = \frac{h}{l}\eta_1 \quad \text{for} \quad 0 < \eta_1 < l,$$

$$u(\eta_1) = \frac{h}{L-l}(L - \eta_1) \quad \text{for} \quad l < \eta_1 < L,$$

$$u(\eta_1 + L) = u(\eta_1).$$

The upper half-plane in the variables $\xi = \xi_1 + i\xi_2$ can be mapped into the domain $\eta_2 > u(\eta_1)$ by the Schwarz–Christoffel mapping [5]. Since the boundary represents an open infinite polygon, the required mapping is obtained by passing to the limit. For an infinite periodic polygon with the vertex angles $\pi\alpha$ and $\pi(2 - \alpha)$, $\alpha < 1$, the Schwarz–Christoffel mapping can be sought in the form

$$\eta(\xi) = A\int_0^{\xi} \frac{\Pi_n(\eta_1 - \pi n - a)^{1-\alpha}}{\Pi_n(\eta_1 - \pi n)^{1-\alpha}} d\eta_1.$$

Here, n and $n - a$ are the images of the vertices, $\pi\alpha$ and $\pi(2 - \alpha)$ are the vertex angles, and a cut is made along the half-line $(-\infty, 0)$. The constants A and $a < \pi$ should be determined from the condition that the vertices of the polygon with the angles $\pi\alpha$ and $\pi(2 - \alpha)$ should coincide with the points nL and $l + ih + nL$, respectively. The formal divergence of the numerator and denominator can be eliminated by dividing the numerator and denominator by infinite constants Π_n and renormalizing the constant A. As a result, the products are transformed into a sine transform:

$$\eta(\xi) = A\int_0^{\xi}\left(\frac{\sin(a - \xi)}{\sin\xi}\right)^{1-\alpha} d\xi. \tag{7}$$

The sides of the step have lengths of h_1 and h_2 that satisfy the relations

$$h_1 = |A|\int_0^{a}\left(\frac{\sin(a - \xi)}{\sin\xi}\right)^{1-\alpha} d\xi,$$

$$h_2 = |A|\int_a^{\pi}\left(\frac{\sin(\xi - a)}{\sin\xi}\right)^{1-\alpha} d\xi,$$

which uniquely determine the quantities $|A|$ and a. The integrand in (7) has a period π. The phase of $A = |A|e^{i\beta}$ is determined by the condition that the averaged boundary line is parallel to axis η_1; from this condition, we find $\beta = (\pi - a)(1 - \alpha)$.

The function $\mathrm{Im}(\xi(\eta))$ is harmonic in the upper half-plane and satisfies the zero condition on the sawtooth boundary. Therefore, $\mathrm{Im}(\xi(\eta)) = \psi(\eta)$ defines the wave function as a function of coordinates η. For large η, we have

$$\mathrm{Im}(\xi(\eta)) \sim -F(\eta_2 - \zeta)$$

in the upper half-plane. According to (7),

$$\eta(\xi) \sim |A|\xi + C.$$

For ζ, we obtain

$$\zeta = |A|\,\mathrm{Im}\int_0^{i\infty}\left[e^{i(\pi - a)(1-\alpha)}\left(\frac{\sin(a - \xi)}{\sin\xi}\right)^{1-\alpha} - 1\right] d\xi.$$

In a particular case of a right-angle sawtooth profile, $\alpha = 1/2$, we have $h = h_1h_2/\sqrt{h_1^2 + h_2^2}$, $\sqrt{h_1^2 + h_2^2} = \pi|A|$. The phase β is determined by the condition $\beta = \arctan(h_2/h_1)$. For ζ, we obtain the expression

$$\zeta = h\,\mathrm{Im}\int_0^{i\infty} e^{i(\pi - a)/2}\left(\sqrt{\frac{\sin(a - \xi)}{\sin\xi}} - 1\right) d\xi - \frac{h}{2}$$

($h/2$ is subtracted to reduce the averaged boundary line to zero).

Figure 2 displays ζ/h_1 as a function of h_1/h_2. The maximal value of ζ is attained in the limit of an equilateral shape of the teeth, when $a = \pi/2$. This value is given by

$$\zeta = h\left(\int_0^{\infty}(\sqrt{\coth\eta_2} - 1)\,d\eta_2 - 1/2\right) = \frac{h\ln 2}{\pi} \approx 0.22h.$$

It should be noted that ζ is relatively small compared with the height of the surface irregularities.

In conclusion, we list the factors considered in this paper that determine the surface contributions to thermodynamics and are reduced to the effective variation in the sample size: (1) the quantization of states, which results in an effective decrease in the sample size by the characteristic electron wavelength; (2) the penetration of the wave function beyond the sample boundary (an increase by the depth of penetration into the adjoining medium); (3) the taking into consideration of the crystalline structure (an increase by a value determined by the lattice constant); and (4) surface irregularities (an increase, relative to the average surface, by a quantity related to the height and width of irregularities).

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research, project nos. 00-02-17658, 02-02-16377, and 01-02-17694.

REFERENCES

1. É. L. Nagaev, Usp. Fiz. Nauk **162** (9), 49 (1992) [Sov. Phys. Usp. **35**, 747 (1992)].
2. M. M. Makhmudian and M. V. Éntin, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 167 (2001) [JETP Lett. **73**, 149 (2001)].
3. É. M. Baskin and M. V. Éntin, Pis'ma Zh. Éksp. Teor. Fiz. **70**, 511 (1999) [JETP Lett. **70**, 520 (1999)].
4. L. S. Braginskiĭ and D. A. Romanov, Fiz. Tverd. Tela (St. Petersburg) **37**, 2122 (1995) [Phys. Solid State **37**, 1154 (1995)].
5. M. A. Lavrent'ev and B. V. Shabat, *Methods of the Theory of Functions of a Complex Variable*, 5th ed. (Nauka, Moscow, 1987; Wissenschaften, Berlin, 1967).

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 180–193.
Translated from Zhurnal Éksperimental’noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 200–213.
Original Russian Text Copyright © 2003 by Sedletsky.

NONLINEAR PHYSICS

The Fourth-Order Nonlinear Schrödinger Equation for the Envelope of Stokes Waves on the Surface of a Finite-Depth Fluid

Yu. V. Sedletsky
Institute of Physics, National Academy of Sciences of Ukraine, Kiev, 03028 Ukraine
e-mail: sedlets@iop.kiev.ua

Received January 14, 2003

Abstract—The method of multiple scales is used to derive the fourth-order nonlinear Schrödinger equation (NSEIV) that describes the amplitude modulations of the fundamental harmonic of Stokes waves on the surface of a medium- and large-depth (compared to the wavelength) fluid layer. The new terms of this equation describe the third-order linear dispersion effect and the nonlinearity dispersion effects. As the nonlinearity and the dispersion decrease, the equation uniformly transforms into the nonlinear Schrödinger equation for Stokes waves on the surface of a finite-depth fluid that was first derived by Hasimoto and Ono. The coefficients of the derived equation are given in an explicit form as functions of *kh* (*h* is the fluid depth, and *k* is the wave number). As *kh* tends to infinity, these coefficients transform into the coefficients of the NSEIV that was first derived by Dysthe for an infinite depth. © *2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

The nonlinear Schrödinger equation (NSE) describes the time evolution of the envelope of a packet of rapidly oscillating low-amplitude waves in a weakly dispersive, nondissipative medium. In recent years, the development of the physics of wave processes modeled by the NSE has shifted toward studying increasingly nonlinear and high-dispersion effects, which corresponds to the inclusion of the next terms in this equation. Most of the studies in this direction pertain to analyzing the possibilities of a further compression and compaction of optical solitons as information bits in optical fibers (see, e.g., [1, 2]). As regards other media, only isolated studies have appeared so far in which equations that include higher order effects were derived: in plasma [3], on the surface of a fluid of infinite [4–7] and finite [8, 9] depth, in magnetics [10], and for systems described by the nonlinear Klein–Gordon equation [11].

Thus, when only one order of smallness (the next after the principal order) for the amplitude of the envelope of the first harmonic A of rapid oscillations was additionally taken into account, the following generalized NSE was derived in [2, 3, 10]:

$$i\frac{\partial A}{\partial T} \pm \varepsilon\left(\frac{1}{2}\frac{\partial^2 A}{\partial X^2} + A|A|^2\right) + i\varepsilon^2\left(\alpha\frac{\partial^3 A}{\partial X^3} + \beta|A|^2\frac{\partial A}{\partial X} + \gamma A\frac{\partial|A|^2}{\partial X}\right) = 0. \tag{1}$$

Here, ε is the formal parameter that characterizes the smallness of the quantities oscillating with amplitude A and the slowness of the variation in A compared to the rapid oscillations. In [4–7], Eq. (1) also includes the term with the product of the amplitudes of the zeroth and first harmonics, and it was supplemented with an equation that expresses the zeroth harmonic in terms of the first harmonic. A system of two equations for the zeroth and first harmonics was obtained in [8], and a system of equations for these two harmonics, including the amplitudes of the preceding orders, was obtained in [9].

An equation of form (1) was first derived in [12] for wave beams. Equation (1) is commonly called the higher order nonlinear Schrödinger equation or the fourth-order NSE (NSEIV). The name “fourth order” stems from the fact that the last three terms in Eq. (1) arise when the terms of order $O(\varepsilon^4)$ are included. In fact, terms of the next order in amplitude arise in the approximation $O(\varepsilon^5)$. This advance was made, for example, in [1, 11]. The second and third terms in Eq. (1) correspond to the classical NSE (the $O(\varepsilon^3)$) approximation). They include the main dispersion contribution—the group-velocity dispersion and the principal nonlinear term—phase self-modulation (the amplitude dependence of the frequency—the Stokes effect). The next terms of order $O(\varepsilon^4)$ have been the subject of research in recent years. The fourth and fifth terms are responsible for the third-order linear dispersion (Airy aberrations) and the pulse profile self-steepening, respectively; the sixth term describes the Raman frequency self-shift effect. Basic publications devoted to the study of soliton and quasi-soliton solutions to Eq. (1) are listed, for example, in [1, 2, 13, 14].

The derivation of NSEIV for the amplitude of the envelope of the fundamental harmonic ultimately consists in eliminating the amplitudes of other harmonics and obtaining the coefficients α, β, and γ in an analytic form. The values and signs of these coefficients as well as their combinations predetermine the properties of the solutions to this equation [13, 14]. Thus, in the ε^3 approximation ($\alpha = \beta = \gamma = 0$), the same or opposite signs of the coefficients of the second and third terms (the choice of + or – in Eq. (1)) are a criterion for the possible existence of solitons under zero boundary conditions. At some values of α, β, and γ, Eq. (1) has N-soliton solutions and can even be integrated by the inverse scattering transform method (for basic and current references, see [1, 2] and [13, 14], respectively).

The goal of this study is to obtain the coefficients of NSEIV for the amplitude of the envelope of weakly nonlinear waves on the surface of a finite-depth fluid (Stokes waves) in an explicit form.

The classical NSE (the $O(\varepsilon^3)$ approximation) for Stokes waves was derived by Zakharov in his pioneering paper [15] using the Hamiltonian method in the three-dimensional case for a narrow spectrum and an infinite depth from the integral equations derived by him. Subsequently, Yuen and Lake [16] obtained the NSE by the averaged Lagrangian method also for an infinite depth. The case of a finite depth was investigated in [17] by the method of multiple scales, in [18] by the Hamiltonian method, and in [19] by the WKB method. The NSE was derived by the method of multiple scales in the two-dimensional case by Hasimoto and Ono [20] and, later, by Stiasnie and Shemer [21] from Zakharov's integral equations. Subsequently, the method of multiple scale was used to generalize this equation in different ways: to the three-dimensional case [22], with an allowance made for the surface tension [23], the bottom irregularities [24], the counter-propagating waves [25], etc.

The NSEIV in the infinite-depth approximation (Dysthe's equation) was derived in [4] by an iterative method, in [26] from Zakharov's integral equations, and in [5] by asymptotically expanding the derivative $\partial A/\partial t$. The sign of one of the terms in Dysthe's equation was corrected in [27], and the same type of error in [26] was corrected in [28]. As a result, Dysthe's equation, which was repeatedly derived by different authors and using different methods, became a classical extension of the NSE for Stokes waves in an infinite-depth fluid. The authors of [29–32] showed that the experimental studies of the time evolution of Stokes wave packets on the surface of deep water are consistent with the numerical calculations based on Dysthe's equation. In addition, this equation was generalized to include the capillarity [28, 6], the flow [7], and the thermocline [33] and was modified for the spatial rather than time evolution [4, 34].

The derivation of NSEIV for a finite-depth fluid and further analysis of the calculated coefficients of the equation and the determination of the dependence of its solutions on the depth and other parameters have the same motivation. Based on the averaged-Lagrangian and harmonic-series methods, Whitham [35] and Benjamin [36], respectively, found that a weakly nonlinear harmonic wave with a small harmonic amplitude perturbation behaved differently if the product of the wave number and the fluid depth was less or more than 1.363 when the NSE and the properties of its soliton solutions were not yet known. In the former case, the wave is modulationally stable; in the latter case, the modulations increase with time. Various evolution scenarios for the modulations were discussed: passage to chaos, return to the original state, a uniform harmonic distribution. After the NSE was derived for a finite depth [20], the number 1.363 appeared for the third time: at this value of kh, the coefficient of the NSE changes its sign to negative, and, according to the properties of this equation, it has soliton solutions for $kh > 1.363$; therefore, the growing modulations can transform into solitons. The possibility of describing the nonlinear evolution of the unstable Benjamin–Feir mode and its FPU-type return with the help of growing-and-decaying solitons [37] and breathers [37, 38] is being explored. On the other hand, the existence of a singular point ($kh = 1.363$) for Stokes waves at which the nonlinear term is equal to zero prompts us to include the next terms of the NSE and, thus, to study the change of signs and relations between the coefficients of NSEIV and the corresponding change in the properties of the solution with kh and to describe the modulational instability and return in terms of the fourth-order NSE. In [13], a wide variety of properties of the solutions to NSEIV, in particular, the evolution of the parameters of the Benjamin–Feir modulational instability, was shown. The fact that the increments and the domain of parameters of this instability decrease when the next terms of the NSE are included was shown for an infinite depth in [4, 5]. Based on the properties of the solutions to Eq. (1), the authors of [4] explained the mechanism of the experimentally found frequency shift.

To solve the formulated problem of deriving the NSEIV for a finite depth, we used the method of multiple scales. The technique used here when introducing the next fast times and coordinates is described in [11]. The difficulties of the method of multiple scales in the problem of allowing for the next terms in the NSE for Stokes waves were previously pointed out in the simpler case of an infinite depth (see the review article [39], Section 5.4). For a finite depth, the NSEIV was derived by the method of multiple scales in [40, 41] and by the WKB method in [8]. In [40], the coefficients of the NSEIV were obtained only at the point $kh = 1.363$ itself, while the expressions derived in [41] for the coefficients do not satisfy the limiting case of an infinite depth. The system of equations derived in [8] was set up for the main component of the first harmonic amplitude (arising in the first order in ε), while we derived an equation for the total amplitude (the sum of the contributions from all approximations to the first harmonic). Another fundamental difference between our study and [8] is that we assumed the motion of a wave-induced flow (zeroth harmonic) with the group velocity of the first harmonic. Without this ansatz, the original system of equations was reduced in [8] only to a system of two equations for the amplitudes of the zeroth harmonic and the first order of the first harmonic, while we derived the closed NSEIV (1) for the total amplitude of the first harmonic of the wave profile.

2. FORMULATION OF THE PROBLEM

We will consider the motion of an ideal incompressible homogeneous fluid in the xy plane. We take the x coordinate along the direction of the motion and the y coordinate along the vertical direction. We assume that the fluid is bounded below by a solid bottom at a depth $y = -h$ and above by a rapidly oscillating free surface $\eta(x, t)$. The evolution equation for the amplitude A_1 of the envelope of the first harmonic of these modulations should be determined.

The fluid velocity potential $\phi(x, y, t)$ satisfies the Laplace equation

$$\begin{gathered}\phi_{xx} + \phi_{yy} = 0, \quad -\infty < x < \infty,\\ -h \le y \le \eta(x, t).\end{gathered} \tag{2}$$

The boundary conditions on the free surface $\eta(x, t)$ are the kinematic condition, implying that the fluid particles on the surface itself move along this surface without leaving it,

$$\eta_t - \phi_y + \eta_x\phi_x = 0, \quad y = \eta(x, t), \tag{3}$$

and the dynamical condition, implying that the fluid pressure in the Cauchy–Bernoulli integral is constant (equal to the atmospheric pressure) on the entire surface,

$$g\eta(x, t) + \phi_t + \frac{1}{2}(\phi_x^2 + \phi_y^2) = 0, \quad y = \eta(x, t). \tag{4}$$

The vertical fluid particle velocity at the bottom is equal to zero,

$$\phi_y = 0, \quad y = -h, \tag{5}$$

and the velocity potential itself is a bounded quantity.

3. THE METHOD OF MULTIPLE SCALES. SUCCESSIVE APPROXIMATIONS

We will use the method of multiple scales, as described for higher approximations in [11]. The variations of ϕ and η with time t will be characterized by a set of rapid oscillations corresponding to time t_0 and slow oscillations described by the time τ that consists of a slow time t_1, a very slow time t_2, and a particularly slow time t_3. In that case,

$$\begin{gathered}\frac{\partial}{\partial t} = \frac{\partial}{\partial t_0} + \varepsilon\frac{\partial}{\partial \tau},\\ \frac{\partial}{\partial \tau} = \frac{\partial}{\partial t_1} + \varepsilon\frac{\partial}{\partial t_2} + \varepsilon^2\frac{\partial}{\partial t_3}.\end{gathered} \tag{6}$$

The variations in space along the x axis will be characterized by a ordinary x_0 coordinate and an extended x_1 coordinate:

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial x_0} + \varepsilon\frac{\partial}{\partial x_1}. \tag{7}$$

We assume the velocity potential and the surface displacement (profile) to be small quantities on the order of ε and retain the terms up to ε^4 inclusive:

$$\phi = \varepsilon\phi_1 + \varepsilon^2\phi_2 + \varepsilon^3\phi_3 + \varepsilon^4\phi_4, \tag{8}$$

$$\eta = \varepsilon\eta_1 + \varepsilon^2\eta_2 + \varepsilon^3\eta_3 + \varepsilon^4\eta_4. \tag{9}$$

We substitute Eqs. (6)–(9) into Eq. (2) and into the boundary conditions (3) and (4), in which we expand ϕ_x, ϕ_y, and ϕ_t in a Taylor power series of η at $y = 0$, and collect the terms with the same powers of ε.

3.1. The $O(\varepsilon^1)$ Approximation

Using the $O(\varepsilon^1)$ approximation, we obtain the results of the linear theory. From the Laplace equation and the boundary condition at the bottom, we have

$$\phi_1 = \phi_1^{(0)} + (\phi_1^{(1)} + \text{c.c.}), \tag{10}$$

where

$$\begin{gathered}\phi_1^{(0)} = \Psi_1,\\ \phi_1^{(1)} = -\frac{i\omega\cosh(k(y + h))}{2k\sinh(kh)}Ae^{i\theta}, \quad \theta = kx_0 - \omega t_0;\end{gathered}$$

from the dynamical boundary condition, we have

$$\eta_1 = \eta_1^{(0)} + \eta_1^{(1)}, \tag{11}$$

where

$$\eta_1^{(0)} = 0, \quad \eta_1^{(1)} = \frac{1}{2}Ae^{i\theta} + \text{c.c.} \tag{12}$$

From the kinematic boundary condition, we obtain a linear dispersion law,

$$\omega^2 - k\sigma g = 0, \quad \sigma \equiv \tanh(kh). \tag{13}$$

We will seek equations for the evolution of the following two new unknown functions of the slow variables x_1, t_1, t_2, and t_3 with time τ: the amplitude of the first harmonic A, which, according to (12), is the envelope of the surface displacement rapidly oscillating with time t_0; and the amplitude of the zeroth harmonic Ψ_1 of the velocity potential (average flow) induced by these rapid oscillations and constant with time t_0.

3.2. The $O(\varepsilon^2)$ Approximation

In the $O(\varepsilon^2)$ approximation, we derive the equation for the potential component ϕ_2,

$$\frac{\partial^2\phi_2}{\partial y^2} + \frac{\partial^2\phi_2}{\partial x_0^2} = -2\frac{\partial^2\phi_1}{\partial x_1 \partial x_0},$$

or, substituting (10) here, the equation

$$\frac{\partial^2\phi_2}{\partial y^2} + \frac{\partial^2\phi_2}{\partial x_0^2} = -\frac{\omega\cosh(k(y+h))}{\sinh(kh)}\frac{\partial A}{\partial x_1}e^{i\theta} + \text{c.c.}$$

We write its solution that satisfies the boundary condition

$$\frac{\partial\phi_2}{\partial y} = 0, \quad y = -h,$$

as the sum of harmonics

$$\phi_2 = \phi_2^{(0)} + (\phi_2^{(1)} + \phi_2^{(2)} + \text{c.c.}), \tag{14}$$

where

$$\phi_2^{(0)} = \Psi_2, \tag{15}$$

$$\begin{aligned}\phi_2^{(1)} &= \frac{1}{\sinh(kh)} \\ &\times\left(D\cosh(k(y+h)) - (y+h)\sinh(k(y+h))\frac{\omega}{2k}\frac{\partial A}{\partial x_1}\right)e^{i\theta}.\end{aligned} \tag{16}$$

We determine the coefficients of the homogeneous solution D from the condition that solution (16) be bounded as the depth tends to infinity [19]. Taking into account the asymptotic relation for the potential $\phi_2^{(1)}$ from (16), we find that for this quantity to be bounded, we must set

$$D = \frac{\omega}{2k}\frac{\partial A}{\partial x_1}hD_1(h),$$

where the arbitrary factor $D_1(h)$ should only satisfy the condition

$$\lim_{h\to\infty} h(D_1(h) - 1) = 0.$$

Thus, the limit of $\phi_2^{(1)}$ is bounded,

$$\lim_{kh\to\infty}\phi_2^{(1)} = -\frac{\omega}{2k}\frac{\partial A}{\partial x_1}ye^{yk}e^{i\theta}.$$

Using the arbitrariness of D_1, we calibrate $\phi_2^{(1)}$ in such a way that the condition

$$\phi_2^{(1)}\Big|_{y=0} = 0$$

is satisfied on the fluid surface.

In this case,

$$D_1(h) = \tanh(kh).$$

Thus, we consider here the case of an arbitrary depth, including an infinite depth. The choice of a boundary condition (the selection rule for the coefficient D) is determined, first, by its naturalness and, second, by the necessity of ensuring the legitimacy of the passage to the limit of an infinite depth in the derived formulas [4]. This remark also applies to the selection rule for the coefficients of the homogeneous solution in the next approximations.

Next, from the dynamical boundary condition at the $O(\varepsilon^2)$ step

$$\begin{aligned}&\frac{\partial\phi_2}{\partial t_0} + \frac{\omega^2}{k\sigma}\eta_2 = -\eta_1\frac{\partial^2\phi_1}{\partial y\partial t_0} \\ &\quad -\frac{1}{2}\left(\frac{\partial\phi_1}{\partial y}\right)^2 - \frac{\partial\phi_1}{\partial t_1} - \frac{1}{2}\left(\frac{\partial\phi_1}{\partial x_0}\right)^2, \\ &\frac{\partial\phi_2}{\partial t_0} + \frac{\omega^2\eta_2}{k\sigma} = \frac{\omega^2(\sigma^2-1)}{4\sigma^2}\bar{A}A - \frac{\partial\Psi_1}{\partial t_1} \\ &\quad + i\frac{\omega}{2\sigma k}\frac{\partial A}{\partial t_1}e^{i\theta} + \frac{1}{8\sigma^2}(3\sigma^2-1)\omega^2A^2e^{2i\theta} + \text{c.c.},\end{aligned} \tag{17}$$

we obtain the corresponding surface profile

$$\eta_2 = \eta_2^{(0)} + \eta_2^{(1)} + \eta_2^{(2)} + \text{c.c.}, \tag{18}$$

which so far has been determined only for the zeroth and first harmonics:

$$\eta_2^{(0)} = \frac{\sigma^2 - 1}{4\sigma} k\bar{A}A - \frac{\sigma k}{\omega^2}\frac{\partial \Psi_1}{\partial t_1}, \qquad (19)$$

$$\eta_2^{(1)} = \frac{i}{2\omega}\frac{\partial A}{\partial t_1} e^{i\theta}. \qquad (20)$$

Finally, from the kinematic boundary condition of the $O(\varepsilon^2)$ approximation

$$\begin{gathered}\frac{\partial \eta_2}{\partial t_0} - \frac{\partial \phi_2}{\partial y} = -\frac{\partial \eta_1}{\partial t_1} + \eta_1 \frac{\partial^2 \phi_1}{\partial y^2} - \frac{\partial \eta_1}{\partial x_0}\frac{\partial \phi_1}{\partial x_0}, \\ \frac{\partial \eta_2}{\partial t_0} - \frac{\partial \phi_2}{\partial y} = -\frac{1}{2}\frac{\partial A}{\partial t_1} e^{i\theta} - \frac{1}{2\sigma} i\omega k A^2 e^{2i\theta} + \text{c.c.},\end{gathered} \qquad (21)$$

we have an identity for the zeroth harmonic and an uncertainty for the first harmonic. This uncertainty can be eliminated only by setting

$$\frac{\partial A}{\partial t_1} + V_g \frac{\partial A}{\partial x_1} = 0, \qquad (22)$$

where

$$V_g \equiv \frac{\partial \omega}{\partial k} = \frac{\omega}{2k}\left(1 + \frac{1 - \sigma^2}{\sigma} kh\right) \qquad (23)$$

is the group velocity. It follows from (22) and the complex-conjugate equation that

$$\frac{\partial (A\bar{A})}{\partial t_1} + V_g \frac{\partial (A\bar{A})}{\partial x_1} = 0. \qquad (24)$$

Thus, the first harmonic and its modulus evolve in the coordinate system of the slow time t_1 and the extended longitudinal x_1 coordinate with the group velocity.

For the second harmonic, the system of equations of the dynamical and kinematic conditions (17) and (21) is not overdetermined and allows the coefficients for the second harmonic $\phi_2^{(2)}$ and $\eta_2^{(2)}$ to be calculated:

$$\eta_2^{(2)} = \frac{3 - \sigma^2}{8\sigma^3} k A^2 e^{2i\theta}, \qquad (25)$$

$$\phi_2^{(2)} = 3i\omega \frac{\sigma^2 - 1}{8\sigma^3} \frac{\cosh(2k(y + h))}{\sinh(2kh)} A^2 e^{2i\theta}. \qquad (26)$$

In (20), we can now take into account the fact that

$$\frac{\partial A}{\partial t_1} = -V_g \frac{\partial A}{\partial x_1}$$

follows from (22), and the derivative $\partial \Psi_1 / \partial t_1$ required in (19) can be determined in the next approximation.

3.3. The $O(\varepsilon^3)$ Approximation

In the $O(\varepsilon^3)$ approximation, we have the equation for the potential component ϕ_3

$$\frac{\partial^2 \phi_3}{\partial x_0^2} + \frac{\partial^2 \phi_3}{\partial y^2} = -\frac{\partial^2 \phi_1}{\partial x_1^2} - 2\frac{\partial^2 \phi_2}{\partial x_1 x_0}.$$

Substituting expressions (10) and (14), we obtain an equation for determining ϕ_3:

$$\begin{gathered}\frac{\partial^2 \phi_3}{\partial x_0^2} + \frac{\partial^2 \phi_3}{\partial y^2} = -\frac{\partial^2 \Psi_1}{\partial x_1^2} \\ + \left\{ i\frac{\omega}{k}\left(k(y + h)\frac{\sinh(k(y + h))}{\sinh(kh)} \right.\right. \\ \left. - \left(kh\sigma - \frac{1}{2}\right)\frac{\cosh(k(y + h))}{\sinh(kh)} \right)\frac{\partial^2 A}{\partial x_1^2} e^{i\theta} \\ \left. + \frac{3\omega k(\sigma^2 - 1)}{\sigma^3}\frac{\cosh(2k(y + h))}{\sinh(2kh)} A \frac{\partial A}{\partial x_1} e^{2i\theta} + \text{c.c.} \right\}.\end{gathered}$$

Its solution that satisfies the boundary condition

$$\frac{\partial \phi_3}{\partial y} = 0, \quad y = -h$$

consists of four harmonics,

$$\phi_3 = \phi_3^{(0)} + \phi_3^{(1)} + \phi_3^{(2)} + \phi_3^{(3)} + \text{c.c.} \qquad (27)$$

Here,

$$\phi_3^{(0)} = -\frac{1}{2}\frac{\partial^2 \Psi_1}{\partial x_1^2}(y + h)^2 + \Psi_3, \qquad (28)$$

$$\begin{gathered}\phi_3^{(1)} = \frac{1}{\sinh(kh)} \\ \times \left[G\cosh(k(y + h)) + [(y + h)^2 \cosh(k(y + h)) \right. \\ \left. - 2h\sigma(y + h)\sinh(k(y + h))] \frac{i\omega}{4k}\frac{\partial^2 A}{\partial x_1^2} \right] e^{i\theta}.\end{gathered} \qquad (29)$$

As noted above, we determine the amplitude of the homogeneous solution G from the condition that the potential $\phi_3^{(1)}$ be bounded as the depth tends to infinity. This condition gives

$$G = \frac{i\omega}{4k}\frac{\partial^2 A}{\partial x_1^2}h^2(2\sigma^2 - 1).$$

In that case,

$$\lim_{kh \to \infty}\phi_3^{(1)} = \frac{i\omega}{4k}\frac{\partial^2 A}{\partial x_1^2}y^2 e^{ky}e^{i\theta}.$$

We find two of the four harmonics of the surface displacement from the dynamical boundary condition:

$$\eta_3 = \eta_3^{(0)} + (\eta_3^{(1)} + \eta_3^{(2)} + \eta_3^{(3)} + \text{c.c.}), \qquad (30)$$

or, more specifically,

$$\begin{aligned}\eta_3^{(0)} = {} & -\frac{k\sigma}{\omega^2}\left(\frac{\partial \Psi_1}{\partial t_2} + \frac{\partial \Psi_2}{\partial t_1}\right) \\ & - \frac{i}{4\sigma}(\sigma^2 - 1)(\sigma kh - 1)\left(A\frac{\partial \bar{A}}{\partial x_1} - \bar{A}\frac{\partial A}{\partial x_1}\right),\end{aligned} \qquad (31)$$

$$\begin{aligned}\eta_3^{(1)} = {} & \left[\frac{i}{2\omega}\frac{\partial A}{\partial t_2} - \frac{k}{2\omega}\left(\frac{k\sigma^2}{\omega}\frac{\partial \Psi_1}{\partial t_1} + \frac{\partial \Psi_1}{\partial x_1}\right)A \right. \\ & \left. + \frac{k^2}{16\sigma^4}(2\sigma^6 - 11\sigma^4 + 9\sigma^2 - 3)A^2\bar{A}\right]e^{i\theta}.\end{aligned} \qquad (32)$$

Let us now take into account the kinematic boundary condition. The terms that constitute the equation for the zeroth harmonic do not contain any indefinite coefficients. Therefore, this equation is the condition that the derivatives of Ψ_1 and A must satisfy:

$$h\frac{\partial^2 \Psi_1}{\partial x_1^2} - \frac{\sigma k}{\omega^2}\frac{\partial^2 \Psi_1}{\partial t_1^2} + \omega\frac{\mu}{8\sigma^2}\frac{\partial(A\bar{A})}{\partial x_1} = 0. \qquad (33)$$

Here,

$$\mu \equiv \frac{2\sigma}{\omega}(2\omega - k(\sigma^2 - 1)V_g) = (\sigma^2 - 1)^2 kh - \sigma(\sigma^2 - 5).$$

Since the product $A\bar{A}$ satisfies (24), we will use the ansatz for which the amplitude of the zeroth harmonic Ψ_1 moves with the group velocity:

$$\frac{\partial \Psi_1}{\partial t_1} = -V_0\frac{\partial \Psi_1}{\partial x_1}, \quad V_0 = V_g. \qquad (34)$$

After the substitution of this expression into (33) and integration, we obtain

$$\frac{\partial \Psi_1}{\partial x_1} = \frac{\omega k\mu}{2\sigma\nu}A\bar{A}, \qquad (35)$$

where we denoted

$$\nu \equiv \frac{4k\sigma}{g}(V_0^2 - gh), \qquad (36)$$

with

$$\nu = ((\sigma + 1)^2 kh - \sigma)((\sigma - 1)^2 kh - \sigma)$$

for $V_0 = V_g$. We see that for deep water,

$$\lim_{kh \to \infty}\frac{\partial \Psi_1}{\partial x_1} = 0.$$

We find the amplitude of the homogeneous solution Ψ_3 from the condition that $\phi_3^{(0)}$ be bounded. Taking into account the asymptotic behavior of the derivative $\partial\Psi_1/\partial x_1$ from (35) and the potential $\phi_3^{(0)}$ from (28) for $kh \longrightarrow \infty$, we set

$$\Psi_3 = -\frac{1}{4}\omega\frac{\partial(A\bar{A})}{\partial x_1}h$$

for the boundedness of $\phi_3^{(0)}$.

It then follows from (28) that for deep water,

$$\lim_{kh \to \infty}\phi_3^{(0)} = \frac{1}{2}\omega\frac{\partial(A\bar{A})}{\partial x_1}\left(y + \frac{1}{8k}\right). \qquad (37)$$

We see from this relation and the expressions $\phi_1^{(0)} = \Psi_1$ and $\phi_2^{(0)} = \Psi_2$ that the following boundary condition for the zeroth harmonic, which is used in [4], is satisfied when $kh \longrightarrow \infty$:

$$\frac{\partial(\phi_1^{(0)}\varepsilon + \phi_2^{(0)}\varepsilon^2 + \phi_3^{(0)}\varepsilon^3)}{\partial y} = \varepsilon^3\frac{\omega}{2}\frac{\partial(A\bar{A})}{\partial x_1}, \quad y = 0$$

(for the case where A is the amplitude of the surface profile, the corresponding expression is given in the last paper from [4]).

The terms of the kinematic condition proportional to $e^{i\theta}$ constitute another evolution equation for the amplitudes A and Ψ_1:

$$\begin{aligned} & i\frac{\partial A}{\partial t_2} + \frac{1}{2}\frac{\partial^2 \omega}{\partial k^2}\frac{\partial^2 A}{\partial x_1^2} \\ & + \frac{1}{16\sigma^4}(2\sigma^6 - 13\sigma^4 + 12\sigma^2 - 9)\omega k^2 A^2\bar{A} \\ & - \left(\frac{k^2}{2\omega}(\sigma^2 - 1)\frac{\partial \Psi_1}{\partial t_1} + k\frac{\partial \Psi_1}{\partial x_1}\right)A = 0,\end{aligned} \qquad (38)$$

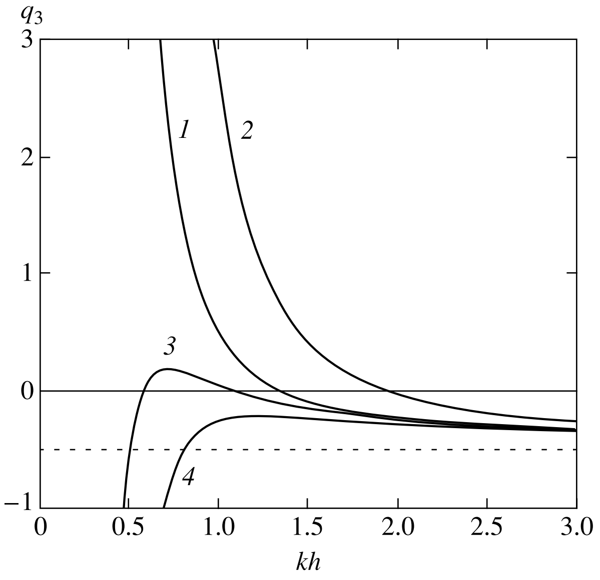


Fig. 1. The coefficient q_3. The curves correspond to different values of V_0 in Eq. (34): *1*—$V_0 = V_g$, *2*—$V_0 = \omega/k$, *3*—$V_0 = -V_g$, and *4*—$V_0 = -V_g + \omega/2k$.

where

$$\frac{\partial^2\omega}{\partial k^2} = \frac{\omega}{4k^2\sigma^2} \times((\sigma^2-1)(3\sigma^2+1)k^2h^2 - 2\sigma(\sigma^2-1)kh - \sigma^2). \tag{39}$$

Taking into account (34), we obtain

$$i\frac{\partial A}{\partial t_2} + \frac{1}{2}\frac{\partial^2\omega}{\partial k^2}\frac{\partial^2 A}{\partial x_1^2} + \frac{1}{16\sigma^4}(2\sigma^6 - 13\sigma^4 + 12\sigma^2 - 9)\omega k^2 A^2\bar{A} - \mu_0\frac{k}{4\sigma}\frac{\partial\Psi_1}{\partial x_1}A = 0, \tag{40}$$

where

$$\mu_0 \equiv \frac{2\sigma}{\omega}(2\omega - k(\sigma^2-1)V_0). \tag{41}$$

Substituting (35) into (40) yields an evolution equation for the amplitude of the first harmonic with time t_2:

$$i\frac{\partial A}{\partial t_2} + \frac{1}{2}\frac{\partial^2\omega}{\partial k^2}\frac{\partial^2 A}{\partial x_1^2} + q_3\omega k^2 A^2\bar{A} = 0, \tag{42}$$

where (provided that $V_0 = V_g$ in (41))

$$q_3 = -\frac{1}{16\nu\sigma^4}((9\sigma^4 - 10\sigma^2 + 9)(\sigma^2-1)^2k^2h^2 + 2\sigma(3\sigma^6 - 23\sigma^4 + 13\sigma^2 - 9)kh - \sigma^2(7\sigma^4 - 38\sigma^2 - 9)). \tag{43}$$

Given (6), the sum of Eq. (22) and Eq. (42) multiplied by ε leads to the nonlinear Schrödinger equation

$$i\left(\frac{\partial A}{\partial\tau} + V_g\frac{\partial A}{\partial x_1}\right) + \varepsilon\left(\frac{1}{2}\frac{\partial^2\omega}{\partial k^2}\frac{\partial^2 A}{\partial x_1^2} + q_3\omega k^2A^2\bar{A}\right) = 0, \tag{44}$$

which was first derived for gravity waves on the surface of a finite-depth fluid layer by Hasimoto and Ono [20]. The goal of this study is to include the next terms (the $O(\varepsilon^4)$ approximation to the Hasimoto–Ono equation).

In Fig. 1, the coefficient q_3 is plotted against the parameter kh. The sign of this coefficient changes from positive to negative at $kh = 1.363$ with increasing k (curve *1*). The NSE equation is known to have soliton solutions under zero boundary conditions only if the product

$$\frac{\partial^2\omega}{\partial k^2}q_3 > 0.$$

For the dispersion law (13),

$$\frac{\partial^2\omega}{\partial k^2} < 0$$

at all k, for example,

$$\frac{\partial^2\omega}{\partial k^2} = -0.1564\frac{\omega}{k^2}$$

at $kh = 1.363$. Therefore, in the $O(\varepsilon^3)$ approximation, envelope solitons are possible only for $kh > 1.363$, in particular, for an infinite depth when $q_3 = -1/2$ (the horizontal dotted line in Fig. 1).

Since (34) is the ansatz that was also used in most of the previous studies, for example, in [20, 23, 40, 41], but only as an assumption, Fig. 1 also shows the plots of q_3 against kh for different velocities V_0 of the amplitude of the zeroth harmonic Ψ_1 in (34). Curve *1* corresponds to the standard condition according to which this velocity is equal to V_g (the group velocity of the first harmonic A). It is in this case that $q_3 = 0$ at $kh = 1.3628$. Curves *3*, *2*, and *4* were constructed, respectively, for $V_0 = -V_g$, $V_0 = V_f$ ($V_f = \omega/k$ is the phase velocity), and $V_0 = -V_g + \omega/2k$ (for this choice, $\eta_2^{(0)}$ in (19) is equal to zero).

The plots show the extent to which the coefficient q_3 and the singular point kh at which q_3 changes its sign are sensitive to the choice of an ansatz for the velocity of the zeroth harmonic. As regards the experiments, most of them (see [30] and references therein, as well as [31, 32]), starting from the classical studies [16], refer to the case of deep water, $kh \gg 1$. In the only experiment for an intermediate depth carried out to date [42], Shemer *et al.* measured the surface profile only at three points, $kh = 0.847$, 1.53, and 4.93. They concluded that the patterns of long-term evolution are qualitatively different for $kh = 0.847$ and 4.93, while the evolution at $kh = 1.53$ is similar to the linear evolution. Of course, it does not unequivocally follow from this result that the pattern qualitatively changes precisely at $kh = 1.36$. The criterion $V_g < V_f$ for which the substitution $\partial/\partial t \longrightarrow V_g\partial/\partial x$ can be made is satisfied for gravity waves with dispersion law (13) for all kh. However, it is only a necessary but not sufficient condition [43].

Below, in this and the next orders, we will use the standard assumption that the velocity of the zeroth harmonic is equal to the velocity of the first harmonic. Note that the next approximation was calculated by the WKB method in [8]. However, since the authors did not consider it possible to use ansatz (34), the evolution equations at each step in ε were left in the form of a system of two coupled equations for the amplitudes A and Ψ_1.

The evolution equations (33) and (42) were derived above in such a way that the solutions of the Laplace equation for the zeroth and first harmonics reconciled with the dynamical boundary condition also satisfied the kinematic boundary condition. The same approach for the second and third harmonics does not lead to new evolution equations, but it allows the as yet undetermined coefficients of these harmonics to be calculated:

$$\begin{aligned}\phi_3^{(2)} = {} & \frac{3\omega}{16k\sigma^4}[4\sigma(\sigma^2-1)k(y+h)\sinh(2k(y+h)) \\ & -(\sigma^2-1)((3\sigma^2+5)hk-3\sigma)\cosh(2k(y+h))] \\ & \times A\frac{\partial A}{\partial x_1}\frac{1}{\sinh(2kh)}e^{2i\theta},\end{aligned} \tag{45}$$

$$\eta_3^{(2)} = i\frac{3(1-\sigma^2)kh+\sigma(\sigma^2-3)}{4\sigma^4}A\frac{\partial A}{\partial x_1}e^{2i\theta}, \tag{46}$$

$$\begin{aligned}\phi_3^{(3)} = {} & -\frac{i\omega k}{128\sigma^6}(\sigma^2-1)(13\sigma^2-9)(3+\sigma^2) \\ & \times A^3\frac{\cosh(3k(y+h))}{\sinh(3kh)}e^{3i\theta},\end{aligned} \tag{47}$$

$$\eta_3^{(3)} = \frac{3}{128\sigma^6}(3-\sigma^2)(\sigma^4+3)k^2A^3e^{3i\theta}. \tag{48}$$

Given ansatz (34) and the evolution equation (40), expression (32) takes the form

$$\begin{aligned}\eta_3^{(1)} = {} & \left[-\frac{k(\sigma^2+1)(kh(\sigma^2-1)-\sigma)}{8\omega\sigma}A\frac{\partial\Psi_1}{\partial x_1}\right. \\ & \left.-\frac{1}{4\omega}\frac{\partial^2\omega}{\partial k^2}\frac{\partial^2 A}{\partial x_1^2}+\frac{k^2}{32\sigma^4}(2\sigma^6-9\sigma^4+6\sigma^2+3)A^2\bar{A}\right]e^{i\theta}.\end{aligned}$$

If we also take into account (35), then

$$\begin{aligned}\eta_3^{(1)} = {} & -\frac{1}{4}\left\{\frac{1}{\omega}\frac{\partial^2\omega}{\partial k^2}\frac{\partial^2 A}{\partial x_1^2}\right. \\ & +\frac{1}{8\sigma^4\nu}[(9\sigma^4-8\sigma^2-3)(\sigma^2-1)^2k^2h^2 \\ & -2\sigma(\sigma^2+1)(\sigma^4-3)kh \\ & \left.+\sigma^2(\sigma^4-16\sigma^2-3)]k^2A^2\bar{A}\right\}e^{i\theta}.\end{aligned} \tag{49}$$

3.4. The $O(\varepsilon^4)$ Approximation

In the order $O(\varepsilon^4)$, we have

$$\phi_4 = \phi_4^{(0)} + (\phi_4^{(1)} + \phi_4^{(2)} + \phi_4^{(3)} + \phi_4^{(4)} + \text{c.c.}), \tag{50}$$

where

$$\phi_4^{(0)} = -\frac{1}{2}\frac{\partial^2\Psi_2}{\partial x_1^2}(y+h)^2 + \Psi_4, \tag{51}$$

$$\begin{aligned}\phi_4^{(1)} = {} & \frac{1}{\sinh(kh)}\left\{F\cosh(k(y+h))\right. \\ & +\frac{\omega}{4k}\left[\frac{1}{3}(y+h)^3\sinh(k(y+h))\right. \\ & -h\sigma(y+h)^2\cosh(k(y+h)) \\ & \left.\left.+h^2(2\sigma^2-1)(y+h)\sinh(k(y+h))\right]\frac{\partial^3 A}{\partial x_1^3}\right\}e^{i\theta}.\end{aligned} \tag{52}$$

We determine the amplitude of the homogeneous solution F, as in the previous iterations, from the condition that the potential $\phi_4^{(1)}$ be bounded when the depth tends to infinity. In this case,

$$F = -\frac{\omega}{12k}\frac{\partial^3 A}{\partial x_1^3}h^3(6\sigma^2-5)\sigma,$$

and for deep water,

$$\lim_{kh\to\infty}\phi_4^{(1)} = \frac{\omega}{12k}\frac{\partial^3 A}{\partial x_1^3}y^3 e^{ky}e^{i\theta}.$$

We find the first two of the five harmonics of the surface displacement from the dynamical boundary condition:

$$\eta_4 = \eta_4^{(0)} + (\eta_4^{(1)} + \eta_4^{(2)} + \eta_4^{(3)} + \eta_4^{(4)} + \text{c.c.}) \tag{53}$$

(see the Appendix). Collecting the terms that contain no harmonics, we obtain the following evolution equation for the amplitude Ψ_2 of the zeroth harmonic in the $O(\varepsilon^4)$ approximation from the kinematic boundary condition:

$$\begin{gathered} h\frac{\partial^2\Psi_2}{\partial x_1^2} - \frac{k\sigma}{\omega^2}\frac{\partial^2\Psi_2}{\partial t_1^2} \\ -i\frac{\omega}{k}\tilde{q}_{40}\left(A\frac{\partial^2\bar{A}}{\partial x_1^2} - \bar{A}\frac{\partial^2 A}{\partial x_1^2}\right) = 0, \end{gathered} \tag{54}$$

where

$$\begin{gathered} \tilde{q}_{40} = \frac{1}{32\sigma^3\nu}((\sigma^2-1)^5k^4h^4 \\ -4\sigma(2\sigma^4+9\sigma^2+5)(\sigma^2-1)^2k^3h^3 \\ +2\sigma^2(\sigma^2-1)(9\sigma^4+16\sigma^2-9)k^2h^2 \\ -4\sigma^3(4\sigma^4-9\sigma^2-7)kh+5\sigma^4(\sigma^2-5)). \end{gathered} \tag{55}$$

Assuming that the potential component Ψ_2 of the zeroth harmonic in the $O(\varepsilon^2)$ approximation, like Ψ_1, moves with the group velocity of the first harmonic,

$$\frac{\partial\Psi_2}{\partial t_1} = -V_g\frac{\partial\Psi_2}{\partial x_1}, \tag{56}$$

and integrating, we find the relation between Ψ_2 and A

$$\frac{\partial\Psi_2}{\partial x_1} = -i\frac{4\omega\sigma}{\nu}\tilde{q}_{40}\left(A\frac{\partial\bar{A}}{\partial x_1} - \bar{A}\frac{\partial A}{\partial x_1}\right). \tag{57}$$

Thus, we see that for deep water,

$$\lim_{kh\to\infty}\frac{\partial\Psi_2}{\partial x_1} = 0.$$

We determine the amplitude of the homogeneous solution Ψ_4 from the condition that $\phi_4^{(0)}$ be bounded. Taking into account the asymptotic behavior of the derivative $\partial\Psi_2/\partial x_1$ from (57) and the potential $\phi_4^{(0)}$ from (51) when $kh \longrightarrow \infty$, for the boundedness of $\phi_4^{(0)}$, we set

$$\Psi_4 = -i\frac{3\omega h}{16k}\frac{\partial}{\partial x_1}\left(A\frac{\partial\bar{A}}{\partial x_1} - \bar{A}\frac{\partial A}{\partial x_1}\right).$$

It the follows from (51) that for deep water,

$$\lim_{kh\to\infty}\phi_4^{(0)} = \frac{i\omega}{8k}\left(3y+\frac{1}{8k}\right)\frac{\partial}{\partial x_1}\left(A\frac{\partial\bar{A}}{\partial x_1} - \bar{A}\frac{\partial A}{\partial x_1}\right). \tag{58}$$

Equating the coefficients of $e^{i\theta}$, we obtain the following evolution equation for the amplitude of the first harmonic in the $O(\varepsilon^4)$ approximation from the kinematic boundary condition:

$$\begin{gathered} i\frac{\partial A}{\partial t_3} - \frac{i}{6}\frac{\partial^3\omega}{\partial k^3}\frac{\partial^3 A}{\partial x_1^3} + i\omega k\tilde{q}_{41}A\bar{A}\frac{\partial A}{\partial x_1} + i\omega k\tilde{q}_{42}A^2\frac{\partial\bar{A}}{\partial x_1} \\ -A\left(\frac{k^2}{2\omega}(\sigma^2-1)\frac{\partial\Psi_2}{\partial t_1} + k\frac{\partial\Psi_2}{\partial x_1}\right) \\ -\frac{k^2}{2\omega}(\sigma^2-1)A\frac{\partial\Psi_1}{\partial t_2} = 0, \end{gathered} \tag{59}$$

where

$$\frac{\partial^3\omega}{\partial k^3} = -\frac{\omega}{8k^3\sigma^3}[(\sigma^2-1)(15\sigma^4-2\sigma^2+3)k^3h^3 \tag{60}$$

$$-3\sigma(\sigma^2-1)(3\sigma^2+1)k^2h^2 - 3\sigma^2(\sigma^2-1)kh - 3\sigma^3],$$

$$\begin{gathered} \tilde{q}_{41} = \frac{1}{16\sigma^5\nu}[(2\sigma^6-11\sigma^4-10\sigma^2+27)(\sigma^2-1)^3k^3h^3 \\ -\sigma(\sigma^2-1)(6\sigma^8-21\sigma^6+9\sigma^4-43\sigma^2+81)k^2h^2 \\ +\sigma^2(6\sigma^8-15\sigma^6-77\sigma^4+71\sigma^2-81)kh \\ -\sigma^3(\sigma^2+1)(2\sigma^4-7\sigma^2-27)], \end{gathered}$$

$$\begin{gathered} \tilde{q}_{42} = -\frac{1}{32\sigma^5\nu}[(4\sigma^6-13\sigma^4+10\sigma^2-9)(\sigma^2-1)^3k^3h^3 \\ -\sigma(\sigma^2-1)(12\sigma^8-51\sigma^6+17\sigma^4-\sigma^2-9)k^2h^2 \\ +\sigma^2(12\sigma^8-67\sigma^6+33\sigma^4-\sigma^2-9)kh \\ -\sigma^3(4\sigma^6-29\sigma^4+42\sigma^2-9)]. \end{gathered}$$

After taking into account ansatz (56) in (59), we obtain

$$\begin{gathered} i\frac{\partial A}{\partial t_3} - \frac{i}{6}\frac{\partial^3\omega}{\partial k^3}\frac{\partial^3 A}{\partial x_1^3} + i\omega k\tilde{q}_{41}A\bar{A}\frac{\partial A}{\partial x_1} + i\omega k\tilde{q}_{42}A^2\frac{\partial\bar{A}}{\partial x_1} \\ -\frac{\mu}{4\sigma}kA\frac{\partial\Psi_2}{\partial x_1} - (\sigma^2-1)\frac{k^2}{2\omega}A\frac{\partial\Psi_1}{\partial t_2} = 0. \end{gathered} \tag{61}$$

If we add Eq. (38) multiplied by ε and Eq. (61) multiplied by ε^2 to Eq. (22), take into account (35), and substitute, according to (6), the sum of the partial derivatives for the functions A and Ψ_1 with the total derivatives

$$\frac{\partial A}{\partial t_1}+\varepsilon\frac{\partial A}{\partial t_2}+\varepsilon^2\frac{\partial A}{\partial t_3} = \frac{\partial A}{\partial \tau}$$

and

$$\frac{\partial \Psi_1}{\partial t_1}+\varepsilon\frac{\partial \Psi_1}{\partial t_2} = \frac{\partial \Psi_1}{\partial \tau},$$

then we obtain the fourth-order (in ε) evolution equation that includes all of the previous iterations

$$\begin{aligned}&i\left(\frac{\partial A}{\partial \tau}+V_g\frac{\partial A}{\partial x_1}\right)+\varepsilon\left(\frac{1}{2}\frac{\partial^2\omega}{\partial k^2}\frac{\partial^2 A}{\partial x_1^2}+q_3A^2\bar{A}\right)\\ &+\varepsilon^2\left(-\frac{i}{6}\frac{\partial^3\omega}{\partial k^3}\frac{\partial^3 A}{\partial x_1^3}+i\omega k\tilde{q}_{41}A\bar{A}\frac{\partial A}{\partial x_1}\right.\\ &\left.+i\omega k\tilde{q}_{42}A^2\frac{\partial \bar{A}}{\partial x_1}-\frac{\mu}{4\sigma}kA\frac{\partial \Psi_2}{\partial x_1}\right) = 0.\end{aligned} \tag{62}$$

The interaction of the amplitude of the first harmonic A with the potential component Ψ_1 of the main flow with the order $O(\varepsilon)$ has already been taken into account in q_3, while its interaction with the component Ψ_2 with the order $O(\varepsilon^2)$ is described by the last term of (62) with (57).

4. AN EQUATION FOR THE TOTAL AMPLITUDE OF THE FIRST HARMONIC OF THE SURFACE PROFILE

Equation (62) was set up for the first order of the amplitude A of the first harmonic. We obtain an equation for the total amplitude A_1 by taking into account the fact that, according to (9), (12), (20), and (49),

$$A_1 = A - i\varepsilon\frac{V_g}{\omega}\frac{\partial A}{\partial x_1}+O(\varepsilon^2). \tag{63}$$

If we determine A from this relation and substitute it into (62), then we derive the following equation for the total amplitude of the first harmonic:

$$\begin{aligned}&i\left(\frac{\partial A_1}{\partial \tau}+V_g\frac{\partial A_1}{\partial x_1}\right)+\varepsilon\left(\frac{1}{2}\frac{\partial^2\omega}{\partial k^2}\frac{\partial^2 A_1}{\partial x_1^2}+q_3A_1^2\bar{A}_1\right)\\ &+\varepsilon^2\left(-\frac{i}{6}\frac{\partial^3\omega}{\partial k^3}\frac{\partial^3 A_1}{\partial x_1^3}+i\omega k\tilde{Q}_{41}A_1\bar{A}_1\frac{\partial A_1}{\partial x_1}\right.\\ &\left.+i\omega k\tilde{Q}_{42}A_1^2\frac{\partial \bar{A}_1}{\partial x_1}-\frac{\mu}{4\sigma}k\frac{\partial \Psi_2}{\partial x_1}A_1\right) = 0,\end{aligned} \tag{64}$$

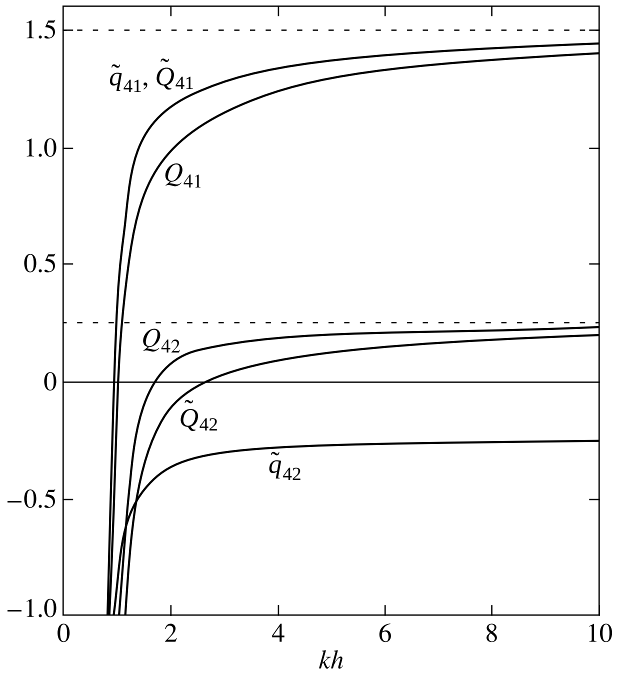


Fig. 2. The nonlinear dispersion coefficients $\tilde{q}_{41}$, $\tilde{q}_{42}$, $\tilde{Q}_{41}$, $\tilde{Q}_{42}$, Q_{41}, and Q_{42}. The horizontal asymptotes correspond to an infinite depth [4].

where

$$\tilde{Q}_{41} = \tilde{q}_{41}, \quad \tilde{Q}_{42} = \tilde{q}_{42}-2\frac{V_g}{\omega}q_3.$$

Using (43) and (23), we obtain

$$\begin{aligned}\tilde{Q}_{42} = \frac{1}{32\nu\sigma^5}&(-(4\sigma^6+5\sigma^4-10\sigma^2+9)(\sigma^2-1)^3k^3h^3\\ &+\sigma(\sigma^2-1)(12\sigma^8-45\sigma^6+71\sigma^4-15\sigma^2+9)k^2h^2\\ &-\sigma^2(12\sigma^8-93\sigma^6+215\sigma^4-111\sigma^2+9)kh\\ &+\sigma^3(4\sigma^6-43\sigma^4+118\sigma^2+9)).\end{aligned}$$

In the limiting case of an infinite depth, an equation of this type was first derived by Dysthe [4]. We see from Fig. 2 that the coefficients $\tilde{Q}_{41}$ and $\tilde{Q}_{42}$ as functions of kh for $kh \longrightarrow \infty$ tend to $\tilde{Q}_{41} = 3/2$ and $\tilde{Q}_{42} = 1/4$, respectively. These are the values of the coefficients of Dysthe's equation if it, like (64), is written for the total amplitude of the first harmonic of the surface displacement (this was done in the last paper from [4] and in [28, 33]) rather than the potential, as in the first paper

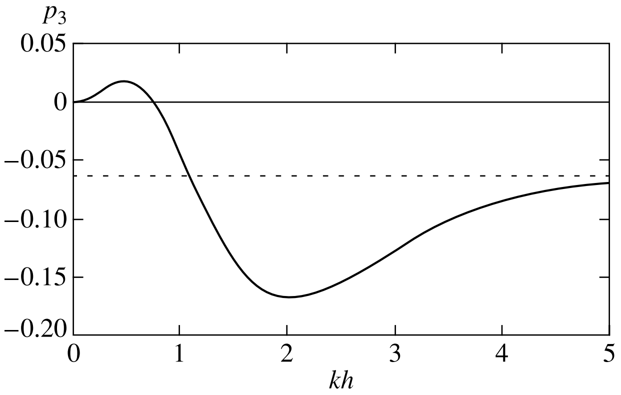


Fig. 3. The cubic dispersion coefficient $p_3 = -(k^3/\omega)(\partial^3\omega/\partial k^3)/6$. The horizontal asymptote corresponds to an infinite depth [4].

from [4]. The remaining coefficients for $kh \longrightarrow \infty$ also tend to their values for an infinite depth [4]:

$$V_g \equiv \frac{\partial\omega}{\partial k} = \frac{1}{2}\frac{\omega}{k}, \quad \frac{1}{2}\frac{\partial^2\omega}{\partial k^2} = -\frac{1}{8}\frac{\omega}{k^2},$$

$$-\frac{1}{6}\frac{\partial^3\omega}{\partial k^3} = -\frac{1}{16}\frac{\omega}{k^3}, \quad q_3 = -\frac{1}{2}, \quad \frac{\mu}{4\sigma} = 1.$$

The following equation (in the notation adopted here) is used for the relationship between the total amplitude A_1 of the first harmonic of the surface profile and the amplitude of the zeroth harmonic $\phi^{(0)}$ of the velocity potential in [4] (precisely for these variables from the last paper in [4]):

$$\frac{\partial\phi^{(0)}}{\partial y} = \frac{\omega}{2}\frac{\partial(A_1\bar{A}_1)}{\partial x_1}, \quad y = 0. \tag{65}$$

We pass to the limit (65) for $kh \longrightarrow \infty$ by taking into account (37) and (58) and the relationship between the total amplitudes and the amplitudes of all approximations for the potential

$$\phi^{(0)} = \phi_1^{(0)}\varepsilon + \phi_2^{(0)}\varepsilon^2 + \phi_3^{(0)}\varepsilon^3 + \phi_4^{(0)}\varepsilon^4$$

and (63) for the surface displacement.

The factor $\partial\Psi_2/\partial x_1$ in (64) can be eliminated if we take into account (57) and (63). Finally, we obtain the fourth-order (in ε) nonlinear Schrödinger equation the last two terms of which include the nonlinearity dispersion:

$$i\left(\frac{\partial A_1}{\partial\tau} + V_g\frac{\partial A_1}{\partial x_1}\right) + \varepsilon\left(\frac{1}{2}\frac{\partial^2\omega}{\partial k^2}\frac{\partial^2 A_1}{\partial x_1^2} + \omega k^2 q_3 A_1^2\bar{A}_1\right)$$
$$+ i\varepsilon^2\left(-\frac{1}{6}\frac{\partial^3\omega}{\partial k^3}\frac{\partial^3 A_1}{\partial x_1^3} + \omega k Q_{41}A_1\bar{A}_1\frac{\partial A_1}{\partial x_1}\right. \tag{66}$$
$$\left. + \omega k Q_{42}A_1^2\frac{\partial\bar{A}_1}{\partial x_1}\right) = 0.$$

Here,

$$Q_{41} = \tilde{Q}_{41} - \frac{\mu}{\nu}\tilde{q}_{40}, \quad Q_{42} = \tilde{Q}_{42} + \frac{\mu}{\nu}\tilde{q}_{40}.$$

Hence, using (55), we obtain

$$Q_{41} = \frac{1}{32\sigma^5\nu^2}$$
$$\times[(3\sigma^6 - 20\sigma^4 - 21\sigma^2 + 54)(\sigma^2 - 1)^5 k^5 h^5$$
$$- \sigma(11\sigma^8 - 99\sigma^6 - 61\sigma^4 + 7\sigma^2 + 270)(\sigma^2 - 1)^3 k^4 h^4$$
$$+ 2\sigma^2(\sigma^2 - 1) \tag{67}$$
$$\times(7\sigma^{10} - 58\sigma^8 + 38\sigma^6 + 52\sigma^4 - 181\sigma^2 + 270)k^3h^3$$
$$- 2\sigma^3(3\sigma^{10} + 18\sigma^8 - 146\sigma^6$$
$$- 172\sigma^4 + 183\sigma^2 - 270)k^2h^2$$
$$- \sigma^4(\sigma^8 - 109\sigma^6 + 517\sigma^4 + 217\sigma^2 + 270)kh$$
$$+ \sigma^5(\sigma^6 - 40\sigma^4 + 193\sigma^2 + 54)]$$

and

$$Q_{42} = \frac{1}{32\sigma^5\nu^2}[(3\sigma^6 + 7\sigma^4 - 11\sigma^2 + 9)(1 - \sigma^2)^5 k^5 h^5$$
$$+ \sigma(11\sigma^8 - 48\sigma^6 + 66\sigma^4 + 8\sigma^2 + 27)(\sigma^2 - 1)^3 k^4 h^4$$
$$- 2\sigma^2(\sigma^2 - 1)$$
$$\times(7\sigma^{10} - 79\sigma^8 + 282\sigma^6 - 154\sigma^4 - \sigma^2 + 9)k^3h^3 \tag{68}$$
$$+ 2\sigma^3(3\sigma^{10} - 63\sigma^8 + 314\sigma^6 - 218\sigma^4 + 19\sigma^2 + 9)k^2h^2$$
$$+ \sigma^4(\sigma^8 + 20\sigma^6 - 158\sigma^4 - 28\sigma^2 - 27)kh$$
$$- \sigma^5(\sigma^6 - 7\sigma^4 + 7\sigma^2 - 9)].$$

Equation (66) and formulas (67) and (68) for Q_{41} and Q_{42} are the main result of our study. Q_{41} and Q_{42} are plotted against kh in Fig. 2. The coefficient $\partial^3\omega/\partial k^3$ is alternating, depending on the fluid depth (Fig. 3). This is important in analyzing the possibility of applying the known cases of the integration of (1) [13, 14] to Eq. (66).

The expressions for the velocity potentials $\phi_4^{(2)}$, $\phi_4^{(3)}$, and $\phi_4^{(4)}$ are given in the Appendix. The expressions for the surface profile $\eta_4^{(2)}$, and $\eta_4^{(3)}$, which are only the concomitant result of our study, are not given here to save space.

5. CONCLUSIONS

A peculiarity of the derivation of the NSEIV for Stokes waves in a fluid layer compared to the same problem in optical fibers is the necessity of taking into account the effect of the low-frequency zeroth harmonic on the motion of the high-frequency first harmonic. This interaction is described by system of equations (33) and (40) in the $O(\varepsilon^3)$ approximation and by system (54) and (59) in the $O(\varepsilon^4)$ approximation. Assuming the motion of the zeroth harmonic with the group velocity of the first harmonic in both approximations, these systems reduce to an equation for one unknown function, the amplitude of the first harmonic A_1 of the free-surface profile. In the $O(\varepsilon^3)$ and $O(\varepsilon^4)$ approximations, this is NSE (44) that was first derived in [20] and NSEIV (66), respectively. The amplitudes Ψ_1 and Ψ_2 of the second unknown function, the velocity potential (to be more precise, their derivatives, which have the physical meaning of the velocity in the first place), are expressed in terms of the total amplitude of the first harmonic A_1 using formulas (35) and (57), in which A_1 may be substituted for A in the $O(\varepsilon^3)$ and $O(\varepsilon^4)$ approximations, respectively. Indeed, we see from relation (63) that Eqs. (44) and (35) required for this in the $O(\varepsilon^3)$ approximation do not change their forms in passing from the first order of the amplitude of the first harmonic A to its total amplitude A_1, as does Eq. (57) in the $O(\varepsilon^4)$ approximation.

After the renormalization of the independent variables and the unknown function, Eq. (66) transforms into Eq. (1). The ± signs are determined by the sign opposite to q_3 (Fig. 1). Thus, we choose – in Eq. (1) if $kh < 1.3628$ and + if $kh > 1.3628$. In the former and latter cases, the properties of the soliton and quasi-soliton solutions to Eq. (1) were described in [14] and [13], respectively. They are determined by the coefficients V_g, $\partial^2\omega/\partial k^2$, $\partial^3\omega/\partial k^3$, q_3, Q_{41}, and Q_{42}, as defined by formulas (23), (39), (60), (43), (67), and (68).

ACKNOWLEDGMENTS

The author is grateful to V.P. Lukomsky and colleagues from the Theoretical Physics Department (Institute of Physics) for fruitful and stimulating discussions.

This study was supported by INTAS (grant no. 99-1637).

APPENDIX

Below, we give expressions for the Stokes expansion components of the potential and the surface displacement in the $O(\varepsilon^4)$ approximation (we denote $\sigma = \tanh(kh)$ and $\nu = ((1+\sigma)^2kh-\sigma)((\sigma-1)^2kh-\sigma)$):

$$\begin{aligned}
\phi_4^{(2)} = i\omega\Bigg\{&\frac{3(1-\sigma^2)}{8\sigma^3}\left(\left(\frac{\partial A}{\partial x_1}\right)^2 + A\frac{\partial^2 A}{\partial x_1^2}\right) \\
&\times\frac{\cosh(2k(y+h))}{\sinh(2kh)}(y+h)^2 - \frac{1}{64\sigma^7}\Bigg(\frac{\sigma^2-1}{k^2} \\
&\times((34\sigma^6+23\sigma^4+60\sigma^2+3)k^2h^2 \\
&-2\sigma(16\sigma^4+23\sigma^2+3)kh-\sigma^2(2\sigma^2-3))A\frac{\partial^2 A}{\partial x_1^2} \\
&+\frac{\sigma^2-1}{k^2}((11\sigma^6+19\sigma^4+45\sigma^2-3)k^2h^2 \\
&-2\sigma(11\sigma^4+22\sigma^2-3)kh+\sigma^2(11\sigma^2-3))\left(\frac{\partial A}{\partial x_1}\right)^2 \\
&+\frac{k^2}{12\nu\sigma^2}((145\sigma^8-682\sigma^6+864\sigma^4-54\sigma^2-81) \\
&\times(\sigma^2-1)^2k^2h^2-2\sigma(\sigma^{10}+471\sigma^8 \\
&-1402\sigma^6+1530\sigma^4-135\sigma^2-81)kh \\
&+\sigma^2(\sigma^8+182\sigma^6+144\sigma^4-54\sigma^2-81))A^3\bar{A}) \\
&\times\frac{\cosh(2k(y+h))}{\sinh(2kh)}+\frac{3(\sigma^2-1)}{16k\sigma^4}(kh(3\sigma^2+5)-3\sigma) \\
&\times\left(A\frac{\partial^2 A}{\partial x_1^2}+\left(\frac{\partial A}{\partial x_1}\right)^2\right)(y+h)\frac{\sinh(2k(y+h))}{\sinh(2kh)}\Bigg\}e^{2i\theta},
\end{aligned}$$

$$\begin{aligned}
\phi_4^{(3)} = {}&\frac{\omega}{128\sigma^7}(\sigma^2-1)(\sigma^2+3) \\
&\times\Bigg(((26\sigma^4+39\sigma^2-45)kh-3\sigma(13\sigma^2-9)) \\
&\times\frac{\cosh(3k(y+h))}{\sinh(3kh)}-3\sigma(13\sigma^2-9)k(y+h) \\
&\times\frac{\sinh(3k(y+h))}{\sinh(3kh)}\Bigg)A^2\frac{\partial A}{\partial x_1}e^{3i\theta},
\end{aligned}$$

$$\phi_4^{(4)} = i\omega k^2$$

$$\times\frac{(\sigma^4-1)(197\sigma^6+747\sigma^4-1269\sigma^2+405)}{768\sigma^9(\sigma^2+5)}$$

$$\times\frac{\cosh(4k(y+h))}{\sinh(4kh)}A^4e^{4i\theta},$$

$$\eta_4^{(4)} = k^3$$

$$\times\frac{21\sigma^{10}+\sigma^8-262\sigma^6+522\sigma^4+81\sigma^2+405}{768\sigma^9(\sigma^2+5)}A^4e^{4i\theta},$$

$$\eta_4^{(0)} = -\frac{k^3}{64\sigma^7\nu^2}((9(\sigma^2-1)^8k^4h^4$$

$$-36\sigma(\sigma^2+1)(\sigma^2-1)^6k^3h^3$$

$$+2\sigma^2(31\sigma^4+18\sigma^2+27)(\sigma^2-1)^4k^2h^2$$

$$-4\sigma^3(13\sigma^6-29\sigma^4-9\sigma^2+9)(\sigma^2-1)^2kh$$

$$+\sigma^4(17\sigma^8-116\sigma^6+254\sigma^4-36\sigma^2+9))A^2\bar{A}^2$$

$$-\frac{1}{16\sigma k\nu}((3+5\sigma)(\sigma-1)(\sigma^2-1)^3k^4h^4$$

$$-2\sigma(\sigma^2-1)(8\sigma^4-3\sigma^3-4\sigma^2-\sigma+8)k^3h^3$$

$$+2\sigma^3(\sigma-1)(3\sigma+1)^2k^2h^2-2\sigma^3(4\sigma^2-\sigma-2)+\sigma^4)$$

$$\times\frac{\partial^2A\bar{A}}{\partial x_1^2}+\frac{1}{4\sigma k}(\sigma^2-1)((3\sigma^2-1)k^2h^2-4\sigma kh+1)$$

$$\times\frac{\partial\bar{A}}{\partial x_1}\frac{\partial A}{\partial x_1}-\frac{\sigma k}{\omega^2}\left(\frac{\partial\Psi_1}{\partial t_3}+\frac{\partial\Psi_2}{\partial t_2}\right),$$

$$\eta_4^{(1)} = -\frac{k^2(\sigma^2+1)}{4\omega^2}A\frac{\partial\Psi_1}{\partial t_2}-\frac{i}{12\omega}\frac{\partial^3\omega}{\partial k^3}\frac{\partial^3A}{\partial x_1^3}$$

$$+\frac{ik}{64\sigma^5\nu^2}((3\sigma^6-16\sigma^4-15\sigma^2-24)(\sigma^2-1)^5k^5h^5$$

$$-\sigma(11\sigma^8-87\sigma^6-183\sigma^4-165\sigma^2-120)$$

$$\times(\sigma^2-1)^3k^4h^4+2\sigma^2(\sigma^2-1)$$

$$\times(7\sigma^{10}-78\sigma^8-92\sigma^6-2\sigma^4-131\sigma^2-120)k^3h^3$$

$$-2\sigma^3(3\sigma^{10}-58\sigma^8-32\sigma^6-102\sigma^4$$

$$+165\sigma^2+120)k^2h^2$$

$$-\sigma^4(\sigma^8+47\sigma^6-101\sigma^4-339\sigma^2-120)kh$$

$$+\sigma^5(\sigma^6+12\sigma^4-121\sigma^2-24))A\bar{A}\frac{\partial A}{\partial x_1}$$

$$-\frac{ik}{64\sigma^5\nu^2}(3(\sigma^6-\sigma^4+3\sigma^2+3)(\sigma^2-1)^5k^5h^5$$

$$-\sigma(11\sigma^8-66\sigma^6+12\sigma^4+20\sigma^2+39)(\sigma^2-1)^3k^4h^4$$

$$+2\sigma^2(\sigma^2-1)$$

$$\times(7\sigma^{10}-73\sigma^8+90\sigma^6+68\sigma^4+19\sigma^2+33)k^3h^3$$

$$-2\sigma^3(3\sigma^{10}-41\sigma^8+94\sigma^6+32\sigma^4-45\sigma^2-27)k^2h^2$$

$$-\sigma^4(\sigma^8-10\sigma^6-52\sigma^4+56\sigma^2+21)kh$$

$$+\sigma^5(\sigma^6-13\sigma^4-\sigma^2+3))A^2\frac{\partial\bar{A}}{\partial x_1}.$$

REFERENCES

1. V. E. Zakharov and E. A. Kuznetsov, Zh. Éksp. Teor. Fiz. **113**, 1892 (1998) [JETP **86**, 1035 (1998)].
2. A. Hasegawa and Y. Kodama, *Solitons in Optical Communication* (Oxford Univ. Press, London, 1995); Y. Kodama and A. Hasegawa, IEEE J. Quantum Electron. **23**, 510 (1987).
3. E. M. Gromov and V. I. Talanov, Zh. Éksp. Teor. Fiz. **110**, 137 (1996) [JETP **83**, 73 (1996)].
4. K. B. Dysthe, Proc. R. Soc. London, Ser. A **369**, 105 (1979); K. Trulsen and K. B. Dysthe, Wave Motion **24**, 281 (1996); K. Trulsen and K. B. Dysthe, J. Fluid Mech. **352**, 359 (1997); K. Trulsen, I. Kliakhandler, K. B. Dysthe, and M. G. Velarde, Phys. Fluids **12**, 2432 (2000).
5. V. P. Lukomskiĭ, Zh. Éksp. Teor. Fiz. **108**, 567 (1995) [JETP **81**, 306 (1995)].
6. F. Dias and K. Kharif, Annu. Rev. Fluid Mech. **31**, 301 (1999).
7. J. R. Stocker and D. H. Peregrine, J. Fluid Mech. **399**, 335 (1999).
8. U. Brinch-Nielsen and I. G. Jonsson, Wave Motion **8**, 455 (1986).
9. D. Clamond and J. P. Germain, Eur. J. Mech. B (Fluids) **18**, 67 (1999).
10. Yu. G. Rapoport, C. E. Zaspel, J. H. Mantha, and V. V. Grimalsky, Phys. Rev. B **65**, 024423-1 (2002).
11. V. P. Lukoms'kiĭ and Yu. V. Sedlets'kiĭ, Zh. Fiz. Dosl. **5**, 107 (2001).
12. A. G. Litvak and V. I. Talanov, Izv. Vyssh. Uchebn. Zaved., Radiofiz. **10**, 539 (1967).

13. V. I. Karpman, J. J. Rasmussen, and A. G. Shagalov, Phys. Rev. E **64**, 026614-1 (2001).
14. A. Mahalingam and K. Porsezian, Phys. Rev. E **64**, 046608-1 (2001).
15. V. E. Zakharov, Prikl. Mekh. Tekh. Fiz., No. 2, 86 (1968).
16. H. C. Yuen and B. M. Lake, Phys. Fluids **18**, 956 (1975); H. Yuen and B. M. Lake, *Nonlinear Dynamics of Deep-Water Gravity Waves* (Academic, New York, 1982; Mir, Moscow, 1987).
17. D. J. Benney and G. J. Roskes, Stud. Appl. Math. **48**, 377 (1969).
18. V. E. Zakharov and V. G. Kharitonov, Prikl. Mekh. Tekh. Fiz., No. 5, 45 (1970).
19. V. H. Chu and C. C. Mei, J. Fluid. Mech. **41**, 873 (1970).
20. H. Hasimoto and H. Ono, J. Phys. Soc. Jpn. **33**, 805 (1972).
21. M. Stiasnie and L. Shemer, J. Fluid Mech. **143**, 47 (1984).
22. A. Devey and K. Stewartson, Proc. R. Soc. London, Ser. A **338**, 101 (1974).
23. T. Kawahara, J. Phys. Soc. Jpn. **38**, 265 (1975).
24. P. Liu and M. Dingemans, Wave Motion **11**, 41 (1989).
25. P. Christodoulides and F. Dias, Phys. Fluids **7**, 3013 (1995).
26. M. Stiassnie, Wave Motion **6**, 431 (1984).
27. P. A. E. M. Janssen, J. Fluid Mech. **126**, 1 (1983).
28. S. J. Hogan, Proc. R. Soc. London, Ser. A **402**, 359 (1985).
29. E. Lo and C. C. Mei, J. Fluid Mech. **150**, 395 (1985).
30. M. P. Tulin and T. Waseda, J. Fluid Mech. **378**, 197 (1999).
31. L. Shemer, Haiyuing Jiao, E. Kit, and Y. Agnon, J. Fluid Mech. **427**, 107 (2001).
32. M. J. Ablowitz, J. Hammack, D. Henderson, and C. M. Schober, Phys. Rev. Lett. **84**, 887 (2000); Physica D (Amsterdam) **152–153**, 416 (2001).
33. Bhattacharyya Sudebi and K. P. Das, J. Aust. Math. Soc. B, Appl. Math. **39**, 214 (1997).
34. E. Kit and L. Shemer, J. Fluid Mech. **450**, 201 (2002).
35. G. B. Whitham, J. Fluid Mech. **27**, 399 (1967).
36. T. B. Benjamin, Proc. R. Soc. London, Ser. A **299**, 59 (1967).
37. M. Tajiri and Y. Watanabe, Phys. Rev. E **57**, 3510 (1998); M. Tajiri and T. Arai, Phys. Rev. E **60**, 2297 (1999); M. Tajiri, T. Arai, and K. Takenahi, Phys. Rev. E **64**, 056622-1 (2001).
38. R. Grimshaw, D. Pelinovsky, E. Pelinovsky, and T. Talipova, Physica D (Amsterdam) **159**, 35 (2001).
39. R. Kh. Zeĭtunyan, Usp. Fiz. Nauk **165**, 1403 (1995) [Phys. Usp. **38**, 1333 (1995)].
40. R. S. Johnson, Proc. R. Soc. London, Ser. A **357**, 131 (1977).
41. T. Kakutani and K. Michihiro, J. Phys. Soc. Jpn. **52**, 4129 (1983).
42. L. Shemer, E. Kit, Haiyuing Jiao, and O. Eitan, J. Waterway, Port, Coastal, Ocean Eng. **124**, 320 (1998).
43. V. E. Zakharov and E. A. Kuznetsov, Usp. Fiz. Nauk **167**, 1137 (1997) [Phys. Usp. **40**, 1087 (1997)].

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 194–203.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 214–223.
Original Russian Text Copyright © 2003 by Ovchinnikov, Sigal.

Multiparameter Family of Collapsing Solutions to the Critical Nonlinear Schrödinger Equation in Dimension *D* = 2

Yu. N. Ovchinnikov[a,b,*] **and I. M. Sigal**[c]
[a]*Max-Planck Institute for Physics of Complex Systems, Dresden, D-01187 Germany*
[b]*Landau Institute for Theoretical Physics, Russian Academy of Sciences, Chernogolovka, Moscow oblast, 142432 Russia*
[c]*Department of Mathematics, University of Toronto, Ontario, Canada, M5S 3G3*
**e-mail: ovc@itp.ac.ru*

Received February 18, 2003

Abstract—We consider the critical nonlinear Schrödinger equation in dimension $D = 2$ and obtain a system consisting of three equations describing the collapse of solutions. The system admits a five-parameter family of solutions. Almost everywhere, except for an exponentially narrow region near the collapse point, the tunneling processes are negligible. The relation between initial data and the condition of occurrence of the collapse is investigated. The separatrix, which divides the collapse domain and expansion regions having no singularities in a finite time interval, is found. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The nonlinear Schrödinger equation with the interaction corresponding to the attraction

$$i\frac{\partial\psi}{\partial t} + \Delta\psi + |\psi|^{2\sigma}\psi = 0 \tag{1}$$

has been studied in many papers. The first results on the collapse were obtained in the late 1960s–early 1970s [1–6]. In the middle of the 1980s, it was shown in [7] that the problem of the collapse can be reduced to the problem on the motion of a particle for an inverted parabolic potential with a nonlinear perturbation and the important role of the tunneling process was discovered. Further progress was achieved in papers [8–16]. The numerical solution to Eq. (1) clearly shows in a wide region outside the barrier the presence of a "tail" leaking from under the barrier and decreasing as $1/\xi$, where $\xi = |\mathbf{r}|/\lambda$ and λ is compression parameter [11, 12] (see figure). The dependence of the compression parameter λ on time in the double logarithm form,

$$\lambda \sim \left(\frac{t^* - t}{\ln\left(\ln\left(\frac{1}{t^* - t}\right)\right)}\right)^{1/2}, \tag{2}$$

was independently found in papers [7, 9] (see also [10–15, 17, 18]). During the collapse, particles are injected into the region outside the barrier. Note that there is no particle flux at a finite physical distance and all particles injected to the overbarrier collapse in a finite time. This is connected with the fact that the physical flux through any fixed surface is always directed to the collapse point. Thus, particles are accumulated in the collapsing region inside the barrier. It is essential for the collapse process that the energy of the particles inside the barrier become infinite as t tends to the collapse point t^*.

Suppose that the initial state is close to the stationary positive solution $\tilde{\psi}$ to Eq. (1), i.e.,

$$\Delta\tilde{\psi} + |\tilde{\psi}|^2\tilde{\psi} = |E_0|\tilde{\psi}. \tag{3}$$

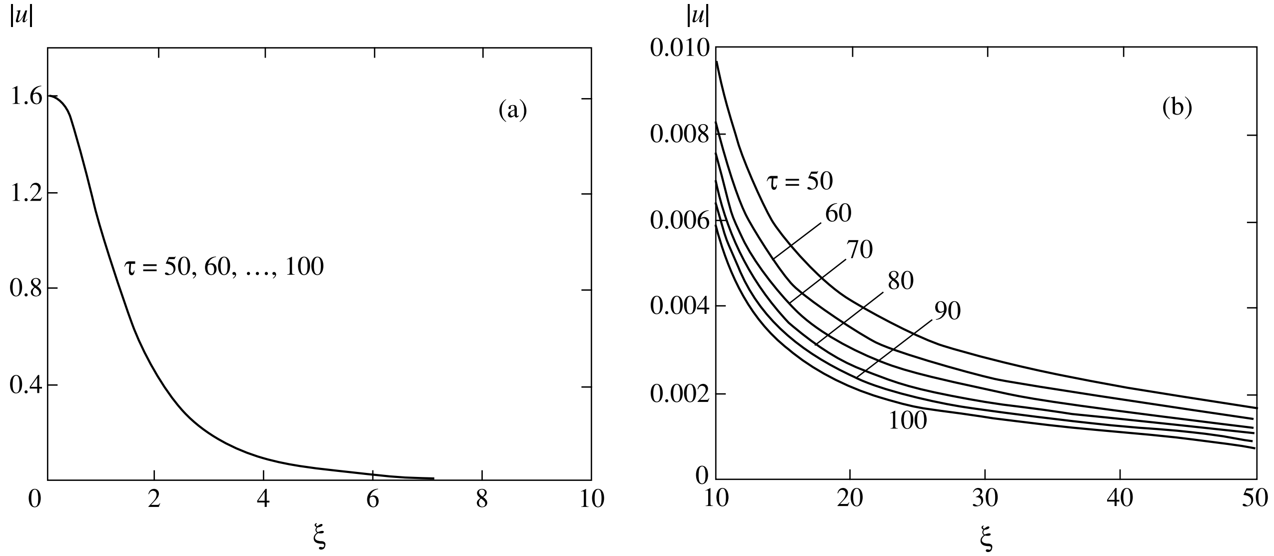
In this case, a small parameter

$$|\lambda\dot{\lambda}| \ll 1 \tag{4}$$

appears, where $\dot{\lambda} = \partial\lambda/\partial t$.

The presence of the small parameter allows one to find wave function ψ in the entire space including the outside barrier region after the right turning point.

Equation (1) admits three exact conservation laws, namely, the particle conservation law, the energy conservation law, and the quasiconformal conservation law [19]. These laws are used below in order to obtain a closed system of equations for quantities $\{\lambda, \mu\}$, where $\lambda(t)$ indicates the degree of compression and the parameter $\mu \equiv \mu(t)$ is a free phase parameter. The system of equations obtained admits a five-parameter family of solutions. This family of solutions contains a separatrix which separates the collapse region from the expansion region.



The evolution of the renormalized profile in critical dimension $D = 2$, $\sigma = 1$, that demonstrate a slow time dependence at large distances [12].

In the problem under consideration, there is an infinitely large number of hidden parameters. This is connected with the fact that not all particles participate in the collapse. A way of decomposing the initial value of function ψ into the collapsing part ψ_{col} and its orthogonal complement $\psi_\perp$ is presented below. Note that the evolution of the compression parameter of form (2) can be realized only in an exponentially narrow region in a neighborhood of the collapse point. The evolution law in the remaining region is different.

2. THE COLLAPSING PART OF THE WAVE FUNCTION

Suppose that wave function ψ at the initial instant of time can be represented in the form given in the Introduction, i.e.,

$$\psi = \psi_{\text{col}} + \psi_\perp, \qquad (5)$$

where the function $\psi_\perp$ is small and decreases quite rapidly as $|\mathbf{r}| \longrightarrow 0$. In this case, function $\psi_\perp$ satisfies the linear Schrödinger equation and has no influence on the dynamics of function ψ_{col}. Below we will neglect the function $\psi_\perp$. Function ψ_{col} determines two quantities conserved in the collapse, viz., the number of collapsing particles and the collapse energy, $\{N_{\text{col}}, E_{\text{col}}\}$.

Assuming that inequality (4) holds, we represent wave function ψ_{col} in the form

$$\psi_{\text{col}} = \left(\frac{1}{\lambda}\tilde{\psi} + \psi_1\right) \times \exp\left[i|E_0|\int\frac{dt}{\lambda^2} + \frac{i\lambda\dot{\lambda}x^2}{4} + i\mu\right], \qquad (6)$$

where $\lambda \equiv \lambda(t)$, $\mu \equiv \mu(t)$, $\dot{\lambda} = \partial\lambda/\partial t$, $x = |\mathbf{r}|/\lambda$, and function $\tilde{\psi}$ is the solution to stationary equation (3), which exponentially decreases at infinity,

$$\frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial}{\partial x}\right)\tilde{\psi} = |E_0|\tilde{\psi} - \tilde{\psi}^3. \qquad (7)$$

In the variables (x, t), Eq. (1) takes the form

$$\begin{aligned}&\frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial}{\partial x}\right)\psi_1 + \tilde{\psi}^2(2\psi_1 + \psi_1^*) - |E_0|\psi_1 + i\lambda^2\frac{\partial\psi_1}{\partial t}\\ &+ i\lambda\dot{\lambda}\psi_1 + \lambda^2\psi_1^2\psi_1^* + \lambda\tilde{\psi}(2\psi_1\psi_1^* + \psi_1^2) - \frac{\lambda^3\ddot{\lambda}}{4}x^2\psi_1 \qquad (8)\\ &- \lambda^2\dot{\mu}\psi_1 - \lambda\dot{\mu}\tilde{\psi} - \frac{\lambda^2\ddot{\lambda}}{4}x^2\tilde{\psi} = 0.\end{aligned}$$

Wave function ψ_{col} has power corrections in the small parameter $\lambda\dot{\lambda}$. These corrections are found below within the framework of perturbation theory with the help of Eq. (8). The "tails" of function ψ_{col}, which are exponentially small in parameter $|\lambda\dot{\lambda}|$, are connected to the tunneling processes in the inverted parabolic potential and will be found separately. The exact conservation laws (the particle conservation law, the energy conservation law, and the quasiconformal conservation law) allow one to obtain a closed system of equations for the quantities $\{N, E, \lambda, \mu\}$, where $\{N, E\}$ are the number of particles and the energy concentrated in the region before the right turning point.

3. PERTURBATION THEORY FOR FUNCTION ψ_1

The presence of the small parameter allows us to employ perturbation theory. The expansion of function ψ_{col} in expression (6) begins with second-order terms, because the first-order terms are separated into the phase factor $\exp(i\lambda\dot{\lambda}x^2/4)$. We represent function ψ_1 in the form

$$\psi_1 = \phi + iZ_3 + \phi_4 + iZ_5 + \dots. \tag{9}$$

From formulas (8) and (9), we find the second-order correction ϕ, namely,

$$\phi = \frac{1}{2|E_0|}\lambda\dot{\mu}\frac{\partial}{\partial x}(x\tilde{\psi}) + \frac{\lambda^2\ddot{\lambda}}{4}\hat{L}^{-1}(x^2\tilde{\psi}), \tag{10}$$

where operator $\hat{L}$ can be written as

$$\hat{L} = \frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial}{\partial x}\right) + 3\tilde{\psi}^2 - |E_0|. \tag{11}$$

The third-order correction Z_3 satisfies the equation

$$\hat{L}_2 Z_3 = -\lambda^2\frac{\partial\phi}{\partial t} - \lambda\dot{\lambda}\phi, \tag{12}$$

where operator $\hat{L}_2$ can be written as

$$\hat{L}_2 = \frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial}{\partial x}\right) + \tilde{\psi}^2 - |E_0|. \tag{13}$$

Since operator $\hat{L}_2$ has the zero mode $\tilde{\psi}$, Eq. (12) has a bounded solution only if the additional condition

$$3\dot{\lambda}\ddot{\lambda} + \lambda\dddot{\lambda} = 0 \tag{14}$$

holds [20]. Equation (14) is not exact. Taking into account the next terms of the expansion, we can obtain a modified equation. In the considered approximation, function Z_3 is equal to

$$Z_3 = \frac{\nu}{\lambda}\tilde{\psi} - \frac{\lambda x^2}{8|E_0|}\tilde{\psi}\frac{\partial}{\partial t}(\lambda^2\dot{\mu}), \tag{15}$$

where function $\nu \equiv \nu(t)$ depends only on time t. The fourth-order correction ϕ_4 can be obtained from Eq. (8) by taking into account formulas (10) and (15), namely,

$$\begin{gathered}\phi_4 = \frac{1}{2|E_0|}\lambda\dot{\nu}\frac{\partial}{\partial x}(x\tilde{\psi}) \\ - \frac{\lambda^2}{8|E_0|}(6\lambda\dot{\lambda}^2\dot{\mu} + 2\lambda^2\ddot{\lambda}\dot{\mu} + 6\lambda^2\dot{\lambda}\ddot{\mu} + \lambda^3\dddot{\mu})\hat{L}^{-1}(x^2\tilde{\psi}) \\ + \frac{\lambda^3\ddot{\lambda}}{4}\hat{L}^{-1}(x^2\phi) + \lambda^2\dot{\mu}\hat{L}^{-1}(\phi) - 3\lambda\hat{L}^{-1}(\tilde{\psi}\phi^2).\end{gathered} \tag{16}$$

4. ENERGY AND NUMBER OF PARTICLES

Recall that the two most important conserved quantities in Eq. (1) are the energy

$$E(\psi) = \int\left(\frac{1}{2}|\nabla\psi|^2 - \frac{1}{2\sigma + 2}|\psi|^{2\sigma + 2}\right)d\mathbf{r} \tag{17}$$

and the number of particles

$$N(\psi) = \int|\psi|^2 d\mathbf{r}. \tag{18}$$

An important role is played by the quantity

$$L = \int d\mathbf{r}|\psi|^2\mathbf{r}^2, \tag{19}$$

which satisfies the quasiconformal conservation law [16, 19]

$$\frac{1}{16}\frac{\partial^2 L}{\partial t^2} = E. \tag{20}$$

Neglecting the correction $\psi_\perp$ in Eq. (5), we obtain two approximate conservation laws

$$E(\psi) \approx E(\psi_{\text{col}}) \equiv E_{\text{col}}$$

and

$$N(\psi) \approx N(\psi_{\text{col}}) \equiv N_{\text{col}}.$$

Passing to the variable $x = |\mathbf{r}|/\lambda$, we split the energy and number of particles into the contributions from the region before and after the largest turning point,

$$x_{\text{t.p.}} = \frac{2\sqrt{|E_0|}}{(-\lambda^3\ddot{\lambda})^{1/2}},$$

$$E_{\text{col}} = E + E_{\text{tail}}$$

and

$$N_{\text{col}} = N + N_{\text{tail}},$$

where E and N are the contributions of the region before the turning point, and E_{tail} and N_{tail} are contributions of the region after the turning point. First we find the number of particles N and energy E concentrated in the collapsing region before the right turning point (we will consider the region after the turning point in the inverted parabolic potential separately). Using Eqs. (6)

and (9) and neglecting exponentially small terms, we obtain

$$N = 2\pi\int_0^\infty dx\,x[\tilde{\psi}^2 + 2\lambda\tilde{\psi}(\phi + \phi_4) + \lambda^2\phi^2], \tag{21}$$

$$\begin{gathered} E = \frac{\pi}{\lambda^2} \\ \times\int_0^\infty dx\,x\left\{-2\lambda|E_0|\tilde{\psi}(\phi + \phi_4) - 2\lambda^2\dot{\lambda}Z_3\frac{\partial}{\partial x}(x\tilde{\psi})\right. \\ \left. + \lambda^2\left(\frac{\partial\phi}{\partial x}\right)^2 + \frac{\lambda^2\dot{\lambda}^2x^2}{4}(\tilde{\psi}^2 + 2\lambda\tilde{\psi}\phi) - 3\lambda^2\tilde{\psi}^2\phi^2\right\}. \end{gathered} \tag{22}$$

Using formulas (A.1), (A.2), and (A.3), we can find two important features, namely, in the approximation considered, parameter ν drops out of expressions (21) and (22) and quantities N and E depend linearly on μ (the second-order terms are also cancelled). As a result, expression (21) for parameters $\{N, E\}$ can be reduced to the form

$$\begin{gathered} N = 2\pi\int_0^\infty dxx\left\{\tilde{\psi}^2 + \frac{\lambda^3}{8|E_0|^2}\right. \\ \times(6\lambda\dot{\lambda}^2\dot{\mu} + 6\lambda^2\ddot{\lambda}\dot{\mu} + 6\lambda^2\dot{\lambda}\ddot{\mu} + \lambda^3\dddot{\mu})(x\tilde{\psi})^2 \\ - \frac{\lambda^3\ddot{\lambda}}{4|E_0|}(x\tilde{\psi})^2 + \left(\frac{\lambda^3\ddot{\lambda}}{4}\right)^3\left[\frac{x^2}{|E_0|}\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)\hat{L}^{-1}(x^2\tilde{\psi})\right. \\ \left.\left. - \frac{3}{|E_0|}\tilde{\psi}\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)(\hat{L}^{-1}(x^2\tilde{\psi}))^2 + (\hat{L}^{-1}(x^2\tilde{\psi}))^2\right]\right\}, \end{gathered} \tag{23}$$

$$\begin{gathered} E = \frac{\partial^2}{\partial t^2}\left(\lambda^2 - \frac{\lambda^4\dot{\mu}}{|E_0|}\right)\frac{\pi}{8}\int_0^\infty dxx(x\tilde{\psi})^2 \\ + \pi\int_0^\infty dxx\left\{\left(\frac{\lambda^2\ddot{\lambda}}{4}\right)^2\left[3\tilde{\psi}\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)(\hat{L}^{-1}(x^2\tilde{\psi}))^2\right.\right. \\ - |E_0|(\hat{L}^{-1}(x^2\tilde{\psi}))^2 - x^2\tilde{\psi}\hat{L}^{-1}(x^2\tilde{\psi}) \\ \left.\left. - x^2\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)\hat{L}^{-1}(x^2\tilde{\psi})\right] + \frac{\lambda^3\dot{\lambda}^2\ddot{\lambda}}{8}x^2\tilde{\psi}\hat{L}^{-1}(x^2\tilde{\psi})\right\}. \end{gathered} \tag{24}$$

Using formulas (A.4) and (A.8), we reduce expressions (23) and (24) to a more compact form:

$$\begin{gathered} N - N_{\text{cr}} = 2\pi\int_0^\infty dx\,x\left\{-\frac{\lambda^3\ddot{\lambda}}{4|E_0|}(x\tilde{\psi})^2\right. \\ - \frac{3}{|E_0|}\left(\frac{\lambda^3\ddot{\lambda}}{4}\right)^2 x^2\tilde{\psi}\hat{L}^{-1}(x^2\tilde{\psi}) \\ \left. + \frac{\lambda^3}{8|E_0|^2}(6\lambda\dot{\lambda}^2 + 6\lambda^2\ddot{\lambda}\dot{\mu} + 6\lambda^2\dot{\lambda}\ddot{\mu} + \lambda^3\dddot{\mu})(x\tilde{\psi})^2\right\}, \end{gathered} \tag{25}$$

$$\begin{gathered} E = \frac{\partial^2}{\partial t^2}\left(\lambda^2 - \frac{\lambda^4\dot{\mu}}{|E_0|}\right)\frac{\pi}{8}\int_0^\infty dxx(x\tilde{\psi})^2 \\ + ((\lambda^2\ddot{\lambda})^2 + \lambda^3\dot{\lambda}^2\ddot{\lambda})\frac{\pi}{8}\int_0^\infty dxx(x^2\tilde{\psi})\hat{L}^{-1}(x^2\tilde{\psi}), \end{gathered} \tag{26}$$

where quantity N_{cr} is equal to the number of particles in the stationary solution (3),

$$N_{\text{cr}} = 2\pi\int_0^\infty dxx\tilde{\psi}^2. \tag{27}$$

This quantity is independent of the choice of energy $|E_0|$. Applying simple transformations to Eq. (23) and using Eq. (24), we can reduce Eq. (23) to the form

$$\begin{gathered} \frac{|E_0|}{2\lambda^2}\left(N - 2\pi\int_0^\infty dx\,x\tilde{\psi}^2\right) + E = \frac{\pi}{4}\int_0^\infty dx \\ \times x\left\{\dot{\lambda}^2(x\tilde{\psi})^2 - \frac{1}{|E_0|}(3\lambda^2\dot{\lambda}^2\dot{\mu} - \lambda^3\ddot{\lambda}\dot{\mu} + \lambda^3\dot{\lambda}\ddot{\mu})(x\tilde{\psi})^2\right. \\ \left. - \left(\frac{\lambda^2\ddot{\lambda}}{2}\right)^2 x^2\tilde{\psi}\hat{L}^{-1}(x^2\tilde{\psi}) + \frac{\lambda^3\dot{\lambda}^2\ddot{\lambda}}{2}x^2\tilde{\psi}\hat{L}^{-1}(x^2\tilde{\psi})\right\}. \end{gathered} \tag{28}$$

Quantities (N, E) in formulas (25) and (26) are the number of particles and the energy concentrated in the collapsing region before the right turning point. Due to the tunneling processes, these quantities are not conserved. The conserved quantities are $\{N_{\text{col}}, E_{\text{col}}\}$ determined for wave function ψ_{col} at the initial instant. To close the system of equations, it is necessary to find the number of particles and the energy in the region after the right turning point.

5. WAVE FUNCTION IN THE REGION AFTER THE TURNING POINT

To investigate the region after the turning point, we represent function $\psi_{\rm col}$ in the form

$$\psi_{\rm col} = \frac{1}{\lambda}\hat{\psi}\exp\left[i|E_0|\int\frac{dt}{\lambda^2} + \frac{i\lambda\dot{\lambda}}{4}x^2\right]. \qquad (29)$$

Substituting expression (29) into formula (1), we obtain the equation for function $\hat{\psi}$ in variables (t, x):

$$\begin{gathered}\frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial}{\partial x}\right)\hat{\psi} + |\hat{\psi}|^2\hat{\psi} - |E_0|\hat{\psi} \\ -\frac{x^2}{4}\lambda^3\ddot{\lambda}\hat{\psi} + i\lambda^2\frac{\partial\hat{\psi}}{\partial t} = 0.\end{gathered} \qquad (30)$$

In the essentially nonlinear region, function $|\hat{\psi}|$ is close to $\tilde{\psi}$ and, outside this region, the nonlinear term is small. Taking this remark into account, we can rewrite Eq. (30) in the form

$$\begin{gathered}\frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial}{\partial x}\right)\hat{\psi} + \tilde{\psi}^2\hat{\psi} - |E_0|\hat{\psi} \\ + a(\tau)x^2\hat{\psi} + i\frac{\partial\hat{\psi}}{\partial\tau} = 0,\end{gathered} \qquad (31)$$

where $a(\tau) \equiv -\lambda^3\ddot{\lambda}/4$. The new time τ is connected to time t by the relation

$$d\tau = \frac{1}{\lambda^2}dt. \qquad (32)$$

Equation (31) describes the motion of a two-dimensional quantum particle in the field of potential barrier $-\tilde{\psi}^2 - a(\tau)x^2$.

We know from quantum mechanics [21] that, in a potential field of this kind, a flow of particles running away from the barrier arises. The right turning point $x_{\rm t.p.}$ can be found from the condition that the potential energy $a(\tau)x^2$ is equal to the total energy $|E_0|$ in Eq. (31). We have

$$x_{\rm t.p.} = \frac{2\sqrt{|E_0|}}{(-\lambda^3\ddot{\lambda})^{1/2}}. \qquad (33)$$

In a neighborhood of the turning point in the region $x < x_{\rm t.p.}$, wave function $\hat{\psi}$ is equal to

$$\begin{gathered}\hat{\psi} = \frac{B|E_0|^{1/8}(-\lambda^3\ddot{\lambda})^{1/8}}{\sqrt{2}(x_{\rm t.p.} - x)^{1/4}}\exp\left\{-\frac{\pi|E_0|}{2\sqrt{-\lambda^3\ddot{\lambda}}}\right. \\ \left.+\frac{2}{3}|E_0|^{1/4}(-\lambda^3\ddot{\lambda})^{1/4}(x_{\rm t.p.} - x)^{3/2}\right\}.\end{gathered} \qquad (34)$$

Constant B in formula (34) is on the order of unity and can be determined from the numerical solution for function $\tilde{\psi}$ in the region $x_{\rm t.p.} \gg x \gg 1$

$$\tilde{\psi}_{x \gg 1} = \frac{B|E_0|^{1/4}}{\sqrt{x}}\exp(-\sqrt{|E_0|}x). \qquad (35)$$

In the region after the turning point $x > x_{\rm t.p.}$, wave function $\hat{\psi}$ is defined by the expression

$$\begin{gathered}\hat{\psi} = \frac{(|E_0|)^{1/4}B}{(x^2 - x_{\rm t.p.}^2)^{1/4}}\left(\frac{x_{\rm t.p.}}{x}\right)^{1/2}\exp\left\{-\frac{\pi|E_0|}{2\sqrt{-\lambda^3\ddot{\lambda}}}\right. \\ +\frac{i\pi}{4} + \frac{ix}{4}(-\lambda^3\ddot{\lambda})^{1/2}(x^2 - x_{\rm t.p.}^2)^{1/2} \\ \left.-\frac{i|E_0|}{(-\lambda^3\ddot{\lambda})^{1/2}}\ln\left(\frac{x + (x^2 - x_{\rm t.p.}^2)^{1/2}}{x_{\rm t.p.}}\right)\right\}.\end{gathered} \qquad (36)$$

Formulas (31) and (36) allow us to find the flux J of particles coming out from under the barrier per real time unit t. We have

$$\begin{aligned}J &= \frac{i}{\lambda^2}2\pi x\left\{\hat{\psi}\frac{\partial\hat{\psi}^*}{\partial x} - \hat{\psi}^*\frac{\partial}{\partial x}\hat{\psi}\right\} \\ &= \frac{4\pi B^2(|E_0|)}{\lambda^2}\exp\left(-\frac{\pi|E_0|}{\sqrt{-\lambda^3\ddot{\lambda}}}\right).\end{aligned} \qquad (37)$$

The particles outgoing from under the barrier continue their motion to the collapse point. The total number of particles in the collapse region remains the same, and function $N(t)$ defined by formula (25) satisfies the equation

$$\frac{\partial N}{\partial t} = -J \qquad (38)$$

with the boundary condition at the initial instant $t_{\rm in}$

$$N(t_{\rm in}) = N_{\rm col}. \qquad (39)$$

Expression (36) for wave function $\hat{\psi}$ is valid in the region $x < x_{\max}$, where $x_{\max}$ is the maximal distance between the collapsing particles and the barrier.

Formula (36) allows us to find all physical quantities related to the existence of overbarrier particles behind the right turning point. For the number of particles N_{tail} and energy E_{tail} concentrated in this region, we obtain

$$
\begin{aligned}
N_{\text{tail}} &= 2\pi B^2(|E_0|)^{1/2} x_{\text{t.p.}} \\
&\times \exp\left(-\frac{\pi|E_0|}{\sqrt{-\lambda^3\ddot{\lambda}}}\right)\ln\left(\frac{2x_{\max}}{x_{\text{t.p.}}}\right),
\end{aligned}
\tag{40}
$$

$$
\begin{aligned}
E_{\text{tail}} &= \frac{\pi}{8\lambda^2}B^2 x_{\text{t.p.}}(|E_0|)^{1/2}\exp\left(-\frac{\pi|E_0|}{\sqrt{-\lambda^3\ddot{\lambda}}}\right) \\
&\times\left\{(\lambda\dot{\lambda}+\sqrt{-\lambda^3\ddot{\lambda}})^2\left(x_{\max}^2-\frac{t_{\text{t.p.}}^2}{2}\right)\right. \\
&\left.+\,x_{\text{t.p.}}^2((\lambda\dot{\lambda})^2+\lambda^3\ddot{\lambda})\ln\left(\frac{2x_{\max}}{x_{\text{t.p.}}}\right)-\lambda\dot{\lambda}\sqrt{-\lambda^3\ddot{\lambda}}\,x_{\text{t.p.}}^2\right\}.
\end{aligned}
$$

Contribution L_{tail} of the region after the right turning point to quantity L can be expressed as

$$
\begin{aligned}
L_{\text{tail}} &= \pi\lambda^2B^2(|E_0|)^{1/2}x_{\text{t.p.}}\exp\left(-\frac{\pi|E_0|}{\sqrt{-\lambda^3\ddot{\lambda}}}\right) \\
&\times\left[\left(x_{\max}^2-\frac{x_{\text{t.p.}}^2}{2}\right)+x_{\text{t.p.}}^2\ln\left(\frac{2x_{\max}}{x_{\text{t.p.}}}\right)\right].
\end{aligned}
\tag{41}
$$

Using the three exact conservation laws and formulas (25), (26), and (40), we obtain the closed system of equations for quantities λ and μ, which contains the truncation parameter $x_{\max}$,

$$
\begin{aligned}
N_{\text{col}}-N_{\text{cr}} &= \frac{\gamma_0}{4|E_0|}(-\lambda^3\ddot{\lambda})-\frac{3\gamma_1}{|E_0|}(\lambda^3\ddot{\lambda})^2 \\
&+\frac{\gamma_0\lambda^2}{8|E_0|^2}(6\lambda^2\dot{\lambda}^2\dot{\mu}+6\lambda^3\ddot{\lambda}\dot{\mu}+6\lambda^3\dot{\lambda}\ddot{\mu}+\lambda^4\dddot{\mu})+N_{\text{tail}}, \\
E_{\text{col}} &= \frac{\gamma_0}{16}\frac{\partial^2}{\partial t^2}\left(\lambda^2-\frac{\lambda^4\dot{\mu}}{|E_0|}\right) \\
&+\gamma_1[(\lambda^2\ddot{\lambda})^2+\lambda^3\dot{\lambda}^2\ddot{\lambda}]+E_{\text{tail}}, \\
E_{\text{col}} &= \frac{1}{16}\frac{\partial^2}{\partial t^2}\left\{\gamma_0\left(\lambda^2-\frac{\lambda^4\dot{\mu}}{|E_0|}\right)+8\gamma_1\lambda^5\ddot{\lambda}+L_{\text{tail}}\right\},
\end{aligned}
\tag{42}
$$

where constants γ_0 and γ_1 are determined in the Appendix (formula (A.9)). We should eliminate the truncation parameter $x_{\max}$. For this purpose, we can use the third equation in system (42). Integrating this equation, we obtain

$$
\begin{aligned}
&[E_{\text{col}}t^2+B_1t+B_0] \\
&= \frac{\gamma_0}{8}\left(\lambda^2-\frac{\lambda^4\dot{\mu}}{|E_0|}\right)+\gamma_1\lambda^5\ddot{\lambda}+\frac{1}{8}L_{\text{tail}}.
\end{aligned}
\tag{43}
$$

Quantities B_0 and B_1 in formula (43) are integration constants. Multiplying Eq. (43) by the quantity $\left(\frac{\dot{\lambda}}{\lambda}+\sqrt{-\frac{\ddot{\lambda}}{\lambda}}\right)^2$, multiplying the first equation of system (42) by the quantity

$$
\frac{x_{\text{t.p.}}^2}{8\lambda^2}(\lambda^3\ddot{\lambda}-\lambda\dot{\lambda}\sqrt{-\lambda^3\ddot{\lambda}}),
$$

and subtracting both equations from the second equation of system (42), we obtain the first equation free of truncation parameter $x_{\max}$:

$$
\begin{aligned}
&E_{\text{col}}-\left(\frac{\dot{\lambda}}{\lambda}+\sqrt{-\frac{\ddot{\lambda}}{\lambda}}\right)^2(E_{\text{col}}t^2+B_1t+B_0) \\
&+(N_{\text{col}}-N_{\text{cr}})\frac{|E_0|}{2\lambda^2}\left(1+\frac{\dot{\lambda}}{\sqrt{-\lambda\ddot{\lambda}}}\right) \\
&= \frac{\gamma_0}{16}\frac{\partial^2}{\partial t^2}\left(\lambda^2-\frac{\lambda^4\dot{\mu}}{|E_0|}\right)+\gamma_1[(\lambda^2\ddot{\lambda})^2+\lambda^3\dot{\lambda}^2\ddot{\lambda}] \\
&-\left(\frac{\dot{\lambda}}{\lambda}+\sqrt{-\frac{\ddot{\lambda}}{\lambda}}\right)^2\left[\frac{\gamma_0}{8}\left(\lambda^2-\frac{\lambda^4\dot{\mu}}{|E_0|}\right)+\gamma_1\lambda^5\ddot{\lambda}\right] \\
&+\frac{|E_0|}{2\lambda^2}\left(1+\frac{\dot{\lambda}}{\sqrt{-\lambda\ddot{\lambda}}}\right)\left[\frac{\gamma_0}{4|E_0|}(-\lambda^3\ddot{\lambda})-\frac{3\gamma_1}{|E_0|}(\lambda^3\ddot{\lambda})^2\right. \\
&\left.+\frac{\gamma_0\lambda^2}{8|E_0|^2}(6\lambda^2\dot{\lambda}^2\dot{\mu}+6\lambda^3\ddot{\lambda}\dot{\mu}+6\lambda^3\dot{\lambda}\ddot{\mu}+\lambda^4\dddot{\mu})\right] \\
&+\pi B^2|E_0|^2\frac{\dot{\lambda}}{\lambda^4\ddot{\lambda}}\exp\left(-\frac{\pi|E_0|}{\sqrt{-\lambda^3\ddot{\lambda}}}\right).
\end{aligned}
\tag{44}
$$

The second equation free of parameter $x_{\max}$ can be obtained by differentiating the first equation of system (41) with respect to time *t*:

$$\frac{\partial}{\partial t}\left\{\frac{\gamma_0}{4|E_0|}(-\lambda^3\ddot{\lambda}) - \frac{3\gamma_1}{|E_0|}(\lambda^3\ddot{\lambda})^2 + \frac{\gamma_0\lambda^2}{8|E_0|^2}(6\lambda^2\dot{\lambda}^2\dot{\mu} + 6\lambda^3\ddot{\lambda}\dot{\mu} + 6\lambda^3\dot{\lambda}\ddot{\mu} + \lambda^4\ddot{\mu})\right\} \quad (45)$$

$$+ \frac{4\pi B^2|E_0|}{\lambda^2}\exp\left(-\frac{\pi|E_0|}{\sqrt{-\lambda^3\ddot{\lambda}}}\right) = 0.$$

The system of two equations (44) and (45) for parameters (λ, μ) describes the total collapse (or expansion) process. The collapse process can be divided into two stages. The first stage includes almost the entire time domain except for an exponentially narrow neighborhood of the collapse point. The second stage is an exponentially narrow neighborhood of the collapse point.

6. COLLAPSE PROCESS OUTSIDE AN EXPONENTIALLY NARROW NEIGHBORHOOD OF THE COLLAPSE POINT

In this region, it is convenient to use system of equations (42), where the terms $\{N_{\text{tail}}, E_{\text{tail}}, L_{\text{tail}}\}$ associated with the tunneling should be omitted. Subtracting the third equation of system (42) from the second equation of this system, we obtain an equation for quantity λ:

$$(\lambda^2\ddot{\lambda})^2 + \lambda^3\dot{\lambda}^2\ddot{\lambda} = \frac{1}{2}\frac{\partial^2}{\partial t^2}(\lambda^5\ddot{\lambda}). \quad (46)$$

Equation (46) can be reduced to the form

$$\frac{\partial}{\partial t}[\lambda(\lambda\dddot{\lambda} + 3\dot{\lambda}\ddot{\lambda})] + 5\dot{\lambda}(\lambda\dddot{\lambda} + 3\dot{\lambda}\ddot{\lambda}) = 0. \quad (47)$$

We can easily integrate Eq. (47) once and obtain

$$\lambda^6[\lambda\dddot{\lambda} + 3\dot{\lambda}\ddot{\lambda}] = \text{const}. \quad (48)$$

The constant in the right-hand side of Eq. (48) can only be zero. As a result, the equation studied in [20] arises. The solution to this equation is

$$\lambda^2 = C_1(t_0 - t)^2 + C/C_1, \quad (49)$$

where C, C_1, and t_0 are integration constants. The value of the constant $C = 0$ corresponds to the separatrix which separates the collapse domain from the extension domain.

We seek quantity $\dot{\mu}$ in the form

$$\dot{\mu} = \frac{\alpha|E_0|}{\lambda^2}. \quad (50)$$

Substituting this value of $\dot{\mu}$ into the first and third equations of system (42), we obtain

$$\begin{aligned} E_{\text{col}} &= \frac{1}{8}\gamma_0 C_1(1-\alpha) + \gamma_1 C C_1, \\ N_{\text{col}} - N_{\text{cr}} &= \frac{\gamma_0 C}{4|E_0|}(-1+2\alpha) - \frac{3\gamma_1}{|E_0|}C^2. \end{aligned} \quad (51)$$

The solvability condition for Eqs. (51) imposes one constraint on the parameters $\{E_{\text{col}}, N_{\text{col}} - N_{\text{cr}}, C, C_1\}$, namely,

$$\frac{4E_{\text{col}}}{C_1} + \frac{|E_0|}{C}(N_{\text{col}} - N_{\text{cr}}) = \frac{\gamma_0}{4} + \gamma_1 C. \quad (52)$$

From Eqs. (51), we find the value of the parameter α

$$\alpha = \frac{4}{\gamma_0}\left\{\gamma_1 C + \frac{2E_{\text{col}}}{C_1} + \frac{|E_0|}{C}(N_{\text{col}} - N_{\text{cr}})\right\}. \quad (53)$$

7. EXPONENTIALLY NARROW NEIGHBORHOOD OF THE COLLAPSE POINT

In the exponentially narrow neighborhood of the collapse point, the processes of tunneling and particle accumulation after the right turning point (in the domain $x > x_{\text{t.p.}}$) become significant. In this domain, as well as in the transition domain, we should use system of equations (44) and (45). We seek a solution to this system for the functions λ and μ in the form

$$\lambda = \frac{\sqrt{t^* - t}}{f}, \quad \dot{\mu} = \frac{\tilde{\chi}|E_0|}{t^* - t}, \quad (54)$$

where t^* is the time of collapse and f and $\tilde{\chi}$ are slow functions of time. Taking into account that the function f is slowly varying, we obtain

$$\begin{aligned} &\frac{1}{2\lambda^2}\left(1 + \frac{\dot{\lambda}}{\sqrt{-\lambda\ddot{\lambda}}}\right) = -f\dot{f}, \quad \lambda^3\ddot{\lambda} = -\frac{1}{4f^4}, \\ &\frac{\partial^2}{\partial t^2}\lambda^2 = \frac{2\dot{f}}{f^3}, \quad \frac{\dot{\lambda}}{\lambda} + \sqrt{-\frac{\ddot{\lambda}}{\lambda}} = -\frac{\dot{f}}{f}. \end{aligned} \quad (55)$$

In the neighborhood of the transition point, the second term on the left-hand side of formula (44) is small. The third term in the left-hand side of formula (44) can be eliminated with the help of the change of variables

$$\tilde{\chi} = -\frac{4|E_0|}{\gamma_0}(N_{\text{col}} - N_{\text{cr}})f^6 + \chi. \quad (56)$$

As a result, system of equations (44) and (45) can be reduced with the help of formulas (55) and (56) to the form

$$\begin{gathered}\frac{\gamma_0}{16}\left\{\frac{\dot f}{f^3}+\frac{\dot\chi}{f^4}-\frac{2\chi\dot f}{f^5}\right\}\\+\frac{2\pi B^2|E_0|^2f^4}{t^*-t}\exp(-2\pi|E_0|f^2) = 0,\\[2ex]-\frac{\gamma_0\dot f}{4f^5}-\frac{\gamma_0\dot\chi}{8f^6}+\frac{3}{4}\frac{\chi\dot f}{f^7}\\+\frac{4\pi B^2|E_0|^2f^2}{t^*-t}\exp(-2\pi|E_0|f^2) = 0.\end{gathered}\tag{57}$$

Multiplying the second equation of system (57) by $f^2/2$ and subtracting it from the first equation of system (57), we obtain an equation without exponential terms, namely,

$$\dot\chi-4\chi\frac{\dot f}{f}+\frac{3}{2}f\dot f = 0.\tag{58}$$

The general solution to Eq. (58) is

$$\chi = \frac{3}{4}f^2+\tilde\alpha f^4,\tag{59}$$

where $\tilde\alpha$ is a constant.

Substituting expression (59) for the function χ into the first formula of system (57), we obtain an equation for the function f, namely,

$$\frac{\dot f}{f^5} = -\frac{16\pi B^2|E_0|^2}{\tilde\alpha\gamma_0(t^*-t)}\exp(-2\pi|E_0|f^2).\tag{60}$$

Solving Eq. (60), we obtain the following equation for the function f:

$$2\pi|E_0|f^2 = \ln\left(\frac{64\pi^2B^2|E_0|^3f^6}{(-\tilde\alpha)\gamma_0}\ln\left(\frac{1}{t^*-t}\right)\right).\tag{61}$$

To find the value of constant $\tilde\alpha$ in Eqs. (59) and (61), we must solve system of equations (44) and (45) in an intermediate domain, where the order of exponential terms is the same as the order of terms appearing in Eq. (46).

8. DECOMPOSITION OF THE INITIAL VALUE OF THE WAVE FUNCTION INTO THE COLLAPSING PART ψ_{col} AND ITS ORTHOGONAL COMPLEMENT $\psi_\perp$

Suppose that wave function ψ, which can be represented in form (5) and (6), is defined at the initial time instant as

$$\psi = \psi_{\text{col}}+\psi_\perp$$

and the norm of the function $\psi_\perp$ is small. First of all, note that there are three free parameters, namely, the energy $|E_0|$ and coordinates (x_0, y_0) of the collapse point. We construct a basis in which the function ψ_{col} (6) can be expanded. Formulas (9), (10), (15), and (16) imply that the basis consists of the functions

$$\begin{gathered}\tilde\psi,\quad \frac{\partial}{\partial x}(x\tilde\psi),\quad \hat L^{-1}(x^2\tilde\psi),\quad x^2\tilde\psi,\\ \hat L^{-1}(x^2\phi),\quad \hat L^{-1}(\phi),\quad \hat L^{-1}(\tilde\psi\phi^2).\end{gathered}\tag{62}$$

The function ψ_{col} should be expanded in basis (62). Then, the norm of the function $\psi_\perp$ should be minimized in all the free parameters, $|E_0|$, (x_0, y_0), and the coefficients of the expansion of the function ψ_{col} in basis (62). Minimizing in the free parameters, we determine the location of the collapse center, the quantity $|E_0|$, and the initial values for system of equations (44) and (45) (taking into account Eqs. (43), (49)–(51), and (53)).

9. CONCLUSIONS

In the adiabatic approximation, a system of equations is obtained which describes the collapse process in the entire time interval until the singular point is reached. Almost everywhere, except for an exponentially narrow region near the collapse point, the tunneling processes in the inverted parabolic potential are negligible. Tunneling results in the particle accumulation in the region after the right turning point. It is essential that the energy in this region become infinitely high as we come near the collapse point.

In the problem under consideration, there are three exact conservation laws, namely, the particle conservation law, total energy conservation law, and quasiconformal conservation law. Using the quasiconformal conservation law, we can eliminate the quantities related to the particle accumulation in the region behind the right turning point. These quantities diverge as we come near the collapse point. As a result, a system of two equations for quantities $\{\lambda, \dot\mu\}$ arises. When solving this system in a neighborhood of the collapse point, a free parameter $\tilde\alpha$ arises. The value of the parameter $\tilde\alpha$ is determined by the conditions of matching the solutions in the transition region where all terms in system of equations (44) and (45) are of the same order. An important specific feature of the problem under consideration is the dropping of the quantity $\dot{\mu}$ from all equations in second-order perturbation theory and the occurrence of this quantity only in the fourth order.

A sufficient condition for the collapse is a negative value of energy E [19]. If the number of particles $N < N_{\mathrm{cr}}$, the solution to Eq. (1) remains finite for any finite time [16]. There exist collapsing solutions for $N = N_{\mathrm{cr}}$ and for any value of energy $E > 0$ [5]. In the adiabatic approximation, we have established an implicit relation between the initial data and the conditions for the collapse arising for a finite time value.

ACKNOWLEDGMENTS

This work was financially supported by the Russian Foundation for Basic Research and the Ministry of Sciences of the Russian Federation. I.M. Sigal acknowledges the financial support of NSERC (grant no. A7901).

APPENDIX

First of all, we write the relations which are used in deducing the equations for the quantities $\{N, E, \lambda, \mu\}$. Multiplying both sides of Eq. (7) by the function $\partial(x^{2n+1}\psi)/\partial x$ and integrating with respect to $x\partial x$ over the interval $(0, \infty)$, we obtain an infinite number of relations for integrals of function $\tilde{\psi}$:

$$\int_0^\infty dxx\{-4n^3x^{2n-2}\tilde{\psi}^2 + (4n+2)|E_0|x^{2n}\tilde{\psi}^2 - (3n+1)x^{2n}\tilde{\psi}^4\} = 0, \tag{A.1}$$

where $n = 0, 1, 2, \ldots$. Some of these equations were obtained and used in paper [20]. One can directly verify that the following equations for the function $\tilde{\psi}$ are valid:

$$\hat{L}\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right) = 2|E_0|\tilde{\psi},$$

$$\hat{L}\tilde{\psi} = 2\tilde{\psi}^3,$$

$$\hat{L}(x^2\tilde{\psi}) = 2x^2\tilde{\psi}^3 + 4\frac{\partial}{\partial x}(x\tilde{\psi}), \tag{A.2}$$

$$\tilde{L}(\tilde{\psi}^3) = 2|E_0|\tilde{\psi}^3 + 6\tilde{\psi}\left(\frac{\partial\tilde{\psi}}{\partial x}\right)^2,$$

$$\tilde{L}(x^2\tilde{\psi}^3) = -2\tilde{\psi}^3 + 2|E_0|x^2\tilde{\psi}^3 + 6\tilde{\psi}\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)^2.$$

Here, the operator $\hat{L}$ is specified by expression (11).

Using Eqs. (A.1) and (A.2), we obtain the following three relations:

$$\int_0^\infty dxx\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)^2 = \int_0^\infty dxx\tilde{\psi}^2 - |E_0|\int_0^\infty dxx(x\tilde{\psi})^2 + \int_0^\infty dxx(x^2\tilde{\psi}^4),$$

$$\int_0^\infty dxx\tilde{\psi}\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)^3 = \frac{|E_0|}{3}\int_0^\infty dxx(x^2\tilde{\psi}^4 + \tilde{\psi}^2), \tag{A.3}$$

$$\int_0^\infty dxxx^2\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)^2 = \int_0^\infty dxx\{5(x\tilde{\psi})^2 - |E_0|x^4\tilde{\psi}^2 + (x\tilde{\psi})^4\}.$$

Let us find quantity I defined by the expression

$$I = \int_0^\infty dxx\left\{3\tilde{\psi}\left(\frac{\partial}{\partial x}(x\tilde{\psi})\right)(\hat{L}^{-1}(x^2\tilde{\psi}))^2 - |E_0|(\hat{L}^{-1}(x^2\tilde{\psi}))^2\right\}. \tag{A.4}$$

Integrating by parts, we reduce the expression for I to the form

$$I = -\int_0^\infty dx\hat{L}^{-1}(x^2\tilde{\psi})[(3\tilde{\psi}^2 - |E_0|)x^2] \times\frac{\partial}{\partial x}(\hat{L}^{-1}(x^2\tilde{\psi})). \tag{A.5}$$

On the other hand, the quantity I is given by

$$I = \frac{1}{2}\int_0^\infty dx\hat{L}^{-1}(x^2\tilde{\psi}) \times\left\{\frac{\partial}{\partial x}\left[x^2\left(\hat{L} - \frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial}{\partial x}\right)\right)\hat{L}^{-1}(x^2\tilde{\psi})\right] - x^2(3\tilde{\psi}^2 - |E_0|)\frac{\partial}{\partial x}\hat{L}^{-1}(x^2\tilde{\psi})\right\}. \tag{A.6}$$

The third term in expression (A.6) is a total derivative and its integral is equal to zero. As a result, the quantity I is defined by the expression

$$I = -\frac{1}{2}\int_0^\infty dx\hat{L}^{-1}(x^2\tilde{\psi})[x^2(3\tilde{\psi}^2 - |E_0|)]\frac{\partial}{\partial x}\hat{L}^{-1}(x^2\tilde{\psi}) + \frac{1}{2}\int_0^\infty dx\hat{L}^{-1}(x^2\tilde{\psi})\frac{\partial}{\partial x}(x^4\tilde{\psi}). \tag{A.7}$$

Formulas (A.5) and (A.7) imply that

$$I = \int_0^\infty dx\hat{L}^{-1}(x^2\tilde{\psi})\frac{\partial}{\partial x}(x^4\tilde{\psi}). \tag{A.8}$$

Let us define two constants γ_0 and γ_1 as

$$\gamma_0 = 2\pi\int_0^\infty dxx(x\tilde{\psi})^2, \qquad \gamma_1 = \frac{\pi}{8}\int_0^\infty dxx\{x^2\tilde{\psi}\hat{L}^{-1}(x^2\tilde{\psi})\}. \tag{A.9}$$

In the quantities γ_0 and γ_1, we can separate the dependence on $|E_0|$ and write these quantities in the form

$$\gamma_0 = \frac{1}{|E_0|}\gamma_{00}, \quad \gamma_1 = \frac{1}{|E_0|^3}\gamma_{11}, \tag{A.10}$$

where γ_{00} and γ_{11} are independent of the choice of $|E_0|$.

REFERENCES

1. V. I. Talanov, Pis'ma Zh. Éksp. Teor. Fiz. **2**, 218 (1965) [JETP Lett. **2**, 138 (1965)].
2. P. L. Keley, Phys. Rev. Lett. **15**, 1005 (1965).
3. A. L. Dyshko, V. N. Lugovoĭ, and A. M. Prokhorov, Pis'ma Zh. Éksp. Teor. Fiz. **6**, 665 (1967) [JETP Lett. **6**, 153 (1967)].
4. E. Dawes and J. H. Marburger, Phys. Rev. **179**, 862 (1969).
5. V. I. Talanov, Pis'ma Zh. Éksp. Teor. Fiz. **11**, 303 (1970) [JETP Lett. **11**, 199 (1970)].
6. V. E. Zakharov, Zh. Éksp. Teor. Fiz. **62**, 1746 (1972) [Sov. Phys. JETP **35**, 908 (1972)].
7. G. M. Fraĭmen, Zh. Éksp. Teor. Fiz. **88**, 390 (1985) [Sov. Phys. JETP **61**, 228 (1985)].
8. V. E. Zakharov and V. F. Shvets, Pis'ma Zh. Éksp. Teor. Fiz. **47**, 227 (1988) [JETP Lett. **47**, 275 (1988)].
9. M. J. Landman, G. C. Papanicolaou, C. Sulem, and P.-L. Sulem, Phys. Rev. A **38**, 3837 (1988).
10. N. E. Kosmatov, V. F. Shvets, and V. E. Zakharov, Physica D (Amsterdam) **52**, 16 (1991).
11. B. J. Le Mesurier, G. C. Papanicolaou, C. Sulem, and P.-L. Sulem, in *Directions in Partial Differential Equations*, Ed. by M. G. Grandall, P. H. Rabinovitz, and R. E. Tuner (Academic, Boston, 1987), p. 159.
12. C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation. Self-Focusing and Wave Collapse* (Springer, New York, 1999).
13. A. I. Smirnov and G. M. Fraiman, Physica D (Amsterdam) **52**, 2 (1991).
14. V. M. Malkin, Phys. Lett. A **151**, 285 (1990).
15. D. Pelinovsky, Physica D (Amsterdam) **119**, 801 (1998).
16. M. I. Weinstein, Commun. Math. Phys. **87**, 567 (1983).
17. G. Perelman, in *Nonlinear Dynamics and Renormalization Group*, Ed. by I. M. Sigal and C. Sulem (American Mathematical Society, Providence, R.I., 2001), CRM Proceedings and Lecture Notes, Vol. 27, p. 147.
18. L. Berge, Phys. Rep. **303**, 259 (1998).
19. V. N. Vlasov, I. A. Petrishchev, and V. I. Talanov, Izv. Vyssh. Uchebn. Zaved., Radiofiz. **14**, 1353 (1971).
20. Yu. N. Ovchinnikov and I. M. Sigal, Pis'ma Zh. Éksp. Teor. Fiz. **75**, 428 (2002) [JETP Lett. **75**, 357 (2002)].
21. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 3: *Quantum Mechanics: Non-Relativistic Theory*, 2nd ed. (Fizmatgiz, Moscow, 1963; Pergamon, New York, 1977).

Translated by E. Pankratiev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 204–215.
Translated from Zhurnal Éksperimental’noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 224–236.
Original Russian Text Copyright © 2003 by Kohn.

MISCELLANEOUS

An Exact Theory of Imaging with a Parabolic Continuously Refractive X-ray Lens

V. G. Kohn
Russian Research Centre Kurchatov Institute, Moscow, 123182 Russia
e-mail: kohn@kurm.polyn.kiae.su

Received January 30, 2003

Abstract—A theory is developed of image formation with an X-ray lens that consists of a large number of elements. Each element has a biconcave parabolic profile and weakly refracts an X-ray beam. Since such a lens can have a relatively large length comparable to the focal length, the thin-lens approximation is inapplicable. An exact expression for the propagator of a continuously refractive lens is derived that describes the transfer of radiation through a refractive parabolic medium. We calculate the image propagator that describes the focusing of a parallel beam and the image transfer (the focusing of a microobject), as well as the Fourier transform of the transmission function for a microobject with a lens, is calculated. The effective aperture of an X-ray lens is completely determined by the absorption of radiation and does not depend on its geometrical cross-sectional sizes. If we write the complex refractive index as $n = 1 - \delta + i\beta$, then the beam diameter at the focus is approximately a factor of $0.8\beta/\delta$ smaller than the diameter of the effective aperture, with the index depending only slightly on the wavelength. A continuously refractive lens has no aberrations in the sense that all of the rays that passed through the lens aperture are focused at a single point. The lens can focus radiation inside it and has the properties of a waveguide; i.e., it can reconstruct the beam structure for some lengths to within the absorption-caused distortions. Nonuniform X-ray absorption in the lens leads to the interesting visualization effect of transparent microobjects when their image is focused. In this case, the phase shift gradient produced by the microobject is imaged. We discuss the properties of the Fourier transform pertaining to the absorption of radiation in the lens.

1. INTRODUCTION

The focusing of electromagnetic radiation by refractive lenses is of great importance both in everyday practice and in scientific research. In particular, the human eye has a lens in its structure that focuses visible light at various distances. A microscope and a telescope extend the vision of the world to mini- and macrosizes, respectively. Because of its high penetrability, hard X-ray radiation with photon energies E from 10 to 50 keV allows the internal structure of microobjects to be studied by nondestructive methods. Clearly, the development of effective refractive lenses for hard X-rays could significantly enhance the possibilities of scientific research into the structure of matter in many fields of science. However, this could not be done for 100 years since the discovery of X-ray radiation mainly for two reasons. First, the refractive index of X-rays is very close to unity. Second, all materials absorb X-ray radiation. For example, for aluminum and $E = 25$ keV, the complex refractive index is $n = 1 - \delta + i\beta$, where $\delta = 8.643 \times 10^{-7}$ and $\beta = 1.747 \times 10^{-9}$.

The problem of weak refraction was first solved in 1996 [1] by using compound lenses made up of a long row of elementary lenses. Each elementary lens has a biconcave shape and a radius of curvature R on the order of 1 mm that is large enough for it to be easily produced. Accordingly, the focal length of such a lens, $F_1 = R/2\delta$, is very large and can reach hundreds of meters. In this case, the focal length of a compound lens with N elements is $F = F_1/N$. Therefore, the focal length can be decreased to a value acceptable for experiments by increasing the number of elements. Fortunately, $\mathrm{Re}\,n < 1$ for X-rays; as a result, the focusing lens is biconcave. In this case, the thickness of the material in the central part of the lens is small compared to the absorption length. Nevertheless, absorption does exist and causes both an overall reduction in the beam intensity and a restriction of the aperture of X-ray compound refractive lenses. The nonuniform absorption of radiation in an X-ray lens is a new property compared to lenses for visible light, which leads to interesting properties of the image, as we show below.

The relatively small aperture of an X-ray lens (fractions of a millimeter) is not a drawback because the X-ray beams generated by synchrotron radiation sources have small cross-sectional sizes and weak divergence. Thus, on third-generation (ESRF, APS, SPring-8) synchrotron radiation sources, the vertical size of the emitting region does not exceed 30 μm, while the distance from the source to the sample is more than 50 m. At present, many papers in which various methods of producing compound refractive lenses for X-rays have been published. The simplest method involves drilling a row of circular holes to obtain a lin-

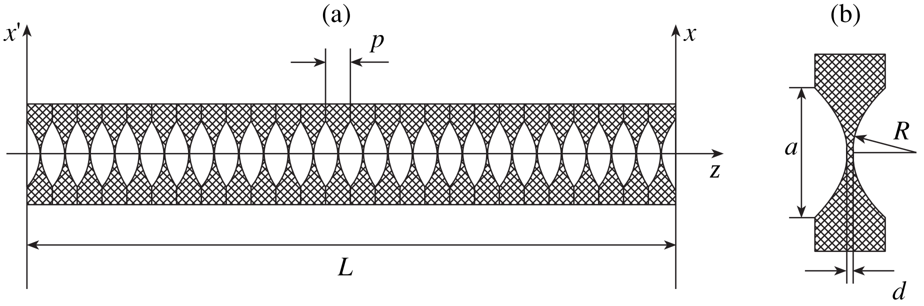


Fig. 1. (a) A compound refractive X-ray lens and (b) the parameters of an individual lens element.

ear focus [1] or two mutually perpendicular rows of holes to obtain a point focus [2–4]. An advantage to such lenses is their extremely low expense; drawbacks are the high surface roughness, the spherical aberrations, and the anomalously large length for a large radius of curvature of the holes. An alternative method consists in forcing air bubbles into a glue [5]. An “alligator” lens in which the parabolic phase shift profile is roughly specified by two rows of perpendicular teeth located at a small angle to the X-ray beam was suggested and tested [6, 7]. Planar lenses in which an accurate parabolic profile with a radius of curvature on the order of 1 μm is produced in a thin layer on a silicon surface were produced by using the complex deep lithography technique [8].

However, compound lenses with circular apertures and parabolic profiles are of greatest interest in terms of their imaging properties. The cross section of such a lens and the parameters of one element are shown in Fig. 1. The elements of a compound parabolic lens are produced by embossing a parabolic profile in aluminum plates [9–12] or in plates of various plastics (see, e.g., [13, 14]). The number of elements in a compound lens can be varied to obtain the required focal length. Presently, lenses with up to several hundred elements are used. A lens with 1000 elements or more can be easily produced. In this case, the length of the compound lens of $L = Np$ increases with N, while its focal length F decreases. Clearly, as long as $L \ll F$, the focal length can be estimated using the thin-lens formula, $F \approx R/2N\delta$. Otherwise, the problem of radiation transfer through a long compound lens with allowance made for the change in the path of rays in the lens itself should be solved. This can be easily done by geometrical optics techniques [15].

The complete solution of the problem must be in the form of an integral equation similar to the Kirchhoff integral. If the change in the transverse structure of the wave field in the thickness p of one element of a compound lens is small, then we may average the density of the lens material over the length p and treat the lens as a homogeneous parabolic medium along the beam direction. A compound lens that satisfies this requirement is called a parabolic continuously refractive (PCR) X-ray lens. The kernel of the integral equation for such a lens is a continuous function of its length and has an analytic form, as was first shown in [16].

Here, we present an exact theory of image formation with a continuously refractive X-ray lens. Apart from deriving the propagator of the lens itself and studying its properties, we calculate the image propagator and analyze the imaging properties of a PCR lens. We show that a continuously refractive lens has no aberrations in the sense that all of the rays emerging from a single point and passing through different parts of the lens aperture converge to a single point in the image. Its effective aperture is completely determined by the absorption of radiation in the lens and decreases with increasing wavelength. At the same time, the beam diameter at the focus is approximately a factor of $0.8\beta/\delta$ smaller than the diameter of the effective aperture, and the numerical coefficient depends on wavelength only slightly. Nonuniform absorption leads to the visualization of transparent microobjects when they are imaged, with the phase shift gradient produced by the microobject being imaged. We also discuss absorption-related properties of the Fourier transform of the transmission function for the object.

2. THE EXACT PROPAGATOR OF A PARABOLIC CONTINUOUSLY REFRACTIVE X-RAY LENS

Let us assume that the synchrotron radiation is premonochromated and has a high degree of spatial coherence. These conditions are satisfied on third-generation synchrotron radiation sources [17]. We choose the optical axis along the z axis (see Fig. 1) and represent the general solution of the Maxwell equation as

$$E(x, y, z) = \exp(ikz)A_t(x, y, z),$$

where $k = \omega/c$ is the wave number in a vacuum. The function $A_t(x, y, z)$ describes the transfer of the transverse dependence of the wave field along the z axis. Since the wavelength $\lambda = 2\pi/k$ is many orders of magnitude smaller than the scale length of the X-ray interaction with the material λ/δ, we can use the paraxial approximation with a high accuracy; i.e., we can disregard the second derivative of A_t with respect to z compared to its first derivative. As a result, substituting this solution into the Maxwell equation yields the parabolic equation for the function $A_t(x, y, z)$

$$\frac{dA_t}{dz} = -ik\eta s(x,y,z)A_t + \frac{i}{2k}\left(\frac{d^2A_t}{dx^2} + \frac{d^2A_t}{dy^2}\right), \quad (1)$$

where $\eta = 1 - n = \delta - i\beta = \delta(1 - i\gamma)$. In the radiation transfer problem, the wave field on the entrance lens surface is assumed to be given, i.e., $A_t(x, y, 0)$ if the z coordinate is measured from the entrance lens surface. Inside an actual compound lens, the function $s(x, y, z)$ is equal to 1 in the regions filled with the lens material and 0 in the voids (see Fig. 1).

Passing to the limit of a PCR lens implies that the thickness p of one element tends to zero, while the number of elements N increases in such a way that the total length of the lens and its aperture do not change. The radius of curvature of the surfaces also increases. In this case, instead of the actual function $s(x, y, z)$, we may use its average value, which does not depend on the longitudinal coordinate,

$$\bar{s}(x,y) = s_0 + \frac{x^2}{pR} + \frac{y^2}{pR}, \quad s_0 = \frac{d}{p}. \quad (2)$$

This dependence holds only within the geometrical lens aperture of diameter

$$a = 2[R(p-d)]^{1/2}$$

(see Fig. 1). We are interested in sufficiently long lenses where the effective working area (effective aperture) of the lens is determined by the absorption of X-rays in the lens material and has a size smaller than the geometrical aperture. In this case, we may ignore the edge effects and formally assume that dependence (2) holds in the entire region of the transverse xy plane concerned.

After the substitution of $\bar{s}$ for s, the general solution of Eq. (1) can be written as the integral equation

$$A_t(x,y,z) = \int dx'dy' P_L^{(t)}(x,y,x',y',z)A_t(x',y',0). \quad (3)$$

The propagator of a PCR lens, i.e., the kernel of Eq. (3), is the solution of Eq. (1) with the initial function

$$P_L^{(t)}(x,y,x',y',0) = \delta(x-x')\delta(y-y'),$$

where $\delta(x)$ is the Dirac delta function. Given the form of the initial function, it is easy to understand that the solution can be sought by the separation of variables,

$$P_L^{(t)}(x,y,x',y',z) = \exp(-ik\eta s_0 z)P_L(x,x',z)P_L(y,y',z). \quad (4)$$

The partial propagator $P_L(x, x', z)$ satisfies the equation

$$\frac{dP_L}{dz} = -ik\frac{x^2}{2z_c^2}P_L + \frac{i}{2k}\frac{d^2P_L}{dx^2},$$
$$P_L(x,x',0) = \delta(x-x'), \quad z_c = \left(\frac{pR}{2\eta}\right)^{1/2}. \quad (5)$$

This equation is formally identical to the Schrödinger equation for a particle in a parabolic potential if the z coordinate is substituted with time. In quantum mechanics, one is usually interested in stationary states and writes the solution as a series each term of which contains the product of the functions of each individual coordinate.

In [16], we obtained a solution in the form of a simple analytic expression by using the Fourier transform and recurrent equations for the coefficients of various powers of x in the argument of the exponent. Below, we derive the same solution in a more straightforward way. Taking into account the reciprocity principle, the solution should be sought in the form of a symmetric function of the x and x' coordinates. In addition, at small longitudinal distances, the solution must be close to the propagator in the empty space, i.e., to the Kirchhoff propagator,

$$P(x-x',z) = \frac{1}{(i\lambda z)^{1/2}}\exp\left[i\pi\frac{(x-x')^2}{\lambda z}\right], \quad (6)$$

which is significant when the lens is illuminated by a point source. On the other hand, it must contain the phase factor characteristic of a thin lens when it is illuminated by a plane wave. Taking into account these considerations, we will seek the solution in the form

$$P_L(x,x',z) = T\left(x,\frac{r}{a}\right)P(x-x',r)T\left(x',\frac{r}{a}\right),$$
$$T(x,z) = \exp\left(-i\pi\frac{x^2}{\lambda z}\right) \quad (7)$$

with the two unknown functions $r(z)$ and $a(z)$.

The initial condition is satisfied if $r(z) \approx z$ and $a(z) \approx z^2$ for $z \longrightarrow 0$. Substituting this form of the solution into the equation and equating the coefficients of the same powers of x and x', we obtain the system of two ordinary differential equations

$$\frac{dr}{dz} = 1 - a, \quad \frac{da}{dz} = \frac{r}{z_c^2}, \quad (8)$$

whose solution can be easily found:

$$r(z) = z_c\sin\frac{z}{z_c}, \quad a(z) = 1 - \frac{dr}{dz} = 1 - \cos\frac{z}{z_c}. \quad (9)$$

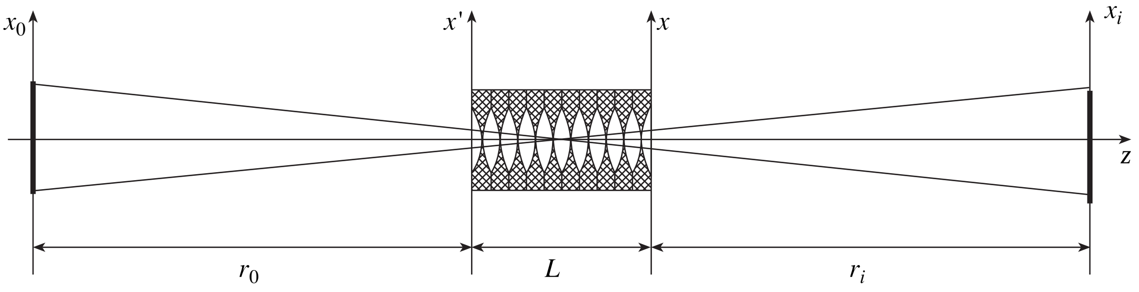


Fig. 2. The experimental scheme for imaging an object with an X-ray PCR lens.

As a result, the propagator of a PCR lens can be written as

$$P_L(x, x', z) = \frac{1}{(i\lambda z_c s_z)^{1/2}}\exp\left[i\pi\frac{(x^2 + x'^2)c_z - 2xx'}{\lambda z_c s_z}\right], \qquad (10)$$

where we denoted

$$s_z = \sin\frac{z}{z_c}, \quad c_z = \cos\frac{z}{z_c}.$$

This expression differs from the formula derived in [16] only by the notation. This notation better takes into account the symmetry properties of the propagator. If the PCR-lens length $L \ll \mathrm{Re}\, z_c$, then

$$P_L(x, x', L) \xrightarrow[L \to 0]{} \exp\left[-i\pi\frac{(x^2 + x'^2 + xx')}{3\lambda F_c}\right]P(x - x', L), \qquad (11)$$

where

$$F_c = \frac{F}{1 - i\gamma} = \frac{z_c^2}{L} = \frac{R}{2N\eta}$$

is the complex focal length of a thin PCR lens. Since we used the relation $F \gg L$ to derive this expression, we may roughly substitute x for x' in the exponential factor. In addition, if the second derivative of the incident wave phase with respect to the transverse coordinates is much smaller than $2\pi/\lambda L$, then the propagator of the empty space can be roughly substituted with the delta function to give

$$P_L(x, x', L) \approx T(x, F_c)\delta(x - x').$$

This expression is commonly used in the thin-lens approximation. Formula (11) gives a more universal approximation for the thin-lens propagator.

Clearly, a PCR lens for which the parameter $\gamma = \beta/\delta$ is at a minimum has the best properties. For this reason, the actual lenses are made of elements with a small atomic number Z (lithium, beryllium, carbon, aluminum). In virtually all interesting cases, $\gamma < 0.005$. If we ignore the absorption, which can be done at least for the rays near the optical axis, then the waveguide properties of a PCR lens follow from the exact expression for the propagator. For

$$L = L_0 = \frac{\pi}{2}(LF)^{1/2} = \frac{\pi}{2}\left(\frac{pR}{2\delta}\right)^{1/2},$$

we obtain $s_z = 1$ and $c_z = 0$. Therefore, a PCR lens makes the Fourier transform of the incident wave. For $L = 2L_0$, the propagator is $\delta(x + x')$; accordingly, the PCR reproduces the incident wave at the exit in an inverse form. For $L = 3L_0$, the lens again makes the Fourier transform but with the opposite sign. Finally, for $L = 2L_0$, the lens faithfully reproduces the incident wave. As the PCR-lens length increases further, these phases are repeated again and again. Since X-rays are absorbed in the PCR lens, both the image and the Fourier transform are produced in a bounded region within the gradually decreasing aperture.

3. THE IMAGE PROPAGATOR WITH AN ARBITRARILY LONG PCR LENS

In an actual experiment, the object being studied, the lens, and the detector are at comparatively large distances from each other, as can be seen from Fig. 2, which also shows the notation for the distances and the coordinate axes. Let us consider the more complex problem of the wave field transfer from the plane immediately behind the object to the detector plane. Clearly, the propagator of this problem is also factorized in the x and y coordinates. Therefore, it will suffice to calculate only the partial image propagator $G(x_i, x_o, r_o, L, r_i)$. It is determined by the convolution of the propagator for a PCR lens with the Kirchhoff propagators that correspond to empty space,

$$G(x_i, x_o) = \int dx dx' P(x_i - x, r_i)P_L(x, x', L)P(x' - x_o, r_o). \qquad (12)$$

Below, to save space, we omit the longitudinal distances in the list of arguments for the image propagator.

We use the following algorithm to calculate the integrals. Let us first consider the extreme case of a thin lens where $P_L(x, x', L)$ is substituted with $T(x, F_c)\delta(x - x')$. Accordingly, the propagator is determined by the single integral

$$\begin{aligned} & G_0(x_i, x_o) \\ & = \int dx P(x_i - x, r_i) T(x, F_c) P(x - x_o, r_o). \end{aligned} \tag{13}$$

This integral reduces to the Fourier transform of the Gaussian function and is again equal to the Gaussian function. It is convenient to write the result as

$$\begin{aligned} & G_0(x_i, x_o) \\ & = T\left(x_i, \frac{r_g}{a_i}\right) P(x_i - x_o, r_g) T\left(x_o, \frac{r_g}{a_o}\right), \end{aligned} \tag{14}$$

where we introduced the parameters

$$r_g = r_o + r_i - \frac{r_o r_i}{F_c}, \quad a_i = \frac{r_o}{F_c}, \quad a_o = \frac{r_i}{F_c}. \tag{15}$$

Expression (14) for the image propagator of a thin lens is identical in form to expression (7) for the propagator of a PCR lens. It immediately follows from this expression that in the limit $\gamma = 0$ and when the condition $r_g = 0$ (the thin-lens formula) is satisfied, the propagator

$$G_0(x_i, x_o) = M^{1/2}\delta(x_i + x_o M),$$

where $M = r_i/r_o$ is the magnification factor. Thus, this expression reproduces the well-known property of a thin lens to focus the image when the lens formula holds:

$$r_o^{-1} + r_i^{-1} = F^{-1}.$$

It is also easy to see that when the condition $a_o = 1$, i.e., $r_i = F$, is satisfied, the term proportional to x_o^2 in the argument of the exponent vanishes and the propagator makes the Fourier transform of the wave field located in the plane immediately behind the object. However, if the object is illuminated by a point source, then the Fourier transform of the transmission function for the object takes place in the focusing plane of the point source. This property also follows from propagator (14), but further calculations are required to prove it (see below).

Substituting expression (7) for the propagator of a PCR lens into integral (12) yields

$$\begin{aligned} G(x_i, x_o) &= \int dx P(x_i - x, r_i) T\left(x, \frac{r_L}{a_L}\right) \\ & \times \int dx' P(x - x', r_L) T\left(x', \frac{r_L}{a_L}\right) P(x' - x_o, r_o), \end{aligned} \tag{16}$$

where $r_L = z_c s_L$ and $a_L = 1 - c_L$. Here, the integral over x' is equivalent to integral (13) but with different expressions for the parameters. Substituting solution (14) for the integral again yields an integral of type (13) with new parameters. As a result, making appropriate algebraic transformations, we derive an exact expression for the image propagator of an arbitrarily long PCR lens in a form similar to the case of a thin lens,

$$G(x_i, x_o) = T\left(x_i, \frac{\tilde{r}_g}{\tilde{a}_i}\right) P(x_i - x_o, \tilde{r}_g) T\left(x_o, \frac{\tilde{r}_g}{\tilde{a}_o}\right), \tag{17}$$

but now the parameters are

$$\begin{aligned} \tilde{r}_g &= (r_o + r_i)c_L + \left(z_c - \frac{r_o r_i}{z_c}\right) s_L, \\ \tilde{a}_i &= 1 - c_L + \frac{r_o}{z_c} s_L, \quad \tilde{a}_o = 1 - c_L + \frac{r_i}{z_c} s_L. \end{aligned} \tag{18}$$

Interestingly, the formulas for the new parameters can also be written in a form similar to the case of a thin lens,

$$\tilde{r}_g = \tilde{r}_o + \tilde{r}_i - \frac{\tilde{r}_o \tilde{r}_i}{\tilde{F}_c}, \quad \tilde{a}_i = \frac{\tilde{r}_o}{\tilde{F}_c}, \quad \tilde{a}_o = \frac{\tilde{r}_i}{\tilde{F}_c}, \tag{19}$$

if we introduce the generalized complex distances

$$\begin{aligned} \tilde{r}_o &= r_o + b_L, \quad \tilde{r}_i = r_i + b_L, \\ \tilde{F}_c &= \frac{z_c}{s_L}, \quad b_L = \tilde{F}_c(1 - c_L). \end{aligned} \tag{20}$$

The result is of great importance because it shows that the imaging properties of an arbitrarily long PCR lens are essentially the same as those of a thin lens. In particular, such a PCR lens has no aberrations in the sense that all ray paths converge at a single point, in contrast, for example, to a lens with a spherical profile. If there were no absorption, then the lens would focus a point source to a point. The blurring of the image point due to the absorption of radiation and the finite aperture is usually attributed to the finite lens resolution. On the other hand, the simple analytic expressions for the generalized distances allow the appropriate corrections that should be made to the experimental scheme to be easily determined.

For example, in the limit of a small lens length compared to the focal length, $L \ll F$, expanding the sine and the cosine in a power series yields

$$\tilde{r}_o = r_o + \frac{1}{2}L, \quad \tilde{r}_i = r_i + \frac{1}{2}L, \quad \tilde{F}_c = F_c + \frac{1}{6}L. \tag{21}$$

This result, which was obtained previously [18] in an approximate and complicated way, is a natural extreme case of the exact theory. It follows from this result that even when the PCR lens has an appreciable length L (tens of centimeters) that satisfies the condition $L \ll F$, it can be treated as a thin lens located in the middle with a sole difference. More specifically, the focal length calculated using the thin-lens formula must be increased by one-sixth of the actual lens length.

On the other hand, at zero distances, the image propagator is identical to the propagator of a PCR lens. Note yet another obvious property of the image propagator: its convolution with the Kirchhoff propagator from a point source $P(x_o - x_s, r_s)$ is described by the same expression (17) in which x_s should be substituted for x_0 and $r_o + r_s$ should be substituted for r_o. The relation

$$\tilde{a}_o + \tilde{a}_i - \tilde{a}_o\tilde{a}_i = \frac{\tilde{r}_g}{\tilde{F}_c} \tag{22}$$

that follows from definitions (19) is used to prove this property.

4. ESTIMATING THE APERTURE AND FOCUS SIZES FOR A PARABOLIC CONTINUOUSLY REFRACTIVE X-RAY LENS

Exact knowledge of such parameters as the size of the effective aperture of an X-ray lens and the size of the focal spot when a plane wave is focused is of considerable practical importance. In the optics of visible light, the aperture is determined by the geometrical sizes of the lens, i.e., by the area through which the rays passing then converge to a focus. For a thin absorbing X-ray lens, it will suffice to consider the intensity distribution of the radiation immediately after the lens when it is illuminated by a plane wave oriented along the optical axis. In this case, the effective aperture is determined by the absorption of radiation in the lens material. Since the total wave intensity in empty space is conserved, simple energy relationships exist between the aperture and focus sizes. For the arbitrarily long lens considered here, this approach does not work, because the incident wave can be partially or completely focused in the lens itself. Therefore, the intensity distribution of the radiation immediately behind the lens does not give us any idea of the actual lens aperture. The aperture of a long PCR lens can be defined in terms of the properties of the propagator $G(x_i, x_o)$. Let us consider a different, simpler approach based on energy considerations. By the effective lens aperture we mean the total intensity of the radiation at the focus that is equal to the intensity of the plane wave that passed through the lens without being absorbed. In this case, we disregard the parasitic absorption in the thin parts of the elementary lenses of thickness d (see Fig. 1).

Below, we restrict our analysis to the case where the lens length is

$$L \le L_0 = \frac{\pi}{2}L_c, \quad L_c = (LF)^{1/2} = \left(\frac{pR}{2\delta}\right)^{1/2}.$$

In this range of lengths, the lens focuses the incident wave in space at a distance $r_i > 0$. The intensity distribution of the radiation in the space behind the lens when it is illuminated by a plane wave can be analyzed to determine the focus sizes. The two transverse coordinates are again factorized, and it will suffice to consider the distribution along the x axis. Clearly, the wave amplitude can be obtained by calculating the convolution of propagator (17) with a coordinate-independent unit function, i.e., by integrating over the x_o coordinate. The integral reduces to the Fourier transform of the Gaussian function, and it can be calculated exactly. As a result, using relation (22) we obtain

$$\begin{aligned} A(x_i) &= \int dx_o G(x_i, x_o) \\ &= \left(\frac{\tilde{F}_c}{\tilde{F}_c - \tilde{r}_i}\right)^{1/2} \exp\left[-i\pi\frac{x_i^2}{\lambda(\tilde{F}_c - \tilde{r}_i)}\right] \\ &= \left(\frac{z_c}{z_c c_L - r_i s_L}\right)^{1/2} \exp\left[-i\pi\frac{x_i^2 s_L}{\lambda(z_c c_L - r_i s_L)}\right]. \end{aligned} \tag{23}$$

Note that the result does not depend on the distance r_o. From a physical point of view, it is clear that this expression can also be derived directly from the image propagator (17) by considering a point near the optical axis at an infinite distance r_o and dividing it by the amplitude of the Kirchhoff propagator at the same distance, because a point source at an infinite distance gives a plane wave in front of the lens aperture.

As follows from (23), a plane wave in front of the lens transforms into a Gaussian wave behind the lens at all distances from the lens. For an arbitrarily long PCR X-ray lens, all parameters are complex. Since the absorption parameter $\gamma \ll 1$, we use the linear (in γ) approximation for qualitative estimation. With the adopted constraint on the lens length, we obtain the relations

$$\begin{gathered} s_L \approx S_L - i\frac{\gamma}{2}u_L C_L, \quad c_L \approx C_L + i\frac{\gamma}{2}u_L S_L, \\ z_c = L_c\left(1 + i\frac{\gamma}{2}\right), \end{gathered} \tag{24}$$

where we introduced the real functions

$$S_L = \sin u_L, \quad C_L = \cos u_L, \quad u_L = L/L_c.$$

The intensity reaches a maximum at the distance r_i behind the lens that satisfies the condition

$$\mathrm{Re}(z_c c_L - r_i s_L) \approx L_c C_L - r_i S_L = 0,$$

i.e., $r_i = F_L C_L$. In this case,

$$|A(x_i)|^2 = \frac{1}{\gamma\alpha_L}\exp\left(-\frac{2\pi}{\gamma\lambda F_L\alpha_L}x_i^2\right),$$
$$\alpha_L = \frac{1}{2}\left(C_L + \frac{u_L}{S_L}\right), \quad F_L = \frac{L_c}{S_L}. \tag{25}$$

Using the derived expression, we obtain

$$A_\gamma = \int dx_i |A(x_i)|^2 = \left(\frac{\lambda F_L}{2\gamma\alpha_L}\right)^{1/2} \tag{26}$$

for the total intensity of the focus (effective aperture) and

$$s_\gamma = 0.664(\gamma\lambda F_L\alpha_L)^{1/2} = 0.47\frac{\lambda F_L}{A_\gamma} \tag{27}$$

for the full width at half maximum (halfwidth) of the intensity peak at the focus.

For a thin lens, $u_L \ll 1$, it follows from these relations that

$$r_i = F, \quad F = \frac{L_c^2}{L}, \quad A_\gamma = 0.707\left(\frac{\lambda F}{\gamma}\right)^{1/2},$$
$$s_\gamma = 0.664(\gamma\lambda F)^{1/2} = 0.47\frac{\lambda F}{A_\gamma}. \tag{28}$$

In the other extreme case, $u_L = \pi/2$, we obtain

$$r_i = 0, \quad A_\gamma = 0.798\left(\frac{\lambda L_c}{\gamma}\right)^{1/2},$$
$$s_\gamma = 0.47(\gamma\lambda L)^{1/2} = 0.47\frac{\lambda L_c}{A_\gamma}. \tag{29}$$

Thus, defining the effective aperture in terms of the total intensity of the focus does not lead to any contradictions. For a self-focusing lens, the generalized focal length is by a factor of 1.57 smaller than the length of the lens itself. Clearly, this is the minimum focal length that can be obtained for a lens with a specified radius of curvature R, thickness of the elementary lens p, and decrement of the refractive index δ. It follows from the derived relations that the linear size of the focus is approximately by a factor of 0.8γ smaller than the linear size of the effective aperture for all lens lengths, with the numerical coefficient depending only slightly on the lens length. Consequently, as the lens length increases, the sizes of the focus and the effective aperture decrease proportionally to each other and the degree of beam compression depends only on the absorption factor γ.

To conclude this section, we give several numerical values for the parameters of Lengeler aluminum lenses [9–12] at a photon energy of 25 keV. These lenses have the following parameters: $p = 1$ mm, $R = 0.2$ mm, $\delta = 8.643 \times 10^{-7}$, $\gamma = 2.02 \times 10^{-3}$, and $L_c = 34$ cm. For the lens of 100 elements that was actually used in experiments, $L = 10$ cm, $u_L = 0.294$, $F_L = 117.4$ cm, and $\alpha_L = 0.986$. In this case, the effective aperture is $A_\gamma = 120$ μm and the lens resolution is $s_\gamma = 0.23$ μm. These parameters smoothly decrease with the increasing number of elements. Thus, for a lens of 300 elements, it can be easily calculated that $L = 30$ cm, $u_L = 0.882$, $F_L = 44$ cm, and $\alpha_L = 0.889$. In this case, the effective aperture is $A_\gamma = 78$ μm and the lens resolution is $s_\gamma = 0.13$ μm. Although the aperture of a long lens decreases, it has a better resolution and can be useful in imaging small objects or their fragments.

5. THE IMAGE OF A POINT SOURCE

Let us consider the imaging properties of an X-ray PCR lens that follow from propagator (17). In the linear (in small parameter γ) approximation, the image of a point source displaced by x_o from the optical axis is focused at the distances that satisfy the condition

$$\mathrm{Re}(\tilde{r}_g) = (r_i + r_o)C_L + \left(L_c - \frac{r_i r_o}{L_c}\right)S_L$$
$$= r_{iL} + r_{oL} - \frac{r_{iL}r_{oL}}{F_L} = 0, \tag{30}$$

where $r_{oL} = r_o + B_L$, $r_{iL} = r_i + B_L$, $B_L = F_L(1 - C_L)$, and F_L is defined in (25). Using the generalized real distances, this condition can be written as the thin-lens formula

$$r_{iL}^{-1} + r_{oL}^{-1} = F_L^{-1}.$$

When condition (30) is satisfied, the parameter $\tilde{r}_g$ becomes purely imaginary, and for the parameters $\tilde{a}_i$ and $\tilde{a}_o$, it will suffice to use only the real part, i.e., to set $\gamma = 0$. In this approximation, the propagator takes the form

$$G(x_i, x_o) = -iM^{1/2}\delta_\sigma(x_i + x_o M), \tag{31}$$

where

$$\delta_\sigma(x) = \frac{1}{\sigma(2\pi)^{1/2}}\exp\left(-\frac{x^2}{2\sigma^2}\right),$$
$$\sigma = \left(\frac{\lambda M}{2\pi}\mathrm{Im}\tilde{r}_g\right)^{1/2}, \quad M = \frac{r_{iL}}{r_{oL}}. \tag{32}$$

Thus, the propagator is a Gaussian function with the maximum centered at $x_i = -x_o M$. In the general case of a long X-ray PCR lens, the magnification factor can differ markedly from the ordinary magnification factor of a thin lens proportional to the ratio r_i/r_o. Accordingly, the image intensity peak is described by the function

$$\begin{aligned} I_{ps}(x_i, x_o) &= |G(x_i, x_o)|^2 \\ &= \frac{M}{2\pi\sigma^2}\exp\left[-\frac{(x_i + x_o M)^2}{\sigma^2}\right] \end{aligned} \tag{33}$$

for which $s_\gamma = 1.665\sigma$. The expression for $\mathrm{Im}\tilde{r}_g$ under the image focusing conditions is simplest in terms of the ordinary distances:

$$\mathrm{Im}\tilde{r}_g = \gamma[L_c S_L + (r_i + r_o)\alpha_L].$$

In terms of the generalized distances, the peak width parameter is

$$\begin{aligned} \sigma &= \left(\gamma\frac{\lambda\alpha_L r_{iL}^2}{2\pi F_L}K_L\right)^{1/2} = \frac{\lambda r_{iL}}{A_\gamma}\left(\frac{K_L}{4\pi}\right)^{1/2}, \\ K_L &= 1 - \frac{2B_L F_L}{r_{iL} r_{oL}}\left(1 - \frac{1 + C_L}{2\alpha_L}\right), \end{aligned} \tag{34}$$

where A_γ is the lens aperture defined in (26).

Let us again consider the extreme cases. For a thin lens where $L \ll F$, the focusing condition is $r_i^{-1} + r_o^{-1} = F^{-1}$, the magnification factor is $M = r_i/r_o$, and $s_\gamma = 0.47\lambda r_i/A_\gamma$. For a self-focusing lens with a length $L = (\pi/2)L_c$, the focusing condition, the magnification, factor, and the halfwidth are given by the expressions

$$\begin{aligned} r_o &= \frac{L_c^2}{r_i}, \quad M = \frac{1 + h_i}{1 + h_i^{-1}}, \\ s_\gamma &= 0.47\frac{\lambda L_c}{A_\gamma}(1 + h_i)\left(1 - \frac{0.726}{2 + h_i + h_i^{-1}}\right), \\ h_i &= \frac{r_i}{L_c}. \end{aligned} \tag{35}$$

In the limit $r_o \longrightarrow \infty$, we again obtain (28). On the other hand, in the limit $r_i \longrightarrow \infty$, both the magnification factor and the image halfwidth indefinitely increase in the two cases, but the ratio of the size of the image of a point to the size of the entire pattern is virtually constant.

Note the following interesting feature of long lenses. As was shown above, a self-imaging lens has a length $L = \pi L_c = 3.14L_c$. On the other hand, a self-focusing lens is by a factor of 2 shorter, while the length of the experimental scheme for imaging without magnification is $(2 + \pi/2)L_c = 3.57L_c$, which is only slightly longer. For this reason, there is little point in using extremely long lenses.

6. THE FOCUSED IMAGE OF AN OBJECT

In an actual X-ray experiment, a relatively thin object illuminated by the wave emitted by a point source located at a distance r_s from the object and having transverse coordinates x_s and y_s is imaged with the lens. In general, the source has finite sizes, but different points of the source are incoherent. Therefore, the intensity should be integrated over the source's coordinates at the final stage of the calculation. In this section, we restrict our analysis to the case where the distance from the source to the object is large and the angular sizes of the source in the object's plane do not exceed the object's characteristic scattering angles. In other words, the coherent image conditions are satisfied. The wave field in the object's image plane referred to the amplitude of the wave incident on the object can be calculated by using the integral

$$\begin{aligned} &A^{(t)}(x_i, y_i) \\ &= i\lambda r_s\int dx_0 dy_0 G_s^{(t)}(x_i, y_i, x_o, y_o)T(x_o, y_o), \end{aligned} \tag{36}$$

where

$$\begin{aligned} &G_s^{(t)}(x_i, y_i, x_o, y_o) \\ &= \exp(-ik\eta s_0 L)G_s(x, x_o)G_s(y_i, y_o), \end{aligned} \tag{37}$$

the function

$$G_s(x_i, x_o) = G(x_i, x_o)P(x_o - x_s, r_s)$$

is the partial image propagator for the object, and the function

$$T(x_o, y_o) = \exp[i\psi(x_o, y_o)]$$

describes the transfer of radiation through the object, i.e., the object's function.

As we have shown above, under the image focusing conditions for the object's points (30), the modulus of $G_s^{(t)}(x_i, y_i, x_o, y_o)$, which is considered as a function of the x_o and y_o coordinates at a given point (x_i, y_i) on the image plane, has a sharp maximum at the point with (x_{oi}, y_{oi}) coordinates, where $x_{oi} = -x_i/M$ and $y_{oi} = -y_i/M$. Here, as in the preceding section, $M = r_{iL}/r_{oL}$. Let us assume that the complex phase of the object's function is a smooth function within the region of the propagator maximum. In the effective domain of integration, the object's function can then be approximated by the expression

$$\begin{aligned} &T(x_o, y_o) \\ &= \exp\left[i\psi_i + i\frac{2\pi}{\lambda}[\xi_{ix}(x_o - x_{oi}) + \xi_{iy}(y_o - y_{oi})]\right] \end{aligned} \tag{38}$$

with the complex parameters

$$\psi_i = \psi(x_{oi}, y_{oi}), \quad \xi_{ix} = \frac{\lambda}{2\pi}\frac{d\psi(x_{oi}, y_{oi})}{dx},$$

$$\xi_{iy} = \frac{\lambda}{2\pi}\frac{d\psi(x_{oi}, y_{oi})}{dy},$$

which depend on the coordinates of the image point.

In this case, the wave field of the image is again factorized,

$$\begin{aligned} A^{(t)}(x_i, y_i) &= \exp(-ik\eta s_0 L + i\psi_i) \\ &\times A^{(x)}(x_i, y_i)A^{(y)}(x_i, y_i), \end{aligned} \qquad (39)$$

and, for example,

$$\begin{aligned} A^{(x)}(x_i, y_i) &= (i\lambda r_s)^{1/2}\int dx_o G(x_i, x_o)P(x_o - x_s, r_s) \\ &\times \exp\left[i\frac{2\pi}{\lambda}\xi_{ix}(x_o - x_{oi})\right]. \end{aligned} \qquad (40)$$

The function $A^{(x)}(x_i, y_i)$ depends on y_i parametrically via ξ_{ix}. Approximation (31) cannot be directly used to calculate the integral, because the expansion in terms of powers of γ has already been made in it and, therefore, it is not accurate enough. The expansion in terms of powers of γ can be made only in the final expressions. On the other hand, an exact result can be easily obtained from the following considerations.

The integral differs only by the phase factor from the convolution of the image propagator with the Kirchhoff propagator if we substitute $\tilde{x}_s = x_s - r_s\xi_{ix}$ for the true coordinate of the source x_s in the latter. As was noted above, the convolution of the image propagator with the Kirchhoff propagator is again equal to the image propagator in which we should substitute $r_o + r_s$ for the distance r_o and $\tilde{x}_s$ for the coordinate x_o. In the expression derived, we should use the image focusing condition

$$r_{iL}^{-1} + r_{oL}^{-1} = F_L^{-1}$$

and we can set $\gamma = 0$ in the preexponential term. As regards the exponent, it will suffice to include the terms of the zero and first powers of γ. Although the algorithm is simple, the calculations are cumbersome because they contain many combinations of many parameters. To obtain an unequivocal result, it is convenient to choose the coordinate of the imaged point on the object, x_{oi}, and the angle of ray exit from the object at this point, θ_o, calculated via the phase gradient as independent transverse coordinates. A useful parameter is also the coordinate of ray entrance into the lens, x_0. These quantities are defined as

$$\begin{gathered} x_{oi} = -\frac{x_i}{M}, \quad \theta_o = \frac{x_{oi} - \tilde{x}_s}{r_s} = \xi_{ix} + \frac{x_{oi} - x_s}{r_s}, \\ x_0 = x_{oi} + r_o\theta_o. \end{gathered} \qquad (41)$$

It is convenient to choose r_o and L_c as independent longitudinal distances. The remaining parameters are expressed in terms of them as

$$\begin{gathered} r_i = Mu, \quad M = \frac{L_c}{v}, \quad u = r_oC_L + L_cS_L, \\ v = r_oS_L - L_cC_L, \quad F_L = \frac{L_c}{S_L}. \end{gathered} \qquad (42)$$

Without giving the intermediate calculations, we immediately write out the result in the linear (in γ) approximation

$$\begin{gathered} A^{(x)}(x_i, y_i) = \frac{1}{iM^{1/2}}\exp\left[i\Phi_x(x_i) - \frac{1}{2}\mu_x(x_i, y_i)\right], \\ \Phi_x(x_i) = \frac{\pi}{\lambda}\left[\frac{(x_{oi} - x_s)^2}{r_s} + \frac{x_{oi}^2}{r_{oL}} + \frac{x_i^2}{r_{iL}}\right], \\ \mu_x(x_i, y_i) = \gamma\frac{2\pi}{\lambda F_L} \\ \times[\alpha_L x_0^2 + (\alpha_L - C_L)(L_c\theta_o)^2 + S_L x_0 L_c\theta_o]. \end{gathered} \qquad (43)$$

When making the expansion in terms of γ, we assumed the conditions $\lambda r_s \gg 2\pi\sigma_o^2$, where $\sigma_o = \sigma/M$ is the width of the maximum of propagator (31) relative to the integration variable x_o, to be satisfied. In other words, we assumed that the phase of the incident spherical wave also smoothly varied in the domain of integration. We can easily consider the general case, but it is of no practical interest. As follows from the derived expression, the phase gradient of the object's function directly affects the image intensity. To qualitatively analyze this effect, let us consider the extreme case of an incident plane wave ($r_s \longrightarrow \infty$) and assume that the object is transparent. If the sample is homogeneous, then $\theta_o = 0$ at all points and the intensity of the radiation in the image plane is described by the Gaussian function

$$I_{im}(x_i) = \frac{1}{M}\exp\left(-\gamma\frac{2\pi}{\lambda F_L}\frac{\alpha_L}{M^2}x_i^2\right). \qquad (44)$$

If, however, a phase gradient exists at some points of the sample, then $\theta_o = \xi_{ix}$ at these points and the intensity will deviate from law (44); depending on different conditions, the intensity can be lower and higher than the background intensity. Thus, a unique opportunity to visualize transparent objects and to obtain the phase contrast through the nonuniform absorption of radiation in the lens emerges. Unlike the ordinary phase contrast observed on synchrotron radiation sources (see, e.g., [19–21]), this contrast has no parasitic oscillations associated with the interference of various rays and directly allows the local phase gradient produced by the object to be determined. Abrupt changes in intensity related to a sharp phase gradient have recently been observed experimentally in the image of a sample with a profiled surface [22].

The physical nature of the visualization of the phase gradient in a sample with an X-ray PCR lens is easiest to understand in terms of geometrical optics. The ray from a source with a coordinate x_s comes to the sample at a point with a coordinate x_{oi}. Since the lens formula holds, all of the rays that emerge from this point at different angles reach a point with a coordinate x_i in the image plane after their passage through the lens. Actually, however, only one path is realized and the ray leaves the sample, making an angle θ_o with the optical (z) axis. Accordingly, the ray in front of the lens has the coordinate x_0 and angle θ_o. The ray path $x = x_t(z)$ in the lens satisfies the condition under which at each point of the path its tangent makes an angle $\theta = (\lambda/2\pi)(d\varphi/dx)$ with the z axis, where $\varphi(x, z)$ is the phase of the wave field in the lens. This phase is equal to the phase of the image propagator (17) if we set $\gamma = 0$, $r_i = 0$, and $x_i = x_t(z)$ in the latter and substitute z for L. On the other hand, $\theta = dx_t/dz$. The equation for the path is particularly simple if we write it via the derivative of θ. As a result,

$$\frac{d\theta}{dz} = \frac{d^2x_t}{dz^2} = -\frac{x_t(z)}{L_c^2}, \quad x_t(z) = x_0C_z + \theta_oL_cS_z, \quad (45)$$

where x_0 and θ_o are the initial point and initial derivative on the path. The initial coordinate and initial angle of the ray as it enters the lens correspond to the parameters defined by (41). Calculating the absorption coefficient integrated over the ray path in the lens,

$$\mu(L) = \gamma\frac{2\pi}{\lambda L_c^2}\int_0^L dz\left(x_0\cos\frac{z}{L_c} + \theta_oL_c\sin\frac{z}{L_c}\right)^2, \quad (46)$$

we obtain a result that matches formula (43). Thus, the contrast is actually related to the change in the ray path and the nonuniform absorption in a PCR lens.

Naturally, the overall image of an object contains both the absorptive and phase contrast and depends on the two components of the total phase gradient along the two coordinate axes. Unfortunately, in solving the inverse problem, the change in intensity alone is not enough to restore the absorption coefficient and the two components of the phase gradient. Additional information can be obtained by moving the object relative to the lens, because different portions of the lens absorb differently.

7. THE FOURIER TRANSFORM OF AN OBJECT WITH A PARABOLIC CONTINUOUSLY REFRACTIVE X-RAY LENS

A thin parabolic lens is known from classical optics to perform the Fourier transform of the function of an object illuminated by a spherical wave from a point source in the focusing plane of the point source. Let us consider this phenomenon for an X-ray PCR lens. We write the Fourier transform of the object's function as

$$T(x_o, y_o) = \int\frac{dqdp}{(2\pi)^2}\tilde{T}(q, p)\exp(iqx_o + ipy_o). \quad (47)$$

In this case, the amplitude of the wave field in the image plane is

$$\begin{aligned} A^{(t)}(x_i, y_i) &= \exp(-ik\eta s_0L) \\ &\times\int\frac{dqdp}{(2\pi)^2}\tilde{T}(q, p)G_f(q, x_s, x_i)G_f(p, y_s, y_i), \end{aligned} \quad (48)$$

where

$$\begin{aligned} &G_f(q, x_s, x_i) \\ &= \int dx_oG(x_i, x_o)P(x_o - x_s, r_s)\exp(iqx_o) \end{aligned} \quad (49)$$

is the partial image propagator for the separate component of the Fourier transform of the object's function. We are interested in the conditions when this propagator is closest to the delta function.

As was noted in the preceding section, the integral can be expressed in terms of the image propagator in which $\tilde{x}_s = x_s - x_q$, $x_q = q\lambda r_s/2\pi$ is substituted for x_s and $\tilde{r}_s = r_o + r_s$ is substituted for r_o. As a result,

$$\begin{aligned} G_f(q, x_s, x_i) &= \frac{1}{(i\lambda\hat{r}_g)^{1/2}}\exp\left[iq\left(x_s - \frac{1}{2}x_q\right)\right] \\ &\times\exp\left[i\frac{\pi}{\lambda\hat{r}_g}(g_ix_i^2 - 2x_i\tilde{x}_s + g_s\tilde{x}_s^2)\right], \end{aligned} \quad (50)$$

where

$$\begin{aligned} \hat{r}_g &= (\tilde{r}_s + r_i)c_L + \left(z_c - \frac{\tilde{r}_sr_i}{z_c}\right)s_L, \\ g_i &= c_L - \frac{\tilde{r}_s}{z_c}s_L, \quad g_s = c_L - \frac{r_i}{z_c}s_L. \end{aligned} \quad (51)$$

Thus, the condition for the Fourier transform is identical to the focusing condition for a point source illuminating the object, i.e., $\mathrm{Re}\,\hat{r}_g = 0$. However, approximate expression (31) cannot be used directly, because the propagator is in the integrand and, hence, the omitted phase factor can play a significant role.

To derive an approximate expression, we represent the complex coefficients as a series in powers of $i\gamma$:

$$\begin{gathered}\hat{r}_g = R_0 + i\gamma R_1 + (i\gamma)^2 R_2, \quad g_i = Q_{i0} + i\gamma Q_{i1}, \\ g_s = Q_{s0} + i\gamma Q_{s1}.\end{gathered} \tag{52}$$

The focusing condition for a point source then has the approximate form $R_0 = 0$. In this case, the following relations hold:

$$r_{iL}^{-1} + r_{sL}^{-1} = F_L^{-1}, \quad Q_{i0} = -\frac{1}{M_s}, \quad Q_{s0} = M_s,$$

where

$$r_{sL} = \tilde{r}_s + B_L = r_{oL} + r_s, \quad M_s = \frac{r_{iL}}{r_{sL}},$$

and the propagator is defined as

$$G_f(q, x_s, x_i) = \frac{1}{i(\gamma\lambda R_1)^{1/2}}\exp[i\varphi(q) - \mu(q)], \tag{53}$$

where

$$\begin{gathered}\mu(q) = \frac{\pi(x_i + M_s\tilde{x}_s)^2}{\gamma\lambda R_1 M_s}, \\ \varphi(q) = q\left(x_s - \frac{1}{2}x_q\right) + \mu(q)\gamma\frac{R_2}{R_1} + \frac{Q_{i1}x_i^2 + Q_{s1}\tilde{x}_s^2}{\lambda R_1}.\end{gathered} \tag{54}$$

Thus, each point x_i in the source's image plane can have an appreciable intensity. It maps a region in the q space of the sample's function centered at point $q = q_i$, where

$$q_i = \frac{2\pi}{\lambda r_s M_s}(x_i - x_s M_s). \tag{55}$$

In addition, because of the absorption in the lens, the Fourier transform of the sample's function is modified by the phase factor; i.e., the convolution of the Fourier transform of the sample's function with some function that depends on the parameters of the lens and the experimental scheme is actually imaged. Nevertheless, if the sample's function is periodic and, hence, has a discrete series of Fourier harmonics whose separation exceeds the width of the propagator maximum, then the image is a system of spots that correspond to individual harmonics of the Fourier transform of the source's function and the size of each spot closely corresponds to the projected size of the source. The separation between the spots depends on the distances used in the experiment and is equal to

$$\Delta x_i = \Delta q\frac{\lambda r_s M_s}{2\pi}.$$

A homogeneous sample is a special case for which only the zero Fourier harmonic exists.

An alternative approach to this problem consists in analyzing expression (36). The Fourier transform is obtained if the propagator $G(x_i, x_o)P(x_o - x_s, r_s)$ does not contain the term proportional to x_o^2 in the phase. This condition can be written as $\mathrm{Re}\,C = 0$, where

$$C = \frac{1 - \tilde{a}_o}{\tilde{r}_g} + \frac{1}{r_s} = 0.$$

It is easy to verify that in the zeroth (in γ) approximation, this condition is equivalent to the focusing condition for a point source written above. In this case, however, a damped exponential of the type $\exp(-\mathrm{Im}(C)x_o^2)$ remains under the integral. In addition, since $\tilde{r}_g$ is a complex quantity, the wave vector of the Fourier transform has a small imaginary part, which depends on the separation between the object and the lens. The manifestation of these features depends on specific conditions and analysis of them is a problem in itself.

REFERENCES

1. A. Snigirev, V. Kohn, I. Snigireva, and B. Lengeler, Nature **384**, 49 (1996).
2. A. Snigirev, B. Filseth, P. Elleaume, *et al.*, Proc. SPIE **3151**, 164 (1997).
3. P. Elleaume, J. Synchrotron Radiat. **5**, 1 (1998).
4. A. Snigirev, V. Kohn, I. Snigireva, *et al.*, Appl. Opt. **37**, 653 (1998).
5. Yu. I. Dudchik and N. N. Kolchevskii, Nucl. Instrum. Methods Phys. Res. A **421**, 361 (1999).
6. B. Cederstrom, R. Cahn, M. Danielsson, *et al.*, Nature **404**, 951 (2000).
7. D. A. Arms, E. M. Dufresne, R. Clarke, *et al.*, Rev. Sci. Instrum. **73**, 1492 (2002).
8. V. V. Aristov, M. V. Grigoriev, S. M. Kuznetsov, *et al.*, Opt. Commun. **177**, 33 (2000).
9. B. Lengeler, C. Schroer, J. Tummler, *et al.*, J. Synchrotron Radiat. **6**, 1153 (1999).
10. C. G. Schroer, T. F. Gunzler, B. Benner, *et al.*, Nucl. Instrum. Methods Phys. Res. A **467–468**, 966 (2001).

11. C. G. Schroer, B. Benner, T. F. Gunzler, *et al.*, Rev. Sci. Instrum. **73**, 1640 (2002).
12. B. Lengeler, C. G. Schroer, B. Benner, *et al.*, J. Synchrotron Radiat. **9**, 119 (2002).
13. M. A. Piestrup, J. T. Cremer, H. R. Beguiristain, *et al.*, Rev. Sci. Instrum. **71**, 4375 (2000).
14. Y. Ohishi, A. Q. R. Baron, M. Ishii, *et al.*, Nucl. Instrum. Methods Phys. Res. A **467–468**, 962 (2001).
15. V. V. Protopopov and K. A. Valiev, Opt. Commun. **51**, 297 (1998).
16. V. G. Kohn, Pis'ma Zh. Éksp. Teor. Fiz. **76**, 701 (2002) [JETP Lett. **76**, 600 (2002)].
17. V. Kohn, I. Snigireva, and A. Snigirev, Phys. Rev. Lett. **85**, 2745 (2000); Opt. Commun. **198**, 293 (2001).
18. V. G. Kohn, I. Snigireva, and A. Snigirev, Poverkhnost, No. 1, 33 (2003).
19. A. Snigirev, I. Snigireva, V. Kohn, *et al.*, Rev. Sci. Instrum. **66**, 5486 (1995).
20. S. W. Wilkins, T. E. Gureyev, D. Gao, *et al.*, Nature **384**, 335 (1996).
21. T. E. Gureyev, S. Mayo, S. W. Wilkins, *et al.*, Phys. Rev. Lett. **86**, 5827 (2001).
22. C. G. Schroer, B. Benner, T. F. Gunzler, *et al.*, Rev. Sci. Instrum. **73**, 1640 (2002).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 24–33.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 28–38.

NUCLEI, PARTICLES, AND THEIR INTERACTION

Measurement of the Relative Probability of the Decay $\phi \longrightarrow \eta'\gamma$ in the Channel $\eta' \longrightarrow \eta\pi^0\pi^0$, $\eta \longrightarrow \gamma\gamma$

V. M. Aulchenko[a], M. N. Achasov[b], S. E. Baru[b], K. I. Beloborodov[b], A. V. Berdyugin[a], A. G. Bogdanchikov[b], A. V. Bozhenok[a], A. D. Bukin[a],*, D. A. Bukin[b], A. V. Vasilev[b], V. B. Golubev[b], T. V. Dimova[b], V. P. Druzhinin[b], V. N. Ivanchenko[b], I. A. Koop[b], A. A. Korol[b], S. V. Koshuba[b], A. P. Lysenko[b], A. V. Otboev[b], E. V. Pakhtusova[b], A. A. Salnikov[b], S. I. Serednyakov[b], V. A. Sidorov[b], Z. K. Silagadze[a], V. V. Shary[a], and Yu. M. Shatunov[b]

[a]*Novosibirsk State University, Novosibirsk, 630090 Russia*
**e-mail: bukin@inp.nsk.su*
[b]*Budker Institute of Nuclear Physics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 630090 Russia*

Received February 25, 2003

Abstract—The process $e^+e^- \longrightarrow \varphi \longrightarrow \eta'\gamma$, $\eta' \longrightarrow \eta\pi^0\pi^0$, $\eta \longrightarrow \gamma\gamma$ is investigated by an SND detector in experiment on a VEPP-2M equipment with colliding electron–positron beams. The analysis of experimental data yields a value of $B(\phi \longrightarrow \eta'\gamma) = (6.7^{+5.0}_{-4.2} \pm 1.5) \times 10^{-5}$ for the decay probability. Taking into account a previous measurement of this quantity by an SND detector in the decay channel $\eta' \longrightarrow \pi^+\pi^-\eta$, one finally obtains $B(\phi \longrightarrow \eta'\gamma) = (6.7^{+2.8}_{-2.4} \pm 0.8) \times 10^{-5}$. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The decay $\phi \longrightarrow \eta'\gamma$ is rare and can hardly be distinguished against the background; therefore, for a long time, the characteristics of this process could not be measured experimentally. The first measurement of the probability B of the decay $\phi \longrightarrow \eta'\gamma$ was carried out experimentally by a CMD-2 detector [1] in 1997:

$$B(\phi \longrightarrow \eta'\gamma) = (12^{+7}_{-5} \pm 2) \times 10^{-5}.$$

In recent years, results for this decay have been obtained simultaneously in several experiments on colliding e^+e^- beams. This fact has aroused increasing interest of theoreticians in a η' meson [2–6] because the structure of η and η' mesons directly determines the intensity of radiative decays of vector mesons. In theoretical studies, the main focus has been on the η–η' mixing and a gluonium admixture in η'.

At present, the world average value of the probability of the decay $\phi \longrightarrow \eta'\gamma$ [7],

$$B(\phi \longrightarrow \eta'\gamma) = (6.7^{+1.5}_{-1.4}) \times 10^{-5},$$

is based on the results of three experiments represented in Table 1.

The most accurate (to date) value, measured by a KLOE detector [11] in the channel $\eta' \longrightarrow \pi^+\pi^-\eta$, $\eta \longrightarrow \gamma\gamma$, was obtained after the publication of the survey [7] and is given by

$$B(\phi \longrightarrow \eta'\gamma) = (6.10 \pm 0.61 \pm 0.43) \times 10^{-5}.$$

The present paper is devoted to the measurement of the probability of the decay $\phi \longrightarrow \eta'\gamma$ in the purely neutral channel $\eta' \longrightarrow \pi^0\pi^0\eta$, $\eta \longrightarrow \gamma\gamma$.

2. EXPERIMENT

The experiment was carried out in 1996 [12] and 1998 [13] using a VEPP-2M electron–positron collider [14] by an SND detector [15]. The photon

Table 1

$B(\phi \longrightarrow \eta'\gamma) \times 10^5$	Decay channel	Experiment
$6.7^{+3.4}_{-2.9} \pm 1.0$	$\eta' \longrightarrow \eta\pi^+\pi^-$, $\eta \longrightarrow \gamma\gamma$	SND, 1999 [8]
$8.2^{+2.1}_{-1.9} \pm 1.1$	$\eta' \longrightarrow \eta\pi^+\pi^-$, $\eta \longrightarrow \gamma\gamma$	CMD-2, 2000 [9]
$4.9^{+2.2}_{-1.8} \pm 0.6$	$\eta' \longrightarrow \eta\pi^+\pi^-$, $\eta \longrightarrow \eta\pi^+\pi^-\pi^0$	CMD-2, 2000 [10]

energy resolution in the calorimeter of the detector was equal to

$$\sigma_E/E = 4.2\%/\sqrt[4]{E(\mathrm{GeV})},$$

and the angular resolution versus the photon energy is given by

$$\sigma_\varphi = \sqrt{(0.82°)^2/E(\mathrm{GeV}) + (0.63°)^2}.$$

We carried out seven scans of the energy domain for a ϕ meson in 1996 and two scans in 1998 in the energy range of 980 meV < $2E$ < 1060 MeV. The integrated luminosity was equal to 13 pb^{-1}, and the total number of ϕ mesons created was 2.1×10^7. The luminosity was measured by elastic scattering of electrons and positrons at large angles and by two-quantum annihilation. The difference between the results obtained by the two methods was no greater than 1%. In this paper, we used the luminosity measured by two-quantum annihilation because this process is the closest to the process being investigated by trigger properties.

3. SELECTION OF EVENTS

The events of the process considered,

$$\phi \longrightarrow \eta'\gamma, \quad \eta' \longrightarrow \eta\pi^0\pi^0, \quad \eta \longrightarrow \gamma\gamma,$$

have seven photons in the final state. To eliminate possible loss in efficiency due to the superposition of excess photons (arising from the beam background, the photon splitting in the calorimeter, the photon radiation by initial particles at a large angle, and the superposition of particles from other events), we selected events with seven or more photons.

The main background process

$$e^+e^- \longrightarrow \phi \longrightarrow \eta\gamma, \quad \eta \longrightarrow 3\pi^0$$

also has exactly seven photons in the final state, and its intensity is 200 times stronger than that of the process considered, which presents the main difficulty in this analysis.

When processing the data, we used the following parameters of events: n_γ, the total number of detected photons; n_{ch}, the total number of charged particles; χ_{2E}, the sum of squares of deviations of experimental angles and photon energies from theoretical ones, normalized by the angular and energy resolution of the calorimeter, which is minimized over all theoretical parameters of particles under the condition that the latter satisfy the energy–momentum conservation law; R_{Emax}, the energy of the hardest photon in the event, normalized by the beam energy; E_{p7}, the energy of the softest of the seven photons (in MeV); and χ_γ, the "quality" of photons, a parameter characterizing the "correctness" of the transverse energy distribution of photons in the calorimeter [16].

We selected events with $n_{ch} = 0$, $n_\gamma \geq 7$, and $\chi_{2E} < 40$. For these events, we carried out kinematic fitting by using the measured values of angles and photon energies: by numerically minimizing the logarithmic likelihood function, we chose the particle parameters that were the closest to the measured values and satisfied the energy–momentum conservation laws and the constraints imposed by the assumption that there exist intermediate metastable particles. The fitting was carried out independently under four different conjectures:

1. The event belongs to the process

$$\phi \longrightarrow X\gamma, \quad X \longrightarrow \eta\pi^0\pi^0, \quad \eta \longrightarrow \gamma\gamma, \quad \pi^0 \longrightarrow \gamma\gamma.$$

A system of seven photons is described by a set of 21 parameters (for example, energy and two angles for each particle). The energy–momentum conservation laws define four relations between these parameters. In addition, the masses of three intermediate unstable particles (η, π^0, π^0) enter another three equations that relate the original parameters. Each kinematic constraint removes one degree of freedom; therefore, 21 – 4 – 3 = 14 free parameters remain in this model to minimize the logarithmic likelihood function (one of these parameters is the mass of the system ($\eta\pi\pi$)). As a result of the reconstruction procedure, we calculate corrected particle parameters; the obtained minimal value of the logarithmic likelihood function $\chi^2_{\eta\pi\pi\gamma}$ is used for the selection of events. In a further selection of events, we also used the parameters $m_{\eta\pi\pi}$ and θ_{cm} obtained as byproducts of the kinematic fitting. The latter parameter is the polar angle of photons with respect to the direction of the η meson in the decay $\eta \longrightarrow \gamma\gamma$ in the rest frame of the η meson.

In spite of the measures undertaken to reject, by a preliminary screening, the largest possible part of the final photon pairs not associated with the π^0 and η meson decays, the reconstruction of an event takes considerable computing time in all models. Therefore, taking into account that the reconstruction model described above is the basic model, we left events with $\chi^2_{\eta\pi\pi\gamma} < 20$ for further analysis (and for reconstruction in other models).

2. The event belongs to the process

$$\eta\gamma \longrightarrow 3\pi^0\gamma \longrightarrow 7\gamma.$$

The total number of free parameters is 21 – 4 – 4 = 13 (four intermediate unstable particles with known masses). As a result of reconstruction, we obtain the minimal value of the logarithmic likelihood function $\chi^2_{\eta\gamma}$.

Table 2. Variation of the number of simulated and experimental events for the main processes under the successive imposition of conditions (1)

	Initial number of events	$n_\eta \geq 7$ $n_{ch} = 0$	$\chi^2_{\eta\pi\pi\gamma} < 20$ $\chi_{2\mathrm{E}} < 40$	$\theta_{cm} > 40°$ $R_{\mathrm{E}max} < 0.7$ $\chi_\gamma < 0$ $\chi_{2\mathrm{E}} < 12$	$E_{p7} > 30$ MeV $\lvert m_{3\pi} - 540\rvert > 40$ $\Delta\chi_1 < -10$ $\Delta\chi_2 < -10$	$\chi^2_{\eta\pi\pi\gamma} < 10$
$\phi \longrightarrow \eta\gamma$	1223844	98350	59644	5557	58	44
$\phi \longrightarrow K_S K_L$	2294939	1861	246	14	0	0
$\phi \longrightarrow K^+K^-$	1371963	35	0	0	0	0
$\phi \longrightarrow \pi^+\pi^-\pi^0$	2185331	1	0	0	0	0
$e^+e^- \longrightarrow \omega\pi^0$	142795	1	0	0	0	0
$\phi \longrightarrow \eta'\gamma$	415000	9398	8568	3411	2167	2030
Experiment	$19.3 \times 10^6\phi$ 12 pb^{-1}	165940	14016	1171	29	23

3. The event belongs to the process

$$\phi \longrightarrow K_{\mathrm{S}}K_{\mathrm{L}}, \quad K_{\mathrm{S}} \longrightarrow \pi^0\pi^0.$$

We consider all possible combinations of five photons and choose a variant that is best described by a given model. Four photons in the calorimeter must arise from the decay of a K_{S} meson; it is required that a K_{L} meson should produce a calorimeter hit with weaker angular and energy binding (the angular and energy distributions of such pseudophotons are obtained on the events $K_{\mathrm{S}}K_{\mathrm{L}}$ that were specially selected from the experimental data). The total number of free parameters in this model is $5 \times 3 - 4 - 3 = 8$. As a result of reconstruction, we obtain the minimal value of the logarithmic likelihood function $\chi^2_{K_{\mathrm{S}}K_{\mathrm{L}}}$.

4. In this model, we assume that, in the reaction

$$e^+e^- \longrightarrow Y + \gamma,$$

either the primary photon has not been detected at all or the measured characteristics of this photon are far from true values. The reconstruction in this model answers the questions of what is the probability that the event involves a particle that decays into three neutral pions and what is the mass of this particle. The kinematics corresponds to the decay

$$Y \longrightarrow 3\pi^0 \longrightarrow 6\gamma.$$

The momentum of the particle Y is not fixed. For a system of seven photons (one photon is assumed to be undetected, but its characteristics are calculated), the total number of free parameters is $7 \times 3 - 4 - 3 = 14$. As a result of kinematic fitting, we obtain a parameter $m_{3\pi}$, the mass of the system $(\pi^0\pi^0\pi^0)$. This reconstruction model is designed for the events $\eta\gamma$ in which the whole event is poorly reconstructed for some reasons.

In all kinematic reconstruction procedures of specific processes, when the number of photons is greater than the number specified in a given model, we chose a combination of photons that minimized the relevant logarithmic likelihood function; other photons were neglected.

To simplify the numerical minimization in a multidimensional space and to facilitate the reconstruction, we analytically determined the minimum in three angular variables corresponding to the rotation of the event as a whole [17]; this reduced the number of variable parameters of the logarithmic likelihood function by three.

Let us present the full list of selection conditions:

$$\begin{gathered}
\eta_\gamma \geq 7, \quad n_{ch} = 0, \\
\chi_{2\mathrm{E}} < 12, \quad \theta_{cm} > 40°, \\
R_{\mathrm{Emax}} < 0.7, \quad E_{p7} > 30 \text{ MeV}, \quad \chi_\gamma < 0, \\
|m_{3\pi} - 540| > 40 \text{ MeV}, \quad \chi^2_{\eta\pi\pi\gamma} < 10, \\
\Delta\chi_1 = \chi^2_{\eta\pi\pi\gamma} - \chi^2_{\eta\gamma} < -10, \\
\Delta\chi_2 = \chi^2_{\eta\pi\pi\gamma} - \chi^2_{K_{\mathrm{S}}K_{\mathrm{L}}} < -10.
\end{gathered} \tag{1}$$

These selection conditions, although they substantially suppress the background processes, preserve a sufficiently high efficiency ϵ of the investigated process: $\epsilon_{\eta'\gamma \to 7\gamma} = 7\%$ (the geometrical component in this efficiency can be estimated as a fraction of a solid

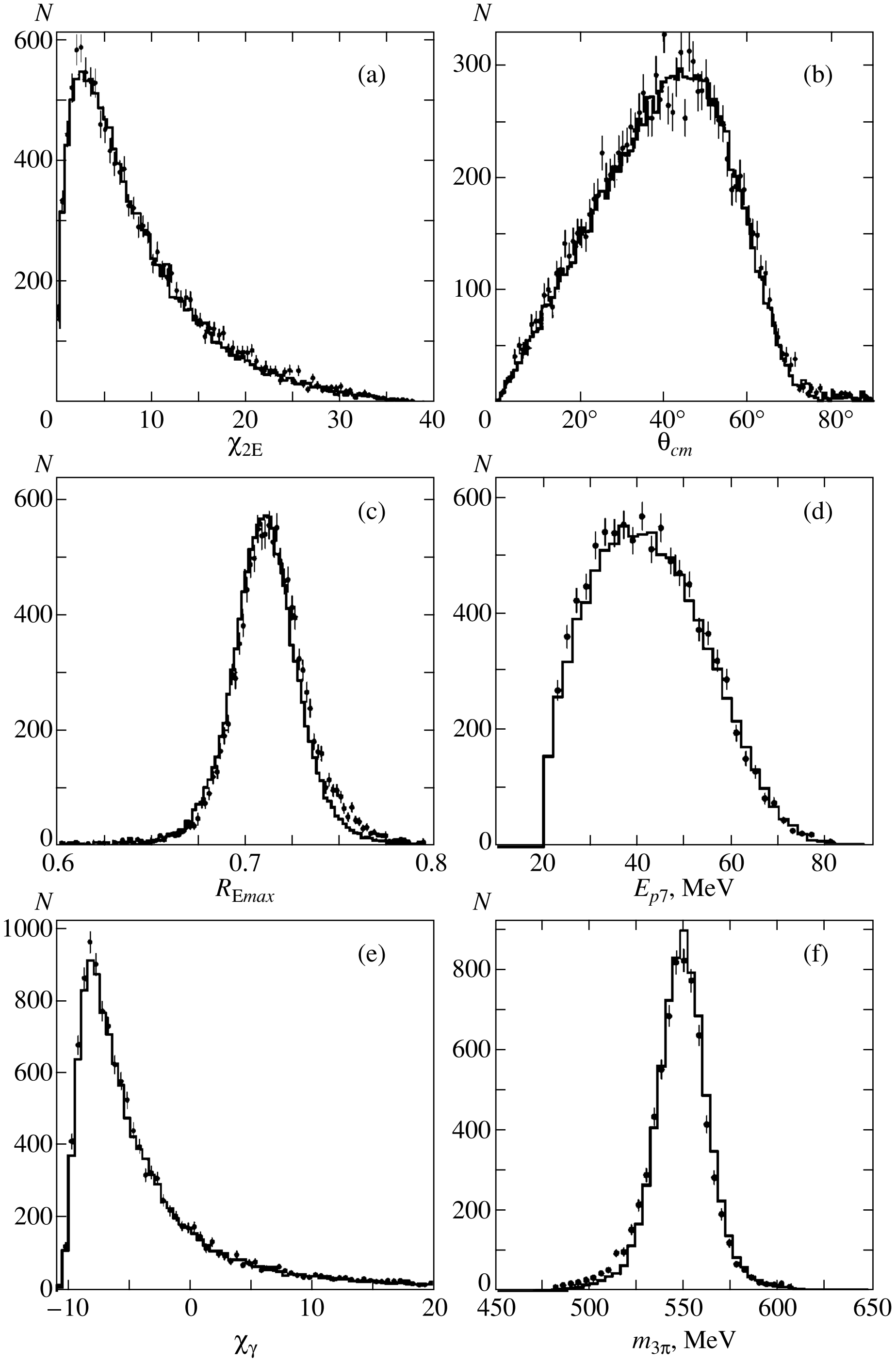


Fig. 1. Distribution of the simulated (histogram) and experimental (dots with error bars) events $\eta\gamma$, selected by conditions (2), in the following parameters: (a) $\chi_{2\mathrm{E}}$, (b) θ_{cm}, (c) $R_{\mathrm{E}max}$, (d) E_{p7}, (e) χ_γ, and (f) $m_{3\pi}$. The distributions are normalized by the integrated luminosity.

angle, covered by the detector, raised to a power of 7, $0.9^7 \approx 50\%$).

The variation in the number of selected simulated and experimental events for various processes under a successive imposition of selection conditions is shown in Table 2.

The ratio of the initial number of simulated events to the number of such events in experiment is different for different processes. For $\phi \longrightarrow \eta'\gamma$, the number of simulated events is five times greater than the number of experimental events because this is the main background process. For the main decay modes of a ϕ meson, in spite of the fact that the initial number of events is very large, the number of events K_SK_L, K^+K^-, and $\pi^+\pi^-\pi^0$ in experiment is 3, 7, and 1.4 times greater than the respective number in simulation. The number of events $\omega\pi^0$ in simulation is 20 times greater than that in experiment.

The table shows that any simulated event for the processes $\phi \longrightarrow K^+K^-$, $\phi \longrightarrow \pi^+\pi^-\pi^0$, and $e^+e^- \longrightarrow \omega\pi^0$ does not even satisfy the preliminary selection conditions

$$\chi^2_{\eta\pi\pi\gamma} < 20, \quad \chi_{2E} < 40\,;$$

this fact indicates that the contribution of these processes to the final result is negligible. Twenty-three experimental events satisfy the full set of conditions (1).

4. NORMALIZATION BY THE PROCESS $\eta\gamma$ AND CHECKING THE CORRECTNESS OF SIMULATION THE MAIN PARAMETERS OF EVENTS

Since there are a large number of background events $\eta\gamma$ and the signatures of these events are close to the process under investigation, this background presents a serious difficulty; on the other hand, these factors provide a good opportunity for checking the agreement between simulation and experiment.

For the latter purpose, we select experimental events that satisfy the conditions under which the contribution of the events $\eta\gamma$ must be dominant:

$$n_\gamma = 7, \quad \chi^2_{\eta\pi\pi\gamma} < 20, \quad \chi_{2\mathrm{E}} < 40. \qquad (2)$$

The total number of experimental events satisfying these conditions is 12457; according to the estimates from simulation, this number includes 0.2% of events $\eta'\gamma$ and 1.8% of events K_SK_L. Such an admixture is very small and is not displayed in the figures presented below.

Figures 1 and 2 represent the distributions of experimental and simulated events $\eta\gamma$ for the main parameters from set (1) (the experimental data are normalized by the integrated luminosity). These distributions display differences between simulation and experiment. The most pronounced differences are displayed by the distribution in $\Delta\chi_2 = \chi^2_{\eta\pi\pi\gamma} - \chi^2_{K_SK_L}$ (Fig. 2d). The distributions for the events $\eta\gamma$ and $\eta'\gamma$ (Fig. 3) are fairly close, so that this selection parameter can be used in the normalization procedure.

The fact that many characteristics of the events $\eta\gamma$ and $\eta'\gamma$ are close to each other allows us to carry out the normalization procedure and introduce a total correction for several sources of systematic errors. Among these effects, we can consider the following: inaccuracies in trigger simulation and in the measurement of integrated luminosity, loss of events due to the superposition of the background charged tracks, a possible inaccurate consideration of the photon conversion before the track system, and inaccurate simulation of the parameters of the event χ_γ, $\chi_{2\mathrm{E}}$, $\Delta\chi_2$. Naturally, such parameters as θ_{cm}, $R_{\mathrm{E}max}$, E_{p7}, and $m_{3\pi}$, with respect to which the distributions of the events $\eta'\gamma$ and $\eta\gamma$ are completely different, must be eliminated from the normalization procedure. For the same reason, we should have also eliminated the parameter $\chi^2_{\eta\pi\pi\gamma}$; however, as was mentioned above, this parameter was used for the preliminary screening of events in order not to hamper data processing by spending computer time for the kinematic fitting of a fortiori background events. However, the selection bound for this parameter in the normalization procedure is set so high that it introduces negligible systematic error. One also cannot ignore the parameter $\Delta\chi_1$ because we use it for distinguishing the event $\eta\gamma$ from $\eta'\gamma$.

For the normalization, we leave only those conditions in set (1) with respect to which the distributions of the events $\eta\gamma$ and $\eta'\gamma$ are close to each other,

$$\chi_{2E} < 12, \quad \chi_\gamma < 0, \quad \Delta\chi_2 < -10, \qquad (3)$$

as well as, as was pointed out above, the condition

$$\chi^2_{\eta\pi\pi\gamma} < 20$$

to optimize the processing time.

Conditions (3) held for 7234 experimental events. Figure 4 shows the distribution of the selected experimental events and simulated events $\eta\gamma$ with respect to the parameter $\Delta\chi_1 = \chi^2_{\eta\pi\pi\gamma} - \chi^2_{\eta\gamma}$ normalized by the integrated luminosity. The arrow in the figure points to the bound $\Delta\chi_1 = -10$ for the selection of the events $\eta'\gamma$; below this bound, there are 73% of these events, while the events $\eta\gamma$ are predominantly (98.7%) above this bound. Using the number $N_{\eta\gamma}$ of experimental events that are above the selection bound $\Delta\chi_1 > -10$ or $\Delta\chi_1 > -5$ and the number $N_\phi = 19.36 \times 10^6$ of produced ϕ mesons corresponding to the integrated luminosity, we can obtain the relative probability of the decay

$$B(\phi \longrightarrow \eta\gamma) = \frac{N_{\eta\gamma}}{N_\phi\,\epsilon_{\eta\gamma}}$$

$$= \begin{cases} \dfrac{7064}{19.36 \times 10^6 \times 2.69 \times 10^{-2}} = 1.36 \pm 0.02\%, \\ \Delta\chi_1 > -10, \\ \dfrac{6838}{19.36 \times 10^6 \times 2.62 \times 10^{-2}} = 1.35 \pm 0.02\%, \\ \Delta\chi_1 > -5. \end{cases} \qquad (4)$$

For these two values of the selection bound with respect to $\Delta\chi_1$, the detection efficiency of the event $\eta\gamma$

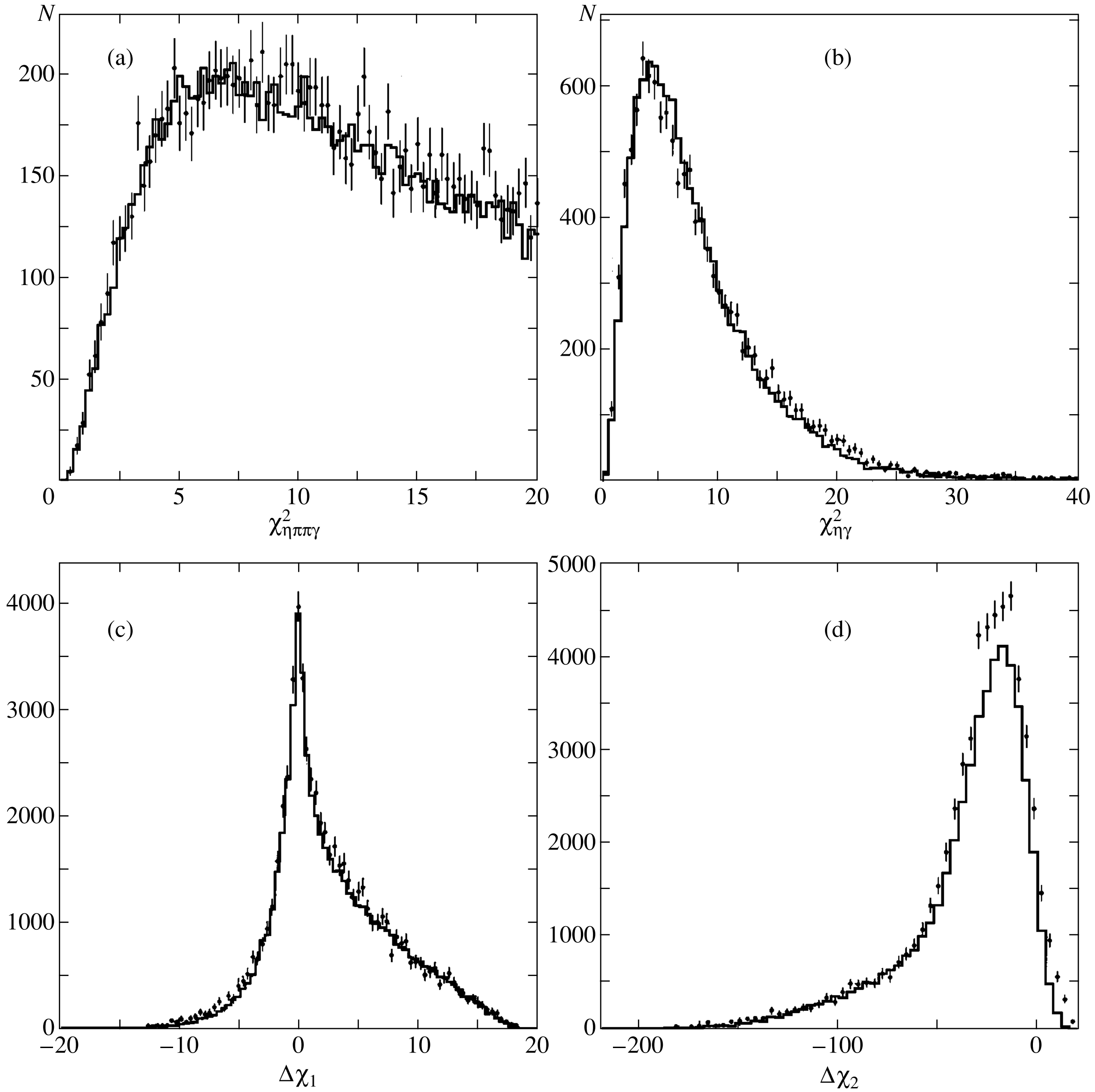


Fig. 2. Same as in Fig. 1 in the following parameters: (a) $\chi^2_{\eta\pi\pi\gamma}$, (b) $\chi^2_{\eta\gamma}$, (c) $\Delta\chi_1 = \chi^2_{\eta\pi\pi\gamma} - \chi^2_{\eta\gamma}$, (d) $\Delta\chi_2 = \chi^2_{\eta\pi\pi\gamma} - \chi^2_{K_SK_L}$.

varies by 3%, whereas the probability $B(\phi \longrightarrow \eta\gamma)$ varies by only 0.7%. This 0.7% difference can be regarded as a systematic calibration error associated with the inaccurate simulation of the parameter $\Delta\chi_1$. A correction to the results obtained by the SND detector from seven-photon events and selection conditions (3) can be found by comparing our result with the tabulated result from [7]:

$$B(\phi \longrightarrow \eta\gamma) = (1.299 \pm 0.026)\%.$$

Hence, the correction factor is

$$q = \frac{1.299 \pm 0.026}{1.36 \pm 0.02} = 0.955 \times (1 \pm 0.025). \quad (5)$$

In addition to the above error of 2.5%, one should include 0.7% (accuracy of simulation the parameter $\Delta\chi_1$) and 0.9% (the accuracy of the tabulated value of $B(\eta \longrightarrow 3\pi^0)$) to the systematic error of normalization. As a result, we obtain

$$2.5\% \oplus 0.7\% \oplus 0.9\% = 2.7\%$$

(the sign "⊕" denotes a quadratic summation of errors).

5. DISTRIBUTION OF THE EVENTS $\eta'\gamma$ IN THE MASS $m_{\eta\pi\pi}$ AND THE DETECTION EFFICIENCY

Until now, we have not used one of the basic reconstruction parameters of the events $\eta'\gamma$, the mass of the

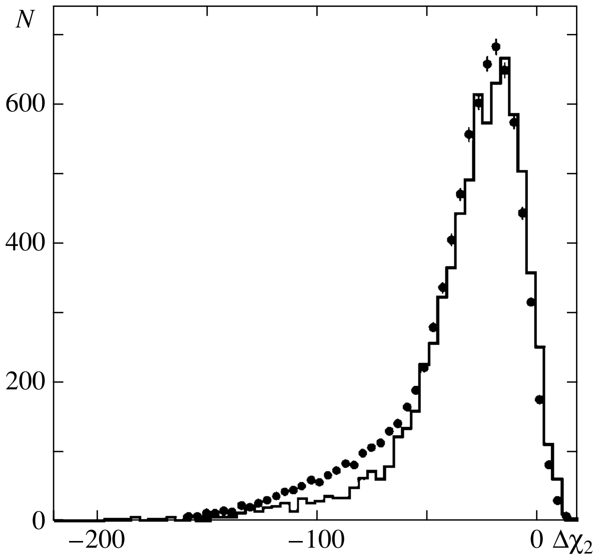


Fig. 3. Distribution of the simulated events $\eta'\gamma$ (histogram) and $\eta\gamma$ (dots with error bars), selected by conditions (2), in the parameter $\Delta\chi_2 = \chi^2_{\eta\pi\pi\gamma} - \chi^2_{K_SK_L}$. The distribution is normalized by the histogram area.

system ($\eta\pi\pi$). The distribution of the simulated events $\eta'\gamma$ in this parameter is shown in Fig. 5. This figure also represents an approximating function depending on the following parameters: the position of the peak, the full width at half maximum (FWHM) of the peak, the asymmetry parameter, and the area below the curve. In addition, about one-third of the events do not fall within the peak domain but form a smooth base approximated by a cubic spline. The emergence of such a base is associated both with the long tails in the energy-deposit distributions of photons in the calorimeter and with incorrect reconstruction of the decay chain

$$\eta' \longrightarrow \eta\pi^0\pi^0, \quad \eta \longrightarrow \gamma\gamma, \quad \pi^0 \longrightarrow \gamma\gamma.$$

When approximating the experimental distribution in the mass $m_{\eta\pi\pi}$, we used the distribution shape for the simulated event $\eta'\gamma$, and the only free parameter was the number of events $\eta'\gamma$, which is equal to the area below the distribution curve. Unfortunately, the distribution of the background events is far from a uniform, or even a linear, distribution, although it does not have such a narrow peak at $m_{\eta\pi\pi} = 960$ MeV.

The detection efficiency of the process $\phi \longrightarrow \eta'\gamma$ obtained by simulation is equal to

$$\epsilon = 5.4 \times 10^{-3} \times (1 \pm 0.020),$$

where the error is only a statistical one. In general, the correction obtained due to normalizing by the process $\eta\gamma$ must be ascribed to the detection efficiency (nearly all effects, except for the measurement error of the integrated luminosity, are associated with the incorrect determination of the detection efficiency by simula-

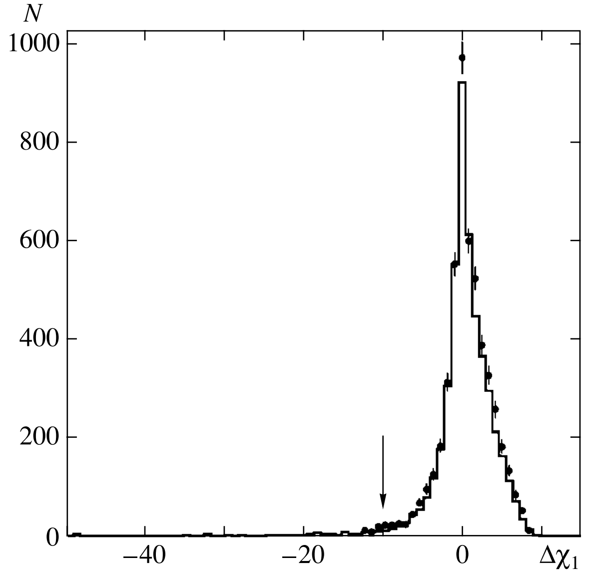


Fig. 4. Distribution of the experimental events $\eta\gamma$ (dots with error bars) and the simulated events $\eta\gamma$ (histogram), selected by conditions (3), in the parameter $\Delta\chi_1 = \chi^2_{\eta\pi\pi\gamma} - \chi^2_{\eta\gamma}$. The distribution is normalized by the integrated luminosity.

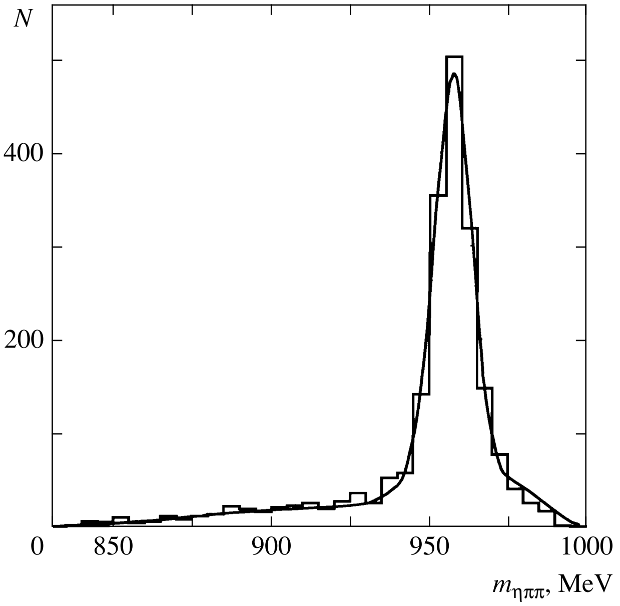


Fig. 5. Distribution of the simulated events $\eta'\gamma$, selected by conditions (1), in the parameter $m_{\eta\pi\pi}$.

tion). Taking into account this correction, we obtain the following value of the detection efficiency:

$$\begin{aligned}\epsilon &= 5.4\times10^{-3}\times(1\pm0.020)/0.955\\ &= 5.7\times10^{-3}\times(1\pm0.020).\end{aligned}\qquad(6)$$

This value includes the known probabilities in the decay chains $\eta' \longrightarrow \eta\pi\pi$ and $\eta \longrightarrow \gamma\gamma$. If we eliminate these probabilities, we obtain the following detection probability for the seven-photon state $\eta'\gamma \longrightarrow 7\gamma$:

$$\epsilon = \frac{0.57\%}{0.209\times0.393} = 6.9\%.$$

6. THE SHAPE OF THE DISTRIBUTION OF BACKGROUND EVENTS IN $m_{\eta\pi\pi}$

From the side of small masses, the distribution of background events in $m_{\eta\pi\pi}$ is bounded by a threshold determined by the sum of masses $m_\eta + 2m_\pi \approx 830$ MeV. From the other side, the distribution is bounded both by the total energy of positron and electron and by the selection conditions $E_{p7} > 30$ and $R_{Emax} < 0.7$. The shape of the background distribution was chosen on the basis of the distribution of the simulated events $\eta\gamma$, which constituted the main background (Fig. 6). To increase the statistical significance, the condition $\Delta\chi_1 < 0$ in the set (1) was loosened. The shape of the background distribution was approximated by cubic splines and was fixed; the verification of the statistical compatibility of this shape of the background distribution for different variants of experimental and simulated samples showed a satisfactory agreement with respect to the criterion χ^2 (in the worst cases with high statistics, the result was $\chi^2/n_D \lesssim 35/25$). Since one cannot completely rely on the simulated results for such large suppression coefficients of the background and there may be a certain background contribution from other processes (most probably, from K_SK_L), the shape of the background was varied when obtaining the final result; this fact has determined the main component of the systematic error.

7. MEASUREMENT OF THE RELATIVE PROBABILITY OF THE DECAY $\phi \longrightarrow \eta'\gamma$

Among 23 experimental events selected by conditions (1) (Fig. 7), there is a large percentage of background events, although the suppression coefficient of the basic background processes is very large. If, when obtaining a result, we were to subtract the remaining background exclusively on the basis of the simulated results, then we would have an additional systematic error that is difficult to evaluate. Fixing the shape of the background distribution in $m_{\eta\pi\pi}$ and of the distribution of the events $\eta'\gamma$ on the basis of simulated data and leaving only the numbers of the events of the effect ($N_{\eta'\gamma}$) and of the background (N_{BG}) free, we obtain the following values for the number of events:

$$N_{\eta'\gamma} = 7.4^{+5.5}_{-4.6},\quad N_{BG} = 15.6^{+6.2}_{-5.4}.\qquad(7)$$

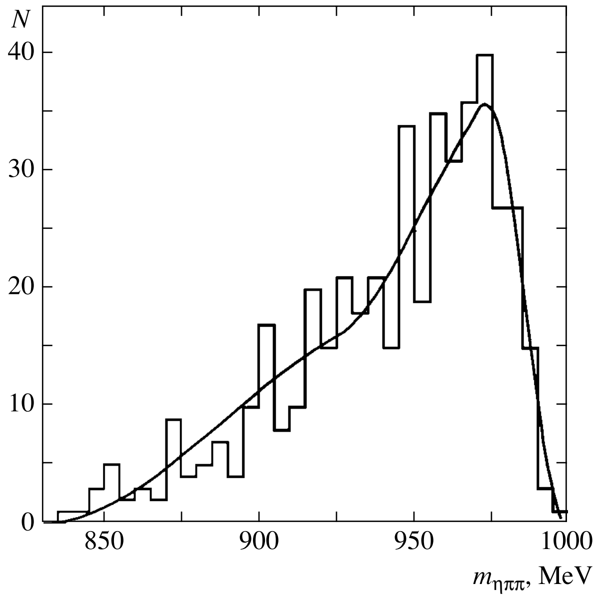


Fig. 6. Distribution, in the parameter $m_{\eta\pi\pi}$, of the simulated events $\eta\gamma$ selected by conditions (1) with the condition $\Delta\chi_1 < 0$ loosened to increase the statistical significance.

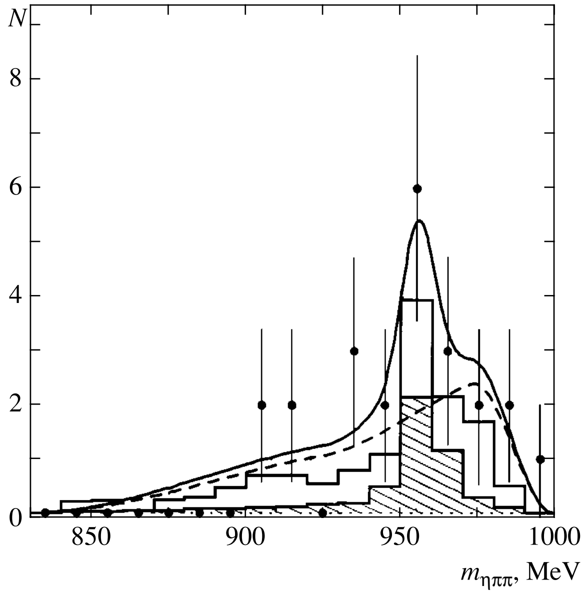


Fig. 7. Fitting of the experimental distribution of events in the parameter $m_{\eta\pi\pi}$ by a theoretical curve with the shape of the background distribution (hatched area) obtained by simulated data.

Table 3

Source of error	Systematic error, %
Error of normalizing by process $\eta\gamma$ (statistics of simulation $\eta\gamma$, error in tabulated values of $B(\phi \longrightarrow \eta\gamma)$ and $B(\eta \longrightarrow 3\pi^0)$, the simulation accuracy of the parameter $\Delta\chi_1$)	2.7
Simulation accuracy of $\chi^2_{\eta\pi\pi\gamma}$ and $\chi^2_{\eta\gamma}$	3.1
Simulation accuracy of the parameter θ_{cm}	3.1
Simulation accuracy of the parameter R_{Emax}	0.2
Simulation accuracy of the parameter $m_{3\pi}$	0.8
Simulation accuracy of the energy E_{p7} of the softest photon	3.7
Error in the tabulated value of $B(\eta' \longrightarrow \eta\pi^0\pi^0)$	5.7
Error in the tabulated value of $B(\eta \longrightarrow \gamma\gamma)$	0.6
Statistical error in the detection efficiency obtained by simulating $\eta'\gamma$	2.0
Uncertainty in the peak width of the distribution in the mass of the system $\eta\pi^0\pi^0$	10
Uncertainty in the shape of distribution of background events in the mass $m_{\eta\pi\pi}$	17
Total	22

Hence, for the relative probability of the decay, we have

$$B(\phi \longrightarrow \eta'\gamma) = \frac{N_{\eta'\gamma}}{N_\phi \epsilon_{\eta'\gamma}}$$
$$= \frac{7.4^{+5.5}_{-4.6}}{19.36 \times 10^6 \times 5.7 \times 10^{-3}} = (6.7^{+5.0}_{-4.2}) \times 10^{-5}. \quad (8)$$

Here, we have given only the statistical error.

8. SYSTEMATIC ERROR

A list of the main systematic errors is given in Table 3. The simulation error of the parameters given in the table was determined by the following scheme: for a given parameter, the correction to the selection bound for the simulated events $\eta\gamma$ was estimated by “pure” experimental events $\eta\gamma$ and simulated events $\eta\gamma$; this correction yields the same detection efficiency as the selection bound in the experimental distribution. The application of the corrected selection bound to the simulated events $\eta'\gamma$ yields a corrected detection efficiency of the events $\eta'\gamma$. The variation of this efficiency was taken as an estimate for the systematic error due to the inaccurate simulation of the distribution in a given parameter.

The error due to the inaccurate simulation of the peak width in the invariant mass $m_{\eta\pi\pi}$ was evaluated by a shift in the result when fitting an approximating curve with the FWHM increased by 15% to the experimental distribution. The error due to the simulation of the peak width was evaluated by an analogous distribution of the events $\eta\gamma$ in the parameter $m_{3\pi}$ (Fig. 3). The discrepancy between simulation and experiment was 12%.

The main contribution to the systematic error comes from the uncertainty in the shape of the background distribution. To check the stability of the result, we performed fitting with different variants of the shape of the background distribution in $m_{\eta\pi\pi}$. As an error estimate, we chose a shift of the result due to the fitting with a linear background.

9. DISCUSSION

Let us present the final result of measuring the probability of the decay $\phi \longrightarrow \eta'\gamma$ in the channel $\eta' \longrightarrow \eta\pi^0\pi^0$, $\eta \longrightarrow \gamma\gamma$:

$$B(\phi \longrightarrow \eta'\gamma) = (6.7^{+5.0}_{-4.2} \pm 1.5) \times 10^{-5}. \quad (9)$$

The accuracy of this quantity is not high as compared with the data obtained by detectors CMD-2 and KLOE in other decay channels of an η' meson. However, the experiment described represents an independent measurement of the decay probability in the purely neutral channel $\eta' \longrightarrow \eta\pi^0\pi^0 \longrightarrow 7\gamma$ and is also important for investigating systematic errors in multiphoton final states. In addition, averaging the probability (9) with the data obtained earlier [8] by an SND detector, we obtain the following result for the total probability determined with the use of the SND detector:

$$B(\phi \longrightarrow \eta'\gamma) = (6.7^{+2.8}_{-2.4} \pm 0.8) \times 10^{-5}, \quad (10)$$

which has an accuracy comparable with that of the data obtained by a CMD-2 detector.

The contributions from a common source in the systematic error are preserved unchanged (2% for the integrated luminosity and 0.6% for $B(\eta \longrightarrow \gamma\gamma)$); other contributions are averaged quadratically.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 00-02-17478.

REFERENCES

1. R. R. Akhmetshin, E. V. Anashkin, M. Arpagaus, *et al.*, Phys. Lett. B **415**, 445 (1997).
2. B. Bagchi, P. Bhattacharyya, S. Sen, and J. Chakrabarti, Phys. Rev. D **60**, 074002 (1999).
3. A. Bramon, R. Escribano, and M. D. Scadron, Phys. Lett. B **503**, 271 (2001).

4. N. Beisert and B. Borasoy, Eur. Phys. J. A **11**, 329 (2001).
5. Tsung-Wen Yeh, Phys. Rev. D **65**, 094019 (2002).
6. Emi Kou, Phys. Rev. D **63**, 054027 (2001).
7. *Review of Particle Physics*, Phys. Rev. D **66**, 1 (2002).
8. V. M. Aulchenko, M. N. Achasov, A. V. Bozhenok, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **69**, 87 (1999) [JETP Lett. **69**, 97 (1999)].
9. R. R. Akhmetshin, E. V. Anashkin, M. Arpagaus, *et al.*, Phys. Lett. B **473**, 337 (2000).
10. R. R. Akhmetshin, E. V. Anashkin, M. Arpagaus, *et al.*, Phys. Lett. B **494**, 26 (2000).
11. A. Aloisio, F. Ambrosino, A. Antonelli, *et al.*, Phys. Lett. B **541**, 45 (2002).
12. M. N. Achasov, M. G. Beck, K. I. Beloborodov, *et al.*, Preprint No. 97-78, INP SO RAN (Inst. of Nuclear Physics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 1997); M. N. Achasov, K. I. Beloborodov, A. V. Berdyugin, *et al.*, in *Proceedings of HADRON-97 Conference*, BNL, Brookhaven (1997), p. 26.
13. M. N. Achasov, V. M. Aulchenko, S. E. Baru, *et al.*, Preprint No. 98-65, INP SO RAN (Inst. of Nuclear Physics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 1998).
14. G. M. Tumaĭkin, in *Proceedings of 10th International Conference on High-Energy Particle Accelerations, Serpukhov* (1977), Vol. 1, p. 443.
15. V. M. Aulchenko, B. O. Baibusinov, T. V. Baier, *et al.*, in *Proceedings of Workshop on Physics and Detectors for DAPHNE, the Frascati Phi-factory, Frascati, Italy* (1991), p. 605; M. N. Achasov, V. M. Aulchenko, S. E. Baru, *et al.*, in *Proceedings of International Workshop on* e^+e^- *Collisions from* ϕ *to* J/Ψ*, Novosibirsk, Russia* (1999).
16. A. V. Bozhenok, V. V. Ivanchenko, and Z. K. Silagadze, Nucl. Instrum. Methods Phys. Res. A **379**, 507 (1996).
17. A. D. Bukin, Preprint No. 97-50, INP SO RAN (Inst. of Nuclear Physics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 1997); A. D. Bukin, Report at Computing in High Energy Physics, CHEP-97, Berlin (1997), http://www.ifh.de/CHEP97/paper/126.ps.

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 34–41.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 39–47.
Original English Text Copyright © 2003 by Amusia, Cherepkov, Chernysheva.

ATOMS, SPECTRA, RADIATION

Elastic Scattering of Slow Positrons on Atoms[¶]

M. Ya. Amusia[a,b], N. A. Cherepkov[c], and L. V. Chernysheva[b,*]
[a]*The Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem, 91904 Israel*
[b]*Ioffe Physicotechnical Institute, Russian Academy of Sciences,*
Politekhnicheskaya ul. 26, St. Petersburg, 194021 Russia
[c]*State University of Aerospace Instrumentation, St. Petersburg, 190000 Russia*
**e-mail: larissa.chernysheva@mail.ioffe.ru*

Submitted August 19, 2002

Abstract—The results of calculations of the elastic scattering cross section of positrons on noble gas and alkali atoms are presented. The calculations are performed within the one-electron Hartree–Fock approximation with multielectron correlations in the so-called random phase approximation with exchange taken into account. Virtual positronium formation is taken into account and proved to be very important. Arguments are presented that the positron polarization potential is repulsive for alkali atoms. The results obtained are in a reasonable agreement with experiment and with some previously reported calculations. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The studies of positron–atom and positron-molecule scattering, in spite of their development over already several decades, is still quite an active area of research (see [1–3] and references therein). The interest in positron slowing and annihilation in gases and other media motivates the investigation of these processes. The photons emitted in the course of annihilation carry extremely valuable information on the electron structure of various objects, from isolated atoms in gases to solid bodies. However, the process of positron collisions on atoms and molecules is also of interest by itself and in comparison to electron collisions on the same objects. The projectile-target interaction mechanisms are most transparent in the collision process at low energies, and we therefore concentrate on this energy region in what follows.

For both the electron and positron scattering, the cross section is determined by the electrostatic and polarization potentials by which the target acts on the projectile. For incoming electrons, however, the exchange with the target electrons is important. It does not exist for positrons at all. At a first glance, the positron scattering process therefore appears to be simpler than that for electrons. Moreover, the positron–atom static potential, being repulsive in general, forces the positron to move into the areas of the target where the potential is the smallest, thus diminishing its overall action. This is opposite to the case of electron–atom (molecule) collision. The contrast looks even stronger if we take into account that as it seems, the polarization potential is attractive for both electrons and positrons. The total potential for electrons must therefore be stronger than for positrons. It thus seems at first glance that any simple approach that is good for electrons should work at least no worse for positrons. For instance, the second-order approximation to the polarization potential is good for electron–atom scattering [4] and can be expected to be at least equally good in describing positron-atom scattering, but this view has proved to be incorrect.

Indeed, the positron that is "pushed" out of the target can interact strongly with a temporarily, or virtually, excited electron that is outside the target. They form a kind of a bound state that can be called the "virtual positronium." We show that this is a very important mechanism that dramatically affects the scattering cross section.

Our aim in this paper is to calculate the slow positron–atom elastic scattering cross sections and to demonstrate the efficiency of a very simple method first proposed in [5] and then developed in [6, 7] that allows taking the virtual positronium formation into account in this process. The latter effect has proved to be extremely important. The suggested method allows us to considerably improve the agreement with experiment and to give a simple qualitative explanation for the large difference between cross sections for noble gases and their neighbors, alkali atoms. Since the introduction of this approach in [5], a number of other calculations have been performed (e.g., [6–10]) based on the idea of the virtual positronium formation but using much more complicated methods (see [8–10]).

2. DETAILS OF CALCULATIONS

The elastic scattering cross section $\sigma(E)$ of a positron with the energy E is determined by the partial

---

[¶] This article was submitted by the authors in English.

scattering phases $\delta_l(E)$, where l is the positron angular momentum, as in [11][1]

$$\sigma(E) = \frac{2\pi}{E}\sum_{l=0}^{\infty}(2l+1)\sin^2\delta_l(E). \tag{1}$$

In the calculations, we limit ourselves to the first four phase shifts $l = 0, 1, 2, 3$, which is sufficient for relatively low positron energies up to 30–40 eV.

The first step of our approach is the Hartree–Fock (HF) calculations, which means the HF approximation for the target atom and the frozen core approximation for the incoming positron, naturally without exchange of the positron and the core electrons. To calculate the HF positron phase shifts $\delta_l^{HF}(E)$, we therefore solve the equation

$$\left[-\frac{\Delta}{2} + \frac{Z}{r} - \int\frac{d\mathbf{r}\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\right]\varphi_{e^+}(\mathbf{r}) = E\varphi_{e^+}(\mathbf{r}) \tag{2}$$

for the positron wave function $\varphi_{e^+}(\mathbf{r})$; here, Z is the nuclear charge and $\rho(r)$ is the atomic electron density. The asymptotic form of the radial part $P_{El}^{e^+}(r)$ of $\varphi_{e^+}(\mathbf{r})$ for large r determines the phase shift,

$$P_{El}^{e^+}(r) = \frac{1}{\sqrt{\pi r}}\sin\left[pr - \frac{\pi l}{2} + \delta_l^{HF}(E)\right], \quad r \longrightarrow \infty. \tag{3}$$

Here, $p = \sqrt{2E}$. The density is obtained by solving the HF equations for the target atom.

A prominent difference between the HF and experimental results at low positron energies in positron-atom elastic scattering exists for almost all atoms considered, for example, He. The next step must therefore be made by taking the polarization interaction into account. This interaction appears in the second order in the positron–electron interaction, in the same way as for the electron–atom scattering (see [4, 12] for details). Assuming that the polarization interaction Σ is weak, we can express the correction $\Delta\delta_l(E)$ to the HF partial positron scattering phase due to the action of Σ as

$$\Delta\delta_l(E) = -\pi\langle El\|\Sigma_l(E)\|El\rangle, \tag{4}$$

where $\Sigma_l(E)$ is the lth component of Σ and El denotes the radial part of the positron wave function $P_{El}^{e^+}(r)$.

In the second order in the Coulomb interaction $V = 1/|\mathbf{r} - \mathbf{r}'|$ between the incoming positron and atomic electrons, the reduced matrix element in the right-hand side of (4) is given by

$$\langle El\|\Sigma_l(E)\|El\rangle = \sum_{L\ge 0}\frac{1}{2L+1}\sum_{i\le F;\,\varepsilon_2,\,l_2>F}\int_0^{\infty}\frac{\langle El,\varepsilon_i l_i\|V_L\|E_1l_1,\varepsilon_2 l_2\rangle\langle E_1l_1,\varepsilon_2l_2\|V_L\|El,\varepsilon_i l_i\rangle}{(2l+1)(E-E_1-\varepsilon_2+\varepsilon_i+i\delta)}dE_1d\varepsilon_2, \tag{5}$$

where E_1l_1 denotes the intermediate positron state, which is determined by solving Eq. (3), and ε_2l_2 and E_il_i stand for the energies and angular momenta of atomic electrons in the virtually excited and ground states, respectively (with their wave functions found in the HF approximation [12]). The reduced Coulomb matrix elements V_L are defined in [12]. The conditions $i \le F$ and $\varepsilon l > F$ indicate occupied and vacant electron states, respectively. Equation (5) corresponds to the lowest order correlation correction in the frame of the random phase approximation with exchange (RPAE). It differs from the expression for the second-order polarization interaction for electron-atom scattering described in [4, 12] because it does not include the exchange between the incoming particle and target electrons. The scattering cross section with the polarization interaction taken into account is determined by Eq. (1), where $\delta_l^{HF}(E)$ is replaced by the phases

$$\delta_l(E) = \delta_l^{HF}(E) + \Delta\delta_l(E) \tag{5a}$$

for $E < I$ with I being the target atom ionization potential. For higher positron energies, the scattering phases given by (4) and (5a) are complex and the imaginary part $\mathrm{Im}\Delta\delta_l(E) \equiv \Delta\delta_l'(E)$ determines the inelastic scattering cross section of the positron by an atom.

The elastic scattering cross section for $E > I$ is given by [4, 12]

$$\sigma(E) = \frac{\pi}{E}\sum_l(2l+1)[\cosh(2\Delta\delta_l'(E)) \\ - \cos(2\delta_l(E))]\exp[-2\Delta\delta_l'(E)], \tag{6}$$

where $\delta_l(E)$ is the real part of the positron l-wave scattering phase shift.

Similarly to Eq. (5), the method described in [4, 12] in some aspects apply well beyond simple second-order perturbation theory in the interelectron interaction. Important higher order corrections are taken into account by calculating the HF wave function of the electron ε_2l_2 in the atomic field with the vacancy i. With this improvement, even the lowest order in the polari-

---

[1] Atomic units are used in this paper, with m and e being the electron mass and charge.

zation interaction gives good results for the elastic scattering of electrons on noble gases. This interaction depends on the projectile energy, is nonlocal, and does not contain free adjustable parameters. Far from the atom, it can be approximated as the polarization potential

$$V_{\text{pol}} = -\frac{\alpha(\varepsilon)}{2r^4}, \tag{7}$$

where $\alpha(\varepsilon)$ is the atom dipole polarizability, ε is the mean excitation energy of the incoming electron in the intermediate state, and r is the distance between the projectile and the center of the target. In phenomenological calculations or estimations for low incoming positron energy E, it is usually assumed that $\varepsilon = 0$. For $0 \le \varepsilon < I_1$, where I_1 is the energy of the first atom excitation level, the dipole polarizability is positive and the polarization potential is therefore attractive. It should be kept in mind, however, that $\alpha(\varepsilon)$ as a function of ε can become negative, at least for alkali and alkali earth atoms, at $\varepsilon > I$, where I is the atomic ionization potential. As a result, the polarization potential in (7) can become repulsive. It is important to note that $\alpha(\varepsilon)$ is complex at $\varepsilon > I$, its imaginary part being proportional to the atom photoionization cross section. The polarization potential can therefore also be complex in principle.

The next step beyond the HF approximation in our analysis of positron–atom scattering consists in taking the polarization interaction into account in the first order, in the same manner as is done for the electron–atom scattering in [4]. The results obtained improve the correspondence with experimental data, but they are still far from being satisfactory. As an illustration, we can use the respective results for any atom, for instance, He (see below).

3. METHOD OF CALCULATIONS

The lack of decisive success after the second-order polarization correction (5) has been taken into account means that something qualitatively important is missing. We believe that as suggested in [5], this is the positronium formation in the intermediate state, that is, the possibility of a temporary binding of the incoming positron and the excited electron (one located far from the atomic core). We assume that, being almost unaffected by the core action, this electron and positron can form a bound state that is almost identical to the free positronium Ps. This alters the energy of the intermediate state, shifting it by the positronium binding energy I_{Ps}, and modifies the wave function of the intermediate state, which is no longer the product of HF wave functions of the positron, the excited electron, and the vacancy created after the virtual excitation of the atomic electron. Instead, the motion of the positron relative to the electron is strongly modified by the binding.

To take the energy shift into account, we must subtract I_{Ps} from the sum of the positron and electron energies $E_1 + \varepsilon_2$ in the denominator of the second-order polarization interaction (5). The modified matrix element is then given by

$$\langle El\|\tilde{\Sigma}_l(E)\|El\rangle = \sum_{L\ge 0}\frac{1}{2L+1}\sum_{i\le F;\,\varepsilon_2,\,l_2>F}\int_0^{\infty}\frac{\langle El,\varepsilon_i l_i\|V_L\|E_1 l_1,\varepsilon_2 l_2\rangle\langle E_1 l_1,\varepsilon_2 l_2\|V_L\|El,\varepsilon_i l_i\rangle}{(2l+1)(E-E_1-\varepsilon_2+I_{\text{Ps}}+\varepsilon_i+i\delta)}dE_1 d\varepsilon_2. \tag{8}$$

The additional phase shifts $\Delta\delta_l(E)$ are determined by Eq. (4) with $\tilde{\Sigma}_l(E)$ instead of $\Sigma_l(E)$.

We note that far from the target atom, Eq. (8) leads to a rather simple expression for the polarization potential,

$$V_{\text{pol}} = -\frac{\alpha(\varepsilon + I_{\text{Ps}})}{2r^4}. \tag{9}$$

According to the discussion at the end of the previous section, it is essential to have in mind that if $I_{\text{Ps}} > I$, then $\alpha(I_{\text{Ps}})$ is a complex quantity, usually with a considerable imaginary part and $\alpha(I_{\text{Ps}})$ can be not only positive but also negative.

It would be much simpler to use (9) (or (7)) instead of (8) (or (5)), but the asymptotic expressions are valid at such large distances from the atom that their contribution to the total phase shift is small. This is why we used Eq. (8) in our calculations. The advantage of our approach is obvious: to describe the positron–atom scattering, we can use almost the same system of computing codes that was used in studying the electron–atom scattering [12].

To properly include the positron–electron interaction mentioned above, instead of simply adding I_{Ps} to the denominator in (8), we must take the modification of the corresponding wave functions into account. This means that the product of the one-positron Hartree and one-electron Hartree–Fock functions $E_1 l_1$, $\varepsilon_2 l_2$ and energies E_1 and ε_2 must be replaced by wave functions and total energies of the interacting or bound positron and electron that move in the atomic field. To find these functions and total energies, the three-body problem must be solved taking into account the interaction between the incoming positron and atomic electron and the vacancy created after the electron virtual excitation. This is very difficult, and simplifications are inevitable. An attractive option is to describe the relative motion of the positron and the electron by a positronium wave function, while considering their center of mass as moving freely, unaffected by the selfconsistent atomic field and the vacancy field. This approximate approach has been developed in [8–10], but the modification of the energy denominator was entirely neglected there. Here, we present the results of a much simpler approach, where only the energy shift due to virtual positronium formation is taken into account and the modification of the positron and electron wave functions is completely neglected [5–7].

4. RESULTS OF CALCULATIONS

Here, we give the results of our calculations for a number of atoms. We start with He.

The results for elastic scattering of positrons by He obtained using Eq. (8) are demonstrated in Fig. 1. It can be seen that the energy shift due to the positronium formation leads to a prominent decrease in the low-energy cross section. The difference is qualitative at $E \leq 1$ eV. As E grows, the influence of the positronium formation becomes smaller, but the deviation from the HF approximation results is huge in the entire region considered, up to $E = 18$ eV. To obtain the cross sections, contributions of the positron *s*, *p*, and *d* partial waves were taken into account. We note that the virtual positronium formation leads to prominent variations in all the partial wave contributions. This is illustrated in Fig. 2, where the results with (Fig. 2a) and without (Fig. 2b) the positronium energy shift are presented.

The cross section of the low energy (e^+ + He)-elastic scattering is much smaller than that of (e^- + He). This can be explained qualitatively as follows: while the selfconsistent field V_{sc} acting on the incoming positron is repulsive, the polarization potential V_{pol}, which behaves as $-\alpha_{\mathrm{He}}(I_{\mathrm{Ps}})/2r^4$ far from the atom, is attractive because $\alpha(I_{\mathrm{Ps}})$ and $\alpha(0)$ for He are positive and of the same order of magnitude as V_{sc}. The contributions of V_{sc} and V_{pol} therefore compensate each other, suppressing the elastic scattering cross section. For the electron scattering, both V_{sc} and V_{pol} are attractive, and instead of compensating, the respective quite big contributions enhance each other.

It is of special interest to compare the (e^+ + He) and (e^+ + Li) scattering because $\alpha_{\mathrm{Li}}(I_{\mathrm{Ps}})$ is negative, complex, and much larger than $\alpha_{\mathrm{He}}(I_{\mathrm{Ps}})$ in absolute value. In accordance with (9), the negative sign of $\alpha_{\mathrm{Li}}(I_{\mathrm{Ps}})$ implies that in this case, the polarization potential is repulsive instead of always being attractive [13] (also see the discussion above). This observation is important

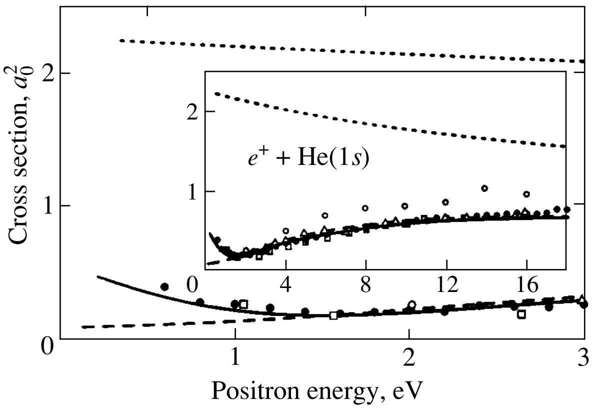


Fig. 1. Cross sections of the elastic positron scattering on He. The dotted line is the HF approximation; the dashed line is the RPAE without positronium formation taken into account; the solid line is the RPAE with the positronium formation taken into account. Experimental data: triangles, [15]; open circles, [16]; open squares, [17]; solid circles, [18]; a_0 is the Bohr radius.

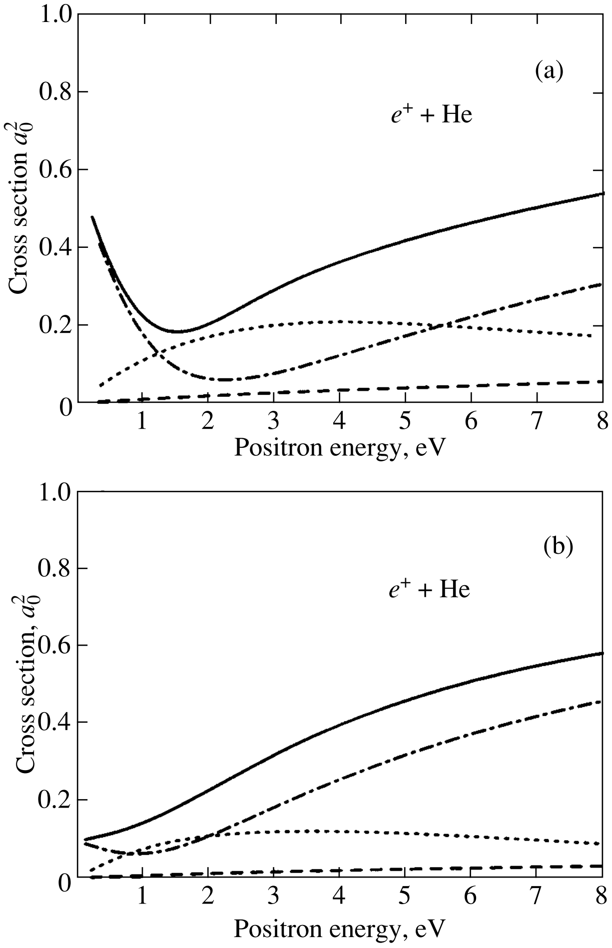


Fig. 2. Partial wave contributions to the cross sections of the elastic positron scattering on He: *a*, with the inclusion of positronium formation; b, without positronium formation. I_{Ps} = 0.5Ry (a), 0 (b). The dotted-and-dashed line is the *s*-wave cross section, the dotted line is the *p*-wave cross section, the dashed line is the *d*-wave cross section, and the solid line is the sum of the *s*, *p* and *d* partial cross sections.

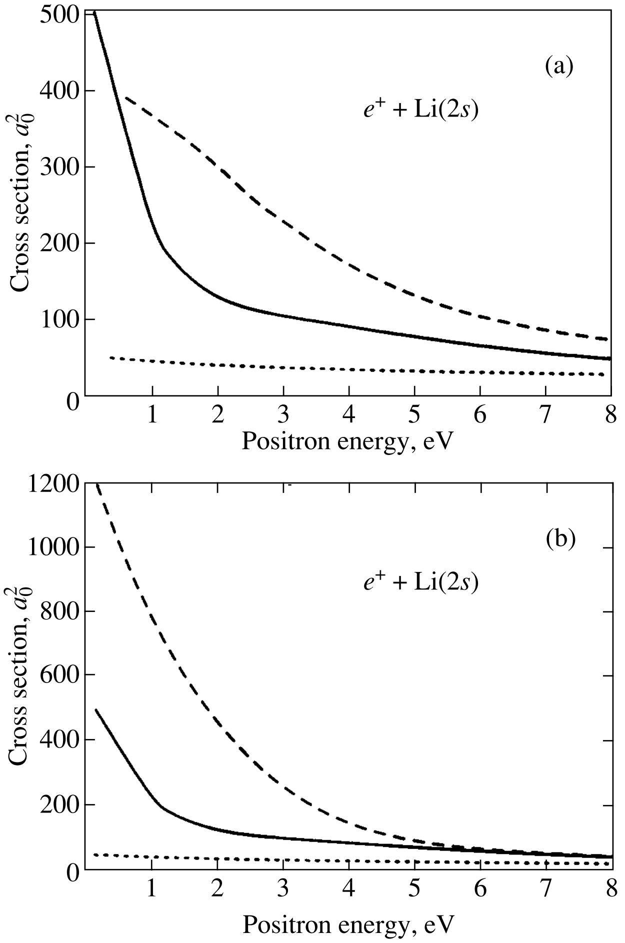


Fig. 3. Cross sections of the elastic positron scattering on Li. The dotted line (a, b) is the HF approximation, the solid line (a, b) is the RPAE with positronium formation taken into account, and the dashed line is a, the result of [2], b, the RPAE without positronium formation.

for different scattering processes in general. Indeed, if the projectile and the target constituent can form a composite particle in the intermediate state, the polarization interaction can easily change its sign, becoming attractive. This was found to occur, for example, in nuclear physics, namely in π-meson–nuclear scattering, where the (π-meson + nucleon) system forms the so-called Δ_{33}-resonance, leading to a change of the sign in the polarization interaction [14]. The imaginary part, depending on its magnitude, can effectively be of either repulsive or attractive nature from the point of view of the elastic scattering. We should therefore expect that because V_{sc} and V_{pol} have the same sign, they contribute constructively and hence lead to an extremely large cross section of the size of, or even bigger than, the (e^- + Li) cross section. The results for (e^+ + Li) cross sections are presented in Fig. 3. It follows that the energy shift accounting for the positronium formation in the virtual state affects the low-energy cross section considerably. In Fig. 3a, we show the results obtained in the HF and random phase approximation with exchange with the positronium formation taken into account and the results derived in [2] using a substantially more complicated method. Although the difference between RPAE and [2] is prominent, the deviation of both of them from the HF approximation is qualitative. Figure 3b clearly demonstrates the magnitude of the effect is of taking positronium formation into account in the virtual state for $E \leq 4$ eV.

Figures 4a, 4b, 4c, and 4d present our results for noble gas atoms Ne, Ar, Kr, and Xe, respectively. In the case of Ne for $E \leq 2$ eV, the role of positronium formation is significant, while the deviation from the HF approximation is quite dramatic. We can see that the first experimental point at about 1 eV demonstrates the essential role of taking the positronium formation into account in the virtual state. The same effect for even higher E is seen for Ar in Fig. 4b. In Fig. 4c (Kr) and Fig. 4d (Xe), the RPAE results (with the positronium formation) are qualitatively different from the HF results. The last experimental points, dark triangles for Kr and dark circles for Xe, are in reasonable agreement with the RPAE results. More accurate data are desirable, however.

The picture of positron scattering on He and Li described above is also qualitatively valid for the Ne–Na pair. Indeed, the (e^+ + Ne) cross section is small, considerably smaller than the geometrical one, while the (e^+ + Na) elastic scattering cross section (Fig. 5a) is very large. Even larger is the cross section for (e^+ + Ne), as can be seen in Fig, 5b. In Fig. 5, we compare our results with the close-coupling calculations from [22]. Because the calculation approaches are substantially different, the difference is not a big surprise, but experimental data are needed. For noble gases heavier than Ne, namely, Ar, Kr, and Xe, the polarization interaction is much larger and the cross sections increase as the atomic number grows. As in the Li–He case, the cross sections for the alkali neighbors are again much larger. The results for Kr and Xe are in qualitative agreement with those obtained in [1] using a much more complicated method.

It is interesting to compare the results for a group of three neighbors, a noble gas, alkali, and alkali-earth atom. As a good example, the groups of atoms He, Li, Be (Fig. 6a) and Ar, K, Ca (Fig. 6b) are considered (Figs. 1–6). We checked the role of the positronium formation for Be and found it very important, as previously. For Be, the cross sections proved to be similar in size and shape to those of Li, which is a consequence of the fact that their polarizabilities $\alpha(I_{\text{Ps}})$ are relatively close. The situation is different for Ca, where the cross section is much smaller at very low E, but then decreases much slower than in K. For Ar, K, and Ca, all three curves are qualitatively similar, but the (e^+ + K)

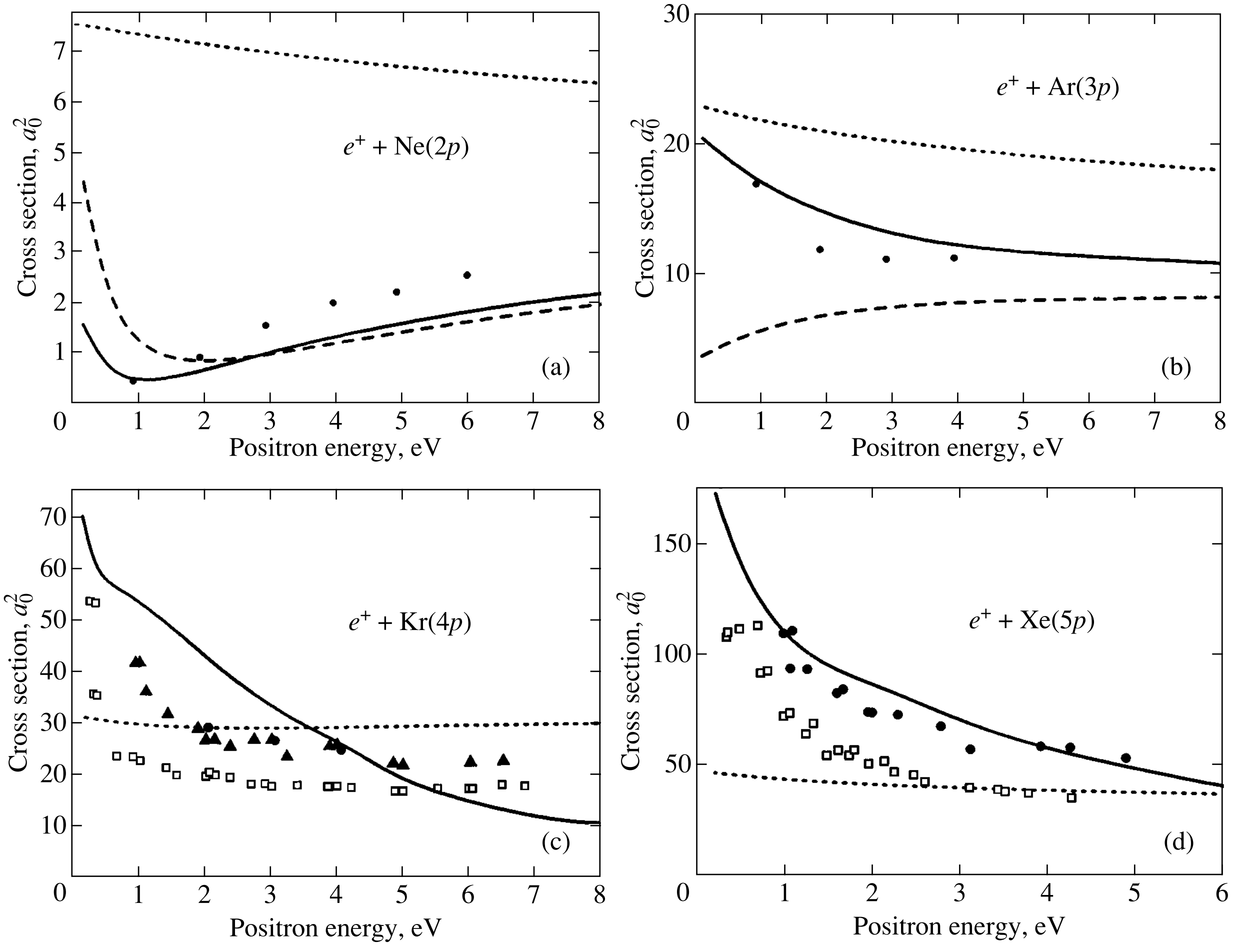


Fig. 4. Calculated cross sections of the elastic positron scattering on noble gases. The dotted line is the HF approximation, the dashed line is the RPAE without taking the positronium formation into account, the solid line is the RPAE with the positronium formation taken into account. a—Ne, experimental points from [19]; b—Ar, experimental points from [19]; c—Kr, experimental points: solid circles, [19]; open squares, [20]; solid triangles, [21]; d, Xe; experimental points: solid circles, [19]; open squares, [17].

elastic scattering cross section is by an order of magnitude larger than that of (e^+ + Ar).

It is of some interest to study the imaginary parts of the elastic scattering phases. They describe the respective partial wave contributions to the cross section of the inelastic process,

$$e^+ + \mathrm{A} \longrightarrow \mathrm{Ps} + \mathrm{A}^+. \qquad (10)$$

The inelastic positron scattering cross section $\sigma_{\mathrm{in}}(E)$ is expressed via the imaginary part of the phase shift $\Delta\delta_l'(E)$ as

$$\sigma_{\mathrm{in}}(E) = \frac{\pi}{E}\sum_{l=0}^{2}(2l+1)[1-\exp(-4\Delta\delta_l'(E))]. \qquad (11)$$

As in calculations of $\sigma(E)$, we limit ourselves to taking the first three partial waves into account (those with $l = 0, 1, 2$). In Fig. 7, we show the results of our calculation of the Ps-formation cross section in positron–Na atom scattering together with the experimental data of [23]. At low energies, there is a strong deviation from the experiment, as in many other calculations [24], but at energies above 3 eV, there is satisfactory agreement. We note that it is assumed in this calculation that any electron obtained by ionization together with the inelastically scattered positron forms a positronium Ps. Obviously, this is an exaggeration: some of the electrons leave the atom without forming a real positronium. This is particularly significant for a small energy E in the cases where the Ps-formation threshold is at $E = 0$ (as in an (e^+ + Na) collision).

We note that using Eq. (8), we can also describe the Wigner–Baz' peculiarities in the elastic scattering cross section near the threshold of inelastic channel (10) [11].

5. SUMMARY AND DISCUSSION OF RESULTS

We have demonstrated that the relatively simple method with both many-electron correlations and vir-

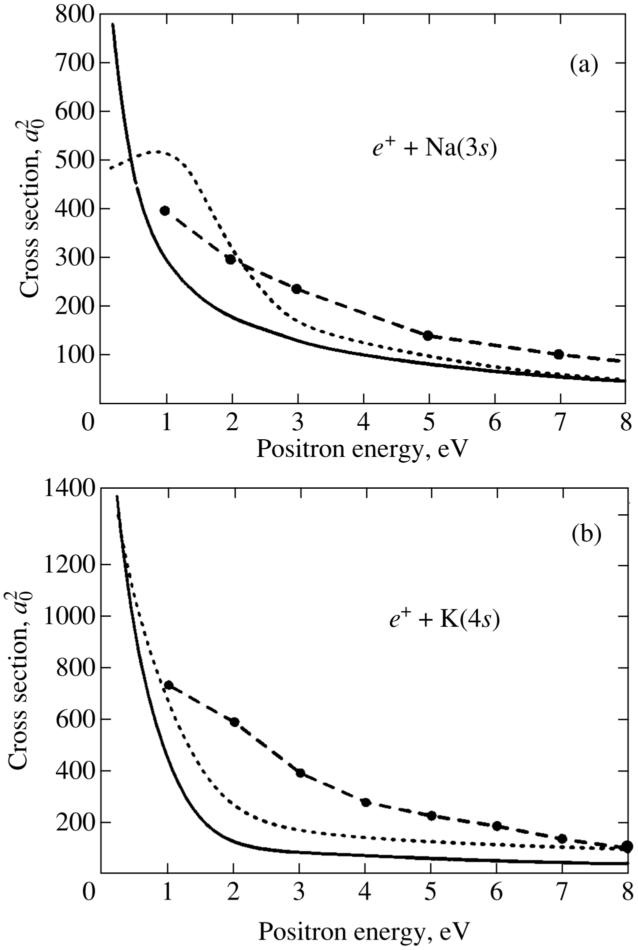


Fig. 5. Calculated cross sections of the elastic positron scattering on Na (a) and K (b). The dotted line is the HF approximation, the solid line is the RPAE with positronium formation taken into account, and the dashed line with heavy dots is the result of close-coupling calculations [22].

tual positronium formation taken into account allows us to obtain relatively good results for the elastic scattering of positrons on different atoms in the Periodic Table, in particular, noble gas and alkali atoms. It also gives an estimate of the Ps-formation cross section. The approach developed in this paper can be applied without any essential difficulty to the study of positron scattering on more complicated targets, such as molecules, clusters, and fullerenes.

It is interesting to know whether the bound states in the (e^+ + A) system can be described within the simple approach developed here. Indeed, given the repulsive nature of V_{sc} and the possibly also repulsive nature of V_{pol}, it is far from trivial that binding can occur at all. We can therefore expect bound states of positrons with those atoms A for which $\alpha_A(I_{\mathrm{Ps}})$ is large (considerably larger than in noble gases) and positive, $\alpha_A(I_{\mathrm{Ps}}) > 0$; i.e., V_{pol} is sufficiently strong and attractive. An interesting and intriguing possibility is that the (e^+ + A) bound state results from the action of the imaginary part $\mathrm{Im}(V_{\mathrm{pol}})$. On the other hand, the binding can originate from the interaction of Ps and A via Van der Waals forces. These are particularly large if A^+ has an electronic structure similar to an atom in the first period of the Periodic Table; i.e., A is in the second period. To detect the possibility of forming a bound state, one must study the magnitude of the scattering phase shift at zero energy: if it reaches π, a bound state is created in the channel under consideration. It must be checked, however, whether this state is stable against the decay through the (Ps + A^+) channel, which requires knowledge of the bound state energy. Finding it is much more complicated than calculating the phase shifts at zero e^+-energy.

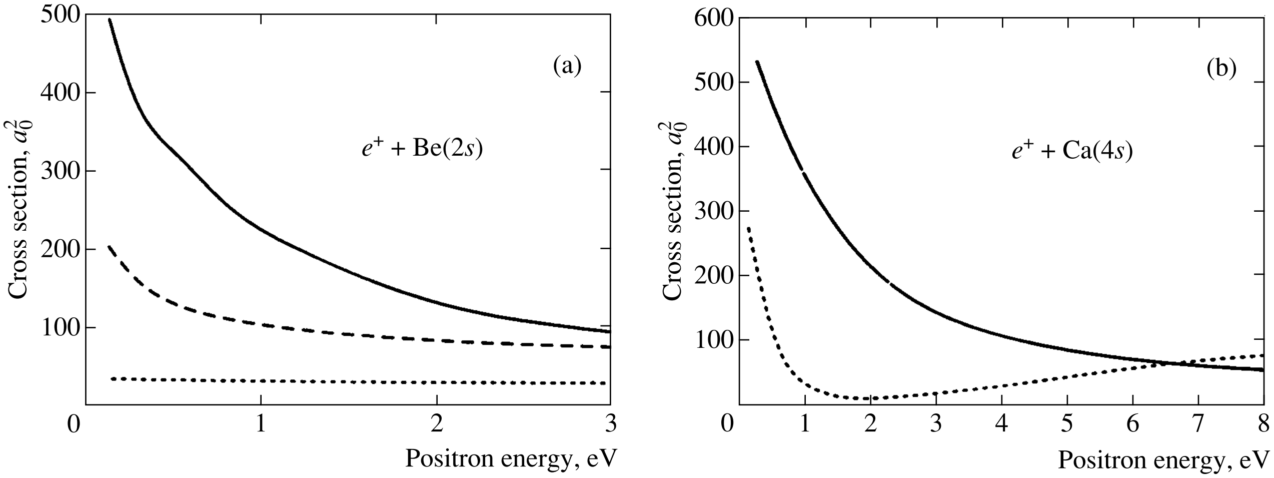


Fig. 6. Calculated cross section of the elastic positron scattering on Be (a) and Ca (b). The dotted line is the HF approximation, the solid line is the RPAE with positronium formation taken into account, and the dashed line is the RPAE without positronium formation.

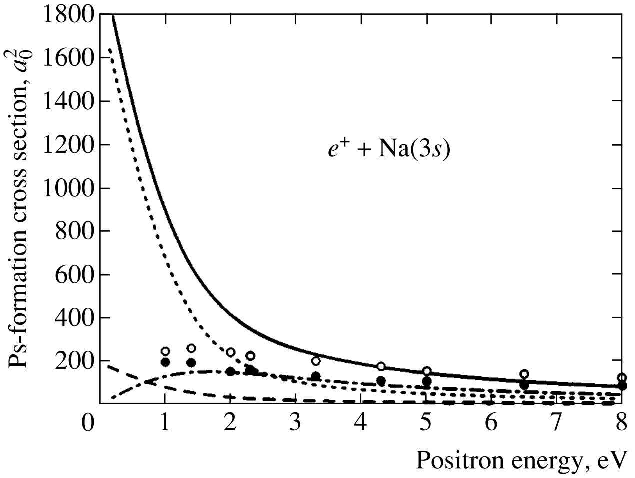


Fig. 7. Ps-formation cross sections for positron–Na scattering in RPAE. The dotted-and-dashed line is the *s*-wave cross section, the dotted line is the *p*-wave cross section, the dashed line is the *d*-wave cross section, and the solid line is the sum of *s*, *p*, and *d* partial cross sections. Experimental points are taken from [23]; open circles show the upper limit; solid circles, the lower limit.

M. Y. A. and L. V. C. acknowledge financial support of the International Science and Technology Center (project no. 1358) and The Hebrew University Intramural Research Fund.

REFERENCES

1. F. A. Gianturco and D. De Fazio, Phys. Rev. A **50**, 4819 (1994).
2. M. T. McAlinden, A. A. Kernoghan, and H. R. J. Walters, J. Phys. B: At. Mol. Opt. Phys. **30**, 1543 (1997).
3. F. A. Gianturco and R. R. Lucchese, Phys. Rev. A **60**, 4567 (1999).
4. M. Ya. Amusia and N. A. Cherepkov, Case Stud. At. Phys. **5**, 47 (1975).
5. M. Ya. Amusia, N. A. Cherepkov, L. V. Chernysheva, and S. G. Shapiro, J. Phys. B **9**, L531 (1976).
6. E. Ficocelli Varracchio, J. Phys. B: At. Mol. Opt. Phys. **23**, L779 (1990).
7. E. Ficocelli Varracchio and L. A. Parcell, J. Phys. B: At. Mol. Opt. Phys. **25**, 3037 (1992).
8. V. A. Dzuba, V. V. Flambaum, G. F. Gribakin, and W. A. King, J. Phys. B.: At. Mol. Opt. Phys. **29**, 3151 (1996).
9. G. F. Gribakin and W. A. King, J. Phys. B: At. Mol. Opt. Phys. **27**, 2639 (1994).
10. G. F. Gribakin and W. A. King, Can. J. Phys. **74**, 449 (1996).
11. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics,* Vol. 3: *Quantum Mechanics: Non-Relativistic Theory,* 2nd ed. (Nauka, Moscow, 1963; Pergamon, New York, 1977).
12. M. Ya. Amusia and L. V. Chernysheva, *Computation of Atomic Processes* (IOPP, Bristol and Philadelphia, 1997), p. 226.
13. M. Ya. Amusia and K. Taulbjerg, in *Abstracts of ICPEAC-XVII, Whistler, Canada* (1995), Vol. 1, p. 364.
14. C. J. Batty, E. Friedman, and A. Gal, Phys. Rep. **287**, 385 (1998).
15. K. F. Canter, P. G. Coleman, T. C. Griffith, and G. R. Heyland, J. Phys. B **6**, L201 (1973).
16. E. Jaduszliwer and D. A. L. Paul, Can. J. Phys. **51**, 1565 (1973).
17. G. Sinapius, W. Raith, and W. G. Wilson, J. Phys. B **13**, 4079 (1980).
18. T. Mizogava, Y. Nakayama, T. Kawaratan, and M. Tosaki, Phys. Rev. A **31**, 2171 (1985).
19. M. S. Dababneh, W. E. Kauppila, J. B. Downing, *et al.*, Phys. Rev. A **22**, 1872 (1980).
20. T. S. Stein and W. E. Kauppila, Adv. At. Mol. Phys. **18**, 53 (1982).
21. M. Charlton, Rep. Prog. Phys. **48**, 737 (1985).
22. R. N. Hewitt, C. J. Noble, and B. H. Bransden, J. Phys. B: At. Mol. Opt. Phys. **26**, 3661 (1993).
23. S. Zhou, S. P. Parikh, W. E. Kauppila, *et al.*, Phys. Rev. Lett. **73**, 236 (1994).
24. T. S. Stein, J. Jiang, W. E. Kauppila, *et al.*, Can. J. Phys. **74**, 313 (1996).

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 42–48.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 48–54.
Original English Text Copyright © 2003 by Smirnov.

**PLASMA,
GASES**

Distribution and Evolution of Electrons in a Cluster Plasma Created by a Laser Pulse¶

M. B. Smirnov
Russian Research Centre Kurchatov Institute, pl. Kurchatova 1, Moscow, 123182 Russia
e-mail: smirnov@imp.kiae.ru

Submitted January 30, 2003

Abstract—We analyze the properties and the character of the evolution of an electron subsystem of a large cluster (with a number of atoms $n \sim 10^4$–10^6) interacting with a short laser pulse of high intensity (10^{17}–10^{19} W/cm^2). As a result of ionization in a strong laser field, cluster atoms are converted into multicharged ions, part of the electrons being formed leaves the cluster, and the other electrons move in a self-consistent field of the charged cluster and the laser wave. It is shown that electron–electron collisions are inessential both during the cluster irradiation by the laser pulse and in the course of cluster expansion; the electron distribution in the cluster therefore does not transform into the Maxwell distribution even during cluster expansion. During cluster expansion, the Coulomb field of a cluster charge acts on cluster ions more strongly than the pressure resulting from electron–ion collisions. In addition, bound electrons remain inside the cluster in the course of its expansion, and cluster expansion therefore does not lead to additional cluster ionization. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

We consider the character of equilibrium in the plasma formed by irradiation of a cluster beam by a strong laser pulse with an intensity greater than 10^{17} W/cm^2, which has been studied experimentally [1–3]. This plasma is used both as a source of neutrons produced with a beam of deuterium clusters [4–6] and for generation of X-rays [7–10]. Under typical experimental conditions, the hierarchy of times of the cluster plasma evolution under consideration is as follows. The typical time of laser pulse duration (τ_1= 30–100 fs) is small compared to the cluster lifetime with respect to Coulomb explosion ($\tau_{\rm exp} \sim$ 100–1000 fs), which is in turn smaller than the typical time of expansion of the uniform plasma being formed, which can reach 1 ns. This hierarchy of times determines the behavior of this plasma.

Along with this time hierarchy, the character of laser energy absorption is important for the yield parameters of the plasma. Excitation of clusters by the laser pulse under consideration proceeds through the electron component of this plasma, but this interaction has a specific character due to the parameters of the laser radiation. First, these processes occur in strong fields, because the radiation intensity exceeds the atomic field strength, which is 3×10^{16} W/cm^2, and we consider higher fields that are available under contemporary experimental conditions [11]. Second, the interaction time is very small, and although it exceeds the typical atomic value, light propagates over a distance of 10 μm within 30 fs. Therefore, in contrast to the classical character of interaction between radiation fields and atomic systems [12–14], other interaction mechanisms are realized in this case [15–19]. Hence, short-time processes occur because of the small duration of the laser–cluster interaction, and overbarrier ionization is the main process of absorption of a strong electromagnetic wave. As a result, the laser radiation energy is consumed by the ionization of cluster atoms and is transmitted to the electron component of clusters. Below, we consider the character of the development of an electron subsystem of clusters under the action of a strong laser pulse.

In the first stage of the ionization process, free electrons in clusters result from overbarrier ionization of cluster atoms [15]. Electrons of zero kinetic energy are formed in this process, and atoms are transformed into multicharged ions that are found in the ground state. The electrons being formed move in the laser and cluster fields and receive energy from these fields. When electrons go outside the cluster, a positive cluster charge arises and interaction between the cluster and laser fields leads to subsequent electron liberation. In addition, collisions of electrons can cause redistribution of the electron energy, and we analyze the role of such processes.

Because the cluster acquires a positive charge under the action of a laser pulse, it expands as a result of the interaction of the cluster field and its ions. In the course of expansion, the electron subsystem can influence the expansion process. In addition, collisions of electrons with multicharged ions that result from ionization of cluster atoms lead to the formation of excited multicharged ions; radiation processes involving multi-

---

¶This article was submitted by the author in English.

charged ions are responsible for the X-ray radiation of this cluster plasma.

Taken together, processes involving the electron subsystem of clusters determine both the cluster ionization rate and the character of energy redistribution of electrons that affects the processes responsible for X-ray emission of the cluster plasma. The goal of this paper is to analyze the character of electron equilibrium in the cluster plasma and the processes involving electrons.

2. IONIZATION OF CLUSTER ATOMS AND CLUSTERS BY A LASER PULSE

2.1. Formation of Multicharged Ions inside Clusters

Under typical conditions of the process under consideration, the photon energy is smaller than the ionization potentials of cluster atoms and ions, and ionization of cluster atoms results from overbarrier transitions of initially bound electrons in the laser wave field. Because of the high electric strength of the electromagnetic wave, this field decreases the barrier from the Coulomb field of the nucleus and an initially bound electron can freely leave this Coulomb field.

In taking into account ionization of individual atoms in a cluster by a laser pulse, we assume its effect on cluster atoms to be identical to that of a constant electric field. This assumption is valid at small values of the Keldysh parameter [15]:

$$\gamma = \frac{\omega\sqrt{2J_Z}}{F},$$

where F is the electric field strength, J_Z is the ionization potential of a multicharged ion with a charge of $Z - 1$, and ω is the laser frequency (we use atomic units in this paper). Therefore, this character of ionization of cluster atomic particles is valid for high electric field strengths F of the laser electromagnetic wave. Hence, we use the Bethe formula [20] for the strength F of the electromagnetic wave at which the barrier disappears for an electron with the ionization potential J_Z,

$$F = \frac{J_Z^2}{4Z}, \tag{1}$$

where Z is the charge of a atomic ion being formed. This formula implies that the overbarrier transition leads to liberation of electrons whose binding energy is less than J_Z.

This criterion can be represented in another form using an analysis of the dynamics of the electron transition from the Coulomb field of the atomic core. Indeed, near the top of the barrier created by the Coulomb field of the atomic core and the constant electric field, we have the Newton equation

Table 1. The charge of cluster atomic ions resulting from cluster irradiation by electromagnetic waves of different intensities at a cluster of 10^6 atoms. The first value is the ion charge at the cluster center; the value in parentheses is the charge at the boundary

	10^{17} W/cm^2	10^{18} W/cm^2	10^{19} W/cm^2
Kr	12(18)	18(26)	24(27)
Xe	11(24)	19(28)	26(43)
Mo	12(14)	14(24)	22(32)
W	12(38)	22(47)	41(56)

$$\frac{d^2r}{dt^2} \approx 2F\frac{r}{r_0} \tag{2}$$

for a classical electron, where $r_0 \sim \sqrt{Z/F}$ is the distance of the top of the barrier from the center of the ion. From this, we find the typical time

$$\tau_{\rm dep} \sim \sqrt{\frac{r_0}{2F}}$$

for the overbarrier electron transition. The requirement that this time is small compared to the period of the electromagnetic wave gives

$$\tau_{\rm dep}^2\omega^2 \sim \frac{2r_0\omega^2}{F} \sim \frac{Z^{1/2}\omega^2}{F^{3/2}} \ll 1. \tag{3}$$

Because of (1), this criterion is identical to the smallness of the Keldysh parameter. This implies that the mechanism of overbarrier ionization of cluster atoms and ions under the action of a strong electromagnetic wave applies to large intensities of the electromagnetic wave. In particular, the right-hand side of criterion (3) gives the approximate value 3×10^{-3} for the laser pulse with $I = 10^{17}$ W/cm^2 and a frequency of 1.5 eV (Ti:sapphire laser).

Charges Z of ions formed by the overbarrier electron transition are given in Table 1. High values of the charge allows us to use a simple formula for the electron binding energy in this case,

$$J_Z = \frac{Z_{ef}^2}{2(n-\delta_{nl})^2}, \tag{4}$$

where Z_{ef} is the effective charge that includes shielding of the nucleus charge by atomic electrons, such that $Z_{ef} \geq Z$, and δ_{nl} is the quantum defect for a given electron shell (usually, $\delta_{nl} < 1$). This formula uses the analogy of multicharged ions with hydrogen-like ions.

We consider one more aspect of this problem. Using the analogy between the action of a strong electromagnetic wave and a constant electric field on an atomic particle, we ignore the absorption of the electromagnetic wave as a result of electron release. In reality, this absorption follows from excitation of the electron subsystem of an individual cluster. We assume that as the electric field strength of the laser wave $F\cos\omega t$ varies from zero, it leads to the release of new electron groups in accordance with Bethe formula (1). Hence, as a result of the overbarrier ionization process, free electrons are formed inside a cluster with zero energy.

We now estimate the excitation time for an individual cluster of n atoms whose electron subsystem acquires the excitation energy roughly equal to nJ_ZZ, where Z is a typical charge of forming multicharged ions. Because the incident energy flux of the laser pulse is $cF^2/8\pi$ and the cluster cross section is equal to its geometric cross section,

$$\pi R^2 = \pi r_W^2 n^{2/3}$$

(where r_W is Wigner–Seitze radius and n is the number of atoms in the cluster), we find the typical time τ during which all the cluster atoms are converted into multicharged ions of charge Z,

$$\tau = n^{1/3}\frac{8J_ZZ}{cr_W^2F^2}. \tag{5}$$

Substituting the electric field strength F from Eq. (1), we obtain

$$\tau\omega = n^{1/3}\frac{128}{c}\frac{\omega Z}{J_Z}\left(\frac{Z/r_W}{J_Z}\right)^2. \tag{6}$$

This value is much smaller than the one for cluster sizes under consideration, and it decreases with the increase of the laser pulse intensity because $J_Z \sim Z^2$. Field ionization in clusters can therefore be considered as an instant process. As a result of this analysis, we find that typical experimental intensities of laser pulses provide fast excitation of the electron component of clusters through ionization of their atoms and ions.

2.2. Ionization of a Cluster as a Whole

Ionization of a cluster under the action of a strong electromagnetic wave is similar to overbarrier ionization of atoms. An electron passes over the barrier and is released. How this differs from the case of an atom is that the electron motion in the cluster field is described by classical laws and the cluster size is restricted. The latter allows a released electron to leave the cluster field. The time of electron displacement by a distance on the order of the cluster radius R is given by

$$t_{\text{esc}} \sim \sqrt{\frac{R^3}{Q}} \tag{7}$$

(the typical electron velocity is approximately $\sqrt{Q/R}$, where Q is the cluster charge). The criterion $t \ll 1/\omega$ then becomes

$$R^3\omega^2 \ll Q. \tag{8}$$

The cluster charge is determined by the Bethe formula for the overbarrier transition of an electron located in the Coulomb field of a charge Q and in a constant electric field of a strength F. Taking the interaction energy of the electron with the electric field of the cluster to be Q/R at the cluster boundary, we obtain from (1) that the charge Q of this cluster is given by

$$Q = 4FR^2. \tag{9}$$

When criterion (8) is valid, this gives

$$R \ll \frac{4F}{\omega^2} \tag{10}$$

for the cluster size.

We now consider one more aspect of this problem. The electron remains in the region of the cluster Coulomb field under the action of the electromagnetic wave, but can return at another stage of variation of the electromagnetic wave strength if the electron trajectory is determined strongly by the fields of the cluster and of the electromagnetic wave. Of course, the large statistical weight of continuum spectrum states for this electron makes electron release favorable. It is important that the Coulomb field of the cluster is not constant in time precisely because it is also created by the motion of bound electrons of the cluster. Fluctuations resulting from the motion of internal cluster electrons lead to randomization of the motion of the electron being transferred and make its departure from the Coulomb cluster field irreversible. This determines the applicability of simple formulas for ionization of the cluster as a whole.

Thus, as a result of the ionization processes both inside the cluster atoms and for the cluster as a whole, a specific plasma is formed such that multicharged atomic ions of the cluster keep a part of the electrons moving inside the cluster. These electrons are locked inside the cluster, whereas a part of the electrons is released and creates a cluster charge. Later, this system decays as a result of cluster expansion caused by Coulomb forces acting on ions. However, the processes of formation of this plasma determine its properties and the character of the subsequent cluster expansion.

We consider one more consequence of cluster charging. Because the charge of cluster ions is not compensated by the electron charge, an additional field arises in the cluster. For simplicity, we use the model where the cluster charge is distributed over the cluster uniformly. Then the electric field of the cluster charge with a strength of

$$F_{cl} = \frac{Qr}{R^3} = 4F\frac{r}{R} \tag{11}$$

acts on an ion located at distance r from the cluster center. This changes the charges of multicharged ions being formed inside the cluster. This problem was examined in [21, 22] in detail. Replacing the electric field strength F of the laser wave in Bethe formula (1) with $F + F_{cl}$, we obtain for the ion charge $Z(r)$ at distance r from the cluster center

$$F\left(1 + 4\frac{r}{R}\right) = \frac{J_Z^2}{4Z}. \tag{12}$$

Thus, this charge is larger near the cluster boundary than at its center. The data in Table 1 pertain to the charge of multicharged ions near the cluster center. We also include in this table the charge near the cluster boundary.

We note that in the above consideration, we assumed that the laser field penetrates the cluster. This is a valid assumption because the skin depth for the laser signal is approximately 100 nm in this case and exceeds the cluster size in the range under consideration (a cluster consisting of 10^6Xe atoms has a radius of 25 nm [23]).

3. ELECTRON DISTRIBUTION IN THE CLUSTER PLASMA

3.1. Relaxation of Electrons in a Cluster

The cluster plasma resulting from a laser pulse expands. There are two forces acting on the cluster that cause it to expand during and after the laser pulse. The first is the pressure due to electrons. Cluster electrons collide with ions and push them outside. The second force is determined by the Coulomb force and depends on the charge distribution in the cluster, which is the distribution of ions and electrons.

The energy distribution of electrons established in the cluster during irradiation is far from equilibrium. The relaxation rate of the electron subsystem is determined by electron–electron collisions. The relaxation time can therefore be defined as the time during which a test electron gains the energy Q/R in collisions with other electrons,

$$\tau_{ee} \sim \frac{Q/R}{\epsilon \nu_{ee}}, \tag{13}$$

where ϵ is the electron energy change and ν_{ee} is the electron–electron collision rate. The electron–electron collision rate is given by [24]

$$\nu_{ee} = N_e\sqrt{2\epsilon}\sigma_t \quad \sigma_t = \frac{4\pi}{\epsilon^2}\ln\Lambda, \tag{14}$$

where $\ln\Lambda$ is the Coulomb logarithm. Substituting Eqs. (14) in Eq. (13), we obtain the relaxation time:

$$\tau_{ee} \sim \frac{Q/R\sqrt{\epsilon}}{4\sqrt{2}\pi N_e \ln\Lambda}. \tag{15}$$

Table 2. The ratio of the relaxation time τ_{ee} for the cluster of 10^6 atoms to the expansion time $\tau_{\exp}$ for different electromagnetic wave intensities

	10^{17} W/cm^2	10^{18} W/cm^2	10^{19} W/cm^2
Kr	53	359	3100
Xe	55	365	3160
Mo	6.1	57	456
W	4.5	33	241

Substituting Eq. (9) in (15), we finally obtain

$$\tau_{ee} \sim \frac{(Q/R)^{3/2} r_W^3}{3\sqrt{2} Z \ln\Lambda} = \frac{8F^{3/2} r_W^{9/2} n^{1/2}}{3\sqrt{2} Z \ln\Lambda}. \tag{16}$$

We now compare the relaxation electron time τ_{ee} with the expansion time. The typical expansion time is given by

$$\tau_{\exp} = \sqrt{\frac{MR^3}{ZQ}} = \frac{1}{2}\sqrt{\frac{MR}{ZF}}, \tag{17}$$

where M is the mass of ion. The ratio of these times is given by

$$\frac{\tau_{ee}}{\tau_{\exp}} = \frac{16}{3\sqrt{2}\ln\Lambda} \frac{F^2 r_W^4}{\sqrt{MZ}} n^{1/3} \gg 1. \tag{18}$$

The ratios $\tau_{ee}/\tau_{\exp}$ for krypton, xenon, molybdenum, and tungsten clusters under typical laser parameters are given in Table 2. Condition (18) is satisfied for all parameters considered, as follows from the data in Table 2.

3.2. Electron Subsystem during Cluster Expansion

We now consider the behavior of electrons in the course of cluster expansion. We use the above fact that the expansion time is large compared to the typical time of establishment of equilibrium in the electron subsystem. This implies that in contrast to [25], the distribution function of electrons is not the Maxwell one, and collisions between electrons during cluster expansion can be ignored. We therefore start with the electron distribution by energy that results from laser ionization of atoms and analyze the evolution of electrons located inside the cluster. For simplicity, we use the model where the positive charge is distributed uniformly inside the cluster, and the self-consistent potential of electrons and multicharged cluster ions is therefore given by

$$U(r) = -\frac{Q}{2R}\left(3 - \frac{r^2}{R^2}\right), \tag{19}$$

where r is the distance from the cluster center, and we consider the interior cluster part $r \le R$. This assumption

Table 3. Typical expansion times (fs) at some parameters of the interaction between a laser wave and clusters for different electromagnetic wave intensities and different *n*

	n	10^{17} W/cm^2	10^{18} W/cm^2	10^{19} W/cm^2
Xe	10^4	92	40	19
	10^5	134	58	28
	10^6	196	84	40
W	10^4	79	33	13
	10^5	116	50	20
	10^6	170	73	30

leads to the electron number density inside the cluster:

$$N_e(r) = Z(r)N_i - \frac{3Q}{4\pi R^3} \tag{20}$$

(for the charge $Z(r)$ of an individual atomic ion of the cluster, see Table 1). Because ions are distributed in the cluster uniformly before and at the first stage of its expansion, the charge of an individual ion depends on the distance from the cluster center as

$$Z(r) = Z_0\left(1 + \frac{2Q}{FR^2}\frac{r}{R}\right)^{1/3}, \tag{21}$$

which corresponds to Eq. (12).

We now consider the character of evolution of the cluster during its expansion. Because the self-consistent cluster potential does not change considerably during the oscillation period of a locked electron, its adiabatic invariant [26, 24]

$$I(t, \epsilon) = \frac{1}{2\pi}\int[2m(\epsilon - U(r))]^{1/2}dr \tag{22}$$

is conserved. The integration happens between two turning points of the locked electron with a certain energy ε. This quantity plays the role of an integral of motion for the electron. The distribution function of locked electrons is then a function of the adiabatic invariant,

$$f = f(I(\varepsilon, t)). \tag{23}$$

To obtain information on the cluster evolution, we assume the cluster charge Q to be constant during the cluster expansion and assume the cluster potential for an individual electron to be given by Eq. (19) when the cluster radius R depends on time. The adiabatic invariant for an individual electron is then equal to

$$I = \frac{1}{4\Omega^2}[\varepsilon - l\Omega], \quad \Omega^2 = \frac{Q}{R^3}, \tag{24}$$

where l is the electron orbital momentum, which ranges from 0 to $\Omega^2 R$.

We show that the electron orbital momentum is conserved during the cluster expansion. Indeed, the electron momentum varies in electron collisions with ions and electrons, and because these collisions lead to small scattering angles, the typical momentum variation during cluster expansion is estimated as

$$\Delta l \sim l\frac{\tau_{\exp}}{\tau_{ei}}, \tag{25}$$

where τ_{ei} is the typical time of electron–ion collisions. Here we take into account that electron–ion collisions are more effective than electron–electron collisions because of the large average charge Z of cluster ions, and the cluster expansion rate is determined by interaction of the cluster charge with each ion. This leads to the following estimates for these parameters:

$$\begin{gathered} \nu_{ei} \sim \frac{1}{\tau_{ei}} \sim \frac{4\pi N_i Z^2 \ln\Lambda}{3v^3} \sim \frac{nZ^2}{l^3}\ln\Lambda, \\ \tau_{\exp} \sim \sqrt{\frac{MR^3}{ZQ}}. \end{gathered} \tag{26}$$

We give typical values of expansion times in Table 3, from which it follows that under typical conditions of cluster evolution,

$$\frac{\Delta l}{l} \sim \tau_{\exp}\nu_{ei} \sim \sqrt{\frac{M}{Z}}\frac{nZ^2}{Q^2}\ln\Lambda \ll 1. \tag{27}$$

Thus, the above analysis shows that under typical conditions of cluster evolution, the orbital momentum l of an individual electron is conserved in the course of cluster expansion, and the electron orbital momentum l can therefore be considered as an integral of motion.

We now consider the character of motion of an individual electron in the course of cluster expansion, using conservation of the adiabatic invariant and the angular electron momentum. The motion of locked electrons consists in rotation around the cluster center and oscillations in the radial direction between two turning points. Under cluster expansion, the turning points of the electron also move. Using the expression for the energy of an individual electron

$$\varepsilon = \frac{\Omega^2 r^2}{2} + \frac{l^2}{2r^2} \tag{28}$$

and the relation between Ω and the adiabatic invariant I in Eq. (24), we obtain that the distances $r_{1,2}$ of the turning points from the center are given by

$$\begin{gathered} \frac{r_{1,2}^2}{R(t)^2} = \frac{4I}{R(t)^2} \\ + \frac{l}{\Omega(t)R(t)^2} \pm 2\left[\frac{4I^2}{R(t)^4} + \frac{2Il}{\Omega(t)(4R(t))}\right]^{1/2}, \end{gathered} \tag{29}$$

where $R(t)$ is the current cluster size. Recalling the relation between Ω and cluster size in Eq. (24), we obtain that

$$\frac{r_{1,2}}{R(t)} \approx \frac{l}{(QR(t))^{1/4}} \pm \frac{\sqrt{2I}}{R(t)}. \quad (30)$$

Because $R(t)$ increases in time, Eq. (30) implies that the electron trajectory remains inside the cluster in the course of cluster expansion and the relative distance of turning points from the cluster center $r_{1,2}/R(t)$ decreases. Because the second term in Eq. (30) decreases faster than the first one, the motion of a locked electron tends to transform into rotation.

The motion of an electron in the cluster field consists of oscillation in the cluster potential $U(r)$ and rotation around the cluster center characterized by the electron angular momentum l. Equation (30) gives the behavior of turning points of a test electron as the cluster expands. We see that turning points of a test electron move from the cluster center slower than the radius increases. This implies that electrons remain inside the cluster during cluster expansion, and cluster expansion does not therefore lead to an additional cluster ionization.

We now consider this problem from another standpoint, introducing the energy E of a test electron such that the energy on the cluster boundary is zero. Assuming that the electron transition outside the cluster boundary means the electron release, we rewrite Eq. (24) for the adiabatic invariant as

$$E = I\Omega^2 + l\Omega - W, \quad (31)$$

where $W = 3\Omega^2R^2/2$ is the cluster well depth. We introduce the reduced electron energy

$$\tilde{E} = \frac{E}{l\Omega} = \left[\frac{I}{R^2} - \frac{3}{2}\right]R^2\Omega + 1. \quad (32)$$

If $\tilde{E}$ becomes positive, the electron is released. However, because $R^2\Omega = \sqrt{QR}$ increases as a result of cluster expansion, the initially negative value $\tilde{E}$ remains negative; i.e., cluster ionization does not occur as a result of cluster expansion in the absence of collisions involving electrons.

Now let us consider the character of cluster expansion that proceeds under the action of the Coulomb field of the cluster charge and under the action of electron collisions. The force acting on a test ion from the cluster charge is given by

$$\mathcal{F}_{cl} = \frac{QZ}{R^2}. \quad (33)$$

The force acting on the test ion in collisions with electrons is

$$\mathcal{F}_{el} \sim \nu_{ei}p \sim \frac{4\pi N_i Z^2}{3v^2}\ln\Lambda, \quad (34)$$

where p is the typical electron momentum. Substituting the minimal electron velocity $v = \sqrt{Zn^{1/3}/R}$ in this formula, we obtain an estimate for the force acting on the cluster boundary due to collisions with electrons,

$$\mathcal{F}_{el} \le \frac{3n^{2/3}Z\ln\Lambda}{R^2}. \quad (35)$$

Comparing the above forces, we have

$$\frac{\mathcal{F}_{cl}}{\mathcal{F}_{el}} = \frac{4}{3}\frac{Fr_W^2}{\ln\Lambda} \gg 1. \quad (36)$$

The cluster expansion rate is therefore determined by the Coulomb force that acts on ions from the cluster charge. Hence, the equation for the evolution of the cluster boundary has the form

$$M\frac{d^2R}{dt^2} = \frac{ZQ}{R^2} \quad (37)$$

and its approximate solution is

$$R \approx \sqrt{2\frac{ZQ}{R_0M}}t. \quad (38)$$

A typical time T of the cluster expansion is

$$T = \sqrt{\frac{MR_0^3}{ZQ}}. \quad (39)$$

Typical cluster expansion times are given in Table 3. We see that the cluster expansion time becomes comparable to the pulse duration under the laser intensities of the order 10^{19} W/cm^2. It may affect cluster evolution during irradiation and after it.

4. CONCLUSIONS

Thus, the analysis of the behavior of an individual cluster in the field of a strong electromagnetic wave shows that electrons locked in the cluster cannot leave it after the laser pulse. Neither electron–electron collisions nor the cluster expansion leads to a considerable additional release of electrons. The pressure created by electron–ion collisions gives a small contribution to the cluster expansion rate. In addition, this analysis demonstrates that under large intensities of the laser pulse, expansion during the irradiation can affect the character of interaction between the laser pulse and the cluster.

The author thanks V. P. Krainov for useful discussions. This work was supported by the Russian Foundation for Basic Research (project no. 01-02-16056) and the U.S. Civilian Research and Development Foundation (grant no. MO-011-0).

REFERENCES

1. A. McPherson, J. Cobble, A. B. Borisov, *et al.*, J. Phys. B **30**, L767 (1997).
2. M. Lezius, S. Dobosz, D. Normand, and M. Schmidt, J. Phys. B **30**, L251 (1997).
3. H. Honda, E. Miura, K. Katsura, *et al.*, Phys. Rev. A **61**, 023201 (2000).
4. T. Ditmire, J. Zweiback, V. P. Yanovsky, *et al.*, Phys. Plasmas **7**, 1993 (2000).
5. G. Grillon, Ph. Balcon, J.-P. Chambaret, *et al.*, Phys. Rev. Lett. **89**, 065005 (2002).
6. J. Zweiback, T. E. Cowan, R. A. Smith, *et al.*, Phys. Rev. Lett. **85**, 3640 (2000).
7. S. Dobosz, M. Schmidt, M. Perdrix, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **68**, 454 (1998) [JETP Lett. **68**, 485 (1998)].
8. J. Abdallah, Jr., A. Ya. Faenon, I. Yu. Skobelev, *et al.*, Phys. Rev. A **63**, 032706 (2001).
9. E. Parra, I. Alexeev, J. Fan, *et al.*, Phys. Rev. E **62**, R5931 (2000).
10. M. Schnurer, S. Ter-Avetisyan, H. Stiel, *et al.*, Eur. Phys. J. D **14**, 331 (2001).
11. G. Mourou, Appl. Phys. B **65**, 205 (1997).
12. E. U. Condon and G. H. Shortley, *Theory of Atomic Spectra*, 2nd ed. (Cambridge Univ. Press, Cambridge, 1980; Inostrannaya Literatura, Moscow, 1949).
13. I. I. Sobel'man, *Atomic Spectra and Radiative Transitions* (Nauka, Moscow, 1977; Springer, Berlin, 1979).
14. V. P. Krainov, H. R. Reiss, and B. M. Smirnov, *Radiative Processes in Atomic Physics* (Wiley, New York, 1997).
15. N. B. Delone and V. P. Krainov, *Multiphoton Processes in Atoms*, 2nd ed. (Springer, Berlin, 2000).
16. J. H. Eberly, J. Javanainen, and K. Rzazewski, Phys. Rep. **204**, 332 (1991).
17. *Atoms in Intense Laser Fields*, Ed. by M. Gavrila (Academic, New York, 1992).
18. M. B. Smirnov and V. P. Kraĭnov, Usp. Fiz. Nauk **170**, 969 (2000) [Phys. Usp. **43**, 901 (2000)].
19. V. P. Krainov and M. B. Smirnov, Phys. Rep. **370**, 237 (2002).
20. H. Bethe and E. Salpeter, *Quantum Mechanics of One- and Two-Electron Atoms*, 2nd ed. (Springer, Berlin, 1975; Fizmatgiz, Moscow, 1960).
21. C. Rose-Petruck, K. J. Schafer, K. R. Wilson, and C. P. J. Barry, Phys. Rev. A **55**, 1182 (1997).
22. M. B. Smirnov and V. P. Krainov, Phys. Rev. A (2003) (in press).
23. V. P. Kraĭnov and M. B. Smirnov, Zh. Éksp. Teor. Fiz. **121**, 867 (2002) [JETP **94**, 745 (2002)].
24. E. M. Lifshitz and L. P. Pitaevskiĭ, *Physical Kinetics* (Nauka, Moscow, 1979; Pergamon Press, Oxford, 1981).
25. H. M. Milchberg, S. J. McNaught, and E. Parra, Phys. Rev. E **64**, 056402 (2001).
26. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 1: *Mechanics*, 3rd ed. (Nauka, Moscow, 1973; Pergamon Press, London, 1980).

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 49–60.
Translated from Zhurnal Éksperimental’noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 55–69.
Original Russian Text Copyright © 2003 by Managadze.

PLASMA, GASES

The Synthesis of Organic Molecules in a Laser Plasma Similar to the Plasma That Emerges in Hypervelocity Collisions of Matter at the Early Evolutionary Stage of the Earth and in Interstellar Clouds

G. G. Managadze
Space Research Institute, Russian Academy of Sciences, ul. Profsoyuznaya 84/32, Moscow, 117997 Russia
e-mail: managadze@space.ru

Received September 5, 2002

Abstract—Ions of organic molecules and polymers as well as multiply ionized hydrocarbons were synthesized and detected with a time-of-flight mass analyzer in laboratory experiments simulating with a laser the plasma processes that accompany a hypervelocity micrometeorite impact on the target surface. A hypervelocity impact of micrometeorites moving at velocities of 80 km s^{-1} on a inorganic target was simulated with a Q-switched laser. The laser provided a power density of 10^9–10^{11} W cm^{-2} in a spot with an impact diameter of 30–150 μm for a pulse duration of 7–10 ns and a laser plasma electron density of 10^5–10^6 K. The ions of organic compounds are shown to be synthesized mostly during the free expansion of a hot laser plasma at the stage of its cooling and recombination if, initially, the plasma was completely atomized and ionized. Molecular ions have high yields only for a carbon target. The results obtained indicate that organic or other polyatomic compounds can be abiogenically synthesized in intense hypervelocity meteorite impacts on the Earth’s surface at the early stage of its formation during meteorite showers and in hypervelocity collisions of dust particles in interstellar molecular clouds. *© 2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

When the impact velocity reaches a critical value V_{cr} of 15–20 km s^{-1} [1], a plasma cloud or a plasma torch is generated in the collision area of the bodies [2]. This requires that the impinging body–target interaction time scale, t_{int}, be shorter than the time scale of heat outflow from the impact area, t_h. An energy that exceeds the binding energy of the atoms in matter and their ionization energy is rapidly absorbed under these conditions. As a result, the impinging body and part of the target turn into a hot plasma. Depending on the physical parameters of the interacting bodies at impact velocities of $V_{\text{sh}} > V_{\text{cr}}$, the generated plasma reaches a temperature of 30–50 eV and is usually completely atomized and ionized at the initial stage. At the stage of adiabatic expansion, the plasma cools down and recombines, and the degree of ionization decreases by a factor of approximately 100. The ion composition of such a plasma corresponds to the elemental composition of the interacting bodies. However, polyatomic ions can also be synthesized in such a plasma during its expansion [3]. Here, our goal is to study this phenomenon or, more specifically, the synthesis of molecular ions during the expansion of a impact plasma.

A hypervelocity impact is the most characteristic phenomenon for the processes in space. Thus, for example, the surfaces of atmosphereless cosmic bodies in the Solar system are continuously subjected to hypervelocity micrometeorite impacts of particles with masses from 10^{-15} to 10^{-12} g. Micrometeorites less than 1 μm in diameter are accelerated by light pressure to velocities of $V \sim 50$ km s^{-1}. Such an impact plays an important role in the formation of the regolith surface layer on atmosphereless objects [4].

Traces from catastrophic hypervelocity impacts of large bodies that took place mostly during the formation of planets have been well preserved in the form of craters on atmosphereless planets and planetary satellites in the Solar System. During this period, the so-called meteorite shower period 4.5–3.5 Gyr ago, the Earth’s surface was also subjected to intense impacts. The presumed impact energy flux was large, comparable to the volcanic heat, about 10^{22} eV m^{-2} per year [5] at collision velocities from 11 to 70 km s^{-1}. Interestingly, life on Earth originated precisely during this period [6, 7].

In interstellar space, the acceleration of dust particles to ultrahigh velocities is also observed in many processes. These processes are mostly associated with the explosions that arise during the formation of novae, but slower processes also contribute to the acceleration of dust particles. Thus, for example, dust particles in gas–dust clouds can be accelerated under light pressure or behind a impact front to relative velocities of 10^2–10^3 km s^{-1} and

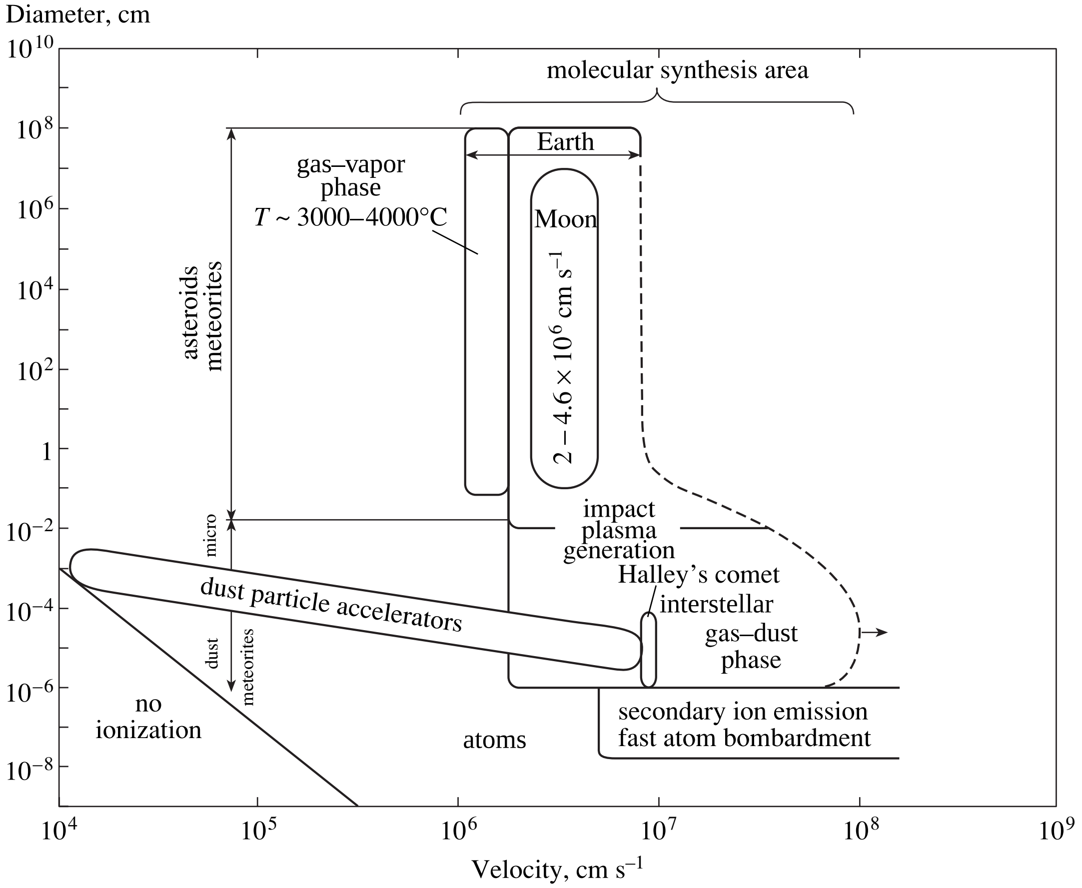


Fig. 1. The overall picture of impact plasma generation in various processes in nature and in a laboratory as a function of the body size and velocity (the diagram is based in part on the data from [28]).

be involved in hypervelocity collisions [8]. Thus, a hypervelocity impact is a common phenomenon in nature virtually at all evolutionary stages of the Universe, from the formation of novae and planetary systems to their death [8–11].

2. FORMULATION OF THE PROBLEM

The physical processes of impact interactions under laboratory conditions are studied using special microparticle accelerators, which are capable of accelerating particles with masses from 10^{-11} to 10^{-17} g to velocities of 30–100 km s^{-1} [12–16]. In laboratory experiments of this class, apart from other results, it was also shown that the microparticle mass composition could be determined from the spectrum of the impact-produced plasma ions [17]. Special-purpose mass spectrometers were developed to study the microparticle mass and isotopic compositions in in situ experiments in space; the impact plasma generated during a hypervelocity impact of a micrometeorite or a dust particle on the target served as the source of ions in these instruments [18]. Thus, for example, the microparticle composition in the gas–dust cloud of Halley's comet was first determined using the PUMA and PIA dust-impact instruments on the Vega and Giotto space missions [19–21]. The microparticle collision velocities in these studies reached 80 km s^{-1}, which provided conditions of $V_{\rm sh} > V_{\rm cr}$ and $t_{\rm int} < t_h$. The overall picture of the collision processes considered above is shown as a plot of the body's size against its velocity in Fig. 1.

The laboratory and space experiments aimed at studying impact interactions or determining the microparticle mass composition should be placed into the group of direct impact experiments in which the hypervelocity microparticle impact was investigated in real time. Such experiments can also be placed into the class of unique or exotic experiments both in cost and in the possibility of realizing initial conditions. However, there are also other, simpler and more accessible methods of studying impact processes in the arsenal of modern science. These include simulations.

Numerical simulations play an important role in the studies of hypervelocity impacts. This is suggested by the results obtained in a series of works on the simulations of spacecraft shield destruction [22] or ionization processes for the dust-impact instruments used in an in situ experiment in the studies of Halley's comet [23]. However, the most promising experimental direction is associated with laboratory simulations of a hypervelocity impact using a laser [24]. It is based on the capabilities of a Q-switched laser to generate a light pulse of duration $t_l \sim 1$–10 ns and to provide a power density of up to 10^{13} W cm^{-2} in a spot 10^2–10^3 μm in diameter. The processes that arise in kinematic and laser impacts can significantly differ in simulations of this class [24]. However, the processes after the plasma generation, in particular, at the stage of adiabatic expansion, cooling, and recombination, must be similar [3].

Our choice of parameters for the laser simulations of impact processes was based on the reliable results obtained during the in situ dust-impact experiments on the Vega mission [18–20]. According to these results, at a impact velocity of about 80 km s^{-1} for microparticles with diameters from 0.01 to 1 μm and masses from 10^{-18} to 10^{-12} g, the ionization multiplicity of the plasma ions was $\alpha \sim 1$ and their energy was $E_i \sim$ 50–70 eV.

The plasma nature of these ions was confirmed in direct numerical simulations [25–27] of the mass-spectrometric cometary dust measurements made with the PUMA-1 and PUMA-2 instruments. It was also shown in these papers that under the above input conditions of the in situ experiment, a plasma was generated for particles with diameters down to 0.01 μm.

To properly determine the impact-simulating laser parameters, let us consider plasma generation during impact. It is well known that, depending on the velocity, the volume of the impact crater is larger than the microparticle volume by approximately a factor of approximately 100–500 [28, 29] and that plasma is generated in a volume that exceeds the microparticle volume by a factor of 5–10. It thus follows that the energy contribution to the plasma generation during the impact does not exceed 1% of the particle energy and that it is spent mainly on the crater formation and the target deformation. This is the reason why the ionization multiplicity and the ion energy are relatively low for the impact plasma, although the power density calculated from the particle diameter is about 10^{13} W cm^{-2} [30].

The dependence of the ionization multiplicity and the ion energy on the power density for a laser plasma has been adequately studied [31–33]. Thus, as the power density increases from 10^9 to 10^{11} W cm^{-2}, the ionization multiplicity for Al increases from 1 to 5 and the particle energy increases from approximately 100 eV to 2 keV. Consequently, the energy contribution to the plasma generation can be accurately determined from these parameters.

Thus, a similarity of the physical processes during the plasma expansion should be ensured when choosing the laser parameters.

The criteria for choosing the laser parameters can be briefly formulated as follows. The laser must ensure, first, the plasma generation, second, identical parameters of the laser and impact plasmas or, more specifically, the ionization multiplicity and the ion energy, and, third, equal numbers of produced ions or equal plasma-generation volumes for these processes.

The first condition constrains t_i, in particular, $t_i < t_l < t_h$, where t_i is the plasma generation time. For a laser with t_l in the range from 0.1 to 10 ns, this condition is satisfied with a comfortable margin.

For the parameters of the impact and laser plasmas to be identical, the identical energy contributions must be ensured. Thus, at a laser power density of $W \sim 10^9$ W cm^{-2}, the plasma ions will be singly ionized and their mean energy will be about 50–70 eV, in close agreement with the impact plasma parameters, at a impact velocity of 80 km s^{-1}.

The number of laser-produced ions at a given power density will depend on the diameter and depth of the laser crater. We can also determine the plasma-generation volume from these quantities and calculate the diameter of the simulated microparticle from these data. Conversely, we can choose the diameter of the laser crater from the particle diameter. Thus, for example, a microparticle impact on a target 10 μm in diameter produces a plasma from an equivalent volume approximately equal to 10^4 μm^3. The plasma-generation volume is the same as that for a laser crater 50 μm in diameter and 3 μm in depth (in a carbon target).

Thus, by varying the power density and the laser crater diameter, we can reproduce the basic parameters for a given microparticle diameter and velocity.

The choice of impact-simulating laser parameters is nothing but the determination of similarity parameters [34] with elements of limited simulation principles [35] without changing the spatial scale. This approach should be used to extend the results obtained for microparticles to large celestial bodies.

The identity of the ion-generation processes in the laser and impact plasmas is evidenced by the fact that such a plasma is used as the source of ions for time-of-flight mass spectrometers, laser [36] and dust-impact [18] ones. In the former instrument, the mass composition of the target is determined when ions are generated in the laser plasma; in the latter instrument, the mass compositions of both the impinging particle and the target are determined through the production of ions in the impact plasma. In these instruments, which are very similar in principle of operation and, often, in design, the hot plasma produced during an ultrafast energy concentration serves as the source of ions.

A laser plasma was commonly used as the source of ions for dust-impact instruments during their testing and laboratory calibrations. In particular, when a mockup of the PUMA flight instrument was tested using a laser source of ions, the effect that caused a decrease in the mass resolution of the instrument at a high plasma density, which corresponded to the falling of a large dust particle on the target, was detected and investigated [21]. The laser simulation results formed

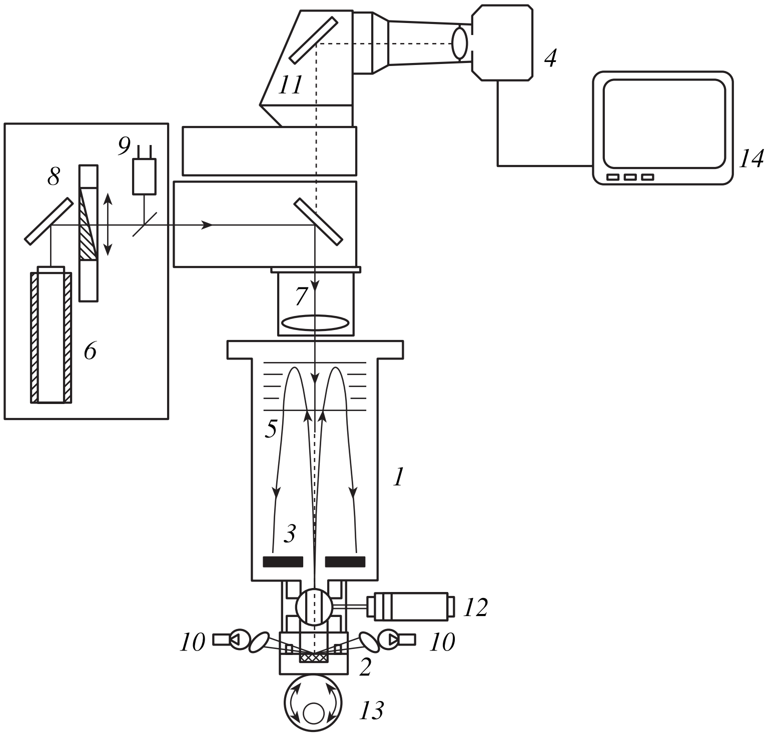


Fig. 2. A scheme of the LASMA time-of-flight laser mass reflectron: *1*—vacuum chamber, *2*—interchangeable carriage with sample, *3*—detector, *4*—TV camera, *5*—reflector, *6*—laser, *7*—focusing lenses, *8*—neutral filter, *9*—radiation power meter, *10*—target illuminator, *11*—microscope, *12*—motorized vacuum valve, *13*—carriage adjusting system, *14*—monitor.

the basis for upgrading the flight instrument; a narrow-energy-window mode was introduced, which ensured that unique mass spectra of large dust particles were obtained.

The synthesis of molecules or molecular ions during the expansion and cooling of a impact plasma [37, 38] is of great interest in analyzing the synthesis of organic materials at the early evolutionary stage of the Earth, the falling of meteorites, and hypervelocity collisions of dust particles in molecular, interstellar, and gas–dust clouds because of their similarity.

3. DESCRIPTION OF THE EXPERIMENT

We used the LASMA time-of-flight laser mass reflectron shown in Fig. 2 [36, 39–42] as the basic instrument for our laser simulations of an impact. It consists of a laser and an analyzer to determine the masses of the laser-generated ions. We used a Q-switched infrared Nd:YAG-glass laser at $\lambda \sim 1.06$ μm, which provided a power density on the order of 10^9–10^{11} W cm^{-2} in a spot 30–100 μm in diameter with a pulse duration of 5–7 ns. These laser impact parameters corresponded to the impact of a microparticle 5–10 μm in diameter at $V_{\mathrm{sh}} \approx 80$ km s^{-1}. The plasma ions flew apart in a vacuum without additional particle acceleration in the so-called free-expansion regime. The ions reflected in the reflector field and fell on the detector. The instrument was axially symmetric: the laser beam, the reflector, the detector, and the target were aligned on the same axis, and the laser radiation was orthogonally brought to the target plane through reflector grids and a central hole in the detector [43, 44]. Since this configuration provided completely symmetric plasma expansion, a high reproducibility of the spectra was achieved. The analytical part of the instrument provided a sensitivity of about 1 ppm for one laser pulse with a mass resolution of 600 at the halfwidth of the mass peak.

We took into account the special role of carbon in the synthesis of organic molecules [6, 7] and its ability to combine into polyatomic structures. Therefore, we used carbon-rich materials as the target, which also increased the yield probability of the synthesized organic molecules.

Specially purified carbon designed for the electrodes of spectroanalytical instruments, in which the impurities did not exceed 0.1 ppm, was used as the target in the first series of experiments. The high purity of the sample was confirmed by measurements for all of the elements except hydrogen, which was detected in the mass spectra. The auxiliary experiments aimed at localizing the source of these ions showed that hydrogen was both in the bulk of the sample and in its surface monomolecular layer. The hydrogen impurity played a positive role.

Singly ionized molecules of methane, acetylene, ethylene, propylene, allene, as well as carbon polymer C_NH_M-type structures containing from 1 to 40 carbon atoms and from 1 to 4 hydrogen atoms were detected and identified in the mass spectra obtained on a carbon target and shown in Fig. 3. These spectra were highly reproducible. Subsequently, such hydrocarbon structures and polyatomic molecules were also obtained with other carbon-containing materials: industrial carbon, sinking oil fraction, and bitumen.

To elucidate the role of Si in the synthesis, we carried out experiments on a high-purity Si target under similar input experimental conditions. Our mass spectra showed that Si was also capable of producing a polymer structure, but the maximum number of Si atoms in the chain did not exceed 11.

Of particular interest were the mass spectra obtained for a sinking oil fraction at a power density in the range 5×10^9–10^{10} W cm^{-2} [45][1] (see Fig. 4). Apart from the mass peaks that correspond to carbon ions with an ionization multiplicity from 1 to 4, they exhibited the peaks of ions that were identified as multiply ionized hydrocarbons: doubly ionized CH_2, CH_3, CH_4; triply ionized CH, CH_3, CH_4, CH_5, C_2H, C_2H_3; and quadruply ionized CH_4, C_2H, C_2H_5, C_3H_8. It may be assumed that by the time of their detection, ions of this type had been observed in a laser plasma for the first time. The generation of such ions is associated with the processes of charge exchange and stripping, which can proceed in a laser plasma. According to theoretical calculations and experimental results, the lifetime is long enough for these ions to be detected with a time-of-flight mass analyzer [46].

To synthesize more complex organic molecules containing the basic gas-forming elements H, C, N, and O, we carried out a series of experiments in which pure inorganic materials containing these elements were mixed with a carbon powder. The NH_4NO_3, $Ni(NO_3)_2$, Ti_2O_3, and $NaNO_3$ materials were used. The main goal of these experiments was to synthesize and detect the organic molecules that could be used in the natural synthesis of amino acids or biopolymers. The results of these measurements are presented in Table 1.

The task proved to be arduous, because the mass peaks of the elements and their isotopes were superimposed on the peaks of the organic materials being studied; thus, for example, the CN^+ and HCN^+ mass peaks were superimposed on the $C_2H_2^+$ (m = 26) and Al^+ (m = 27) peaks. The absence of nitrogen-containing salts and oxides of special purity as well as materials containing the stable C^{13} and N^{15} isotopes required to shift the mass peaks of the organic materials made it much more difficult to obtain “pure” peaks. To detect molecular peaks, we examined those areas of the mass scale where there were no peaks from the elements or where atypical strong violations of the isotopic ratios were observed. Several such spectra are shown in Fig. 5. In this series of experiments, together with the peaks of the atomic ions of H, C, N, O, Na, Ti, Ni and the above $C_NH_M^+$-type carbon structures, we obtained the mass peaks corresponding to the NH_3, OH^+ or NH^+, H_2O^+ or NH_4^+, H_3O^+, CO^+, N_2^+ molecular ions using the amino acids that were synthesized in the laboratory for the first time [47, 48]. The combination of these materials with hydrocarbons could primarily provide the synthesis of organic materials whose mass peaks were also observed experimentally: CH_2NH^+ (m = 29), H_2CO^+ (m = 30), $C_2H_2O^+$ (m = 42), $HNCO^+$ (m = 43), $NHCHO^+$ (m = 44), and $CH_2O_2^+$ (m = 46). We cannot rule out the possibility that some of these mass peaks also include the N_2O^+, NO_2^+, CO^+ materials of the same mass whose contributions could not be determined here.

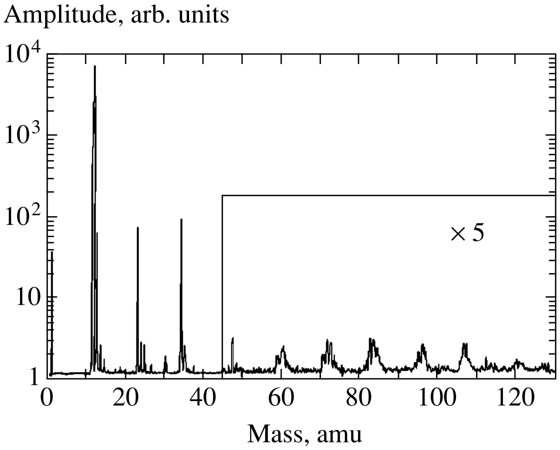


Fig. 3. The mass spectrum of polyatomic ions for a carbon target.

To confirm the plasma nature of the polyatomic ions, we carried out an experiment in which a laser acted on a mixture of carbon and silicon powders as well as carbon and tungsten powders. In such a mixture, the synthesis of polyatomic molecules containing Si and C or W and C at the powder mixing stage can be ruled out. We detected the mass peaks of polyatomic ions shown in Fig. 6 that correspond to Si_m^+ up to m = 11 and $Si_mC_n^+$, $Si_mC_nH_p^+$, where m was varied from 1 to 7 and n and p were varied from 1 to 3. Ions WC, WC_2, WC_3, and W_2C were also obtained in this series of experiments. The presence of such ions can be explained only by their synthesis in a plasma torch during the plasma expansion and cooling and suggests that organic materials could also be synthesized in processes of this kind.

---

[1] The spectra were taken back in 1990 and included in the 1992 report.

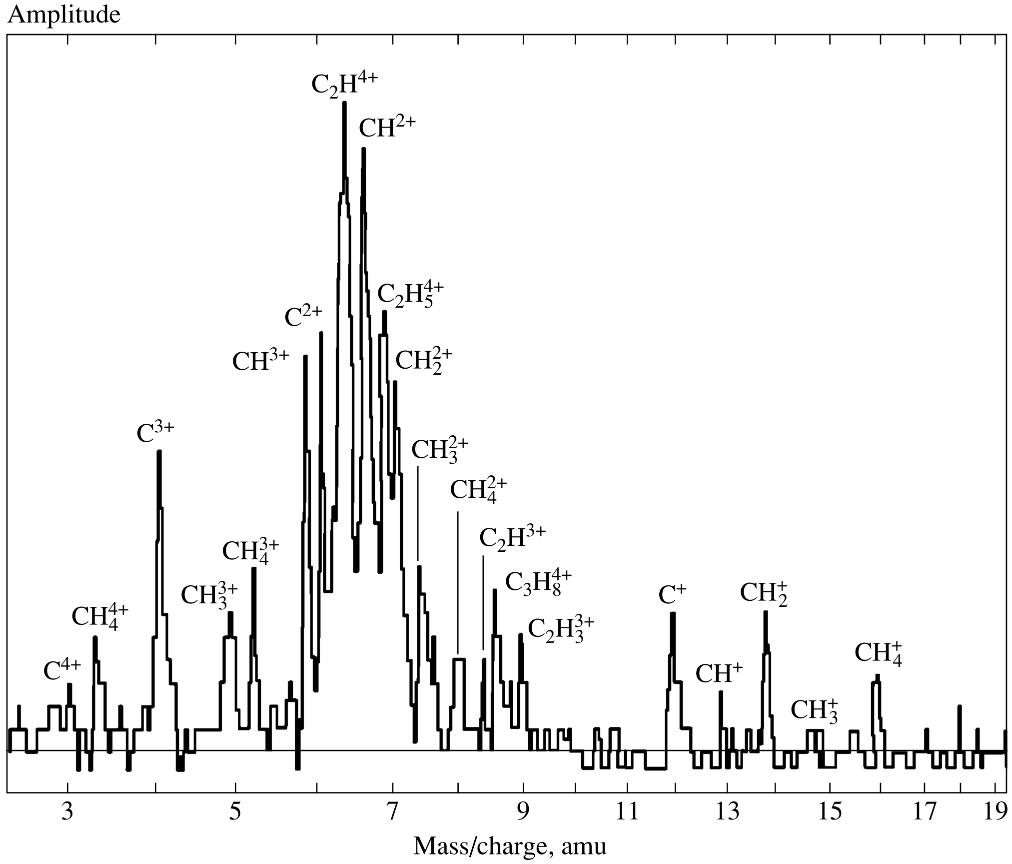


Fig. 4. The mass spectrum of multiply ionized carbon and hydrocarbon atoms.

The relative yield of the synthesized polyatomic ions was determined from our mass spectra. For light elements (from H to Ca), it ranged from 0.1 to 1% of the number of atomic ions.

The yield efficiency of the polyatomic ions for an energy expenditure of about 1 eV was determined from the condition that the energy required for the dissociation and ionization of one atom is 25–30 eV. Therefore, for the formation of one molecular ion, $\eta = 10^{-3}(30)^{-1} = 3 \times 10^{-5}$ molecules will be synthesized for a relative yield of about 0.1% and an energy expenditure of 1 eV.

The relative yield of (0.1–1)% must also be retained for the neutral component. The ionization coefficient decreases from 1% to 0.1% in a recombining plasma after its cooling and, consequently, the contribution of neutral atoms becomes substantial. The studies carried out in [49] show that the neutral component of the laser plasma has the same nature as the ion component. Therefore, one might expect with a high probability that the molecules synthesized during the plasma expansion in the same ratio are also present in the neutral component of the recombining plasma. This implies that the total number of synthesized molecular ions and neutral molecules will be comparable to the number of atomic ions.

Thus, our laser simulations of a hypervelocity impact showed that organic compounds could be synthesized from inorganic materials. The laser power density for most of the experiments was 10^9 W cm^{-2} for a laser spot diameter of 30–50 μm. According to the similarity criteria, the parameters of the generated plasma were identical to those of the impact plasma at a microparticle velocity of 80 km s^{-1} and a diameter of 5–10 μm.

4. DISCUSSION

Let us consider the most important results and compare them with the data obtained by other authors.

First, consider the results obtained during the in situ experiment aimed at studying Halley's comet, in which organic molecules were detected with the PUMA instrument [50]. The authors of this paper explain the appearance of molecular ions by the impact absorption of the organic materials frozen on the dust particle surface at the time of their impact on the target. In light of the foregoing, judging by the input experimental conditions, these results can be explained by the synthesis of organic molecules during the hypervelocity impact of a dust particle that contains elements forming organic materials, including those frozen on the particle surface. In this case, the impact synthesis is also suggested by the similarity of the organic materials obtained in [5] and in our study. Therefore, it would be of interest to reconsider, possibly simultaneously, these data because of their great importance, by taking into account the simulation results.

In [51, 52], the laser-produced fullerenes and metallofullerenes were controlled with a mass spectrometer. An important feature of these studies is that the laser acted on a carbon target in two regimes: in the Q-switched regime with $t_{\text{int}} \approx 8$ ns and in the free-generation regime with $t_{\text{int}} \approx 230$ μs at a wavelength of 1.06 μm. In this case, the laser power density for the three presented typical mass spectra was 10^5 ($t_{\text{int}} \approx 230$ μs), 5×10^9, and 1.5×10^{10} W cm^{-2} ($t_{\text{int}} \approx 8$ ns).

According to the results obtained and their interpretation, fullerenes were produced both in the gaseous phase at a low power density of 10^5 W cm^{-2} and in the plasma phase at a high power density of 1.5×10^{10} W cm^{-2}. However, we clearly see from our spectra that the generation of fullerenes is different in nature. In the former case, a low-temperature gas-vapor phase takes place; in the latter case, a powerful plasma torch is undoubtedly generated and fullerenes together with more complex molecular ions are intensely synthesized during the plasma expansion and cooling.

The spectrum taken at a power density of 5×10^9 W cm^{-2} closely matches the spectra presented here both in the structure of the hydrocarbon compounds and in the number of carbon atoms in the polyatomic chains.

The above comparison of the experimental results leads us to conclude that at low power densities and for $t_{\text{int}} > t_h$, carbon clusters and polyatomic ions evaporate from the surface and then combine into more complex structures, being in the gas-vapor phase. A qualitative jump occurs when the power density is increased to about 10^9 W cm^{-2} and for $t_{\text{int}} < t_h$, because in this case, after the complete atomization and ionization of the material, its atoms and ions are associated into polyatomic structures during the plasma expansion and cooling.

The most interesting result of [51] is that its authors clearly showed the complication of the molecules with increasing power density. Thus, for example, when the power density was increased from 5×10^9 to 15×10^9 W cm^{-2}, the number of carbon atoms in the polymer structures increased from 31 to 300. A similar effect may also be expected when increasing the spatial size of the impact area.

Table 1. The mass peaks of the experimentally detected polyatomic ions and their interpretation

Mass	Atomic ions	Molecule based on			Name
		C	N	O	
12	C	–			
13	C	CH			
14	N	CH_2			
15	–	CH_3	NH		
16	O	CH_4	NH_2		methane
17	–	–	NH_3	OH	ammonia
18	–	–	NH_4	H_2O	
19	F	–		H_3O	
24	Mg	C_2			
25	Mg	C_2H			
26	Mg	C_2H_2	CN?		acetylene
27	Al	C_2H_3	HCN?		
28	Si	–	N_2		
29	Si	–	CH_2NH		carbamide
30	Si	CH_2O		NO	formaldehyde
34	S	–	–	H_2O_2	
36		C_3			
37	Cl	C_3H			
38	–	C_3H_2			
39	K	C_3H_4			
40	Ca	C_3H_4	C_2H_2N?		allene
41	K	–	C_2H_3N		
42	–			C_2H_2O	ketene
43	–		NHCO		
44	Ca	C_2H_4O	HNCHO		acetaldehyde
46	Ti	C_2H_6O	NO_2	CH_2O_2	ethanol/formic acid
48	Ti	C_4			
49	Ti	C_4H			
50	V	C_4H_2			
51	V	C_4H_3			
60		C_5			detected
72		C_6			also
84		C_7			C_nH_m
96		C_8			n up to 10
108		C_9			m up to 14
120		C_{10}			

There are several factors that directly or indirectly confirm the synthesis of molecules in a plasma torch. In particular, the free-expansion experimental configuration rules out the fall of ions on the detector from the peripheral zone of the laser impact, where the power density may be too low for complete atomization but high enough for the desorption of a small number of clusters or molecular ions. Therefore, we can assume that in the free-expansion regime or in the regime when there is no external, artificially applied accelerating electric field, only completely atomized plasma ions

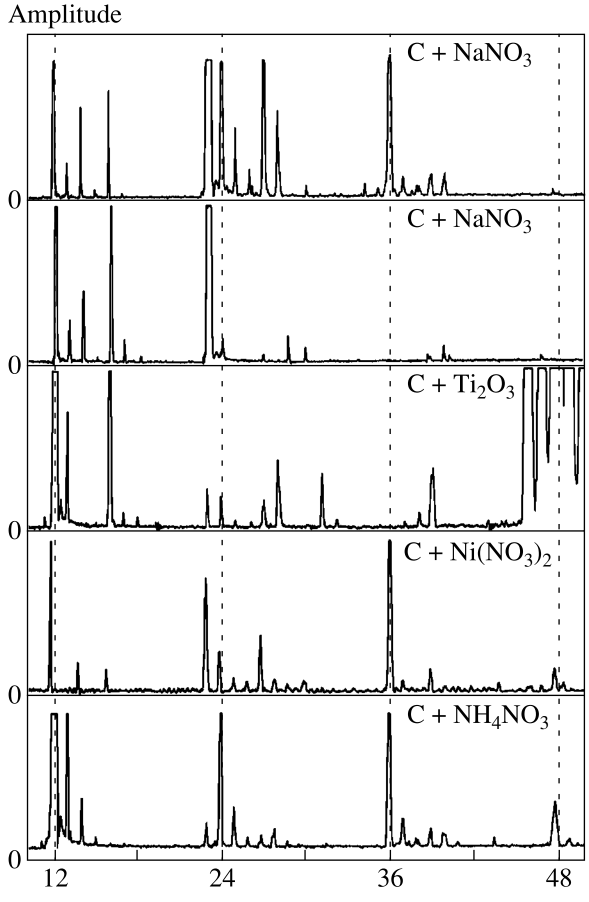


Fig. 5. The mass spectra for targets containing a mixture of a carbon powder with the various salts and oxides indicated in the spectra.

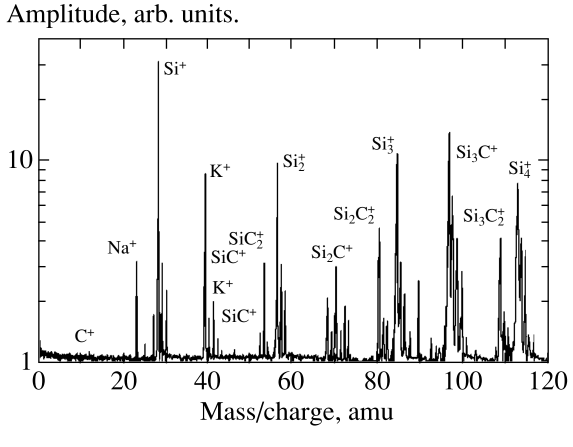


Fig. 6. The mass spectrum for polyatomic ions of a mechanical mixture of C and Si powders.

from the central hot zone of the plasma formation are involved in the ionization and subsequent acceleration.

An important argument for the plasma synthesis is that the experimentally observed multiply ionized hydrocarbons cannot be produced in thermal processes during successive ionization. They can be produced only when a multiply ionized carbon atom combines, for example, with neutral hydrogen. This is because the hydrocarbon binding energy $E_h \ll E_{\mathrm{C}}^{2+}$, where E_{C}^{2+} is the carbon double ionization energy, being ~35 eV.

However, the experiments with two-component powders most clearly show the plasma nature of the produced molecular ions. In this case, the plasma torch is the most likely area of the combination of the components to form more complex two-component molecules.

Thus, it follows from our results and the results of other authors that laser simulations of hypervelocity impacts allow the impact-plasma processes, including the plasma expansion, to be reproduced, and molecular ions can be synthesized in such a hot, nonequilibrium medium.

5. POSSIBLE APPLICATIONS OF THE RESULTS

Below, we use the results of our simulations to consider the two most important processes associated with the synthesis of organic molecules on Earth at the early stage of its formation and in interstellar gas–dust clouds.

Let us first consider the possible abiogenic synthesis of organic compounds on Earth in impacts during meteorite showers. The following two important prebiological processes are of interest in this connection: the formation of biologically important organic compounds from inorganic compounds and elements, for example, amino acids or their precursors; and the formation of short organic polymer chains.

Several physical mechanisms that are moderately simple in initial conditions were devised [6] and realized in laboratory experiments [5, 47, 48] for the former process. For the latter process concerning the formation of polymer chains, only the complex polymerization mechanism of organic compounds on a clay surface is considered [6, 53].

The formation of simple organic compounds in an atmosphere containing methane (CH_4), hydrogen (H_2), ammonia (NH_3), and water vapor under solar ultraviolet radiation or under the flux of cosmic rays, including high-energy charged particles, hydrogen and helium atoms, and electrons, were considered among the most probable processes and media in which these compounds could be synthesized on the Earth's surface [5, 53]. The so-called strongly reduced atmosphere [54] was usually considered to produce amino acids on Earth. In this medium, amino acids could also be synthesized through electric discharges, the Earth's radioactivity, and heating of the medium by volcanic activity [55] or impact-wave processes [5, 56–58].

Several early laboratory experiments that simulated the above processes confirm the formation of organic materials [47, 48].

Matsu and Abe [56] investigated the possible effects of a hypervelocity meteorite impact but only for the impact degassing of a planetesimal and the generation of a hot atmosphere. According to the authors of this hypothesis, such a impact-induced atmosphere could contain the CO and CO_2 compounds needed for the synthesis of amino acids.

In the experiments described in [59, 60], a gaseous mixture of CO, N_2, and H_2O was subjected to an arc discharge in a magnetic field. As a result of this discharge, amino acids together with uracil and cytosine were detected in the hydrolysate of the product.

Experiments that simulated the effects of the impact waves generated when meteorites entered the Earth's primordial rarefied atmosphere have been carried out [58]. In these experiments, several amino acids were synthesized from a gaseous mixture of CH_4, CH_6, HN_3, and H_2O under the effect of impact waves in laboratory conditions.

The authors of [61, 62] were the first to show that organic materials could be produced when meteorites and rocks were intensely heated and evaporated at temperatures of 3000–4000°C. In the opinion of the authors, this heating of a material to produce a dense gas-vapor cloud could take place during the fall of a meteorite moving at a velocity of 10–15 km/s and could result both in the decomposition of the material into simple compounds and in the synthesis of new compounds when it cooled down. According to [62], under these collision conditions, up to about 30% of the material of the colliding bodies passes into a gaseous phase. The process considered above may be arbitrarily classified as the synthesis of organic molecules at subcritical impact velocities; it is of interest in its own right.

To determine the possible role of meteorite impacts, let us examine Table 2 [5], which includes data on the energy sources for the chemical evolution at the early stage of the Earth's formation. The last row of the table is entitled "Meteorite impacts," and the presumed flux is estimated to be ~10^{22} eV m^{-2} $year^{-1}$. If we use the results obtained in this study for the yield energy, about 3×10^{-5} molecules eV^{-1}, then $AB \approx 3 \times 10^{17}$ molecules m^{-2} $year^{-1}$. This value is comparable to an analogous value for an X-ray impact and the impact of cosmic rays, which were recognized to be highly efficient energy sources for the formation of organic compounds.

The velocity of the molecular ions synthesized during the laser simulation of a hypervelocity impact is lower than the particle escape velocity from the Earth. Therefore, the molecules synthesized in the plasma of a meteorite impact after its expansion must return to the Earth's surface, settling down and accumulating. In this case, the synthesis proceeds irrespective of the initial parameters of the gaseous medium and irrespective of the presence of an atmosphere and its type.

Table 2. Basic energy sources for the chemical evolution of matter at the early formation stage of the Earth's surface

Energy source	Presumed flux, *A*, eV m^{-2} $year^{-1}$	Glycine yield energy, B^a, molecules eV^{-1}	*AB*, molecules m^{-2} $year^{-1}$
Solar radiation			
total	6.8×10^{28}	~0	–
λ < 200 mm	2.2×10^{25}	~0	–
λ < 150 mm	9.1×10^{20}	~0	–
λ < 110 mm	4.2×10^{22}	$<8 \times 10^{-5\,b}$	$<3 \times 10^{17}$
electric discharge	1.8×10^{22}	7×10^{-9}	1×10^{10}
	$\sim 1.0 \times 10^{24}$		$\sim 7 \times 10^{15}$
Volcanic heat	3.4×10^{22}	~0	–
Radioactivity^c	2.0×10^{23}	~0	–
Cosmic rays	2.9×10^{21}	2×10^{-1}	6×10^{17}
Meteorite impacts	1.0×10^{22}	$3 \times 10^{-5\,d}$	$3 \times 10^{17\,d}$

Note: [a] The yield energy for a mixture of CO and N_2 (1 : 1) as the source material.

[b] For an X-ray energy of 1.5 keV.

[c] for the Earth's crust thickness of 0–1.0 km.

[d] Our results.

Thus, the simulations of a hypervelocity impact presented here indicate that organic molecules, in particular, the precursors of amino acids and short polymers could be synthesized on Earth before the appearance of a dense atmosphere under meteorite impacts if the collision velocity exceeds V_{cr}. All elements for the synthesis of organic molecules are assumed to be in the composition of the interacting bodies. This assumption is valid, because 90% of the meteorites colliding with the Earth are rocky and because there is a high percentage of carbonaceous chondrites among them with all of the necessary materials for the synthesis of organic materials available in their composition.

Let us consider an alternative possibility of the impact synthesis of molecules, including organic molecules in the interstellar medium.

The interstellar gas is mainly composed of hydrogen (~70%) and helium (28%) atoms and ions. Atoms and ions of other elements and molecules account for less than 1% and are mainly contained in dust particles.

In recent years, about a hundred molecules or molecular ions, including organic ones, have been discovered and identified in molecular gas–dust clouds by radioastronomical methods. Organic compounds, among which HCN, CH_2NH, and CH_3NH are known as the source material for the formation of amino acids, dominate in clouds of this type [7].

Two processes are currently believed to be responsible for the formation of molecules: chemical association reactions in a gaseous medium and particle surface reactions in clouds [9].

Dust in interstellar clouds is made up of either small crystals or, possibly, amorphous structures composed of silicate, graphite, and metal oxides about 10^{-2} μm in size—small or large particles up to 1 μm in size produced by the freezing of small H_2O, CO, and HN particles on the surface. Dust particles emerge mostly in the slowly outflowing atmospheres of red giants. However, there are also other hypotheses about the formation, growth, acceleration, and destruction of dust particles. Thus, for example, dust particles can be accelerated to velocities of about 10^3 km s^{-1} when they pass through the impact front produced during cloud–cloud collisions or under the radiation pressure of stars [8].

Collisions of dust particles are a major mechanism of their destruction [10], and about 10% of the dust particles can be destroyed [55] during a hypervelocity impact. However, molecules and polyatomic ions will be synthesized in these processes.

This indirectly follows from the in situ experiment aimed at studying Halley's comet, in which plasma ejections and plasma ion spectra were recorded during the impacts of microparticles 0.01 μm in diameter. The numerical simulations [25–27] of a laboratory dust-impact experiment [11–13] and the laser simulations [3] considered above show that the target material whose mass is a factor of 5–10 [30] larger than the mass of the impinging particle is drawn into the plasma generation and that the synthesized polyatomic ions account for no less than (1–0.1)% of the total number of atomic ions. This ratio is also valid for neutral atoms whose number is larger by a factor of approximately 100 than the number of ions. It thus follows that the absolute yield of the molecules and molecular ions in hypervelocity collisions of dust particles can be significant and play an important role in this process, although its probability is low.

Using laser simulations, we can show that when a dust particle 0.01 μm in diameter collides with a submicron particle, the number of molecules and molecular ions with a mean mass of about 20 amu is approximately equal to 10^3. Estimates indicate that in this time, i.e., in the particle collision time, on the order of 5×10^6 hydrogen molecules can be produced on the particle surface via the sticking of hydrogen ions. However, the number of organic molecules formed via the sticking of atoms to dust particles will be three to four orders of magnitude smaller, because the abundance of such elements as C, N, O, S, and P is three to four orders of magnitude lower than that of hydrogen, and these elements are concentrated mainly in dust particles.

Thus, the formation of organic molecules in collisions of dust particles about 0.01 μm in diameter will be comparable to or higher in efficiency than the universally accepted formation of molecules via the sticking of atoms to dust particles. As the particle diameter increases, the yield of the molecules also increases. Consequently, the synthesis of organic molecules in impact processes in the interstellar medium can play an important role and be more efficient than the processes suggested previously [9].

Returning to the laboratory simulations, it should be noted that we compared the chemical composition of the polyatomic ions produced in our simulations with that of the molecules and polyatomic ions detected in interstellar gas–dust clouds. We considered only molecules containing C, N, O, H, and Si. About 45% of the molecular ions considered above that were produced in the laboratory experiments were also observed in the form of ions and molecules in molecular clouds [63].

The experimental results presented above and their analysis lead us to make the following assumptions. First, the material from which the planets were formed could initially be enriched in the organic molecules synthesized through hypervelocity collisions. It may well be that the Solar system was also formed from matter enriched in organic material.

Second, organic materials could also be intensely synthesized in hypervelocity impacts on Earth at the early stage of its evolution and be accumulated on the surface, being shielded from the decomposition under radiation by the layer of rock ejected from an impact crater. Consequently, the synthesis of organic materials could begin much earlier, even before the appearance of a dense atmosphere and liquid water.

With the appearance of an atmosphere and water, the abundance of the organic materials in the surface layer could play a crucial role in the origin of life on Earth and its evolution.

Currently, an important and independent experimental confirmation of the concept of molecular synthesis in the plasma generated during a hypervelocity impact in the direct dust particle acceleration experiments presented here has been found. After the above results of our laser simulations of impact processes and their interpretation were reported at the European Geophysical Society 27th General Assembly [64], we simultaneously analyzed and compared our results with the results of laboratory calibrations of the onboard dust-impact mass spectrometer. This instrument is designed to study the chemical composition of micrometeorites [65] and was developed at the Max-Planck Institute for future space missions. The mass spectra of the dust-impact experiments revealed hitherto noninterpreted mass peaks of the rhodium carbide dimer obtained during the impact of a carbon-containing microparticle moving at a velocity of about 16 km s^{-1} on a target of pure rhodium. These spectra also exhibited the mass peaks of rhodium, carbon, and carbon clusters up to C_{16}. The relative yield of the dimer was no less than 10%. The only plausible explanation for the presence of a mass peak of such a compound in the spectra related to the synthesis of the dimer in plasma during a hypervelocity impact was offered.

Analysis of published papers suggests that no experimental or theoretical studies of the synthesis of organic materials in the plasma of a hypervelocity impact had been carried out previously. Such studies have been performed only in terms of laser simulations of this process in [3, 64, 66] and here. Nor have we found any numerical simulations devoted to this problem. Therefore, the results presented should facilitate the study of the molecular synthesis, both in laser plasma and in the plasma generated in impact processes, as a possible and efficient natural source for the synthesis of molecules, including organic materials and polymers, at the early stage of the Earth's formation and in interstellar clouds during hypervelocity impacts.

ACKNOWLEDGMENTS

I thank my colleagues A.A. Rukhadze, L.M. Mukhin, W.B. Brinckerhoff, E.N. Evlanov, Yu.G. Malama, M.B. Guseva, N.N. Anuchina, N.G. Bochkarev, N.V. Voshchinnikov, M.V. Gerasimov, L.A. Kuznetsova, V.I. Moroz, and G.N. Zastenker for discussions of the questions considered here. I am also grateful to A.E. Chumikov and N.G. Managadze for help in carrying out some of the experiments and data reduction

REFERENCES

1. Ya. B. Zel'dovich and Yu. P. Raĭzer, *Physics of Impact Waves and High-Temperature Hydrodynamic Phenomena*, 2nd ed. (Nauka, Moscow, 1966; Academic, New York, 1966).
2. J. F. Friichtenicht and J. C. Slattery, *Ionization Associated with Hypervelocity Impact*, NASA TN D-2091 (1963).
3. G. G. Managadze, Preprint No. PR-2037, IKI RAN (Inst. for Space Research, Russian Academy of Sciences, Moscow, 2001).
4. R. Z. Sagdeev, G. G. Managadze, A. V. Tur, and V. V. Yanovskiĭ, in *Proceedings of International Conference on the FOBOS Project* (Akad. Nauk SSSR, Moscow, 1986), p. 129.
5. K. Kobayashi and T. Saito, in *The Role of Radiation in the Origin and Evolution of the Life*, Ed. by M. Akabosh, N. Full, and R. Neverro-Gonzales (Kyoto Univ. Press, Kyoto, 2000), p. 25.
6. H. J. Goldschmidt and T. Owen, *The Search for Life in the Universe* (Benjamin, Menlo Park, 1980; Mir, Moscow, 1983).
7. S. Miller and L. Orgel, *The Origins of Life on Earth* (Prentice Hall, Englewood Cliffs, N.J., 1974).
8. L. Spitzer, Jr., *Physical Processes in Interstellar Medium* (Wiley, New York, 1978; Mir, Moscow, 1981).
9. N. G. Bochkarev, *Fundamentals of Physics of the Interstellar Medium* (Mosk. Gos. Univ., Moscow, 1992).
10. A. G. G. M. Tielens, C. F. McKee, G. G. Seab, and D. J. Holltnbach, Astrophys. J. **431**, 321 (1994).
11. N. V. Voshchinnikov, Itogi Nauki Tekh., Ser.: Issled. Kosm. Prostranstva **25**, 98 (1986).
12. W. Knabe and F. R. Kruger, Z. Naturforsch. A **37**, 1335 (1982).
13. P. R. Ratcliff, M. J. Burchell, M. J. Cole, *et al.*, Int. J. Impact Eng. **20**, 663 (1997).
14. R. Roybal, S. Stein, S. M'iglionico, and J. Shively, Int. J. Impact Eng. **17**, 707 (1995).
15. P. R. Ratcliff, F. Gogu, E. Grun, and R. Srama, Adv. Space Res. **17** (12), 111 (1996).
16. E. M. Drobyshevski, B. G. Zhukov, V. A. Sakharov, *et al.*, Int. J. Impact Eng. **17**, 285 (1995).
17. B. K. Dallman, E. Grun, and J. Kissel, Planet. Space Sci. **25**, 135 (1977).
18. J. Kissel, R. Z. Sagdeev, J. L. Bertaux, *et al.*, Nature **321** (6067), 280 (1986).
19. D. A. Mendis, in *Exploration of Halley's Comet*, Ed. by M. Growing, F. Praderie, and R. Reinhard (Springer, Berlin, 1988), p. 939.
20. R. Reinhard, in *Exploration of Halley's Comet*, Ed. by M. Growing, F. Praderie, and R. Reinhard (Springer, Berlin, 1988), p. 950.
21. R. Z. Sagdeev, J. Kissel, E. N. Evlanov, *et al.*, Astron. Astrophys. **187**, 179 (1987).
22. R. Z. Sagdeev, S. I. Anisimov, A. A. Galeev, *et al.*, Adv. Space Res. **2** (12), 133 (1983).
23. K. Hornung, Yu. G. Malama, and K. Thoma, Adv. Space Res. **17** (12), 77 (1996).
24. V. V. Kostin, V. E. Fortov, I. K. Krasyuk, *et al.*, Teplofiz. Vys. Temp. **35**, 962 (1997).
25. A. V. Petrovtsev, V. Yu. Politov, and A. T. Sapozhnikov, Preprint No. 135, RFYaTs–VNIITF (All-Russia Research Inst. of Technical Physics, Russian Federal Nuclear Center, Snezhinsk, 1998).
26. E. N. Avronin, N. N. Anuchina, V. V. Gadzhieva, *et al.*, Preprint No. 177, IPM AN SSSR (Inst. of Applied Mathematics, USSR Academy of Sciences, Moscow, 1985).
27. E. N. Avronin, N. N. Anuchina, V. V. Gadzhieva, *et al.*, Fiz. Goreniya Vzryva **32**, 117 (1996).
28. J. Kissel and F. R. Krueger, Appl. Phys. A **42**, 69 (1987).
29. D. Guring, in *High-Velocity Impact Phenomena*, Ed. by R. Kinslow (Academic, New York, 1970; Mir, Moscow, 1973), p. 468.
30. K. Hornung, Yu. Malama, and Kh. Kestenboim, Astrophys. Space Sci. **279**, 256 (2000).
31. O. A. Bykovskiĭ and I. N. Nevolin, *Laser Mass Spectrometry* (Énergoizdat, Moscow, 1985).
32. G. F. Tonon, Acad. Sci., Paris **262**, 1413 (1965).
33. N. G. Basov, V. A. Boĭko, and Yu. A. Drozhbin, Dokl. Akad. Nauk SSSR **192**, 1248 (1970) [Sov. Phys. Dokl. **15**, 576 (1970)].
34. G. G. Managadze and I. M. Podgornyĭ, Geomagn. Aeron. **8**, 609 (1968).
35. G. G. Managadze and I. M. Podgornyĭ, Dokl. Akad. Nauk SSSR **180**, 1333 (1968) [Sov. Phys. Dokl. **13**, 593 (1968)].
36. G. G. Managadze and I. Y. Shutyaev, in *Laser Ionization Mass Analysis*, Ed. by A. Vertes, R. Gijbels, and

F. Adams (Wiley, New York, 1993), p. 505, Chem. Anal. Ser., Vol. 124.

37. Yu. A. Bykovskiĭ, G. I. Zhuravlev, V. I. Belousov, *et al.*, in *Production and Analysis of Materials of Special Purity*, Ed. by A. D. Zorin (Nauka, Moscow, 1979), p. 276.
38. C. R. Phipps and R. W. Dreyfus, Chem. Anal. (N.Y.) **124**, 369 (1993).
39. G. G. Managadze and N. G. Managadze, Zh. Tekh. Fiz. **69** (10), 138 (1999) [Tech. Phys. **44**, 1253 (1999)].
40. N. G. Managadze, G. G. Managadze, and A. Ziegler, in *Proceedings of 45th Conference of ASMS* (Palm Springs, USA, 1997), p. 1243.
41. G. G. Managadze and N. G. Managadze, Preprint No. Pr-1962, IKI RAN (Inst. for Space Research, Russian Academy of Sciences, Moscow, 1997), p. 22.
42. W. B. Brinckerhoff, G. G. Managadze, R. W. McEntire, *et al.*, Rev. Sci. Instrum. **71**, 536 (2000).
43. G. G. Managadze, RF Patent No. 1732396 (1992), Byull. Izobret., No. 17 (1992).
44. G. G. Managadze and N. G. Managadze, RF Patent No. 2096861 (1997), Byull. Izobret., No. 32 (1997).
45. G. G. Managadze, *Universal Multi-Purpose Transportable Mass-Spectrometric Complex*, Report (APTI, Washington, 1992).
46. S. R. Andrews, F. M. Harris, and D. E. Parry, Chem. Phys. **166**, 69 (1992).
47. H. C. Urey, Proc. Natl. Acad. Sci. USA **38**, 351 (1952).
48. S. L. Miller, in *Mineral Deposits and Involution of the Biosphere*, Ed. by H. D. Holland and M. Schidlowski (Springer, Berlin, 1982), p. 155.
49. Yu. A. Bykovskiĭ, N. M. Vasil'ev, I. D. Laptev, *et al.*, Zh. Tekh. Fiz. **44**, 2023 (1974) [Sov. Phys. Tech. Phys. **19**, 1258 (1974)].
50. J. Kissel and F. R. Krueger, Nature **326** (6115), 755 (1987).
51. R. Zhang, Y. Achiba, J. K. Fisher, *et al.*, J. Phys. Chem. **103**, 9450 (1999).
52. L. Laska, J. Krasa, L. Juha, *et al.*, Carbon **34**, 363 (1996).
53. C. Sagan and B. N. Khare, Science **173**, 417 (1971).
54. V. I. Moroz and L. M. Mukhin, Kosm. Issled. **5** (6), 901 (1977).
55. L. M. Mukhin, Nature **251**, 50 (1974).
56. T. Matsu and Y. Abe, Nature **319**, 303 (1986).
57. S. L. Miller and H. C. Urey, Science **130**, 245 (1959).
58. A. Bar-Nun, N. Bar-Nun, S. H. Bauer, and C. Sagan, Science **168**, 470 (1970).
59. S. Miyakawa, K. Kobayashi, and A. B. Sawaoka, Jpn. J. Appl. Phys. **36**, 4481 (1997).
60. S. Miyakawaet, K. Kobayashi, and A. B. Sawaoka, J. Am. Chem. Soc. **121**, 8144 (1999).
61. L. M. Mukhin, M. V. Gerasimov, and E. N. Safonova, Nature **340**, 46 (1989).
62. M. V. Gerasimov, L. M. Mukhin, and É. N. Safonova, Izv. Akad. Nauk SSSR, Ser. Geol., No. 4, 119 (1991).
63. V. S. Strel'nitskiĭ, in *Little Encyclopedia: Physics of Space* (Sovetskaya Éntsiklopediya, Moscow, 1989), p. 415.
64. G. G. Managadze, in *Proceedings of 27th General Assembly of the European Geophysical Society, Nice* (2002), Abstract EGS02-A-06871, p. 334.
65. M. Stubig, R. Srama, E. Grun, and G. Schafer, in *Proceedings of 27th General Assembly of the European Geophysical Society, Nice* (2002), Abstract EGS02-A-01364, p. 274.
66. G. G. Managadze, Geophys. Res. Abstr. **3**, 7595 (2001).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 61–69.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 70–79.
Original Russian Text Copyright © 2003 by Garanin, Amelicheva, Burenkov, Ivanova, Sofronov.

**PLASMA,
GASES**

Relaxation of a 2D MHD Flow across a Magnetic Field (2D Hydrodynamic Flow) in a Bounded Region

S. F. Garanin*, O. A. Amelicheva, O. M. Burenkov, G. G. Ivanova, and V. N. Sofronov

*All-Russia Research Institute of Experimental Physics, Russian Federal Nuclear Center,
Sarov, Nizhegorodskaya oblast, 607190 Russia*

**e-mail: sfgar@vniief.ru*

Received November 25, 2002

Abstract—The problem on magnetohydrodynamic (MHD) flow of a solitary vortex across a magnetic field in a volume confined by rigid walls is solved numerically for large Reynolds numbers (including magnetic Reynolds numbers) and small Alfven-Mach numbers M_A. In this case, the MHD problem is reduced to that of two-dimensional hydrodynamic turbulence. It is shown that sound is not generated by a turbulent medium for small values of M_A; consequently, this kinetic energy dissipation channel is closed in this case. Calculations show that, in contrast to 3D turbulence, kinetic energy dissipation for 2D turbulence occurs, as expected, over time periods on the order of L^2/ν (L is the characteristic size of the system and ν is the kinematic viscosity). In our calculations with numerical viscosity $\nu \sim v\Delta x$ (Δx is the unit cell size), this corresponds to time values on the order of $\sim(L/\Delta x)(L/v)$. In the kinetic energy spectra for a turbulent flow in a bounded region in the inertial interval (lying between the energy-carrying and viscosity regions), the values of $E(k)$ decrease with increasing wave numbers k at a higher rate than in proportion to k^{-3}. The volume distribution of vorticity becomes narrower with time (the characteristic values of curl**v** decrease) and is blurred; for large time periods, the distribution approximately retains its shape as well as asymmetry with respect to positive and negative values, which is associated with the asymmetry of the initial conditions. *© 2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

Two-dimensional magnetohydrodynamic (MHD) flows of plasma across a magnetic field play a significant role in many dynamic plasma systems with a magnetized plasma. By way of an example, we can consider a magnetic compression (MAGO) system [1–3] providing an approach to controlled nuclear fusion, which consists of two stages.

1. First, a magnetized hot plasma suitable for subsequent compression (with an azimuth magnetic field of about 0.1 MG, a concentration on the order of 10^{18} cm^{-3}, a temperature of about 300 eV, and a low impurity concentration since impurities may increase radiative loss) is created in compartment *2* of a special toroidal MAGO chamber consisting of two compartments connected through a narrow annular nozzle (Fig. 1).

2. Then the plasma is compressed with the help of high-power magnetic drivers (such as magnetic explosion generators) quasiadiabatically (at rates on the order of 1 cm/μs) to parameters corresponding to the Lawson criterion.

At both stages, the plasma flow occurs in the *rz* plane perpendicular to the azimuth magnetic field.

Plasma is heated at the first stage when it is expelled by a magnetic piston from compartment *1* to compartment *2*. The initially cold plasma in the nozzle region is accelerated to supersonic velocities (exceeding the Alfven velocity) and is heated during deceleration in collisionless shock waves formed at the nozzle exit [4] and as a result of anomalous viscous heating in near-electrode layers [5]. In this way, the magnetic energy of the plasma is transformed into kinetic and then thermal energy. After the passage of the plasma to the second compartment and leveling out of the total pressure in the two compartment, a comparatively calm plasma with $\beta \approx 1$ (β is the ratio of the thermal pressure to the

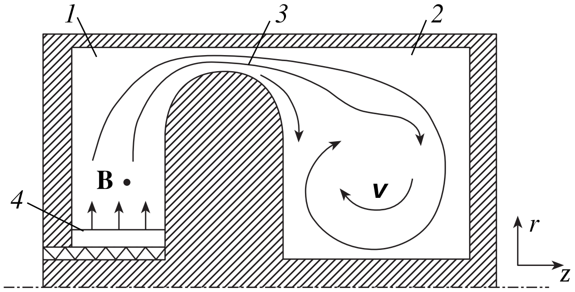


Fig. 1. Schematic diagram of a MAGO plasma chamber: **B** is the magnetic field, *1* and *2* are compartments, *3* is the annular nozzle, and *4* is the magnetic piston.

magnetic pressure) and with essentially subsonic velocities is formed in compartment *2*.

It is this plasma that is intended for compression in a quasispherical or cylindrical manner by moving the walls to the second compartment of the chamber (e.g., by displacing the cylindrical outer wall of the chamber shown in Fig. 1 in the inward direction along the radius). In spite of comparatively low velocities of the plasma and the small value of its kinetic energy as compared to the magnetic energy, these velocities are still higher than the compression velocities; such a motion of the plasma is interesting in several aspects. First, the plasma flow may affect the convective cooling of the plasma, carrying heat from hot plasma regions to cold walls. Second, the flowing plasma can be contaminated by impurities washed away from the walls. This erosion of the wall material may become especially strong when the plasma is compressed for initiating thermonuclear fusion [1–3] since the substance of the walls in this case is obviously in the plasma state and is easily miscible with the hydrogen plasma. Consequently, it is important to know the evolution of the hydrodynamic flow after the plasma heating stage and the lifetime of this flow.

Classical transport coefficients of a hot magnetized plasma [6] such as viscosity and magnetic diffusion coefficients are small due to magnetization. For this reason plasma flows are characterized by large Reynolds numbers and magnetic Reynolds numbers and become turbulent, as is almost always the case with large Reynolds numbers. Since the plasma flow across the magnetic field occurs at moderate Alfven-Mach numbers M_A, MHD instabilities also develop across the magnetic field [7] and the arising turbulence is two-dimensional by nature [8]. In direct 2D MHD computations (see, for example, [2, 9]) made in a specific special geometry with specific boundary and initial conditions taking into account the rich variety of physical effects, it is difficult to obtain answers to fundamental questions concerning the relaxation of a 2D flow. This is due to the fact that (i) we have to single out phenomena of interest against the background of all other factors and (ii) the inclusion of all factors inevitably lowers the computational potential used to study the 2D flow as well as phenomena directly related to the flow. Consequently, it is expedient to consider separately a 2D turbulent MHD flow without taking into account insignificant phenomena in this flow. Since the MHD problem is reduced in this case to the problem of 2D hydrodynamic turbulence, its main features can be analyzed both in the MHD formulation and in the 2D hydrodynamic formulation.

Following [10], we carry out the numerical simulation of flows in a bounded region for large Reynolds numbers to determine the parameters of these flows and the kinetic energy dissipation rates. In our computations, possible dissipation mechanisms are viscosity[1] associated with the difference scheme and generation of sound with its subsequent nonlinear damping in shock waves.

If the generation of sound were possible, it could be an important channel of kinetic energy dissipation for a 2D turbulent flow since other mechanisms are slow in this case in view of the smallness of transport coefficients (viscosity and magnetic diffusion) and since the energy loss rate due to generation of sound (see [11] and Section 2) should be determined only by Mach numbers M (which are smaller than unity for real flows after the heating stage in the MAGO chamber, but not too small, $M \approx 0.4$). In order to explore the possibility of generation of acoustic waves by turbulence in a bounded volume for small Mach numbers M (we denote by M the total Mach number, which virtually coincides with M_A in our computations), it is sufficient to analyze this possibility in a simplified formulation (for a 1D problem) in which a turbulent flow plays the role of the driving force producing compressions and expansions in the volume. The results of this investigation are described in Section 2.

The relaxation of a 2D flow is investigated for the evolution of a solitary circular vortex in a square box. Since the initial conditions have to be “forgotten” to a certain extent in the course of flow evolution, these conditions are not very significant for determining parameters appearing at late stages of turbulence (e.g., the exponent of the energy spectrum in the inertial interval). It would be interesting, however, to establish the characteristic times of a transition to turbulence (the order of magnitude of these time intervals have to be independent of the initial conditions and the shape of the region in question) and their ratio to the times of kinetic energy decay due to viscosity.

2. GENERATION OF SOUND BY TURBULENCE IN A BOUNDED REGION

A possible channel for kinetic energy dissipation in turbulent flow is the generation of acoustic waves (or magnetoacoustic waves in the MHD case) and their subsequent shock-wave damping. The theory of sound generation by turbulence occupying a finite volume in an unbounded medium was developed by Lighthill [7]. This theory is applicable when the size of the medium is large with characteristic acoustic wavelengths excited in a turbulent flow. If turbulence occupies a volume with a characteristic size L and characteristic velocities are v, the characteristic wavelengths are on the order of cL/v (c is the velocity of sound, $c \gg v$) are

---

[1] For a magnetized plasma with $(\omega\tau)_i \gg 1$, energy dissipation due to viscosity is stronger than ohmic dissipation by a factor m_i/m_e (for small M_A, the relative contribution of ohmic dissipation is even smaller). However, the specific mechanism of dissipation (when it is small and is manifested more strongly on a short scale) is immaterial for general properties of a turbulent flow.

much larger than the sizes of the volume occupied by turbulence. In this case, the Lighthill theory is inapplicable, and the question arises whether or not acoustic waves are generated and what the intensity of these waves is. In other words, will turbulent motion be dissipated along this channel and what will the rate of this dissipation be?

Since the scale of pressure pulsations in a turbulent flow is on the order of ρv^2, we can answer these questions by considering a 1D flow excited by such pulsations of pressure in the region with a size on the order of L with a characteristic pulsation time on the order of L/v. Since $v \ll c$, such pulsations lead to plasma displacements on the order of $(v/c)^2L$ with characteristic velocities on the order of $(v/c)^2v$. For this reason, we analyze the flow bounded by a rigid wall on one side and with a piston performing preset oscillations with an amplitude on the order of $(v/c)^2L$ and a characteristic times of about L/v on the other side. We sought the answers to the following questions: will shock waves be excited in the region for small values of v and will the piston perform work on the average? To make the motion of the piston smooth at the initial instant, the equation of motion was chosen in the form of the sum of two sine functions,

$$x = L\left(\frac{v}{c}\right)^2\left(0.55\sin\frac{2\pi t}{1.1T} - 0.45\sin\frac{2\pi t}{0.9T}\right), \qquad (1)$$

where $T = 2L/v$. For obtaining reliable computation results, we took a large number of meshes over a large time period on the order of $200T$.

Figure 2 shows the velocity profiles obtained in computations with Mach numbers $M \equiv v/c = 0.5$ and $M = 0.33$ at instant $t = 1000L/c$. These profiles are typical of both computations and show that shock waves are formed in the region for $M = 0.5$ and are not formed for $M = 0.33$. Computations show that shock waves are formed in the region for $M \geq 0.5$ and the piston performs work on the average, while no shock waves are formed for $M \leq 0.33$ and the piston does zero work on the average. Thus, we can conclude that in the case of 2D turbulence, when a small fraction of energy is contained in a small-scale region (in the 2D turbulence spectrum, $E(k)$ values for large wave numbers k decrease faster than k^{-3}; see below), this dissipation channel does not exist for small Mach numbers.

For 3D turbulence in a bounded region, this dissipation channel still exists due to a noticeable fraction of energy in the short-wave spectral region, although it is strongly suppressed for small values of M. Indeed, the characteristic frequencies of pulsations in turbulence are on the order of kv. Consequently in the case of 3D turbulence, when $E(k) \sim k^{-5/3}$ for large values of k and, accordingly, the velocities $v_k \sim v(kL)^{-1/3}$, the frequencies increase with k and the applicability condition $kv_k \sim c/L$ for the Lighthill theory holds for large values of k. Since the amount of energy emitted in the form of sound per unit mass of a turbulent medium per unit time is given by

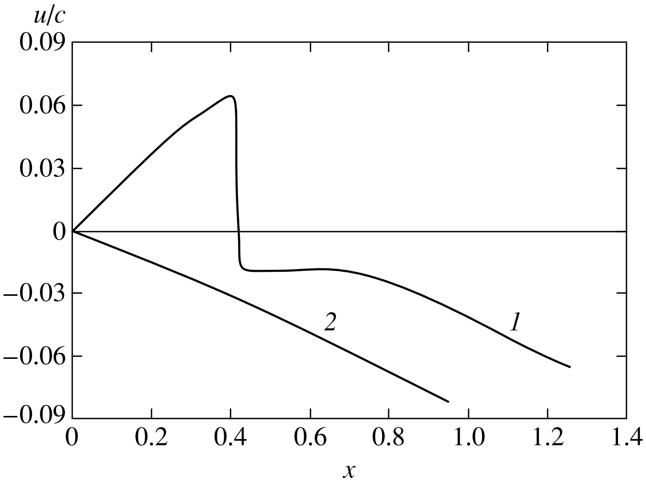


Fig. 2. Velocity profiles u in the gas created by the piston oscillating according to law (1) for Mach numbers $M \equiv v/c = 0.5$ (*1*) and 0.33 (*2*) at instant $t = 1000L/c$.

$$\varepsilon_s \sim \frac{v^8}{c^5 L},$$

the same estimate will be valid in the range of application of the theory upon the substitution $v \longrightarrow v_k$, $L \longrightarrow 1/k$; this gives the following estimate for a bounded volume:

$$\varepsilon_s \sim \frac{v^{10.5}}{c^{7.5} L}.$$

Thus, the intensity of sound generation by a 3D turbulent medium is additionally suppressed in a bounded medium by a factor of $(v/c)^{2.5}$ as compared to that in an unbounded medium.

3. 2D COMPUTATION OF RELAXATION OF A VORTEX FLOW

3.1. Formulation of the Problem

In order to find the characteristics of 2D flows in a bounded region, we considered the evolution of a circular vortex in a square region. We assumed that an azimuth velocity depending on the radius according to the law[2]

$$v_\varphi = \begin{cases} r, & 0 < r < 1, \\ 2 - r, & 1 < r < 2, \\ 0, & 2 < r, \end{cases} \qquad (2)$$

---

[2] It should be noted that, in accordance with MHD analysis, 2D flows emerging in the MAGO system at the stage following dynamic heating are quite complex and turbulent from the very outset; for this reason, the evolution of such flows does not contain the stage of instability development, which is typical of a flow with initial conditions (2). However, the general properties of a developed 2D turbulent flow, which are independent of the initial conditions, must be applicable to 2D MAGO flows also.

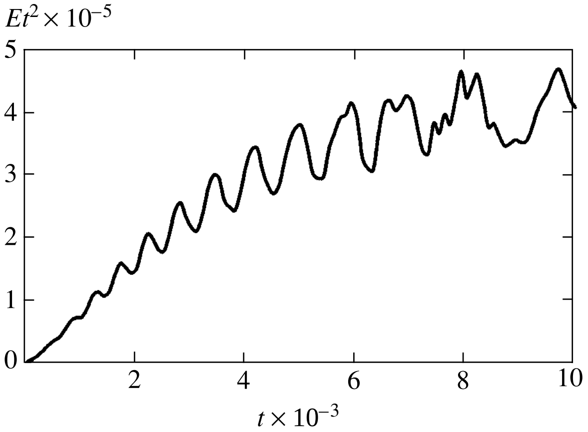


Fig. 3. Time dependence of quantity Et^2 for computations made on a 100×100 grid (energy is measured in units of its initial value).

is set at the initial instant in the region $-3 < x < 3$, $-3 < y < 3$ with constant density and velocity of sound. The velocity of sound c was assumed to be equal to 2.5 since the initial Mach number in these computations can be set at $M = 0.4$. Such a velocity distribution is unstable [7] for $1 < r < 2$ and its 2D evolution will lead to turbulence. The boundaries of the region were assumed to be rigid and perfectly slipping.

In a purely hydrodynamic formulation, the problem is invariant to rotations through an angle of $\pi/2$ and computations could be made only for a quarter of the region. However, the computations were made in the MHD formulation in the rz geometry with azimuth magnetic field; a 2D flow was simulated by moving the region to a large but finite radius (the ratio of the radius to the region size was equal to 10). This led to perturbations breaking the symmetry of the problem so that the computations were made in the entire region.

Computations were based on a 2D MHD code in the ideal formulation, i.e., with a magnetic field frozen into the substance. The equations of ideal magnetohydrodynamics [12] were integrated using a rectangular Euler (i.e., fixed spatial) grid and an explicit conservative difference scheme of the second order of accuracy with respect to space and time. The spatial approximation of convective terms had the first order of accuracy. The main dissipative process in calculations was the viscosity associated with the difference scheme and being on the order of $\nu \sim v\Delta x$ (Δx is the spatial mesh spacing of the grid). However, the origin of this viscosity is insignificant for determining the turbulent flow parameters on large and intermediate scales (on energy-carrying and inertial intervals of wave vectors) for small viscosity values. It is important that viscosity is significant on small scales and ensures a certain level of kinetic energy dissipation.

In our computations, the velocity of sound at the initial instant was assumed to be a purely Alfven velocity. Subsequently, the internal energy and the total velocity of sound increased to a certain extent due to kinetic energy dissipation. However, this increase was insignificant since we considered practically incompressible liquids (for $M = 0.4$, the kinetic energy in the region was about 2% of the magnetic energy).

Comparing computations with different numbers of meshes, we can determine the characteristic Reynolds number of the flow under investigation. For the computation in question with a number of meshes equal to 400×400, the Reynolds number defined as

$$\mathrm{Re} \approx \frac{2v}{Ld\ln E/dt}$$

(v is the mean square velocity in the region; for the region in question, we assume that $L = 3$ and $d\ln E/dt$ is the decrement of kinetic energy damping) amounts approximately to 300.

3.2. Results of Computation

Since the energy damping in our case was determined by computational viscosity on the order of $\nu \sim v\Delta x$, the kinetic energy E for large time periods, for which the characteristic sizes of the flow are determined by the region size L, have to decrease in accordance with the equation

$$\frac{dE}{dt} \approx E\frac{\nu}{L^2} \approx E\frac{v\Delta x}{L^2}. \tag{3}$$

Since $E \propto v^2$, it follows from Eq. (3) that the velocity decreases as $v \propto L^2/\Delta xt$ and the energy, accordingly, decreases as $E \propto 1/t^2$ for large time intervals. In order to verify this, calculations were made on a relatively coarse 100×100 grid up to large values of time. The time dependence of the quantity Et^2 used in this computation is shown in Fig. 3 (energy is measured in units of its initial value, and time is measured in units of the ratio of the unit size to the unit velocity). It can be seen that the energy corresponding to large values of time indeed decreases as $1/t^2$.

Figure 4 shows the $|v|$ isolines of the flow calculated on a 400×400 grid at time instants $t = 0$ (initial instant) and $t = 30$, 200, and 500. It can be seen from Fig. 4 that the pattern of the flow acquires a complex form typical of turbulence upon the development of instability.

Figure 5 shows the time dependence of the kinetic energy

$$E = \frac{1}{2}\int v^2 dxdy,$$

the enstrophy

$$H = \frac{1}{2}\int (\mathrm{curl}\,\mathbf{v})^2 dxdy,$$

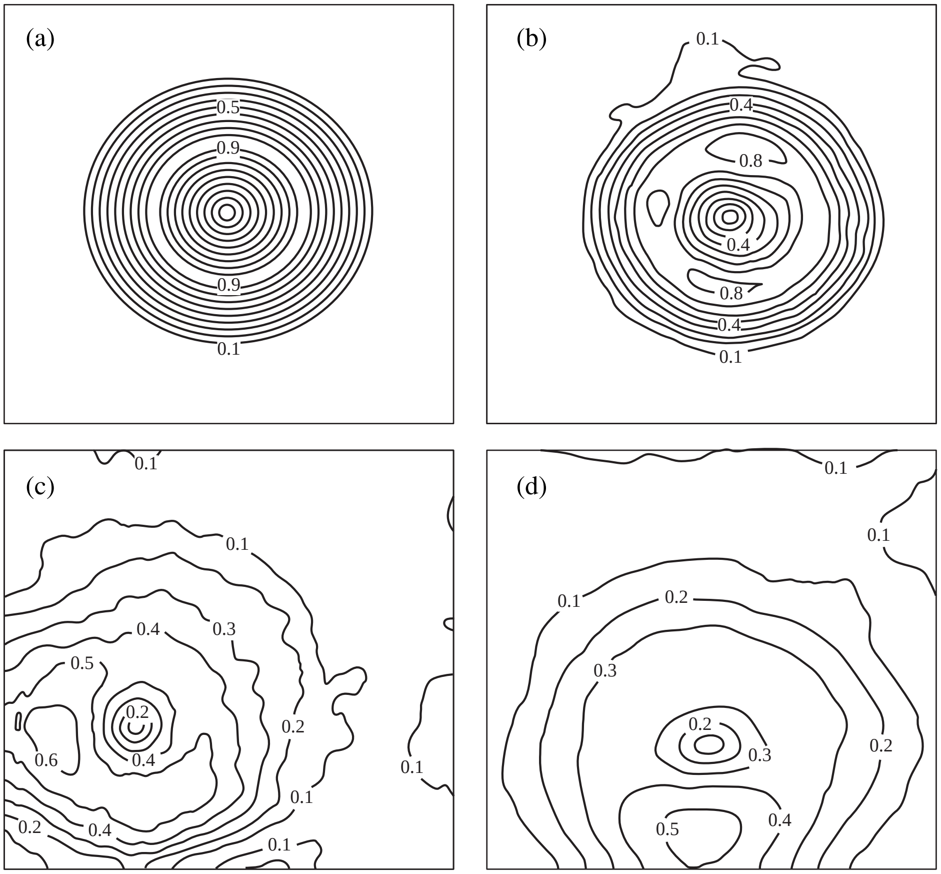


Fig. 4. Isolines of velocity modulus at instants $t = 0$ (a), 30 (b), 200 (c), and 500 (d).

and the angular momentum

$$M = \int v_\varphi r dx dy$$

of the region (radius r and azimuth velocity v_φ are measured from the center of the region). It can be seen that the energy decreases over quite long periods of time determined by the computational viscosity; the energy decay time in the system is on the order of L^2/ν and is proportional to the Reynolds number. This result was obtained for the flows computed on all grids considered by us (100×100, 200×200, and 400×400) and is quite obvious. Small short-period oscillations of kinetic energy are associated with the propagation of acoustic (magnetoacoustic) waves excited for these Mach numbers and do not lead to kinetic energy dissipation (see Section 2).

While using the obtained result on slow damping of 2D turbulence, which is determined, in contrast to 3D turbulence, by the Reynolds numbers, it is important to describe correctly the physical mechanisms leading to dissipation in plasma flows in a magnetic field (for the MAGO system). Such mechanisms include the longitudinal physical viscosity [6], which may be mainly determined (in the case of a magnetized plasma) by the establishing of equilibrium between the longitudinal and transverse degrees of freedom of ions, and strongly magnetized shear viscosity [6]. In many cases, physical viscosity was disregarded in computations made for an MAGO system (see, for example, [2, 9]); nevertheless, these computations led to rapid attenuation of the flow. This apparently indicates considerable numerical computational errors in the description of the plasma at the stage following dynamic heating.

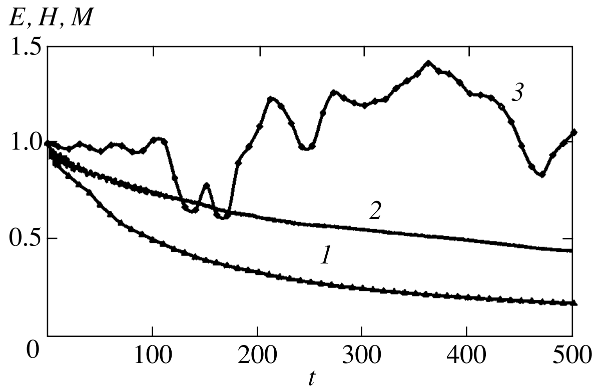


Fig. 5. Time dependences of H (*1*), E (*2*), and M (*3*). All the quantities are measured in units of their initial values.

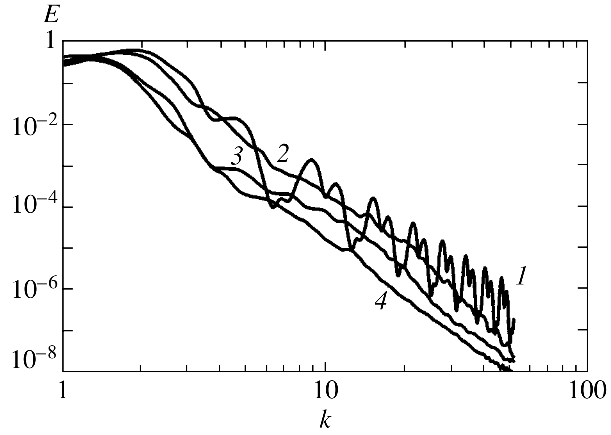


Fig. 6. Spectra of 2D turbulence kinetic energy in a bounded region at instants $t = 0$ (*1*), 50 (*2*), 300 (*3*), and 500 (*4*).

It can be seen from Fig. 5 that the enstrophy decreases slightly faster than energy, which points to a certain increase in the characteristic scales of the flow. The presence of the enstrophy flux in addition to a decrease in the energy could lead to the dependence $E(k) \propto k^{-3}$ [8] for the kinetic energy spectrum in the range of large wave numbers in a certain time interval. However, this rearrangement of characteristic scales of the flow occurs for considerably decreasing kinetic energy in view of insufficiently large Reynolds numbers. For this reason, the above dependence actually is not observed for the energy spectrum (see below). The drop in the initial enstrophy and the dependence $E(k) \propto k^{-3}$ for the energy spectrum could probably be obtained in a certain time interval on the curves describing the time dependence of enstrophy on simulating a flow with large values of Re, e.g., using a finer grid.

Since the system geometry is not invariant to rotations through an arbitrary angle, the angular momentum in the system is not conserved (see Fig. 5). The angular momentum remains virtually unchanged for small time intervals, indicating an increase in the vortex size since the kinetic energy decreases on the corresponding time interval. With increasing time, the angular momentum experiences more significant changes associated with interaction with the walls. The turbulization of the flow and the formation of other large vortices even led to sign reversal for the angular momentum in calculations with different initial conditions [10]. In the case considered here, angular momentum performs irregular oscillations with a large period, preserving approximately its characteristic value in spite of the fact that the kinetic energy decreases by more than half during the computation time.

The kinetic energy spectra of the flow in question calculated on a 400 × 400 grid at different instants are shown in Fig. 6. Since we study the flow in a bounded square box, the expansion of the velocity field into the spectrum was carried out using the formulas

$$v_x = \sum_{m=1,\,n=0}^{N} a^x_{mn} \sin\frac{m\pi x}{L}\cos\frac{n\pi y}{L},$$

$$v_y = \sum_{m=0,\,n=1}^{N} a^y_{mn} \cos\frac{m\pi x}{L}\sin\frac{n\pi y}{L},$$

$$k^2 = \frac{(m^2+n^2)\pi^2}{L^2},$$

$$E(k)dk = \sum_{m,\,n\neq 0}[(a^x_{mn})^2 + (a^y_{mn})^2] + 2\sum_{m}[(a^x_{m0})^2 + (a^y_{0n})^2].$$

Here, N is the number of meshes in a certain direction in the grid, x and y are measured from the boundaries of the region, and summation in the last formula is carried out over the values of m and n falling in the given interval of dk values. The total energy in Fig. 6 is measured in its initial value. Energy values $E(k)$ are given only for $k < 50$ since the effect of the finite number of meshes in the grid starts manifesting itself and the error in the evaluation of Fourier integrals increases (this can be judged from the behavior of the spectrum calculated for $k \approx 50$, especially for $t = 500$, when small oscillations appear on the curves).

At the initial instant, for large values of k, the energy decreases in proportion to k^{-4} and exhibits oscillations associated with singularities of function $v_\varphi(r)$ defined by formula (2). As turbulence develops, the curve describing the energy spectrum on the segment corresponding to the inertial interval (approximately from $k \approx 3$ to $k \approx 15$) preserves a practically constant slope corresponding to the dependence $E(k) \propto k^{-n}$, where $n = 4.2 \pm 0.4$ for different instants. Thus, the slope of the curve describing the spectrum is slightly larger than that required for a direct enstrophy flux towards smaller scales (which requires the k^{-3} dependence [8]). Figure 6 also shows that the energy condensation predicted for bounded regions [8] is observed for small values of k: the energy spectrum is narrowed with time and energy values decrease quite sharply for large values of k; the main fraction of energy is concentrated in harmonics with minimum possible wave numbers in a bounded system.

We have analyzed the flow in the region with ideal slippery walls. However, the behavior of the energy spectrum in the case of 2D turbulence for large values of wave vectors k should also be preserved for a flow in the region with zero velocity at the boundaries (adhe-

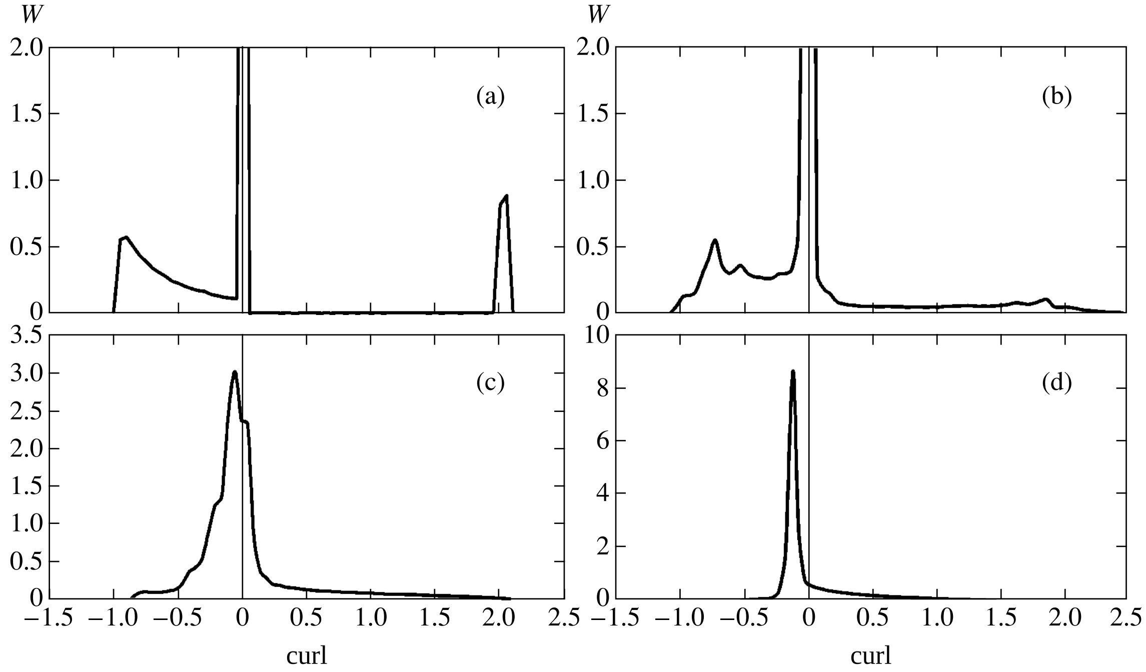


Fig. 7. Evolution of volume distribution of vorticity for consecutive time intervals $t = 0$ (a), 50 (b), 200 (c), and 500 (d).

sion condition) in view of the absence of energy transfer to the region of large k. Attenuation of the hydrodynamic flow in the 2D case occurs over time periods on the order of L^2/ν. The drag forces emerging in a flow through a tube or during streamlining of bodies have to be of the same order of magnitude as for a laminar flow and have to differ from the drag forces in the latter flow with arbitrarily large Reynolds numbers only in a constant factor.

It should also be noted that heat transfer processes and mixing of substances with 2D turbulence differ from those in the case of 3D turbulence. In the 2D case, when $E(k)$ decreases in the inertial region faster than according to the k^{-3} law, the relative velocity v of two particles separated by a small distance λ from each other is determined by large scales of the flow and is proportional to this distance:

$$v \propto v_0 \frac{\lambda}{L},$$

where v_0 and L are the characteristic velocity and spatial scale of the flow. Consequently, the time variation of this distance follows the law $\lambda_1 \exp(v_0 t/L)$ (λ_1 is the initial spacing). Thus, the time required for the particles to move apart to a large distance λ_2 is a logarithmic function of the initial spacing,

$$t \propto \frac{L}{v_0} \ln \frac{\lambda_2}{\lambda_1},$$

in contrast to 3D turbulence, for which even two infinitely close particles in the inertial region move apart over a finite time interval to a finite distance [11]. Thus, heat transfer and mixing in the 2D case are slightly retarded as compared with time $t \sim L/v_0$ of these processes typical of the 3D case (if we take the main scale L of the flow for λ_2 and the viscous scale for λ_1, which is equal to $\lambda_1 \sim \sqrt{L\nu/v_0}$ in the 2D case, the retardation factor is $\ln\sqrt{\mathrm{Re}}$). Accordingly, temperature pulsation in the 2D case (as well as pulsations in the concentrations of substances being mixed) also differ from the pulsations in the 3D case; repeating the derivation of distribution of these pulsations [11], we find that the pulsations in the inertial region have to be distributed in the 2D case in proportion to $\ln\lambda$; i.e., they have to be almost independent of the spacing between particles, while these pulsations in the 3D case are proportional to $\lambda^{1/3}$.

As applied to flows of a magnetized plasma, this should result in less uniform distributions of impurities and temperature in the plasma as compared to the 3D case, which must be taken into account in the description of plasma radiation by impurities and plasma cooling. For example, the model of homogeneous mixing used for describing 3D flows may be inapplicable in the 2D case.

For a 2D flow of an ideal fluid, the vorticity

$$(\mathrm{curl}\,\mathbf{v})_z = -\frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x}$$

have to be conserved in each Lagrangian part of the liquid by virtue of the Thomson theorem [11]. However, the vorticity of each particle in a 2D turbulent flow changes since in this case enstrophy cascades towards large wave numbers take place, and the enstrophy in the entire region decreases even when the kinetic energy is conserved. It is interesting to analyze this phenomenon together with the evolution of the volume distribution of vorticity.

For conducted computations, we obtained the volume distribution of vorticity $W(\mathrm{curl})$, where curl ≡ $(\mathrm{curl}\mathbf{v})_z$; i.e., W is the fraction of the total volume dV/V, in which the value of the curl lies in a preset interval $d\mathrm{curl}$ so that

$$W = \frac{dV}{Vd\mathrm{curl}}.$$

Figure 7 illustrates the evolution of this distribution for time instants $t = 0$, 50, 200, and 500.

It can be seen from Fig. 7 that the distribution becomes narrower with time. First, at $t < 50$, the distribution exhibits a certain diffusion both for positive and negative values. The valleys between the initial peaks, which correspond to the region of increasing velocity in Fig. 4a for a positive value of the curl, to the region of decreasing velocity for its negative value, and to the region of stationary region for zero curl, are blurred. Then the peaks for negative and positive values of vorticity disappear, and only a single peak remains in the vicinity of zero. For large values of time, the vorticity distribution preserves its shape with a peak for small negative values, a wing rapidly decreasing towards negative values, and a wing decreasing slowly towards positive values. However, the ratio of the widths of these wings changes with time as t varies from 200 to 500. This is probably associated with the fact that time intervals $t = 200$–500 are not large enough for attaining a steady state of 2D turbulence. A comparison of curves *1* and *2* in Fig. 5 shows that the ratio of quantities H and E still varies in this temporal range. It should be noted that the asymmetry in the vorticity distribution, which stems from the initial conditions, is preserved. Thus, the initial conditions of the problem are not "forgotten" completely, which is also confirmed by the angular momentum dynamics (see Fig. 5). This quantity preserves its sign despite considerable oscillations (under other initial conditions, the angular momentum preserved its sign on the average over a large time period [10]).

4. CONCLUSIONS

Let us formulate the general conclusions concerning 2D turbulence in a bounded region.

1. In the case of 2D turbulence in a bounded region for small Mach numbers M, turbulence does not generate sound; consequently, this dissipation channel is closed completely for small M.

2. The results of 2D computations should be applied with care to 3D turbulent flows in view of considerable difference in the properties of 2D and 3D turbulent flows. In 3D flows, kinetic energy attenuation, heat transfer, and mixing of substances may occur at much higher rates than in the 2D case. For example, the kinetic energy attenuation for a flow in a bounded region in the 3D case have to occur over time periods on the order of $\alpha L/v$, where α is a small factor characterizing the intensity of turbulent pulsations, while analogous time periods in 2D calculations are on the order of L^2/ν.

3. The turbulence energy spectrum for large periods of time becomes narrower, indicating energy condensation for small k. For large values of k, the energy slopes more rapidly with increasing k than by the k^{-3} law.

4. The volume distribution of vorticity for large time periods approximately preserves its shape and exhibits asymmetry relative to positive and negative values due to asymmetry in the initial conditions. In this sense, the initial conditions of the problem are not "forgotten" completely.

In addition, the following conclusions can be drawn, which are essential for description of 2D turbulent flows of a magnetized plasma across the magnetic field in regions bounded by material walls (e.g., for an MAGO system).

(1) Since the time of 2D turbulence energy dissipation is not associated with energy transfer on a small scale, it is necessary for calculating plasma flows in a magnetic field to correctly describe the physical mechanisms leading to dissipation (longitudinal physical viscosity [6], which can mainly be determined in the case of a magnetized plasma by establishing equilibrium between the longitudinal and transverse degrees of freedom of ions, and strongly magnetized shear viscosity [6]).

(2) The distributions of impurities and temperature of the plasma may be less uniform as compared to the 3D case, which may be important for describing the radiation of plasma by impurities and its cooling.

ACKNOWLEDGMENTS

The authors are grateful to N.V. Zmitrenko, V.I. Mamyshev, V.F. Tishkin, and V.B. Yakubov for cooperation at the initial stage of the study and for fruitful discussions.

REFERENCES

1. A. M. Buĭko, G. I. Volkov, S. F. Garanin, *et al.*, Dokl. Akad. Nauk **344**, 323 (1995) [Phys.–Dokl. **40**, 459 (1995)].
2. I. R. Lindemuth, R. E. Reinovsky, V. K. Chernyshev, *et al.*, Phys. Rev. Lett. **75**, 1953 (1995).

3. S. F. Garanin, IEEE Trans. Plasma Sci. **26**, 1230 (1998).
4. S. F. Garanin, A. I. Golubev, and N. A. Ismailova, Fiz. Plazmy (Moscow) **26**, 426 (2000) [Plasma Phys. Rep. **26**, 397 (2000)].
5. S. F. Garanin, Fiz. Plazmy (Moscow) **26**, 309 (2000) [Plasma Phys. Rep. **26**, 283 (2000)].
6. S. I. Braginskiĭ, in *Reviews of Plasma Physics*, Ed. by M. A. Leontovich (Atomizdat, Moscow, 1963; Consultants Bureau, New York, 1963), Vol. 1, p. 183.
7. S. F. Garanin and S. D. Kuznetsov, Fiz. Plazmy (Moscow) **22**, 743 (1996) [Plasma Phys. Rep. **22**, 674 (1996)].
8. S. D. Danilov and D. Gurariĭ, Usp. Fiz. Nauk **170**, 921 (2000) [Phys. Usp. **43**, 863 (2000)].
9. A. M. Buyko, S. F. Garanin, G. G. Ivanova, *et al.*, in *Digest of Technical Papers of 12th IEEE International Pulsed Power Conference*, *Monterey, California*, Ed. by C. Stallings and H. Kirbie (1999), Vol. 2, p. 1052.
10. S. F. Garanin, O. M. Burenkov, G. G. Ivanova, *et al.*, in *Digest of Technical Papers of Pulsed Power Plasma Science-2001*, *Las Vegas, Nevada*, Ed. by R. Reinovsky and M. Newton (2001), Vol. 1, p. 512.
11. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 6: *Fluid Mechanics*, 3rd ed. (Nauka, Moscow, 1986; Pergamon, New York, 1987).
12. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 8: *Electrodynamics of Continuous Media*, 2nd ed. (Nauka, Moscow, 1982; Pergamon Press, Oxford, 1984).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 70–77.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 80–88.
Original Russian Text Copyright © 2003 by Anshukova, Bulychev, Golovashkin, Ivanova, Krynetskiĭ, Minakov, Rusakov.

SOLIDS
Structure

Anomalous Low-Temperature Behavior of the Thermal Characteristics of MgB_2

N. V. Anshukova[a], B. M. Bulychev[b], A. I. Golovashkin[a],*, L. I. Ivanova[c], I. B. Krynetskiĭ[b], A. A. Minakov[d], and A. P. Rusakov[c]
[a]*Lebedev Institute of Physics, Russian Academy of Sciences, Moscow, 119991 Russia*
[b]*Moscow State University, Moscow, 119899 Russia*
[c]*Moscow Institute of Steel and Alloys, Moscow, 117936 Russia*
[d]*Institute of General Physics, Russian Academy of Sciences, Moscow, 119991 Russia*
**e-mail: golov@sci.lebedev.ru*

Received December 17, 2002

Abstract—We have studied the behavior of the thermal expansion coefficient $\alpha(T)$ (in a zero magnetic field and at $H \approx 4$ T), the heat capacity $C(T)$, and the thermal conductivity $\kappa(T)$ of magnesium boride (MgB_2) in the vicinity of T_c and at lower temperatures. It was established that MgB_2, like oxide-based high-temperature superconductors, exhibits a negative thermal expansion coefficient at low temperatures. The anomaly of $\alpha(T)$ in MgB_2 is significantly affected by the magnetic field. It was established that, in addition to the well-known superconducting transition at $T_c \approx 40$ K, MgB_2 exhibits an anomalous behavior of both heat capacity and thermal conductivity in the region of $T \approx 10$–12 K. The anomalies of $C(T)$ and $\kappa(T)$ take place in the same temperature interval where the thermal expansion coefficient of MgB_2 becomes negative. The low-temperature anomalies are related to the presence of a second group of charge carriers in MgB_2 and to an increase in the density of the Bose condensate corresponding to these carriers at $T_{c2} \approx 10$–12 K. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The unexpected high-temperature superconductivity recently reported by Nagamatsu *et al.* [1] for the well-known compound magnesium boride (MgB_2) has stimulated extensive investigation into the properties of this substance. Presently, a considerable number of papers devoted to MgB_2 can be found in the literature [2]. The interest in this compound is related to a high critical temperature ($T_c \approx 40$ K) in combination with relatively simple structure, high conductivity, and large critical fields and currents (including those in strong magnetic fields). The critical current densities observed in MgB_2 already exceed 10^7 A/cm^2 at a critical fields of up to 40 T [2]. Magnesium boride, in contrast to cuprate high-temperature superconductors (HTSCs), exhibits less anisotropic properties. Possessing large coherence length (correlation length), this material is very attractive for HTSC electronics.

In the present-day stage of research in the field, it is important to establish whether the critical temperature of superconductors of this class can be increased further. In order to elucidate this question, it is necessary to understand whether MgB_2 belongs to the traditional superconductors described by the Bardeen–Cooper–Schrieffer (BCS) theory or this compound is close to oxide HTSCs. However, the results of investigations reported so far provide no unambiguous answer concerning the nature of superconductivity in MgB_2.

Indeed, the critical temperature of MgB_2 is close to a limiting value predicted by the BCS theory, or even exceeds this estimate, which can be considered as indicative of an unusual mechanism of superconductivity in this compound. On the other hand, a high carrier density ($N \approx 1.5 \times 10^{23}$ cm^{-3} [3]) can be treated as evidence in favor of the usual superconductivity, since the characteristic carrier density in oxide HTSCs is $N \approx (3–5) \times 10^{21}$ cm^{-3}. However, the energy band structure calculations performed for MgB_2 showed that the observed carrier density is due to the presence of two groups of carriers [4] related to different regions of the Fermi surface (formed by different electron states of boron). If the quasi-two-dimensional p_{xy} states of boron with a carrier density $N_{xy} \le 10^{22}$ cm^{-3} play the same role as do the quasi- two-dimensional states of oxygen in the CuO_2 planes of cuprate HTSCs, then it is probable that the superconductivity observed in MgB_2 at $T \approx 40$ K is caused by only one of these groups. Both the results of theoretical calculations [5, 6] and the experimental data for heat capacity [7–13] and thermal conductivity [12–15] point to the possibility that two superconducting gaps may exist in MgB_2. One of these gaps corresponds to $T_c \approx 40$ K and the other (also arising at $T \approx 40$ K) corresponds to $T_{c2} \approx 10$–12 K, where it exhibits a sharp growth. However, the final judgment requires additional experiments.

As is known, oxide HTSCs exhibit several characteristic anomalies in their thermal properties. In particular, the thermal expansion of high-quality samples of such HTSCs shows an anomaly in the form of negative thermal expansion coefficient $\alpha(T)$ at low temperatures [16]. In addition, the temperature variation of α in this region is significantly affected by the applied magnetic field [17]. No such anomalies are observed in usual superconductors. However, our recent preliminary results [18] showed that analogous low-temperature anomalies take place in MgB_2 as well. Thus, it is still difficult to unambiguously classify MgB_2 as a usual superconductor.

In this study, the behavior of the thermal expansion coefficient $\alpha(T)$, the heat capacity $C(T)$, and the thermal conductivity $\kappa(T)$ of MgB_2 was studied in the vicinity of T_c and at lower temperatures. In addition, we studied the effect of an external magnetic field ($H \approx 4$ T) on the behavior of $\alpha(T)$. It was established that all three thermal characteristics—heat capacity, thermal conductivity, and thermal expansion—exhibit an anomalous low-temperature behavior (in the region of $T \approx 10$–13 K). The anomalies of $C(T)$ and $\kappa(T)$ take place in the same temperature interval where the thermal expansion coefficient of MgB_2 becomes negative.

2. EXPERIMENTAL METHODS

The thermal expansion of samples was studied by dilatometry. The relative sample length variation, $\Delta L/L$ (L is the sample length), was determined with the aid of a strain gauge possessing a sensitivity of about 10^{-7} [19]. The magnetic field was generated by a superconducting solenoid and applied to the samples in the direction of straining. The dilatometer was calibrated by measuring temperature dependences of the thermal expansion of rare earth oxides.

The temperature variation of heat capacity and thermal conductivity was studied by modulation calorimetry [20, 21] at a continuous temperature sweep rate of about 1 K/min and a temperature modulation frequency of 20 Hz, or (for checking) under quasiisothermal conditions at various modulation frequencies in the interval from 0.05 to 160 Hz. The amplitude of the thermal flux modulation was 0.1, 0.45, and 0.7 mW at temperatures within 5–10, 10–20, and 20–50 K, respectively. The amplitude of the corresponding sample temperature oscillations ranged within 0.002 to 0.07 K. The modulated thermal flux $P(T) = P_0\cos\omega t$ supplied to one side of a disk sample induced decaying temperature waves $T(t) = \mathrm{Re}[T_0\exp(i\omega t \pm kz]$. The heat capacity and thermal conductivity of the sample were determined from the results of measurements of the amplitudes (T_{01}, T_{02}) and phases (φ_1, φ_2) of the sample temperature oscillations $T_{01}\sin(\omega t + \varphi_1)$ and $T_{02}\sin(\omega t + \varphi_2)$ on both sides of the disk. The temperature dependences of the heat capacity and thermal conductivity of the sample were monitored with a temperature resolution of 0.01 K. The method of two-channel modulation calorimetry is described in more detail in [20, 21].

3. SAMPLES

The samples were prepared by hot pressing from the initial MgB_2 powder synthesized via a reaction of metallic magnesium with boron. The process was conducted under standard conditions (950–1000°C, 4 h, atmospheric pressure) and yielded a homogeneous target material. The disk samples were hot pressed at 50 kbar and 950–1000°C. Variation of the pressure during sintering led to 3% changes in the sample density. According to X-ray diffraction data, the density of samples synthesized at a maximum pressure amounted to 97–98% of the ideal value.

The X-ray diffraction patterns of synthesized MgB_2 samples obtained on a DRON-4 diffractometer coincided with the reference patterns [2]. In addition, the quality of samples was checked by measurements of the electrical and magnetic properties, which also corresponded to standard values [2]. The Meissner effect exceeded 44%. The measurements of thermal expansion, heat capacity, and thermal conductivity were performed on disk samples with a diameter of 2.8–3.2 mm a height of 1–5 mm.

4. EXPERIMENTAL RESULTS

Figure 1a presents the plot of $\Delta L/L$ versus temperature for Mg_2B in the absence of an applied magnetic field ($H = 0$). For comparison, Fig. 1b shows analogous experimental data for $YBa_2Cu_3O_{7-x}$ [22], $Bi_2Sr_2CaCu_2O_8$ [23], $La_{2-x}Sr_xCuO_4$ ($x = 0.1$) [24], and $Ba_{1-x}K_xBiO_3$ ($x = 0.13$) [24]. A decrease in the lattice parameters with increasing temperature (in the low-temperature range) was also reported for $YBa_2Cu_4O_8$ [25]. In MgB_2, the $\Delta L/L$ value is negative in the temperature interval $7\ \mathrm{K} \le T \le 16.5$ K. Thus, the thermal expansion coefficient $\alpha = (1/L)dL/dT$ of MgB_2 is negative for $T =$ 7–11 K. As can be seen from Fig. 1b, the oxide HTSCs also exhibit regions of negative thermal expansion $\alpha(T)$. Therefore, MgB_2 shows the same anomalous low-temperature behavior ($\alpha < 0$) as the oxide HTSCs studied in [22–24] (Fig. 1b).

Figure 2a shows the influence of a magnetic field ($H = 36$ kOe) on the temperature dependence of $\Delta L/L$ in MgB_2. For comparison, Fig. 2b illustrates the effects of magnetic field on the thermal expansion of $Ba_{0.6}K_{0.4}BiO_3$ and $La_{1.9}Sr_{0.1}CuO_4$ [24]. As can be seen, application of the magnetic field ($H \approx 40$ kOe) produces

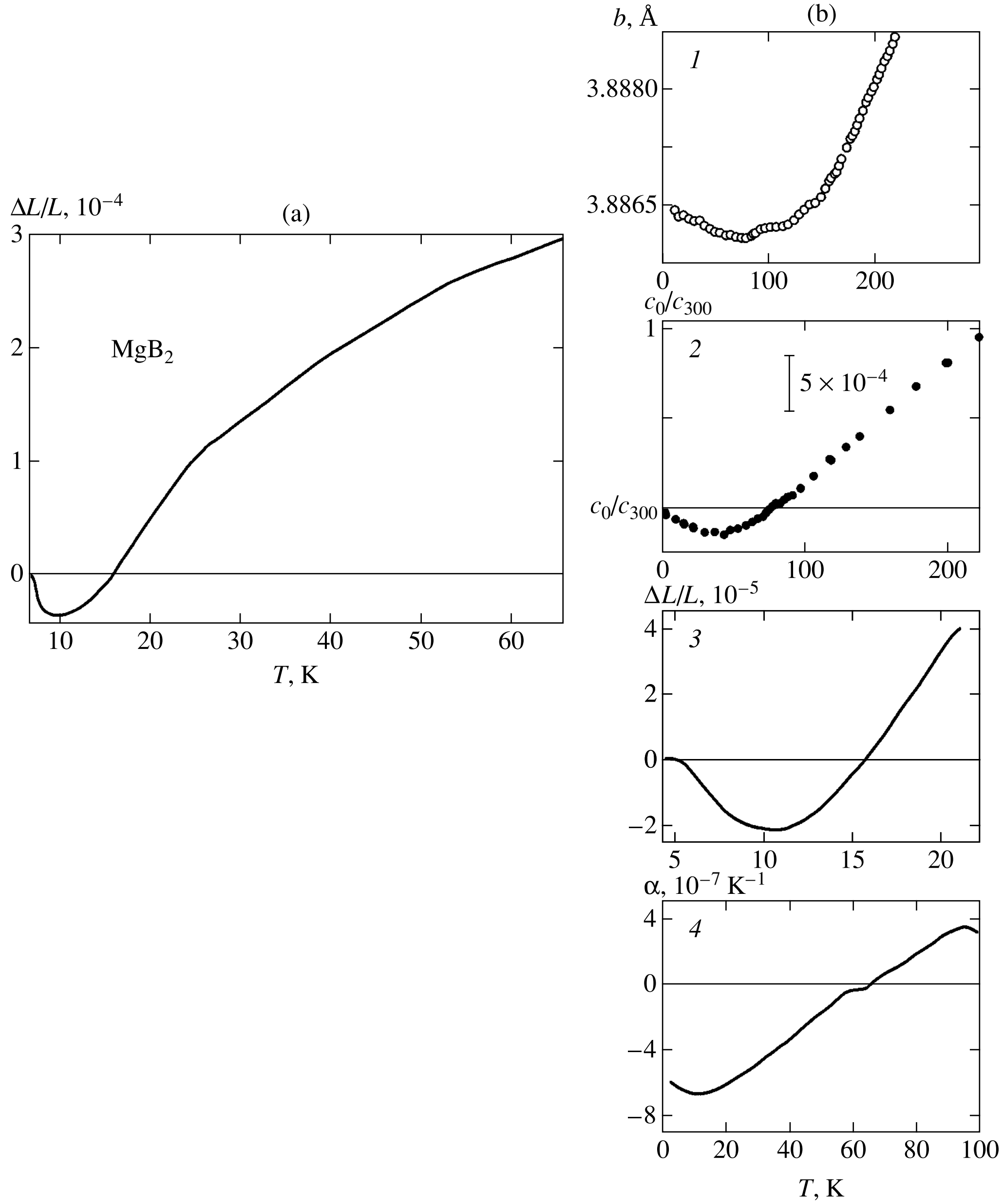


Fig. 1. Temperature dependence of the linear thermal expansion coefficient $\Delta L/L$ in MgB_2 (a) in comparison with analogous data for other HTSC systems (b): (*1*) $YBa_2Cu_3O_{7-x}$ (b is the lattice parameter along the b axis [22]); (*2*) $Bi_2Sr_2CaCu_2O_8$ (c is the lattice parameter along the c axis; c_0 and c_{300} are the c values at $T = 0$ and 300 K, respectively [23]); (*3*) $La_{2-x}Sr_xCuO_4$ ($x = 0.1$; *ab* is the plane [17]); (4) $Ba_{1-x}K_xBiO_3$ ($x = 0.13$; α is the thermal expansion coefficient [17]).

an anomalous action upon $\alpha(T)$ in these oxide HTSCs at low temperatures.

Figure 3 shows the plots of $\Delta L/L$ versus H for MgB_2 measured at various temperatures. As can be seen, the curve observed at $T = 12.1$ K (i.e., in the region of $T \leq 16.5$ K, where $\Delta L/L$ is negative) is qualitatively different from the curves measured at $T > 16.5$ (only three plots for $T = 18.8$, 28.2, and 37.5 K are presented, for example). The behavior of $\Delta L/L$ observed at $T > 16.5$ K can be explained by the magnetostriction effect; on the contrary, the sign of $\Delta L/L$ at $T = 12.1$ K is opposite and magnetostriction cannot account for the observed variation of $\Delta L/L$. In $Ba_{0.66}K_{0.34}BiO_3$, magnetostriction was reported [26] to decrease $\Delta L/L$ with increasing H (up to 5 T) at low temperatures. We have also observed such dependences for MgB_2 at $T > 16.5$ K. Therefore, in addition to magnetostriction, a stronger opposite effect similar to that in oxide HTSCs is operative in MgB_2 at $T < 16.5$ K.

At still higher temperatures, the curves of $\alpha(T)$ and $\alpha(H)$ in MgB_2 resemble the temperature dependences observed in usual metals. Similar behavior was

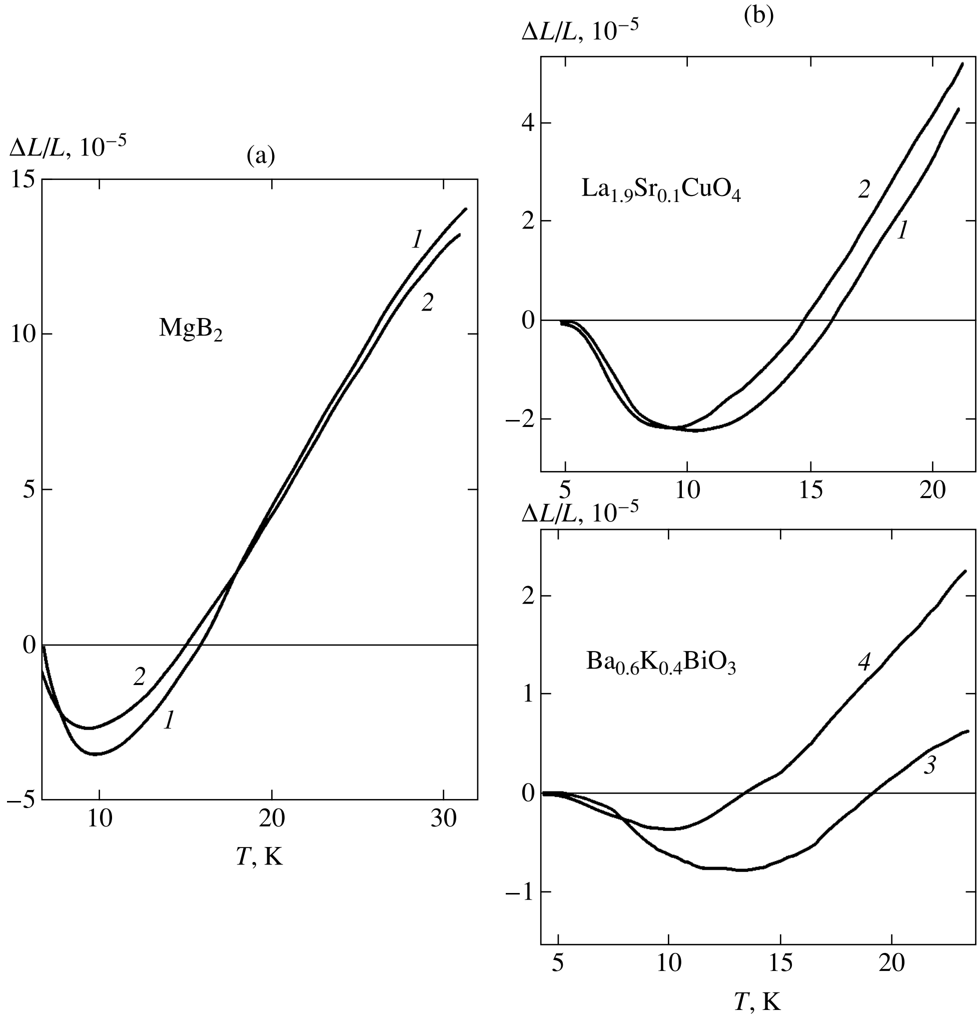


Fig. 2. The influence of a magnetic field on the temperature dependence of thermal expansion: (a) MgB_2 at $H = 0$ (*1*) and 3.6 T (*2*); (b) $Ba_{0.6}K_{0.4}BiO_3$ and $La_{1.9}Sr_{0.1}CuO_4$ at $H = 0$ (*1, 3*) and ~4 T (*2, 4*) [17].

observed in samples of the $Ba_{1-x}K_xBiO_3$ system with metallic conductivity ($x > 0.4$) [17].

Figure 4 presents the temperature dependence of the heat capacity plotted as C/T versus T in the temperature interval from 5 to 45 K for one of the MgB_2 samples studied (sample 1). The curve exhibits two clearly manifested features, at $T \approx 38$–40 K and $T \approx 10$ K, as more sharply revealed by differential plots in the insets. The feature at $T \approx 38$–40 K is attributed to a transition to the superconducting state. Figure 5 shows the analogous C/T versus T plot in the 5–50 K interval for another MgB_2 sample (sample 2). As can be seen, both the aforementioned features (see the insets in Fig. 5) are observed for this sample as well. The two samples exhibit a generally similar behavior, somewhat differing only in magnitude of the heat capacity jumps. This is explained by slightly (to within 3%) different densities of these samples.

The thermal conductivity $\kappa(T)$ of MgB_2 samples studied was significantly dependent on the conditions of preparation (temperature and pressure of the synthesis). At low temperatures, the value of κ in MgB_2 is small and close, for example, to that in Nb_3Sn. Figure 6 shows the temperature dependence of the thermal conductivity of MgB_2 (sample 1) measured in the temperature interval from 5 to 45 K. As can be seen, the curve of $\kappa(T)$ exhibits both anomalies indicated above for $C(T)$, although the anomaly in the region of $T \approx$ 38–40 K is less pronounced (the dashed straight line in Fig. 6 is drawn only for contrast). The higher the thermal conductivity level, the less pronounced this anomaly. The low-temperature anomaly at $T \approx 10$–12 K is more clearly manifested. In Fig. 7, this feature is depicted for both samples 1 and 2 on a greater scale against the low-temperature interpolation curves shown by dashed curves. In all cases, for both T_c and $T \approx$ 10–12 K, the decay of $\kappa(T)$ with decreasing temperature slows down in the vicinity of the phase transition.

Thus, there is an evident coincidence of the three anomalies observed in our experiments for the thermal

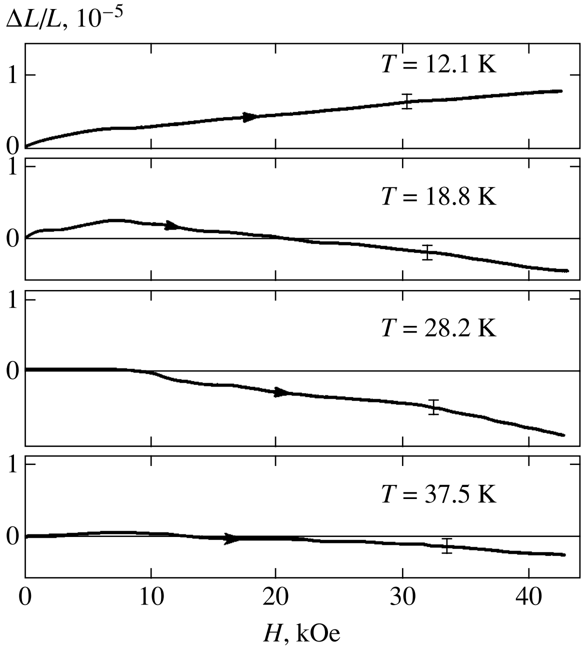


Fig. 3. Plots of the linear thermal expansion $\Delta L/L$ versus H for MgB_2 measured at various temperatures. Bars indicate the experimental uncertainty. Arrows indicate the direction of variation of the magnetic field strength in the course of measurements.

properties of MgB_2 at $T \approx 10$–12 K, including the heat capacity $C(T)$, the thermal conductivity $\kappa(T)$, and the thermal expansion coefficient $\alpha(T)$ [18].

5. DISCUSSION

1. The heat capacity jump ΔC is usually estimated in the presence of a strong magnetic field breaking the superconducting state. Under these conditions, the entropy, free energy, and the superconducting transition parameters (e.g., within the framework of the BCS theory [7]) can be evaluated as well. However, for estimating only ΔC, we can use a difference between the experimental $C(T)$ values and the low-temperature heat capacity interpolation, provided that the maximum of ΔC at $T < T_c$ still occurs close to T_c and the interpolation formula for $T > T_c$ is fitted to experiment in the vicinity of T_c. In our case, there was a good agreement between the interpolation using the extended Debye model $C/T = \gamma + \beta_2 T^2 + \beta_4 T^4$ and the experimental data for MgB_2 in the temperature interval 40 K < T < 50 K.

However, the difference between the above interpolation and that based on a simpler expression $C/T = \gamma + \beta_2 T^2$ in the region of the heat capacity jump was as small as 1–1.5%, and at higher temperatures this difference was negligibly small. Figure 8 demonstrates how the interpolation curve $C/T = 4.394 + 1.065 \times 10^{-2} T^2$ with the coefficients determined by least squares fits to the experimental data for MgB_2 sample 1 at 40 K ≤ T ≤ 45 K. The applicability of such simple interpolation formulas to description of the experimental data is related to the fact that the temperature in the interpolation region ($T \approx 40$ K) is much smaller than the Debye temperature for MgB_2 ($\Theta \approx 900$–1000 K [7– 11]). This simple interpolation is reliable only within a rather narrow temperature interval of $\Delta T \approx 15$–20 K [7–11].

Insets *1* in Figs. 4 and 5 show the heat capacity jumps $\Delta C/T$ determined by subtracting interpolated values from the experimental data in the regions of T_c for MgB_2 samples 1 and 2, respectively. As can be seen, the transition to a superconducting state in both samples begins at $T \approx 40$ K. Therefore, the critical temperature determined at the onset of the heat capacity jump

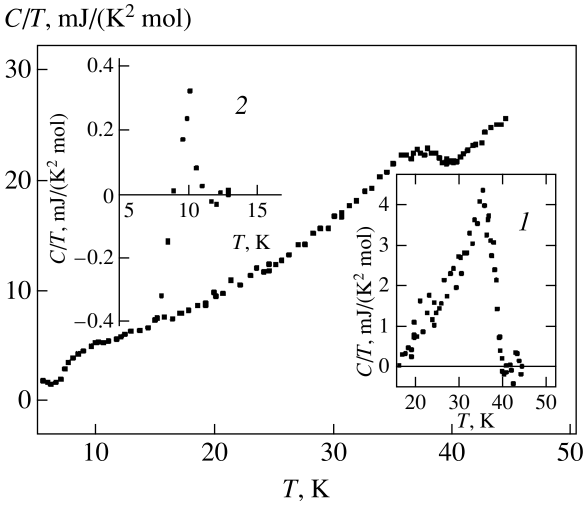


Fig. 4. The temperature dependence of the heat capacity of MgB_2 plotted as C/T versus T (sample 1). The insets show the heat capacity jumps at (*1*) T_c and (*2*) $T \approx 10$ K.

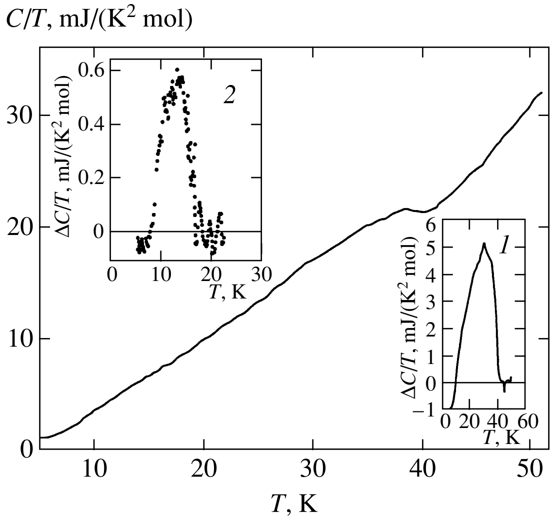


Fig. 5. The temperature dependence of the heat capacity of MgB_2 (sample 2). The insets are as in Fig. 4.

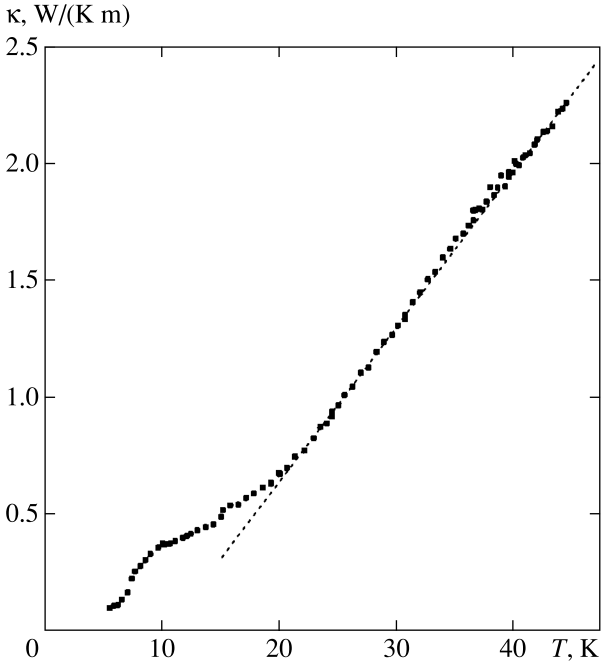


Fig. 6. The temperature dependence of the thermal conductivity of MgB_2 (sample 1) measured in the temperature interval from 5 to 45 K. Dashed line is drawn for contrasting the feature observed at the superconducting transition temperature T_c.

is $T_c \approx 40$ K. The peaks of the heat capacity jumps occur rather close to a boundary of the interpolation region ($T \approx 40$ K), which justifies applicability of the aforementioned simple interpolation (Debye formula) for estimation of the ΔC jump in the vicinity of T_c. For the MgB_2 samples studied, we obtained $\Delta C \approx 145$–152 mJ/(K mol). Although these estimates are somewhat greater than the other published values [7–12], the general agreement is quite satisfactory.

The second anomaly on the $C(T)/T$ curves was observed for both MgB_2 samples in the region of $T \approx$ 10–15 K. The $C(T)/T$ curve at temperatures above this anomaly is well approximated by the Debye formula. As can be seen from the differential heat capacities plotted in the insets in Figs. 4 and 5, determined by subtracting interpolated values from the experimental data in the regions of anomalies, both samples exhibit additional heat capacity jumps ΔC_2. The peaks of these features are observed at $T \approx$ 10–12 K. This behavior is indicative of a phase transition taking place in MgB_2 at $T = T_{c2} \approx$ 10–12 K.

There is a certain theoretical basis [4–6] to believe that this temperature corresponds to a sharp growth in the second (smaller) superconducting gap corresponding to the second group of charge carriers mentioned above. Below the temperature corresponding to the peak of this anomaly, the experimental heat capacity exhibits a sharp decrease similar to that observed in usual superconductors in the course of the superconducting transition. According to our estimates, $\Delta C_2 \approx$ 3.2 mJ/(K mol) for sample 1, and $\Delta C_2 \approx 7.6$ mJ/(K mol) for sample 2. It should be noted that the number of experimental points obtained in this temperature inter-

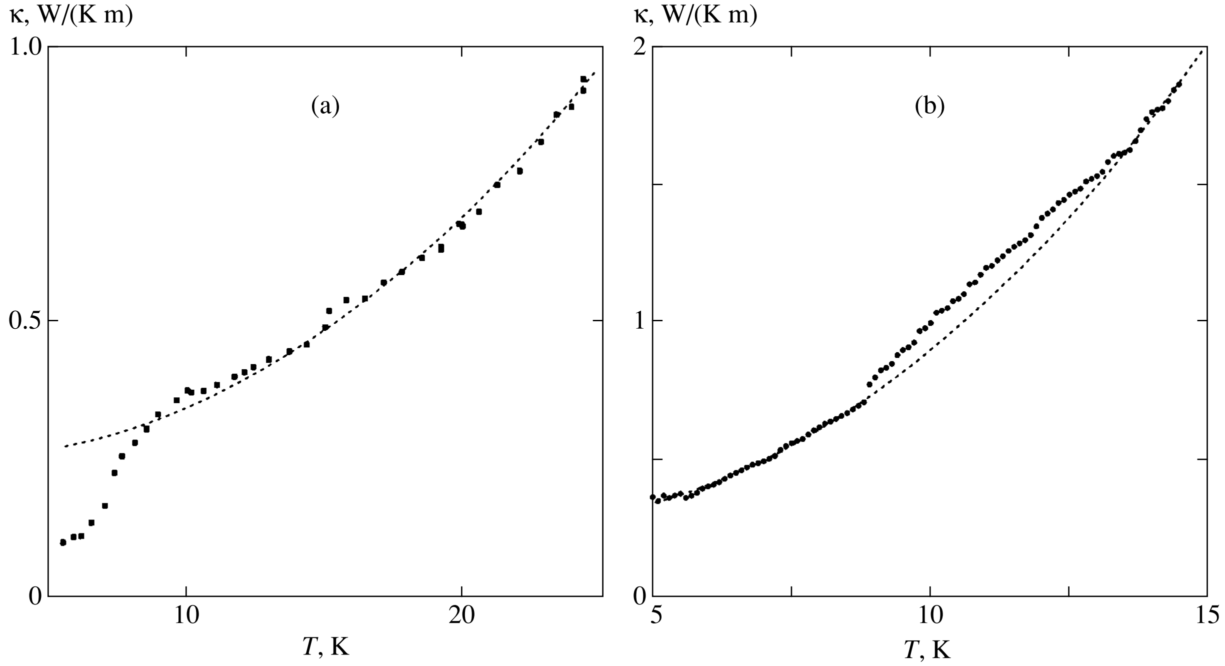


Fig. 7. The low-temperature anomaly in the thermal conductivity of MgB_2 at $T \approx$ 10–12 as observed for samples (a) 1 and (b) 2. Dashed curves show result of interpolation according to the Debye model.

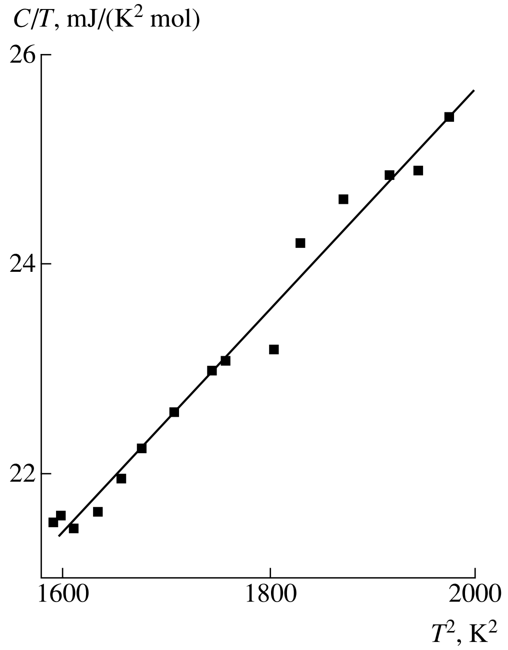


Fig. 8. The plot of C/T versus T^2 demonstrating how the interpolation curve $C/T = 4.394 + 1.065 \times 10^{-2}T^2$ (solid line) with the coefficients determined by least squares fits to the experimental data (black squares) for MgB_2 (sample 1) at 40 K < T < 45 K.

val for sample 1 was relatively small, so that the above ΔC_2 value can be considered as a lower limit.

In our previous measurements of the heat capacity of MgB_2 [7–12], the second (low-temperature) feature on the $C(T)$ curve was not as clearly revealed. Nevertheless, Golubov *et al.* [6] showed that the theory based on a two-band model (with different superconducting gaps on π and σ sheets of the Fermi surface) provides for a better agreement with those experiments than does the single-band theory. Estimates show that the second (smaller) gap must correspond to the critical temperature T_{c2}. Our new results that clearly reveal the second feature in $C(T)$ at $T \approx$ 10–12 fully confirm this hypothesis.

2. In this study, we have observed anomalies in the heat capacity $C(T)$, the thermal conductivity $\kappa(T)$, and the thermal expansion coefficient $\alpha(T)$ of MgB_2 in the region of $T \approx$ 10–12 K. Since this coincidence can hardly be accidental, there must be a certain common reason accounting for the anomalous behavior of three different quantities in this temperature interval. We believe that the temperature $T \approx$ 10–12 K corresponds to a sharp increase in the small superconducting gap corresponding to the second group of charge carriers. In this case, the anomaly in $C(T)$ is obvious. An increase in the thermal conductivity $\kappa(T)$ in the vicinity of a superconducting transition was also frequently observed in various alloys and compounds [27] and attributed to a decrease in the phonon scattering on electrons or holes upon their pairing.

As can be seen from Fig. 6, a small increase in $\kappa(T)$ on the background of generally decreasing thermal conductivity is also manifested at $T \approx$ 38–40 K, that is, in the region of the main critical temperature T_c. Since the concentration of "frozen" charge carriers responsible for this transition ($N \leq 10^{22}$ cm^{-3}) is significantly smaller than the total carrier density ($N \approx 1.5 \times 10^{23}$ cm^{-3} [3]), we may expect a significantly more pronounced effect in the region of the second transition at $T \approx$ 10–12 K, where the main body of charge carriers is subject to pairing.

3. The above approach to interpretation of the experimental data, based on the notion about two superconducting gaps corresponding to two groups of charge carriers, allows us to estimate the values of coefficient γ for each group. The value of γ determined from the temperature dependence of the heat capacity at $T > T_c$ is essentially a sum of the values for each group of charge carriers: $\gamma = \gamma_1 + \gamma_2$, where γ_1 refers to the first group (determining the behavior at $T_c \approx$ 40 K) and γ_2 refers to the second group related to the anomaly at $T_{c2} \approx$ 10–12 K. It was found that γ = 4.39 mJ/(K^2 mol) for sample 1 and 3.99 mJ/(K^2 mol) for sample 2.

Assuming that the contribution due to the electron heat capacity of the first group at temperature below 15 K is negligibly small (as evidenced by a nearly linear plot of C/T versus T^2 in this region above the second anomaly), we can independently estimate the γ_1 and γ_2 values. This yields γ_1 = 1.48 mJ/(K^2 mol) for sample 1 and 1.44 mJ/(K^2 mol) for sample 2; and γ_2 = 2.91 mJ/(K^2 mol) for sample 1 and 2.55 mJ/(K^2 mol) for sample 2. Thus, according to our experimental data for MgB_2, the average coefficients are γ_1 = 1.4–1.5 mJ/(K^2 mol) and γ_2 = 2.55–2.9 mJ/(K^2 mol). Since γ is proportional to the electron density of states (DOS) on the Fermi level, the ratio $\gamma_2/\gamma_1 \approx 2$ characterizes the DOS ratio for the two groups of charge carriers in MgB_2.

4. We can also estimate the ratio of the heat capacity jump at T_c to the γ value. According to the BCS theory, $\Delta C/\gamma T_c = 1.43$. Taking this value for $\Delta C/T_c$ at the peak maximum (Figs. 4 and 5), we obtain $\Delta C/\gamma_1 T_c \approx 2.89$ and 2.57 for samples 1 and 2, respectively. These values are indicative of a strong coupling in the first group of carriers (with lower density) in MgB_2. This conclusion agrees with the results of calculations [28] performed using a model of two groups of carriers, where it was demonstrated that the electron–phonon interaction in MgB_2 is not weak. For the second (low-temperature) transition, the $\Delta C_2/\gamma_2 T_{c2}$ ratio determined directly from Figs. 4 and 5 does not exceed 0.3. Such a small value of this ratio may indicate that the superconducting gap appears only on some parts of the Fermi surface corresponding to the second group of charge carriers.

6. CONCLUSIONS

We have experimentally established that MgB_2, like oxide-based high-temperature superconductors, exhibits a negative thermal expansion coefficient at low temperatures. The anomaly in the thermal expansion is significantly affected by the magnetic field. An anomalous behavior was also observed for the heat capacity and the thermal conductivity of MgB_2 in the region of $T \approx$ 10–12 K. Thus, the anomalies of the three quantities take place in the same temperature interval. These low-temperature anomalies are explained by the presence of a second group of charge carriers and by an increase in the density of the Bose condensate corresponding to these carriers at $T_{c2} \approx 10$–12 K. This conclusion is consistent with the results of investigations of the other properties of MgB_2.

ACKNOWLEDGMENTS

The authors are grateful to Ya. G. Ponomarev for his help in conducting this investigation and to E. G. Maksimov for valuable remarks.

This study was supported by the Russian Foundation for Basic Research (project no. 01-02-16395) and by the Federal Targeted Scientific-Technological Program “Important Directions in the Physics of Condensed Media” (Subprogram “Superconductivity”).

REFERENCES

1. J. Nagamatsu, N. Nakagawa, T. Nuranaka, *et al.*, Nature **410**, 63 (2001).
2. C. Buzea and T. Yamashita, Supercond. Sci. Technol. **14**, R115 (2001).
3. S. L. Bud’ko, C. Petrovic, G. Lapertot, *et al.*, E-print archives, cond-mat/0102413.
4. J. Kortus, I. I. Mazin, K. D. Belashchenko, *et al.*, Phys. Rev. Lett. **86**, 4656 (2001).
5. A. Y. Liu, I. I. Mazin, and J. Kortus, Phys. Rev. Lett. **87**, 087005 (2001).
6. A. A. Golubov, J. Kortus, O. V. Dolgov, *et al.*, J. Phys.: Condens. Matter **14**, 1353 (2002).
7. Y. Wang, T. Plackowski, and A. Junod, Physica C (Amsterdam) **355**, 179 (2001).
8. F. Bouquet, R. A. Fisher, N. E. Phillips, *et al.*, Phys. Rev. Lett. **87**, 047001 (2001).
9. S. L. Bud’ko, G. Lapertot, C. Petrovic, *et al.*, Phys. Rev. Lett. **86**, 1877 (2001).
10. H. D. Yang, J.-Y. Lin, H. H. Li, *et al.*, Phys. Rev. Lett. **87**, 167003 (2001).
11. Ch. Wälti, E. Felder, C. Degen, *et al.*, Phys. Rev. B **64**, 172515 (2001).
12. E. Bauer, Ch. Paul, St. Berger, *et al.*, J. Phys.: Condens. Matter **13**, L487 (2001).
13. N. V. Anshukova, B. M. Bulychev, A. I. Golovashkin, *et al.*, Kratk. Soobshch. Fiz., No. 4, 24 (2002).
14. A. V. Sologubenko, J. Jun, S. M. Kazakov, *et al.*, cond-mat/0111273; cond-mat/0112191; cond-mat/0201517.
15. M. Schneider, D. Lipp, A. Gladun, *et al.*, Physica C (Amsterdam) **363**, 6 (2001).
16. N. V. Anshukova, A. I. Golovashkin, L. I. Ivanova, and A. P. Rusakov, Usp. Fiz. Nauk **167**, 887 (1997) [Phys. Usp. **40**, 843 (1997)].
17. N. V. Anshukova, A. I. Golovashkin, L. I. Ivanova, *et al.*, Pis’ma Zh. Éksp. Teor. Fiz. **71**, 550 (2000) [JETP Lett. **71**, 377 (2000)].
18. N. V. Anshukova, B. M. Bulychev, A. I. Golovashkin, *et al.*, Kratk. Soobshch. Fiz., No. 7, 16 (2001).
19. N. V. Anshukova, Yu. V. Boguslavskiĭ, A. I. Golovashkin, *et al.*, Fiz. Tverd. Tela (St. Petersburg) **35**, 1415 (1993) [Phys. Solid State **35**, 714 (1993)].
20. A. A. Minakov, Yu. V. Boguslavsky, and C. Schick, Thermochim. Acta **317**, 117 (1998).
21. A. A. Minakov, S. A. Adamovsky, and C. Schick, Thermochim. Acta **377**, 173 (2001).
22. H. You, U. Welp, and Y. Fang, Phys. Rev. B **43**, 3660 (1991).
23. Z. J. Yang, M. Yewondwossen, D. W. Lawther, *et al.*, J. Supercond. **8**, 223 (1995).
24. N. V. Anshukova, A. I. Golovashkin, L. I. Ivanova, *et al.*, Pis’ma Zh. Éksp. Teor. Fiz. **71**, 550 (2000) [JETP Lett. **71**, 377 (2000)].
25. O. V. Alexandrov, M. Frencois, T. Graf, and K. Ivon, Physica C (Amsterdam) **170**, 56 (1990).
26. V. V. Eremenko, V. A. Sirenko, G. Shimak, *et al.*, Fiz. Tverd. Tela (St. Petersburg) **40**, 1199 (1998) [Phys. Solid State **40**, 1091 (1998)].
27. R. Berman, *Thermal Conduction in Solids* (Clarendon Press, Oxford, 1976; Mir, Moscow, 1979).
28. Y. Kong, O. V. Dolgov, O. Jepsen, and O. K. Andersen, Phys. Rev. B **64**, 020501 (2001).

Translated by P. Pozdeev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 78–84.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 89–95.
Original Russian Text Copyright © 2003 by Zobov, Popov.

On the Coordinate of a Singular Point of Time Correlation Functions for the System of Nuclear Magnetic Moments of a Crystal

V. E. Zobov[a],* and M. A. Popov[b]

[a] *Kirenskiĭ Institute of Physics, Siberian Division, Russian Academy of Sciences, Akademgorodok, Krasnoyarsk, 660036 Russia*
[b] *Krasnoyarsk State University, Krasnoyarsk, 660041 Russia*
**e-mail: rsa@iph.krasn.ru*

Received December 25, 2002

Abstract—The hypothesis concerning the existence of singular points on the imaginary time axis for a correlation function of a system with the dipole–dipole interaction of nuclear spins of a crystal is verified. Within the framework of the self-consistent fluctuating field theory taking into account the principal corrections related to the correlation of local fields, a result for this coordinate is obtained in terms of the ratios of lattice sums. Experimental values of this coordinate are calculated from the wings of the nuclear magnetic resonance absorption spectrum of a BaF_2 crystal for the magnetic field directions along the three crystallographic axes. Good agreement of the theoretical and experimental results justifies this hypothesis. © *2003 MAIK “Nauka/Interperiodica”.*

1. INTRODUCTION

Nuclear magnetic systems of crystals are convenient objects for studying nonequilibrium statistical physics of many-particle systems. The point is that, first, the exact form of the interaction (dipole–dipole) is known; second, the magnetic system is well isolated from the lattice; and third, one can control the system state with the help of a resonant radio-frequency field and observe it using nuclear magnetic resonance (NMR) methods. An important characteristic of these systems is the rate of attaining equilibrium between the subsystems in the presence of a large mismatch of resonant frequencies determined by the wings of the spectra of the correlation functions. This fact stimulated the study of such systems. In a number of experimental studies, it was found that the frequency dependence of the wings of the spectra can be described by a simple exponential function (see, e.g., [1–3] and the analysis of other experiments in [4]) instead of the expected Gaussian function [5]. The peculiarity of this shape of the spectrum wing is that the corresponding correlation function must have a singular point on the imaginary time axis. In turn, this may indicate a new type of collective effects in such systems. Unfortunately, low accuracy of the registration of a weak signal on the spectrum wing makes the interpretation of its shape ambiguous.

Theoretical investigations [6] confirmed the possibility of the existence of singular points on the imaginary time axis for correlation functions of rigid spin lattices at high temperatures, at least for lattices of large dimension d. The divergence of the form of the spectrum wing from the Gauss distribution is caused by the time fluctuations of the local magnetic field on the spin due to the flips of the neighboring spins creating this field. In turn, these flips are caused by the internal interaction between the spins (dipole–dipole or exchange interaction). The coordinate of a singular point can be easily calculated [6–9] in the approximation of a self-consistent fluctuating local field for lattices of large dimension, i.e., in the case when the correlation of local fields can be neglected. The problem on the variation of this coordinate with decreasing space dimension remains so far unsolved.

In our earlier publications [10, 11], we found the first terms of the expansion in the inverse dimensionality of space for the coordinate of a singular point of the autocorrelation function (ACF) of the Heisenberg model with an isotropic interaction of nearest neighbors. The experimental data [1–3] were obtained for nuclear magnetic systems of crystals with the dipole–dipole interaction. This interaction is characterized by the anisotropy and necessitates the inclusion of distant neighbors. Both these factors are taken into account in this paper when determining the coordinate of a singular point.

In the approximation of a self-consistent fluctuating field described in Section 2, a simple nonlinear equation for the ACF taking into account the axial symmetry of the dipole–dipole interaction in the spin space is written out the coordinate of a singular point of the solution to this equation is determined, and a formula for the variation of this coordinate with a small variation in the coefficients of the power series in the time for the ACF is derived. In Section 3, we calculate the first corrections to the singular point coordinate due to the local field correlation arising when the space dimension is decreased. In Section 4, theoretical results are compared with experimental data.

2. EQUATIONS FOR AUTOCORRELATION FUNCTIONS

We consider a system of nuclear magnetic moments with spin $I = 1/2$ that form a perfect lattice of dimension d. The spin dynamics in a strong constant magnetic field is determined by the secular part of the dipole–dipole interaction with the Hamiltonian [5]

$$\mathcal{H} = \sum_{i \neq j} b_{ij}\left[I_i^z I_j^z - \frac{1}{2}(I_i^x I_j^x + I_i^y I_j^y)\right], \qquad (1)$$

where

$$b_{ij} = \frac{\gamma^2 \hbar (1 - 3\cos^2\theta_{ij})}{2r_{ij}^3},$$

θ_{ij} is angle between the internuclear vector $\mathbf{r}_{ij}$ and the constant magnetic field, and I_i^α is the α component ($\alpha = x, y, z$) of the vector spin operator at the ith site. For a high temperature, the time-dependent correlation functions of two spins located at the ith and jth lattice sites are defined by the expression [5]

$$\Gamma_{\alpha ij}(t) = \frac{\mathrm{Tr}\{\exp(i\mathcal{H}t)I_i^\alpha \exp(-i\mathcal{H})I_j^\alpha\}}{\mathrm{Tr}\{(I_i^\alpha)^2\}}. \qquad (2)$$

We obtain the cross correlation function for $i \neq j$ and the autocorrelation function for $i = j$. By virtue of the translation symmetry of the lattice, we omit subscript ii on the ACF. Taking into account the axial symmetry of the Hamiltonian with respect to spin components, we use the notation $\Gamma_x(t) = \Gamma_y(t) = X(t)$. Autocorrelation functions (2) can be expanded into power series,

$$\Gamma_\alpha(t) = \sum_{n=0}^{\infty} \frac{(-1)^n M_{2n\alpha} t^{2n}}{(2n)!}, \qquad (3)$$

where the nth coefficient of the expansion is determined via the $2n$-fold commutator

$$M_{2n\alpha} = \frac{\mathrm{Tr}\{[\mathcal{H}, \ldots[\mathcal{H}, I_i^\alpha]\ldots]I_i^\alpha\}}{\mathrm{Tr}\{(I_i^\alpha)^2\}}. \qquad (4)$$

It is known [5] that $M_{2n\alpha}$ is the $2n$-order moment of the spectral density of the corresponding ACF.

Up to now, exact equations for ACFs have not been obtained because of the complexity of description of a many-particle system with strong interactions. Many approximate versions of the equations have been proposed. In particular, the following system of nonlinear integral equations for ACF was derived in [12, 13]:

$$\frac{d}{dt}\Gamma_\alpha(t) = -\int_0^t G_\alpha(t - t_1)\Gamma_\alpha(t_1)dt. \qquad (5)$$

The kernels of these equations are represented in the form of a series in irreducible dressed skeleton diagrams with an increasing number of vertices. Each term of the series can be expressed in terms of the multiple time integral of the product of functions $\Gamma_x(t)$ and $\Gamma_z(t)$. All contributions corresponding to the diagrams with two and four vertices are determined.

System of equations (5) has been investigated in [6] in the approximation of lattices of an infinite dimension that corresponds to the approximation of a self-consistent fluctuating local field. In this limit, the equations for ACF correspond to the averaged precession of the magnetic moment in a three-dimensional Gaussian random local field whose correlation functions are expressed via the spin ACFs as

$$\langle \omega_\alpha(t)\omega_\alpha(0)\rangle = \langle \omega_\alpha^2\rangle \Gamma_\alpha(t), \qquad (6)$$

where

$$\langle \omega_x^2\rangle = \langle \omega_y^2\rangle = \frac{S_1}{4}, \quad \langle \omega_z^1\rangle = S_1, \quad S_1 = \sum_j b_{ij}^2.$$

For series $G_{\alpha 0}(t)$, majorizing series are found and the existence of singular points of the ACFs on the imaginary time axis is established. In the neighborhood of the nearest singular point with coordinate τ_0, the principal part has the form

$$\Gamma_\alpha(t) \approx A_\alpha(\tau_0 + it)^{-2}. \qquad (7)$$

The coordinate of the singular point estimated by the spectral moments on the order from two to ten is

$$\tau_0 = \frac{2.77}{M_{2x}^{1/2}}, \qquad (8)$$

where $M_{2x} = 5S_1/4$ is the second moment of the spectrum of function $\Gamma_x(t)$.

If the space has a finite dimension, in the series for the kernel, one should take into account the additional terms

$$\delta G_\alpha(t) = G_\alpha(t) - G_{\alpha 0}(t)$$

that contain coupling loops and multiple interactions of neighboring spins. These corrections reflecting the correlation of the local fields vanish in the limit as $d \longrightarrow \infty$. It is natural to expect that, if the dimension of the space is sufficiently large, the relative variations of moments $M_{2n\alpha}$ of the ACFs and the coordinates of the singular point τ_c with additional terms taken into account are small (on the order of $\varepsilon \sim 1/d$); i.e.,

$$M_{2n\alpha} = M^{(0)}_{2n\alpha} - \varepsilon M^{(1)}_{2n\alpha} + \dots, \quad \tau_c = \tau_0 + \varepsilon\delta\tau_c + \dots$$

The coordinate τ_c of the singular point (equal to the convergence radius of the power series in time) can be determined as the limit of the ratio between the moments,

$$\begin{aligned}\tau_c^2 &= \lim_{n\to\infty}\frac{M_{2(n-1)\alpha}}{M_{2n\alpha}}2n(n-1)\\ &= \tau_0^2\lim_{n\to\infty}\frac{1-\varepsilon M^{(1)}_{2(n-1)\alpha}/M^{(0)}_{2(n-1)\alpha}+\dots}{1-\varepsilon M^{(1)}_{2n\alpha}/M^{(0)}_{2n\alpha}+\dots},\end{aligned}$$

which gives

$$2\frac{\delta\tau_c}{\tau_0} = \lim_{n\to\infty}\left(\frac{M^{(1)}_{2n\alpha}}{M^{(0)}_{2n\alpha}} - \frac{M^{(1)}_{2(n-1)\alpha}}{M^{(0)}_{2(n-1)\alpha}}\right). \tag{9}$$

Due to the complexity of series $G_{\alpha 0}(t)$, even the first corrections $M^{(1)}_{2n}$ (linear in ε) can hardly be found. Therefore, to estimate $\delta\tau_c$, we take the approximate version of the equation for $\Gamma_\alpha(t)$ that allows one to determine high-order moments. In the case of dipole–dipole interaction (1), as a result of the efforts of many authors [3, 4, 8, 9, 14–18], it has been established that a good approximation is obtained if the (longitudinal) interaction between the *z* components of the spin is fully taken into account and the *xx* and *yy* (transverse) interactions are taken into account to the minimal required extent. For the ACF of the *x* component of the spin, we take the Anderson–Weiss equation [19] describing the spin precession in a Gaussian longitudinal field, renormalize this field to ensure the correct value of the second moment M_{2x}, and determine its correlation function (6) via the ACF of the *z* component of the spin [4, 8. 9]:

$$X(t) = \exp\left\{\int_0^t\int_0^{t_1}\Gamma_z(t_2)\,dt_1dt_2\right\}. \tag{10}$$

Here and below, in the formulas we pass to the imaginary dimensionless time $t \longrightarrow it(5S_1/4)^{-1/2}$, preserving its previous notation. After this substitution, the argument of the exponential function in expression (10) becomes positive and the coefficient M_{2x} of the integral becomes equal to unity. For the ACF of the *z* component, we use the equation

$$\Gamma_z(t) = 1 + \frac{2}{5}\int_0^t\int_0^{t_1}X^2(t_2)dt_1dt_2. \tag{11}$$

It will be shown below that, for the location of the singular point on the imaginary time axis, this equation gives an insignificant difference as compared with the equation used before [4, 8, 9],

$$\Gamma_z(t) = 1 + \frac{2}{5}\int_0^t\int_0^{t_1}X^2(t_2)\Gamma_z(t_1-t_2)dt_1dt_2, \tag{12}$$

but simplifies the calculations.

Since the kernel of Eq. (11) is $X^2(t)$, we denote the latter by $Y(t)$ and obtain, using formula (10), the differential equation

$$\begin{aligned}\frac{d}{dt}Y(t) &= 2Y(t)\int_0^t\Gamma_z(t_1)dt_1\\ &= 2Y(t)\left\{t + \frac{2}{5}\int_0^t\int_0^{t_1}\int_0^{t_2}Y(t_3)dt_1dt_2dt_3\right\}.\end{aligned} \tag{13}$$

Substituting $Y(t)$ in the form of the series

$$Y(t) = \sum_{n=0}^{\infty}t^{2n}Y_{2n} \tag{14}$$

into Eq. (13) and equating the coefficients of the same powers of time, we obtain the recurrence equation

$$Y_{2n} = \frac{1}{n}Y_{2n-2} + \frac{2}{5n}\sum_{k=0}^{n-2}\frac{Y_{2k}Y_{2(n-k-2)}}{(2k+1)(2k+2)(2k+3)}. \tag{15}$$

The coordinate of the singular point (equal to the convergence radius of series (14)) can be determined (taking into account the order of pole (7)) as the limit of the ratio

$$\tau_0^2 = \lim_{n\to\infty}\frac{Y_{2n-2}(2n+2)(2n+3)}{Y_{2n}(2n-1)2n}. \tag{16}$$

Solving recurrence equation (15) and using formula (16), we find

$$\tau_0 = \frac{2.69}{M_{2x}^{1/2}},$$

whereas, using Eq. (12) with the convolution, we obtain

$$\tau_0 = \frac{2.68}{M_{2x}^{1/2}}.$$

Both these values differ from a more precise value (8) by 3%. This difference exerts no substantial influence on the values of corrections $\delta\tau_c$; therefore, for the calculation of these corrections, we will use the simplest equation.

3. CALCULATION OF THE CORRECTIONS TO THE COORDINATE OF THE SINGULAR POINT OF THE CORRELATION FUNCTION

Let us consider function $\Gamma_x(t)$ as the generating function of the lattice patterns formed by bonds b_{ij}. This can be justified by the structure of expression (4) for moments, where each commutator with $\mathcal{H}$ adds a bond b_{ij} to the constructed pattern. In more detail, the rules of constructing the patterns are considered in [6, 10, 11]. The solution to the system of equations (10) and (11) of the zeroth-order approximation is the generating function of the root trees constructed from the double bonds. On these trees, the double bonds caused by the *zz* interactions (*z* fields in Eq. (10)) alternate with the bonds caused by the *xx* and *yy* interactions (the variations of *z* fields described by Eq. (11)). An arbitrary number of branches which have no intersections can emerge from any node of this tree. This can hold as $d \longrightarrow \infty$. When constructing the patterns on finite-dimensional lattices, there is a probability that the branches intersect either directly near the node from which they emerge, which results in a multiple interaction of the neighbors, or near a far node, forming a loop of bonds. For a tree with such a fragment, the weighting factor obtained during the construction of the tree by calculating the initial multiple commutators (4) does not coincide with the factor obtained in the case of intersection due to the mechanical overlap of branches constructed independently by Eqs. (10) and (11) and located on a real lattice. Therefore, the latter should be eliminated and replaced by trees with a correct weight.

To theoretically estimate the variation of the coordinate of the singular point of the ACF caused by the described variations of the moments, we consider the dimension *d* of the space as a variable. If *d* is taken sufficiently large, then the intersection probability is low and we can consider only the simplest intersections, namely, the quadruple interaction of neighbors and loops in the form of a triangle formed by four bonds. Such contributions are contained even in the fourth moment of the ACF. The intersections in which more bonds are involved give a higher order of smallness in $1/d$ [9, 10]. In the case of the dipole–dipole interaction and $d = 3$, their smallness is confirmed by the ratio of the values of the lattice sums [2, 20, 21].

3.1. Quadruple Interaction of Neighbors

As is shown in [3], in the case when the number of neighbors is not very large, to improve the accuracy of the main approximate equation, one should replace the Anderson–Weiss function (10) by the product

$$P_i(t) = \prod_j F_{ij}(t), \tag{17}$$

$$F_{ij}(t) = 1 + \frac{b_{ij}^2}{S_1}\int_0^t\int_0^{t_1} F_{ij}(t_2)\Gamma_{zj/i}(t_1 - t_2)dt_1dt_2, \tag{18}$$

where $\Gamma_{zj/i}(t)$ is the ACF of *z* component of the *j*th spin disregarding the interaction with the *i*th spin on which the field is considered,

$$\Gamma_{zj/i}(t) = 1 + \frac{2}{5}\sum_{k \neq i,\, j}\frac{b_{jk}^2}{S_1} \times \int_0^t\int_0^{t_1}\frac{P_j(t_2)P_k(t_2)}{F_{ij}(t_2)F_{jk}^2(t_2)}\Gamma_{zj/i}(t_1 - t_2)dt_1dt_2. \tag{19}$$

In addition, recurrence of the interaction of the *j*th and *k*th spins is excluded in the kernel of this integral equation.

We differentiate the square of function $P_{ij}(t)$ (17) with respect to time:

$$\frac{d}{dt}P_i^2(t) = 2P_i^2(t)\sum_{j \neq i}\frac{1}{F_{ij}(t)}\frac{d}{dt}F_{ij}(t). \tag{20}$$

Taking into account relations (17)–(20), one can easily verify that, neglecting the contribution of separate interaction b_{ij}^2 in comparison with sum S_1, we return to the equation of zeroth-order approximation (13). To within the first correction from these contributions, we find that

$$Y(t) = P_i^2(t) = Y_0(t) - \frac{S_2}{S_1^2}Y_1(t).$$

Here,

$$S_2 = \sum_j b_{ij}^4$$

and $Y_1(t)$ is the solution to the equation

$$\begin{aligned}\frac{d}{dt}Y_1(t) = {} & 2Y_1(t)\int_0^t\Gamma_z(t_1)dt_1 + \frac{4}{5}Y_0(t)\int_0^t\int_0^{t_1}\int_0^{t_2}Y_1(t_3)dt_1dt_2dt_3 + C\frac{dY_0(t)}{dt}\int_0^t\int_0^{t_1}\Gamma_z(t_2)dt_1dt_2 \\ & + \frac{4}{5}DY_0(t)\int_0^t\int_0^{t_1}\int_0^{t_2}Y_0(t_3)dt_1dt_2dt_3 + \frac{8}{5}AY_0(t)\int_0^t\int_0^{t_1}\int_0^{t_2}Y_0(t_3)dt_1dt_2dt_3\int_0^{t_3}\int_0^{t_1}\Gamma_z(t_5)dt_4dt_5 \\ & - 2BY_0(t)\int_0^t\Gamma_z(t-t_1)dt_1\int_0^{t_1}\int_0^{t_2}\Gamma_z(t_3)dt_2dt_3,\end{aligned} \tag{21}$$

where $\Gamma_z(t)$ is determined via $Y_0(t)$ with the help of Eq. (11). The coefficients A, B, C, and D are introduced into Eq. (21) in order to separate the contributions of different kinds. For $A = 3/2$, $C = 1$, $D = 1$, and $B = 0$, we obtain the contribution corresponding to the exclusion of the quadruple interaction of neighbors due to the intersection of the tree branches emerging from the same node. For $A = 0$, $C = 0$, $D = 0$, and $B = 1$, we obtain the contribution of the quadruple interaction with the correct weight, which, therefore, appears with the opposite sign. This contribution stems from the second iteration in Eq. (18).

3.2. Triangle Composed of Four Bonds

There are two reasons for the formation of the simplest triangle-like loops: first, the contribution of the cross correlation function $\Gamma_{zjk}(t)$ to the correlator of the local field on the separated spin in expressions (10) and (17) and, second, the result of the simultaneous action of the field of the third spin on the two spins bonded by the transverse (flip–flop) interaction in the kernels of integral equations (11) and (19) for $\Gamma_z(t)$. In the one-loop approximation, we obtain

$$Y(t) = Y_0(t) - \frac{S_3}{S_1^2}Y_1(t),$$

where

$$S_3 = \sum_{i,\,j} b_{ij}^2 b_{ik} b_{jk},$$

$Y_0(t)$ is the zeroth-order approximation (13), and, for the first-order correction $Y_1(t)$, we obtain equations of the form (21) with the following values of the parameters: $A = 1$, $B = 0$, $C = 0$, and $D = 1$.

The structure of Eq. (21) becomes clear when interpreted in terms of generating functions of root trees having an embedded fragment with a branch intersection. The summands on the right-hand side with coefficients A, B, C, and D represent the contribution of the corresponding fragment with the intersection joined to the tree root. If, however, the intersection occurs at a far node of the tree, then the necessary chain of bonds from the root to the fragment is composed via iterations with the help of the first two terms on the right-hand side of the equation. Recall that we operate in the approximation linear in intersections; i.e., we assume that there is at most one intersection on the tree. By virtue of this assumption and the translation invariance of the lattice sites, the form of the expression for the fragment is independent of its location in the tree. However, its contribution to $Y_1(t)$ obviously depends on the length of the chain leading to this site.

3.3. Calculation of Corrections

Substituting series (14) for functions $Y_0(t)$ and $Y_1(t)$ into Eq. (21), we obtain the following recurrence equation for coefficients $Y_{2n}^{(1)}$ of function $Y_1(t)$:

$$\begin{aligned}Y_{2n}^{(1)} = {} & \frac{1}{n}Y_{2n-2}^{(1)} + \frac{2}{5n}\sum_{k=0}^{n-2}\frac{Y_{2k}^{(0)}Y_{2(n-k-2)}^{(1)} + Y_{2k}^{(1)}Y_{2(n-k-2)}^{(0)}}{(2k+1)(2k+2)(2k+3)} + D\left(Y_{2n}^{(0)} - \frac{1}{n}Y_{2n-2}^{(0)}\right) + \frac{C}{2}\left(1-\frac{1}{n}\right)Y_{2n-2}^{(0)} \\ & + \frac{C}{5n}\sum_{k=0}^{n-2}\frac{2(n-k-2)Y_{2k}^{(0)}Y_{2(n-k-2)}^{(0)}}{(2k+1)(2k+2)(2k+3)(2k+4)} + \frac{2A}{5n}\sum_{k=0}^{n-3}\frac{Y_{2k}^{(0)}Y_{2(n-k-3)}^{(0)}}{(2k+3)(2k+4)(2k+5)} \\ & + \frac{8A}{25n}\sum_{k=0}^{n-4}\sum_{p=0}^{n-k-4}\frac{Y_{2p}^{(0)}Y_{2k}^{(0)}Y_{2(n-k-p-4)}^{(0)}}{(2p+1)(2p+2)(2p+3)(2p+4)(2k+2p+5)(2k+2p+6)(2k+2p+7)}\end{aligned} \tag{22}$$

$$-\frac{B}{6n}Y^{(0)}_{2n-4}-\frac{4B}{5n}\sum_{k=0}^{n-3}\frac{(2k)!}{(2k+5)!}Y^{(0)}_{2k}Y^{(0)}_{2(n-k-3)}-\frac{4B}{25n}\sum_{k=0}^{n-4}\sum_{p=0}^{k}\frac{(2k-2p)!(2p)!}{(2k+7)!}Y^{(0)}_{2p}Y^{(0)}_{2(k-p)}Y^{(0)}_{2(n-k-4)}.$$

Using this equation, we determine coefficients $Y^{(1)}_{2n}$ for different contributions; then, by formula (9), we find the corresponding corrections to the coordinate of the singular point. The correction corresponding to the exclusion of the forbidden quadruple interaction of neighbors ($A = 3/2$, $B = 0$, $C = 1$, $D = 1$) is equal to

$$\frac{S_2}{S_1^2}\frac{\delta\tau_c}{\tau_0} = 1.507\frac{S_2}{S_1^2}, \tag{23}$$

the correction corresponding to the addition of the allowed quadruple interaction of neighbors ($A = 0$, $B = 1$, $C = 0$, $D = 0$) is

$$\frac{S_2}{S_1^2}\frac{\delta\tau_c}{\tau_0} = -0.124\frac{S_2}{S_1^2}, \tag{24}$$

and the correction corresponding to the inclusion of correlation of fields in the form of a triangle composed of four bonds ($A = 1$, $B = 0$, $C = 0$, $D = 1$) is

$$\frac{S_3}{S_1^2}\frac{\delta\tau_c}{\tau_0} = 0.55\frac{S_3}{S_1^2}. \tag{25}$$

4. COMPARISON WITH EXPERIMENT

In a neighborhood of the nearest singular point, principal part (7) determines the wing of the ACF spectrum,

$$g(\omega) \approx A_\alpha|\omega|\exp(-|\omega|\tau_c). \tag{26}$$

Moreover, since the singularities of all time correlation functions of the spin system under consideration must be located at the same point, the argument of the exponential function in formula (26) is the same (including the NMR absorption spectrum, i.e., the Fourier transform of the correlation function of the total spin of the system [5]).

The wing of the NMR absorption spectrum was experimentally investigated in [3] for a single crystal of BaF_2 with a magnetic field directed along the crystallographic axes [111], [110], and [100]. The frequency dependence of the spectrum for detuning from the spectrum center exceeding $2.2\sqrt{M_2}$ (for [100], even $2.1\sqrt{M_2}$), where $M_2 = 9S_1/4$, is well described by exponential function (26) (this is justified by the fact that the experimental points shown in Fig. 3 in [3] in the semilogarithmic coordinates lie on a straight line). The values of argument τ_e of the exponential function in formula (26), at which the best agreement with the experiment is achieved in the detuning interval from $2.2\sqrt{M_2}$ to $3\sqrt{M_2}$, are given in the table in the form of the ratio to the limiting theoretical value of τ_0 (8). The mean square error increases from 0.5% in orientation [111] to 2% in orientation [100] due to a decrease of the signal-to-noise ratio with increasing NMR line width. However, the actual accuracy of determining τ_e is lower, first, because of the systematic distortions in the wing introduced by the spectrometer, and second, since the simple dependence (26) is attained in the limit as $\omega \longrightarrow \infty$, i.e., in the spectral region that is inaccessible because of the noise. As the center of the spectrum is approached, the deviation of its shape from dependence (26) becomes noticeable.

Let us return to theoretical results. Collecting corrections (23)–(25), we obtain

$$\frac{\tau_c}{\tau_0} = 1 + 1.38\frac{S_2}{S_1^2} + 0.55\frac{S_3}{S_1^2}. \tag{27}$$

Substituting the values of the lattice sums for a simple cubic lattice from [2], we find the values of this ratio for the three main orientations of the magnetic field which are presented in the table. It is difficult to estimate the accuracy of these values, because expansion (27) is asymptotic in $1/d$. We estimate the error caused by the replacement of complete equation (5) with simplified equations (10) and (11), as well as the error in determining the convergence radius of the series by its coefficients, in the range of 2–3%.

The results presented in the table show good agreement between the theoretical and experimental values of the coordinates of singular points of the correlation functions for all three orientations of the field. It should be emphasized that the orientation dependence of the second moment, which is the frequency scale of the spectrum, does not affect the ratios presented in the table. Their values depend not on the mean square of local fields, but on the extent of correlation of these fields, which is expressed in formula (27) in terms of the ratio of different lattice sums. On the one hand, the coincidence of two independent estimates for the coordinate shows that the errors whose values were difficult

The ratios of the experimental τ_e and theoretical τ_c values of coordinates of the singular points of the correlation functions to limiting value τ_0 (8) for three directions of the magnetic field

Field direction	τ_e/τ_0	τ_c/τ_0
[111]	1.10	1.14
[110]	1.24	1.25
[100]	1.33	1.34

to estimate are small. On the other hand, the values of the coordinate ratios obtained may indicate that, with the decrease in the space dimension from $d = \infty$ to $d = 3$, the singular point moves but does not go to infinity. A final conclusion can be made after increasing the accuracy of theoretical calculations and experimental measurements.

Note in conclusion that result (27) can be applied to experiments performed on other crystals and for other field orientations after substituting the corresponding values of the lattice sums.

ACKNOWLEDGMENTS

The authors are grateful to Yu.N. Ivanov and A.I. Livshits for the presented experimental NMR spectra.

This study was supported by the Russian Foundation for Basic Research (project no. 02-02-17463).

REFERENCES

1. D. A. McArthur, E. L. Hahn, and R. E. Walstedt, Phys. Rev. **188**, 609 (1969).
2. H. T. Stokes and D. C. Ailion, Phys. Rev. B **15**, 1271 (1977).
3. V. E. Zobov, M. A. Popov, Yu. N. Ivanov, and A. I. Livshits, Zh. Éksp. Teor. Fiz. **115**, 285 (1999) [JETP **88**, 157 (1999)].
4. V. E. Zobov, A. A. Lundin, and O. E. Rodionova, Zh. Éksp. Teor. Fiz. **120**, 619 (2001) [JETP **93**, 542 (2001)].
5. A. Abragam, *The Principles of Nuclear Magnetism* (Clarendon Press, Oxford, 1961; Inostrannaya Literatura, Moscow, 1963).
6. V. E. Zobov, Teor. Mat. Fiz. **77**, 426 (1988); Teor. Mat. Fiz. **84**, 111 (1990).
7. M. Blume and J. Hubbard, Phys. Rev. B **1**, 3815 (1970).
8. A. A. Lundin, A. V. Makarenko, and V. E. Zobov, J. Phys.: Condens. Matter **2**, 10131 (1990).
9. V. E. Zobov and A. A. Lundin, Zh. Éksp. Teor. Fiz. **106**, 1097 (1994) [JETP **79**, 595 (1994)].
10. V. E. Zobov, Teor. Mat. Fiz. **123**, 116 (2000).
11. V. E. Zobov and M. A. Popov, Teor. Mat. Fiz. **131**, 491 (2002).
12. P. Résibois and M. De Leener, Phys. Rev. **152**, 305 (1966).
13. P. Borckmans and D. Walgraef, Physica (Amsterdam) **35**, 80 (1967); Phys. Rev. **167**, 282 (1968).
14. P. Borckmans and D. Walgraef, Phys. Rev. B **7**, 563 (1973).
15. G. Sauermann and M. Wiegand, Physica B (Amsterdam) **103**, 309 (1981).
16. G. E. Karnaukh, A. A. Lundin, B. N. Provotorov, and K. T. Summanen, Zh. Éksp. Teor. Fiz. **91**, 2229 (1986) [Sov. Phys. JETP **64**, 1324 (1986)].
17. M. I. Bulgakov, A. D. Gul'ko, F. S. Dzheparov, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **58**, 614 (1993) [JETP Lett. **58**, 592 (1993)].
18. B. N. Provotorov, T. P. Kulagina, and G. E. Karnaukh, Zh. Éksp. Teor. Fiz. **113**, 967 (1998) [JETP **86**, 527 (1998)].
19. P. W. Anderson and P. R. Weiss, Rev. Mod. Phys. **25**, 269 (1953).
20. W. F. Wurzbach and S. Gade, Phys. Rev. B **6**, 1724 (1972).
21. V. E. Zobov and M. A. Popov, Zh. Éksp. Teor. Fiz. **103**, 2129 (1993) [JETP **76**, 1062 (1993)]; Zh. Éksp. Teor. Fiz. **108**, 1450 (1995) [JETP **81**, 795 (1995)].

Translated by E. Pankratiev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 85–92.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 96–104.
Original Russian Text Copyright © 2003 by Zvezdin, Plokhov.

Nonlinear Dynamics of Semiclassical Spin in a Time-Dependent Magnetic Field

A. K. Zvezdin[a],* and D. I. Plokhov[b],**

[a]*Institute of General Physics, Russian Academy of Sciences, ul. Vavilova 38, Moscow, 119991 Russia*
[b]*Moscow State University, Vorob'evy gory, Moscow, 119899 Russia*
**e-mail: zvezdin@gagarinclub.ru*
***e-mail: plokhov@nm.ru*

Received April 14, 2002

Abstract—The dynamics of magnetic nanoclusters (or molecules) with a large spin in a magnetic field whose strength varies in proportion to time is analyzed. Such a field breaks the symmetry relative to rotations through 2π, as well as clockwise and counterclockwise rotations, and induces a number of new coherent quantum effects in the spin dynamics, such as the formation of a band energy spectrum with continuous spin states or the emergence of "Bloch" oscillations in spin precession and interband Zener tunneling. Bloch oscillations are manifested in experiment as equidistant identical jumps on the magnetization curve. The interband Zener tunneling gives rise to additional jumps and peaks on the susceptibility of the system. *© 2003 MAIK "Nauka/Interperiodica".*

1. The interest in problems related to the dynamics of spin systems has received considerable impetus in recent years [1–11]. This is mainly due to the recent discovery of macroscopic quantum tunneling of magnetization and molecular bistability and quantum hysteresis, as well as a new type of magnetic oscillations associated with the Berry phase. These mesoscopic effects were discovered in so-called systems with a giant spin and in systems of magnetic Mn_{12} and Fe_8 nanoclusters with the ground-state spin equal to 10. The problems are related to the macroscopic quantum coherence, quantum measurements in spin systems, and the mechanisms of destruction of quantum correlations due to the interaction with the environment, especially upon a transition from quantum objects to macroscopic objects. These problems are of practical interest for magnetic nanoelectronics (spintronics) and quantum informatics. Such nanoclusters with a giant spin may be used as bistable elements for future-generation magnetic storage devices. These nanoclusters are also of interest to specialists in quantum computers for possible implementation of cubits [12–14].

This study is aimed at analysis of the dynamics of an anisotropic quantum system with a large spin, placed in a magnetic field that increases (decreases) with time. Such a field generates a torque acting on the spin and inducing its precession, thus revealing new features in the dynamics of the spin system. Here, we develop the ideas formulated by one of the authors for magnetic clusters [15, 16], metal rings and ring-shaped molecules [17].

2. Let us consider a quantum-mechanical system (ion, molecule, cluster, etc.) experiencing the action of a magnetic field that increases (decreases) at a constant rate. We write the Hamiltonian of the system in the form

$$H = g\mu_B \mathbf{J}\cdot\mathbf{B}(t) + V_{CF}, \qquad (1)$$

where g is the Lande factor, μ_B is the Bohr magneton, and V_{CF} is the crystal field operator. We assume that the total angular momentum $J \gg 1$; consequently, we will describe the dynamics of this system using a semiclassical approximation. We also assume that $g = 2$. The crystal field $V_{CF} = V_{CF}^1 + V_{CF}^2$, where V_{CF}^1 is a field with easy-plane symmetry, while V_{CF}^2 creates anisotropy in the easy plane. Here $|V_{CF}^2| \ll |V_{CF}^1|$. We also assume that the Cartesian z axis is perpendicular to the easy plane and coincides in direction with the magnetic field vector $\mathbf{B}$.

3. In order to describe the spin dynamics, we will use the coherent quantum states $|\theta, \varphi\rangle$ [18], where θ and φ are the polar and azimuth angles of the angular momentum. We write the Lagrangian of the system in the form

$$L = \frac{M}{\gamma}(1-\cos\theta)\dot{\varphi} - U_a(\theta, \varphi) + MB(t)\cos\theta, \qquad (2)$$

where M is the magnetic moment of a particle and $\gamma = e/mc$. Formula (2) can be derived from Hamiltonian (1) by using the standard technique of coherent quantum states. The first term in Eq. (2), which is known as the Wess–Zumino term, reflects the nonorthogonality of coherent states at different instants; the second term is the anisotropy energy, while the last term is just the

Zeeman energy. In the simplest case (which, however, is sufficient for clarifying the basic aspects of the problem), the (orthorhombic) anisotropy energy can be chosen in the form

$$\begin{gathered} U_a(\theta, \varphi) = -K_1 \sin^2\theta - K_2 \sin^2\theta \cos 2\varphi, \\ 0 < K_2 \ll K_1. \end{gathered} \tag{3}$$

We will also consider other types of anisotropy, viz., the tetragonal anisotropy

$$\begin{gathered} U_a(\theta, \varphi) = -K_1 \sin^2\theta - K_2 \sin^4\theta - K_3 \sin^6\theta \\ - K_4 \sin^4\theta \cos 4\varphi \end{gathered} \tag{4}$$

and the hexagonal anisotropy

$$\begin{gathered} U_a(\theta, \varphi) = -K_1 \sin^2\theta - K_2 \sin^4\theta - K_3 \sin^6\theta \\ - K_4 \sin^6\theta \cos 6\varphi. \end{gathered} \tag{5}$$

The anisotropy constants K_i are chosen so that they ensure easy-plane anisotropy.

4. The partition function of a quantum system can be represented as a functional integral in the Euclidean space ($\tau = it$),

$$Z = \int \exp\left(\frac{1}{\hbar}\int_0^{\hbar\beta} L(\theta, \varphi) d\tau\right) D\cos\theta D\varphi. \tag{6}$$

Here, $\beta = 1/k_B T$, $\theta = \theta(\tau)$ and $\varphi = \varphi(\tau)$. Since $K_2 \ll K_1$, we assume that $\theta - \pi/2 \ll 1$ in moderate magnetic fields. In this case, expression (6) for the partition function can easily be integrated with respect to variable θ. The computational procedure is as follows. The action

$$S = \int_0^{\hbar\beta} L(\theta, \varphi) d\tau$$

attains its minimal value if

$$\cos\theta \approx \frac{M}{2K_1}\left(B - \frac{i\dot{\varphi}}{\gamma}\right). \tag{7}$$

Substituting expression (7) into (6), expanding the action into a series in the vicinity of the minimum to within the second-order terms, and evaluating the Gaussian integral appearing in this case, we obtain

$$Z = \int \exp\left(\frac{1}{\hbar}\int_0^{\hbar\beta} L_{\text{eff}}(\varphi) d\tau\right) D\varphi. \tag{8}$$

Here, we have omitted the insignificant preexponential factor to simplify the notation.

The effective Lagrangian appearing in expression (8) can be written to within the total time derivative in the form

$$\begin{gathered} L_{\text{eff}} = \frac{I\dot{\varphi}^2}{2} - U_a(\varphi) + \gamma I \dot{B} \varphi, \\ U_a(\varphi) = -K_2 \cos 2\varphi. \end{gathered} \tag{9}$$

Here, we have introduced the notation $I = M^2/2K_1\gamma^2$ for the effective moment of inertia of the particle. Analogous formulas are valid for tetragonal (4) and hexagonal (5) anisotropy. Constant K_4 plays the role of quantity K_2. Anisotropy constants K_2 and K_3 lead to insignificant redefinition of quantity K_1 in Eq. (9). In the subsequent analysis, we will not explicitly indicate these distinctions while writing all the formulas for the simplest case (3) unless this leads to misunderstanding.

It should be emphasized that variable φ is defined here not on the set S^1 ($0 \le \varphi \le 2\pi$) as is usually done in the theory of angular momentum, but on the set $\mathbb{R}^1$ of all real numbers. In the given problem, the latter set is a trivial fibration of space S^1, playing the role of the basis of space $\mathbb{R}^1$. This is due to the fact that the presence of field $B_z(t) = at$ breaks the system symmetry relative to the transformation $\varphi \longrightarrow \varphi + 2\pi n$, where n is an integer. Indeed, this symmetry breaking is obvious since the magnetic field $B_z(t)$ varying with time generates a steady-state vortex electric field $E_\varphi = -\dot{B}R/2\pi c$, where R is the radius of the circle of rotation. Consequently, rotations $\varphi \longrightarrow \varphi + 2\pi n$, as well as clockwise and counterclockwise rotations, are nonequivalent.

Systems with a potential energy of the "washboard" type,

$$U(x) = U_0(x) + cx,$$

where $U_0(x)$ is a periodic function and c is a certain constant (the potential energy appearing in expression (9) is precisely of this type), were analyzed earlier. An electron moving in a crystal in a constant external electric field [19–21] or a Josephson junction carrying a direct current possesses [22–24] energy of this form. For this reason, we can expect that the spin dynamics will exhibit some properties similar to those of the above-mentioned systems. The formation of the band energy spectrum, Bloch oscillations, and the interband Zener tunnel effects are examples of such characteristic features.

5. We can proceed to analysis of quantum effects in the same way as it was done by Anderson [25]. We will treat the macroscopic generalized coordinate and momentum of the system as operators. The generalized momentum corresponding to coordinate φ is given by

$$P_\varphi = \frac{\partial L}{\partial \dot{\varphi}} = I(\dot{\varphi} - \gamma B).$$

This expression should be supplemented with an arbitrary constant α since Lagrangian (9) is defined to within the total derivative $\alpha\dot{\varphi}$. In quantum mechanics, constant α is defined in terms of the Berry phase [5, 26]. For our purposes (in particular, for considering Bloch oscillations in spin precession), it is sufficient to analyze the case of an integral spin, i.e., to set $\alpha = 0$ [26]. Then the system Hamiltonian can be written in the form

$$H = \frac{1}{2I}(P_\varphi + \gamma IB)^2 + U_a(\varphi), \qquad (10)$$

where $P_\varphi = -i\hbar\partial/\partial\varphi$. The gauge transformation $\psi \longrightarrow \psi\exp(i\gamma IB\varphi/\hbar)$ leads to the Schrödinger equation of the type

$$i\hbar\frac{\partial\psi}{\partial t} = H\psi, \quad H = \frac{P_\varphi^2}{2I} + U_a(\varphi) - \gamma I\dot{B}\varphi. \qquad (11)$$

It is useful to consider precession for $U_a(\varphi) = U_0 = \text{const}$. In this case, Eq. (11) can be written in the form

$$\left[\frac{d^2}{d\varphi^2} + \frac{2I}{\hbar^2}\gamma I\dot{B}\varphi + \frac{2I}{\hbar^2}(E - U_0)\right]\psi(\varphi) = 0. \qquad (12)$$

Point $\varphi_T = (U_0 - E)/\gamma I\dot{B}$ is a classical turning point. Precession takes place for $\varphi \geq \varphi_T$, while for $\varphi < \varphi_T$ it is impossible from the classical point of view.

Introducing for convenience a new dimensionless variable

$$\xi = \frac{\varphi - \varphi_T}{l_0}, \quad l_0 = \left(\frac{\hbar^2}{2I^2\gamma\dot{B}}\right)^{1/3},$$

we can write Eq. (12) in the form

$$\left(\frac{d^2}{d\xi^2} + \xi\right)\bar{\psi}(\xi) = 0.$$

Here, $\bar{\psi}(\xi) = \psi(\varphi_T + l_0\xi)$. The solution to this equation has the form

$$\bar{\psi}(\xi) = N\mathrm{Ai}(-\xi), \quad N = \left(\frac{2I^{1/2}}{(\gamma\dot{B})^{1/2}\hbar^2}\right)^{1/3}, \qquad (13)$$

where $\mathrm{Ai}(x)$ is the Airy function. In the region $\varphi \geq \varphi_T$, the wave function oscillates at a frequency increasing with φ. In this case, the energy of the spin increases continuously under the action of the torque. For $|\varphi - \varphi_T| \gg l_0$, function (13) asymptotically approaches the function

$$\psi(\varphi) = N\left(\frac{l_0}{\varphi - \varphi_T}\right)^{1/4}\sin\left[\frac{2}{3}\left(\frac{\varphi - \varphi_T}{l_0}\right)^{3/2} + \frac{\pi}{4}\right],$$

which readily leads to the following expression for the period of precession:

$$\Delta\varphi = \frac{2\pi\hbar}{I\sqrt{2\gamma\dot{B}(\varphi - \varphi_T)}}.$$

6. In order to analyze the spin dynamics in the general case, we first consider the properties of Hamiltonian (11) for $\dot{B} = 0$. Before passing to an analysis of the equation with $U_a(\varphi) \neq \text{const}$, we consider the boundary conditions to this equation, i.e., the behavior of function $\psi(\varphi)$ upon a change in φ by 2π. The eigenstates of Hamiltonian (11) are the Bloch functions

$$\psi_n(\varphi + \pi) = e^{i\pi m}\psi_n(\varphi), \qquad (14)$$

where m is an arbitrary real number ($m \in \mathbb{R}^1$) and n is the energy band number. In analogy with the term "charge states" used for characterizing such states in the theory of the Josephson effect, we can introduce the term "continuous spin states." Parameter m in this case can be naturally referred to as quasispin (cf. the quasimomentum of a band electron). It is also interesting to compare these excitations with anions [27].

It is well known in quantum mechanics that the projection of the spin moment onto a preferred direction is quantized. Quantized spin states are defined in space S^1 ($0 \leq \varphi < 2\pi$); the spin moment quantization is naturally related to the symmetry of the quantum problem to the rotation of the coordinate system through angle 2π around the z axis or, in other words, with the boundary conditions $\psi(\varphi + 2\pi) = \pm\psi(\varphi)$. On the other hand, as noted above and as follows directly from the form of Hamiltonian (11), the symmetry relative to the shift $\varphi \longrightarrow \varphi + 2\pi$ is broken in the field $B_z(t) = at$; consequently, the boundary conditions $\psi(\varphi + 2\pi) = \pm\psi(\varphi)$ are violated together with the standard quantization of the spin moment. Indeed, the states described by the wave functions $\psi(\varphi)$ and $\psi(\varphi + 2\pi)$ are physically distinguishable since the spin gains energy from the torque over the precession period. It was noted above that this requires the use of an expanded (many-sheeted or foliated) space $\mathbb{R}^1$ ($-\infty < \varphi < \infty$) for describing the spin dynamics.

A similar situation is realized in a Josephson junction and in a quantum dot under the Coulomb blocking conditions, where the phase of the wave function plays the role of angle φ and the charge or number of particles plays the role of the spin component S_z. This question is considered in detail in [28].

Let $U_a(\varphi) = -K_2\cos2\varphi$, where K_2 is a constant. Then Schrödinger equation (11) is reduced to the Mathew equation; the theory of the latter equation implies [29] that the energy spectrum of Hamiltonian (10) has a band structure. This means that eigenvalues $E_n(m)$ are functions defined in the corresponding Brillouin zones. For $K_2 \approx 0$, the band structure corresponds to the free electron approximation, $E_n(m) = \hbar^2m^2/2I$, with forbidden bands at the boundaries of the Brillouin zones

($m_B = \pm 1, \pm 2, \ldots$), which are narrow to the extent that the value of K_2 differs from zero.

The wave function can be represented in the form $\psi(\varphi) = u(\varphi)\exp(-iEt/\hbar)$. In this case, Schrödinger equation (11) for Hamiltonian (10) can be written in the form (Mathew equation)

$$u'' + (\mu^2 - 2b^2\cos 2\tilde{\varphi})u = 0, \qquad (15)$$

where $\mu^2 = 2IE/\hbar^2$ and $b^2 = IK_2/\hbar^2$. Here, we have used a new variable $\tilde{\varphi} = \varphi + \pi/2$. The tilde will be omitted below. The form of the spectrum in the lowest band for $|m| \ll 1$ can be described by the formula $E_0(m) = \hbar^2 m^2/2I$; for $|m| \approx 1$, we have

$$E_0(m) = \frac{\hbar^2}{2I}\left\{1 - \sqrt{(|m|-1)^2 + \left(\frac{IK_2}{2\hbar^2}\right)^2}\right\}^2. \qquad (16)$$

The width of the allowed band with number n ($n = 0$ corresponds to the principal band) is defined as

$$\Delta E_n = \frac{\hbar^2}{I}\left(n + \frac{1}{2}\right). \qquad (17)$$

Forbidden bandwidths decrease rapidly (for $K_2 \ll K_1$) upon an increase in the band number:

$$\delta E_{n,\,n+1} = K_2\left(\frac{K_2/K_1}{n+1}\right)^n. \qquad (18)$$

In particular, it can easily be seen that the width of the first forbidden band is equal to K_2. Equations (16)–(18) define the energy spectrum of Hamiltonian (10) quite accurately.

7. Let us now consider the effects emerging in magnetic field $B_z(t)$. The last term $\gamma I\dot{B}\varphi$ in Eq. (11) plays the same role as the energy eEx of a Bloch electron in an external electric field.

Let us consider the dynamics of momentum p_φ in the case when the magnetic field varies adiabatically slowly ($|\gamma I\dot{B}| \ll K_2$). It is convenient to use, instead of $\dot{B}$, the quantity $j_m = \dot{B}/4\pi$, which can be referred to as "magnetic current" since quantity j_m generates a time-independent vortex electric field in the same way as the electric current generates a magnetic field. In order to describe the spin dynamics under the action of magnetic current $j_m = \dot{B}/4\pi$, we consider a wave packet consisting of Bloch functions (14). Let us suppose that $\bar{m} = \langle m\rangle$ and $\bar{\varphi} = \langle\varphi\rangle$ are the mean values of quasispin and the coordinate of the packet center, and the values of Δm and $\Delta\varphi$ ($\Delta m\Delta\varphi \sim 1$) specify the corresponding indeterminacies. Under the action of magnetic current j_m, the wave packet formed at instant $t = 0$ is displaced to the boundary of the (e.g., right) Brillouin zone, is reflected from it, and the group velocity reverses its sign. Then the packet propagates to the left boundary of the Brillouin zone, is reflected from it, and so on. In this case, the dispersion Δm and the packet width $\Delta\varphi$ change periodically. This process is known as Bloch oscillations. Mathematically, this process is described by the following equations for mean values $\bar{m}$ and $\bar{\varphi}$:

$$\dot{\bar{m}} = \frac{\gamma I\dot{B}}{\hbar}, \quad \dot{\bar{\varphi}} = \frac{1}{\hbar}\frac{dE_n(\bar{m})}{d\bar{m}}. \qquad (19)$$

In this (adiabatic) process, the system remains in a state with a preset n and the observed physical quantities (e.g., magnetic moment) are oscillating functions of time with frequency,

$$f_{Bl} = \frac{\gamma I}{2\hbar}\frac{B_1}{\tau}. \qquad (20)$$

If the external field has a harmonic component in addition to the linear term, i.e.,

$$B(t) = \frac{B_1}{\tau}t + B_2\sin(2\pi f t),$$

resonances at frequencies $f = rf_{Bl}$ are possible, where r is a rational number (Stark resonances).

8. An increase in the magnetic current ($|\gamma I\dot{B}| \geq K_2$) gives rise to the Zener tunnel effect [30, 31]. The essence of the effect is that the system overcomes the potential barrier separating two neighboring bands of allowed energy values under the action of magnetic current j_m. Let us determine the probability of this process.

The energy spectrum of Hamiltonian (11) for each fixed value of φ and for moderate rates of magnetic field variation is defined, to a fairly high degree of accuracy, by the relation

$$E_n(m, \varphi) = E_n(m) - \gamma I\dot{B}\varphi, \qquad (21)$$

where $E_n(m)$ is the energy spectrum of Hamiltonian (11) for zero magnetic current (see item 6). Since we can determine the "local" value of quasispin $m = m(\varphi)$ for each φ, we obtain, in accordance with the well-known Bloch theorem [32], the following expression for eigenstates instead of relation (14):

$$\psi_n(\varphi) = \exp\left(i\int_{\varphi_0}^{\varphi} m(\xi)d\xi\right)u_n(\varphi); \qquad (22)$$

here, $u_n(\varphi + \pi) = u_n(\varphi)$. In the limits of the first forbidden band, we have

$$m(\varphi) = 1 + i\beta(\varphi), \qquad (23)$$

where $\beta(\varphi)$ is a certain real-valued function that must be determined. Assuming that function $u_n(\varphi)$ varies insignificantly as compared to the exponential function under condition (23), we can obviously write the tunneling probability in the form

$$P = \exp\left(-2\int_{\varphi_1}^{\varphi_2}|\beta(\xi)|\,d\xi\right), \tag{24}$$

where $\beta(\varphi_1) = \beta(\varphi_2) = 0$. Carrying out elementary transformations, we obtain from Eqs. (16), (21), and (23)

$$\beta(\varphi) = \frac{I}{\hbar^2}\sqrt{\frac{K_2^2}{4} - (\gamma I\dot{B}\varphi)^2};$$

using relation (24), we obtain

$$P = \exp\left(-\frac{\pi}{8}\frac{K_2^2}{\mu_B\hbar\dot{B}}\right). \tag{25}$$

The tunneling probability per unit time amounts to

$$P_t = f_{Bl}\exp\left(-\frac{\pi}{8}\frac{K_2^2}{\mu_B\hbar\dot{B}}\right), \quad f_{Bl} = \frac{\gamma I}{2\hbar}\frac{B_1}{\tau}. \tag{26}$$

The formulas derived above can easily be generalized to the case of a Zener transition between adjacent excited bands with numbers n and $n + 1$ (the quantity K_2 in this case has the meaning of the width $\delta E_{n,\,n+1}$ of the corresponding forbidden band (18). For small values of the anisotropy constant K_2, the tunneling probabilities are practically equal to unity even for small values of n, which indicates the existence of free precession.

9. Let us now consider the behavior of the average magnetic moment of the spin system in question. Its component along the z axis is given by (7):

$$M_z = M\cos\theta \approx M\left(\frac{\pi}{2} - \theta\right) = \frac{M^2}{2K_1}\left(B_z - \frac{\dot{\varphi}}{\gamma}\right).$$

Averaging with the appropriate wave function, we obtain

$$\langle M_z\rangle = \chi_\perp\left(B_z - \frac{\langle\dot{\varphi}\rangle}{\gamma}\right), \tag{27}$$

where $\chi_\perp = M^2/2K_1$.

Let us first consider the limiting case of free precession, $K_2 = 0$. In the simplest case (3), Eqs. (19) give $\langle\dot{\varphi}\rangle = \gamma[B_z(t) + c]$, where c is a constant determined by the initial conditions. Substituting this expression into Eq. (27), we obtain $\langle M_z\rangle = -\chi_\perp c = \text{const}$. Thus, a distinguishing feature of free precession is that the accelerated spin precession under the action of an increasing (decreasing) magnetic field screens the contribution from the paramagnetic susceptibility of an ion ($\chi_\perp B_z$) so that the average magnetic moment is independent of the field.

The situation changes significantly in the case of Bloch oscillations. The $\langle M_z\rangle(B_z)$ dependence is the sum of "conventional" linear dependence ($\chi_\perp B_z$) and periodic curve with a period of $\Delta B = \dot{B}/f_{Bl} = 2b$, $b = \hbar/\gamma I$. Magnetization experiences jumps by $\Delta M_z = 2g\mu_B$ for applied field values $B_z = b + k\Delta B$, $k \in N$. Thus, the curve describing the $\langle M_z\rangle(B_z)$ dependence resembles a ladder. For $K_2 = 0$, the shape of the steps is strictly rectangular. If, however, $K_2 \neq 0$, the front edges of the steps are slightly blurred, but their height and position remain unchanged.

In the general case, the $\langle M_z\rangle(B_z)$ dependence exhibits typical features of both limiting processes. Indeed, it was shown above (18) that for $K_2/K_1 \ll 1$, the forbidden bandwidth is a rapidly decreasing function of the band number. Consequently, in the first approximation (which will be referred to as a one-band approximation), we can neglect Bloch oscillations in the first excited and next bands. This means that precession under the action of field $B_z(t)$ can be regarded as free in all bands except the principal band. On the whole, precession can be visualized as follows. A wave packet that formed at the initial instant is slightly blurred as it reaches the boundary of the Brillouin zone, is partly reflected from it, and partly tunnels into the next band, in which it precesses freely.

As a result of tunneling of the wave packet to the next energy bands, the magnetization jump ΔM_z becomes a function of the jump number k and is defined as

$$\Delta M_z = 2g\mu_B(1-p)^{k-1},$$

where p is the probability of a tunnel transition from the principal energy band to the first excited band. This probability is defined by formula (25).

In the two-band approximation, we take into account Bloch oscillations occurring in the first excited band. Their influence on the shape of the magnetization curve is manifested in the formation of additional jumps at instants when the applied magnetic field assumes values of $B_z = k\Delta B$, $k \in \mathbb{N}$. The magnitudes of the main and additional jumps are mixed up in view of the possibility of a tunnel transition from the first excited energy band back to the principal band; as a result, the general formula for the height of the magnetization steps is extremely cumbersome. On a qualitative level, we can state that the heights of the main jumps decrease with increasing jump number (however, the rate of this decrease is lower than in the one-band approximation), while the heights of additional jumps increase during the few first Bloch periods and

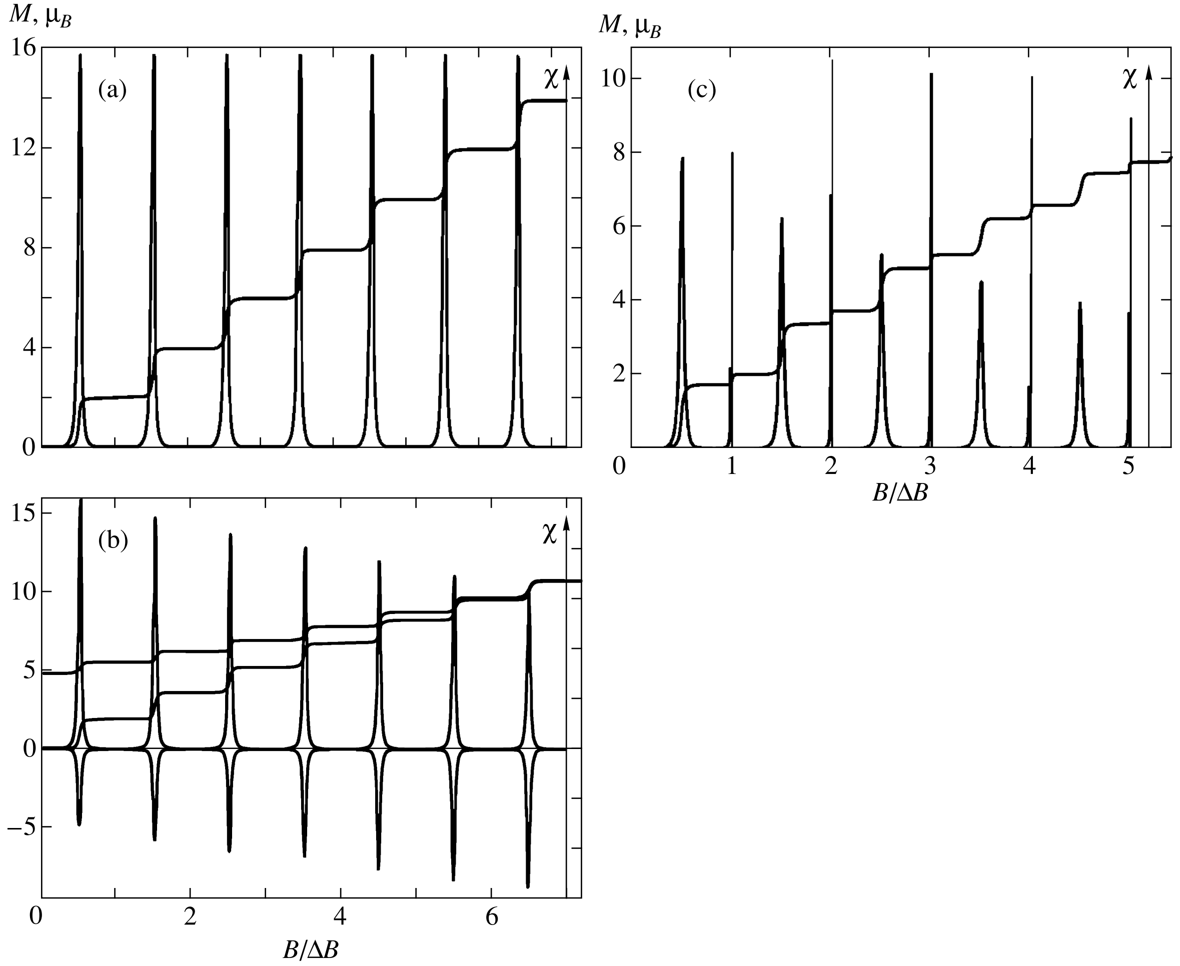


Fig. 1. Magnetization curve and magnetic susceptibility peaks as functions of the field for orthorhombic anisotropy, $K_2/K_1 = 0.01$, for (a) $B_1/\tau = 0.5 \times 10^{11}$ Oe/s (the Zener tunneling probability is negligibly low); (b) $B_1/\tau = 1.5 \times 10^{11}$ Oe/s (the Zener tunneling in the system leads to hysteresis, but the spin precesses freely in the excited energy bands), and (c) $B_1/\tau = 2.0 \times 10^{11}$ Oe/s (peak heights are reduced by half; Bloch oscillations take place not only in the principal band but also in the first excited energy band).

start decreasing only after attainment of a certain maximal value.

Figures 1 and 2 show the dependences of magnetization and susceptibility on the external magnetic field for various types of anisotropy and various rates of magnetic field variation. The field increases linearly with time from zero to a certain value during seven and a half Bloch periods and then decreases at the same rate back to zero. In calculations, one must be sure that the condition $\theta - \pi/2 \ll 1$ (see item 4) is not violated. This is possible only for a large (not less than ten) value of the spin moment of a nanocluster. The magnetization is recalculated per cluster and is measured in units of Bohr magneton. The magnetic field itself is measured in units of increment ΔB over a Bloch period. The calculations are carried out in the framework of the one-band model taking into account the Zener tunneling whose intensity increases with the rate of variation of the magnetic field. The susceptibility peaks for the reverse direction of the magnetic field are inverted in the figure for convenience and better visualization.

An important distinguishing feature of free precession for systems with tetragonal or hexagonal anisotropy is that accelerated spin precession under the action of an increasing (decreasing) magnetic field screens the contribution of the paramagnetic susceptibility of an ion ($\chi_\perp B_z$) incompletely, which is manifested in the absence of horizontal segments on the magnetization curve (see item 10 below).

The results obtained in the framework of the two-band model are shown in Fig. 1c. It is assumed that the magnetic field increases linearly with time from zero to a certain value during five Bloch periods and then decreases back to zero at the same rate.

All of the above-mentioned features in the behavior of the spin system in question are reflected in the figures and can easily be observed.

10. Finally, we will make a number of important remarks concerning the above features of systems of large-spin particles with tetragonal (4) or hexagonal (5) anisotropy.

In these cases (for which the potential energy is $U_{at}(\varphi) = -K_4\cos4\varphi_t$ or $U_{ah}(\varphi) = -K_4\cos6\varphi_h$), Schrödinger equation (11) can be written in a form coinciding with Eq. (15) if we introduce the notation

$$\mu_t^2 = \frac{IE_t}{2\hbar^2}, \quad b_t^2 = \frac{IK_4}{4\hbar^2}, \quad \tilde{\varphi} = 2\varphi_t + \frac{\pi}{2} = \varphi + \frac{\pi}{2}$$

and

$$\mu_h^2 = \frac{2IE_h}{9\hbar^2}, \quad b_h^2 = \frac{IK_4}{9\hbar^2}, \quad \tilde{\varphi} = 3\varphi_h + \frac{\pi}{2} = \varphi + \frac{\pi}{2}.$$

Indices "t" and "h" are used to distinguish between the quantities pertaining to tetragonal and hexagonal anisotropy from those corresponding to the simplest case and bearing no special index. A comparison of the parameters introduced above with parameters μ and b from Eq. (15) for identical values of constants K_1 enables us to draw the following conclusions.

1. The energy spectrum of the system with tetragonal and hexagonal anisotropy has the form $E_{nt}(m) = 4E_n(m)$ and $E_{nh}(m) = 9E_n(m)$, respectively. Here, $E_n(m)$ corresponds to the simplest case (3).

2. The width of the first forbidden band is K_4.

3. The frequency of Bloch oscillations is defined by formula (20) as before.

4. The horizontal segments on the magnetization curve typical of the simplest case disappear.

Let us consider statement 4 in greater detail. Since the procedure for evaluation of integral (6) remains unchanged upon a transition to another type of anisotropy, relation (7) holds; consequently, in the case of tetragonal (hexagonal) anisotropy, expression (27) for the average magnetic moment component along the z axis remains in force. Since $\varphi_t = \varphi/2$ and $\varphi_h = \varphi/3$, we have

$$M_t(B) = \frac{\chi_\perp}{2}B + M(B) \tag{28}$$

and

$$M_h(B) = \frac{2\chi_\perp}{3}B + M(B), \tag{29}$$

where $M(B)$ is the magnetization of a system whose anisotropy energy has the simplest form. The above statement obviously follows from formulas (28) and (29).

11. Let us obtain some numerical estimates. For spin $S = 10\hbar$, $K_1 \approx 10$ cm^{-1}, and $B_1/\tau \approx 10^{10}$ Oe/s, the frequency of Bloch oscillations is $f_{Bl} \approx 10^7$ Hz. As the rate of magnetic field buildup increases to values of $B_1/\tau \sim (10^{11}–10^{12})$ Oe/s, the Zener tunnel effect becomes appreciable if we set $K_2/K_1 \approx 10^{-2}$.

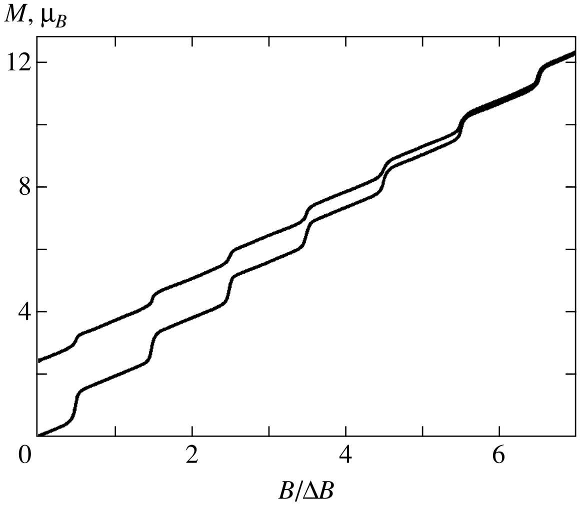


Fig. 2. Magnetization curve for tetragonal anisotropy, $K_4/K_1 = 0.01$, $B_1/\tau = 1.5 \times 10^{11}$ Oe/s. Bloch oscillations take place in the principal energy band. The Zener tunneling occurring in the system leads to hysteresis.

The above analysis was carried out for zero temperature. However, thermal fluctuations (at $T \neq 0$), as well as the interaction with dissipative environment, suppress the coherent quantum effects in question and require special treatment. Here, we only indicate the applicability limits of the approximation $T = 0$: $T \ll K_1$ and $t_s \gg 1/f_{Bl}$, where t_s is the spin relaxation time. For $K_1 \approx 10$ cm^{-1} and $B_1/\tau \approx 10^{10}$ Oe/s, we obtain $T \ll 2$ K and $t_s \gg 10^{-7}$ s. These limits appear as easily attainable in contemporary low-temperature experiments.

Thus, we have proved that a magnetic field increasing (decreasing) with time induces new coherent quantum effects in the dynamics of an anisotropic spin system. These effects include the formation of the band energy spectrum with continuous spin states, quasi-Bloch oscillations, and interband Zener tunneling, which are manifested in the form of magnetization jumps and susceptibility peaks in the spin system under investigation.

ACKNOWLEDGMENTS

One of the authors (A.K.Z.) sincerely thanks B. Barbara and R. Giraud for fruitful discussions of related problems and for their warm hospitality at the Néel Laboratory (Grenoble). The authors are also grateful to the participants of the Yu.V. Kopaev seminar (Lebedev Physical Institute, Russian Academy of Sciences) for discussions and to A.M. Ignatov, who indicated a similarity of spin excitations in question to anions.

This study was supported by the Russian Foundation for Basic Research (project no. 02-02-17389), INTAS (grant no. 99-01839), and the Federal Special Program "Physics of Solid-State Nanostructures."

REFERENCES

1. B. Barbara, L. Thomas, F. Lionti, *et al.*, J. Magn. Magn. Mater. **200**, 167 (1999).
2. B. Barbara and L. Günther, Phys. World **12**, 35 (1999).
3. J. R. Friedman, M. P. Sarachick, J. Tejada, *et al.*, Phys. Rev. Lett. **76**, 3830 (1996).
4. L. Thomas, F. Lionti, R. Ballou, *et al.*, Nature **383**, 145 (1996).
5. E. M. Chudnovsky and J. Tejada, *Macroscopic Quantum Tunneling of the Magnetic Moment* (Cambridge Univ. Press, Cambridge, 1998).
6. V. V. Dobrovitskii and A. K. Zvezdin, Europhys. Lett. **38**, 377 (1997).
7. L. Günther, Europhys. Lett. **39**, 1 (1997).
8. A. Garg, Europhys. Lett. **22**, 205 (1993).
9. I. Tupitsyn and B. Barbara, in *Magneto-Science from Molecules to Materials*, Ed. by M. Drillon and J. Miller (Wiley–VCH, Hamburg, 2001), p. 109.
10. A. K. Zvezdin, V. V. Dobrovitskii, B. N. Harmon, *et al.*, Phys. Rev. B **58**, 14733 (1998).
11. R. Giraud, W. Wernsdorfer, A. Tkachuk, *et al.*, Phys. Rev. Lett. **87**, 057203 (2001).
12. D. P. Loss and D. P. Di Vincenzo, Phys. Rev. A **57**, 120 (1998).
13. V. V. Dobrovitskiĭ, A. K. Zvezdin, and A. F. Popkov, Usp. Fiz. Nauk **166**, 439 (1996) [Phys. Usp. **39**, 407 (1996)].
14. J. Tejada, E. M. Chudnovsky, E. del Barco, *et al.*, Nanotechnology **12**, 181 (2001).
15. A. K. Zvezdin, Kratk. Soobshch. Fiz., No. 12, 13 (1999).
16. A. K. Zvezdin, E-print archives, cond-mat/0004074.
17. A. K. Zvezdin, Kratk. Soobshch. Fiz., No. 11, 3 (2000).
18. A. M. Perelomov, *Generalized Coherent States and Their Applications* (Nauka, Moscow, 1987; Springer, New York, 1986).
19. F. Bloch, Phys. Rev. Lett. **137**, A787 (1965).
20. F. Bloch, Phys. Rev. Lett. **166**, 415 (1968).
21. M. A. Omar, *Elementary Solid State Physics* (Addison-Wesley, Reading, Ma, London, 1975).
22. K. K. Likharev, *Introduction to the Dynamics of Josephson Junctions* (Nauka, Moscow, 1985).
23. D. V. Averin, A. B. Zorin, and K. K. Likharev, Zh. Éksp. Teor. Fiz. **88**, 692 (1985) [Sov. Phys. JETP **61**, 407 (1985)].
24. G. Schön and A. D. Zaikin, Phys. Rep. **198**, 237 (1990).
25. P. W. Anderson, *Lectures on the Many-Body Problem* (Academic, New York, 1964), Vol. 2, p. 113.
26. V. A. Kalatsky, E. Müller-Hartmann, V. L. Pokrovsky, *et al.*, Phys. Rev. Lett. **80**, 1304 (1998).
27. S. Rao, E-print archives, hep-th/9209066 (2001).
28. A. I. Larkin, K. K. Likharev, and Yu. N. Ovchinnikov, Physica B (Amsterdam) **126**, 414 (1984).
29. *Handbook of Mathematical Functions*, Ed. by M. Abramowitz and I. A. Stegun, 2nd ed. (Dover, New York, 1971; Mir, Moscow, 1979).
30. C. Zener, Proc. R. Soc. London, Ser. A **145**, 523 (1934).
31. Y. Gefen, E. Ben-Jacob, and A. O. Caldeira, Phys. Rev. B **36**, 2770 (1987).
32. N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Holt, Rinehart, and Winston, New York, 1976; Mir, Moscow, 1979).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 97, No. 1, 2003, pp. 93–100.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 1, 2003, pp. 105–113.

SOLIDS
Structure

Heat Capacity Anomaly in the Vicinity of the Tricritical and Upper Critical Points

V. P. Voronov[a], **E. E. Gorodetskii**[a],*, **and A. T. Berestov**[b]
[a]*Institute of Oil and Gas Problems, Russian Academy of Sciences, Moscow, 119991 Russia*
[b]*Moscow State Institute of Electronic Engineering (Technical University), Zelenograd, Moscow oblast, 103498 Russia*
**e-mail: gorod@ogri.ru*

Received December 11, 2002

Abstract—The anomalous behavior of the isochoric heat capacity of a mixture of methane, pentane and heptane is studied experimentally in the vicinity of the liquid–vapor critical point in the cases when (a) the critical temperature T_c approaches the tricritical point T_{TCP} and (b) the critical temperature approaches the upper critical end point T_U. It is shown that in all cases, the singular part of the heat capacity of the mixture has the form $C_{\text{sing}} = A|\tau|^{-\alpha}$, where $\tau = (T - T_c)/T_c$ and $\alpha \approx 0.11$. When $T_c \longrightarrow T_U$, amplitude A of the heat capacity anomaly is found to be approximately constant. At the same time, the amplitude of the anomaly tends to zero in the vicinity of the tricritical point: $A \propto |\tau_c|^{\varepsilon}$, where $\tau_c = (T_c - T_{TCP})/T_{TCP}$ and $\varepsilon = 1.6$–1.7. The inevitable vanishing of this mode of the heat capacity anomaly leads to a negative value of the critical index $\tilde{\alpha}$ characterizing the heat capacity anomaly at the tricritical point, while the tricritical point theory and the isomorphism hypothesis predict $\tilde{\alpha} = 0.5$. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The variety of phase behavior of solutions makes it possible to use them for studying anomalous properties of substances in the vicinity of various singular points. It was shown in our recent publication [1] that the three-phase equilibrium of the methane + heptane mixture makes it possible to measure heat capacity in the vicinity of the vapor–liquid critical point at a constant value of chemical potential of the impurity and to verify the isomorphism hypothesis experimentally. This study is devoted to experimental investigation of the heat capacity behavior in the vicinity of the liquid–vapor critical point for ternary carbon-containing mixtures in situations when this critical point approaches either the tricritical point or upper critical end point. It should be emphasized that we are speaking here and below of changes in the heat capacity in the vicinity of liquid–vapor critical points in the presence of a noncritical liquid phase enriched with heavy components.

It is well known [2] that high-molecular hydrocarbons of the C_nH_{2n+2} series are poorly soluble in methane. This leads to the emergence of three-phase equilibrium lines bounded by the upper and lower critical end points and to discontinuities on the curves describing how the critical parameters of mixtures of methane with saturated hydrocarbons (starting with hexane) depend on the concentration of the heavy component (Fig. 1). At the upper critical end point (U), one of the liquid phases and the gaseous phase become identical; at the lower critical end point (L), the two liquid phases become identical. Apart from the upper and lower critical end points attained simultaneously with the disappearance of the noncritical phase, there are curves containing the upper and lower critical points in the system, which are attained in the presence of a noncritical phase. Figure 2 shows a typical phase diagram with the upper liquid–vapor critical point in the presence of a liquid noncritical phase.

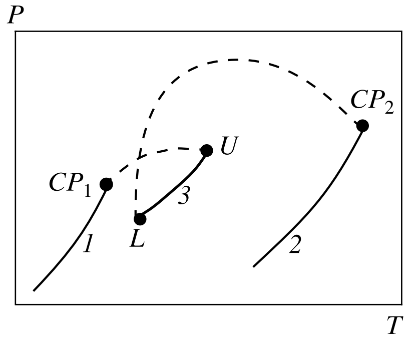


Fig. 1. Schematic phase diagram of a binary mixture with a three-phase equilibrium curve. Solid curves *1* and *2* correspond to coexistence of pure components; CP_1 and CP_2 are liquid–vapor critical point for pure components; dashed curves correspond to critical points of a binary mixture; accordingly, U and L are the upper and lower critical end points; solid curve *3* corresponds to the liquid–liquid–vapor three-phase equilibrium.

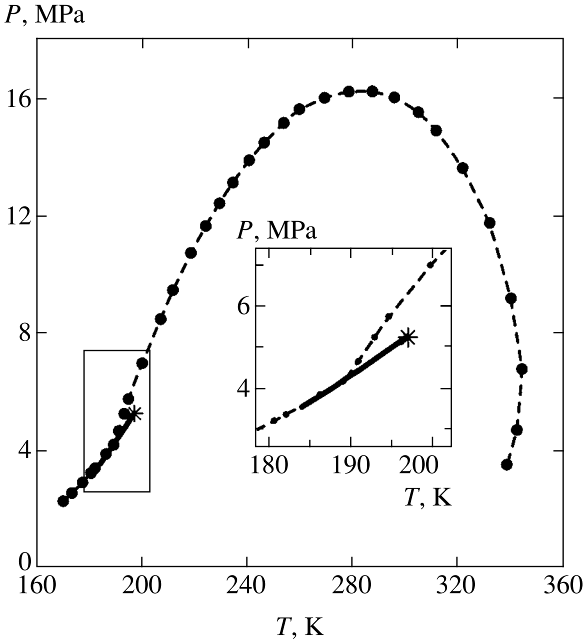


Fig. 2. Phase diagram for a ternary mixture of methane + 0.034 of a molar fraction of (pentane + 0.43 of a molar fraction of heptane). The experimental points connected by the dashed curve describe the boundary of the two-phase equilibrium region. The solid curve describes three-phase equilibrium; the asterisk marks the upper critical point.

The three-phase equilibrium curve for the methane + heptane mixture studied by us earlier starts from the heptane-enriched phase crystallization curve; for this reason, the lower critical end point is unattainable in this case. The experimentally observed temperature interval of coexistence of three phases in this mixture is equal to 23 K. The temperature interval between the upper (T_U) and lower (T_L) critical end points in the methane + hexane mixture is $\Delta T_{UL} = T_U - T_L = 13.45$ K [3]. In the methane + pentane system, there is no three-phase equilibrium at all; for this reason, the curve of critical points is continuous. Thus, the temperature interval ΔT_{UL} of three-phase equilibrium for binary mixtures of methane with normal alkanes is obviously the smaller, the lower the molecular mass of the heavy component. An admixture with a molecular mass such that $\Delta T_{UL} = 0$ could exist in principle. At such a point, all three phases are critical and the point is called the tricritical point. In binary mixtures for which the molecular mass of impurities varies discretely, a tricritical point could be realized only by chance. However, a continuous variation of the molecular mass of the heavy component can be easily achieved for ternary mixtures.

We introduce the total molecular concentration X of high-molecular impurities and concentration Z of the heaviest component in the mixture:

$$X = \frac{n_2 + n_3}{n_1 + n_2 + n_3}, \quad Z = \frac{n_3}{n_2 + n_3}.$$

Here, n_1, n_2, n_3 are the numbers of moles of the solution components in the increasing order of their molecular mass. The effective molecular mass of the heavy impurity is

$$M_{\text{eff}}(Z) = (1 - Z)M_2 + ZM_3,$$

where M_2 and M_3 are the molecular masses of the lighter and the heavier components. A decrease in Z reduces the effective molecular mass of the impurity, and the system may be quite close to the tricritical point for a reasonable fixed value of X. On the other hand, a decrease in X for a fixed Z brings the liquid–vapor critical point to the upper critical end point. This study is aimed at analysis of the behavior of heat capacity in these two cases.

The object of investigations was a ternary mixture of methane + X molar fractions of (pentane + Z molar fractions of heptane). Such a choice of the mixture was dictated by the fact that a decrease of concentration Z from unity to zero continuously transforms the system from the methane + heptane binary mixture with a discontinuous critical locus to a methane + pentane binary mixture with a continuous critical locus. We studied the phase behavior and the isochoric heat capacity of the mixture in the following two cases:

(i) the total concentration X of heavy impurities is fixed, while the heptane concentration Z is varied: $X = 0.0345$; $Z = 1.0$, 0.65, 0.43, 0.05, 0.019, and 0.0 molar fraction of heptane;

(ii) the heptane concentration Z is fixed, while the total concentration X of heavy impurities is varied: $Z = 0.43$; $X = 0.0546$, 0.0345, 0.0195, and 0.0099 molar fraction of the solution.

In case (i), the choice of concentration $X = 0.0345$ of a molar fraction is dictated by the previously established fact [4] that the behavior of the isochoric heat capacity $C_{\rho, X}$ of a binary mixture of methane + 0.0345 of a molar fraction of pentane is completely similar to the behavior of heat capacity $C_{P, X}$ at the tricritical point of a ^{3}He–^{4}He mixture [5]. This led to the conclusion that the critical point of the liquid–vapor mixture methane + 0.0345 of a molar fraction of pentane is close to the tricritical point. In this sense, a decrease in the heptane concentration Z in a ternary mixture of methane + 0.0345 of a molar fraction of (pentane + Z molar fractions of heptane) is a way for studying the dependence of the heat capacity anomaly on the proximity to the tricritical point.

In the case (ii), the choice of the fixed value of the heptane concentration $Z = 0.43$ of a molar fraction is determined by the fact that a three-phase equilibrium region necessarily exists in such a ternary mixture. A decrease in concentration X is a way for studying the dependence of heat capacity anomaly on the proximity to the upper critical end point.

2. EXPERIMENT

An analysis of the phase behavior and isochoric heat capacity of hydrocarbon mixtures was carried out on a precision adiabatic scanning calorimeter with a cell volume of 14.9 cm^3. The cell is connected to a strain gauge intended for measuring pressure and to the filling system through a thin capillary. The mixture with the required composition was prepared directly in the measuring cell. The filling system and the cell were preliminarily evacuated, after which the cell was filled with a solution of heavy components (pentane + Z molar fractions of heptane) with a preset Z, which was prepared in a separate vessel. Then the required amount of methane was added. The concentration and density of the mixture were determined by weighing. The procedure of filling was normally carried out at a cell temperature of 170–180 K.

The schematic of the calorimeter is described in [6]. In order to maintain adiabatic conditions, the cell is surrounded by two copper screens whose temperature strictly “followed” the cell temperature. Heat transfer through the capillary was prevented by connecting it to a copper ring via a heat pipeline in the immediate vicinity of the cell; the ring temperature was maintained at the level of the cell temperature. The temperature of the upper part of the capillary was kept approximately 1 K higher to prevent condensation of the mixture in it. The temperature was measured with a platinum thermometer with a nominal resistance of 100 Ω, which was placed in a pocket in the cell. The mixture was stirred with a magnetic agitator driven by a solenoid with a period of 10 s. The process of measuring the enthalpy (amount of heat supplied to the cell) described in [7] was executed and controlled by a computer system. In addition to the enthalpy, we measured the temperature and pressure in the cell. The heat capacity was calculated by numerical differentiation of the enthalpy with respect to temperature.

Figure 2 shows a typical phase diagram for the ternary hydrocarbon mixture studied by us ($X = 0.0345$ and $Z = 0.43$ of a molar fraction). Experimental points connected by the dashed curve form the boundary of the two-phase state region. The region of three-phase equilibrium in the studied mixtures turned out to be very narrow and appears in the figure as a solid line terminating at the upper critical point (asterisk in Fig. 2). We did not investigate the neighborhood of the lower critical point; for this reason, we assume that the temperature interval $\Delta T_{3\text{ph}}$ in which three-phase equilibrium exists coincides with the temperature interval of three-phase equilibrium on the isochore corresponding to the upper critical point. It will be proved below that the values of ΔT_{UL} and $\Delta T_{3\text{ph}}$ in the vicinity of the tricritical point are approximately proportional to each other.

Figure 3 shows the temperature dependences of heat capacity on different isochores for the same mixture as in Fig. 2. The jumps on the curves in Fig. 3 correspond to phase transitions in the mixture. It can be seen from the figure that, at a high density (curve *1*), the mixture passes directly from the two-phase to the one-phase state upon heating. As the density decreases, there appears a temperature region in which three phases coexist (curves *2*–*5*). The transitions from the two-phase to the three-phase state correspond to separation of the phase enriched with the light component into liquid and vapor. Further heating leads first to a transition of the system from the three-phase to the two-phase state and then to the one-phase state (on isochores *4* and *5*, transitions from the two-phase to the one-phase state lie beyond the temperature interval presented in the figure). The critical isochore (curve 5) exhibits an infinite increase in the heat capacity of the phase enriched with the light component when the temperature approaches the value corresponding to the liquid–vapor critical point. This heat capacity anomaly is due to the fact that the liquid–vapor critical point in this case is realized in the presence of the third (noncritical) phase, ensuring the constancy of chemical potentials of heavy impurities in critical phases [1]. In accordance with the isomorphism hypothesis, the type of the heat capacity anomaly C_{ρ,μ_2,μ_3} at constant values of the chemical potentials of impurity components at the liquid–vapor critical point of the mixture coincides with the type of the heat capacity anomaly C_ρ in the vicinity of the liquid–vapor critical point of one-component liquids. In all the mixtures under study, we observe precisely this behavior of heat capacity. Experimental investigation

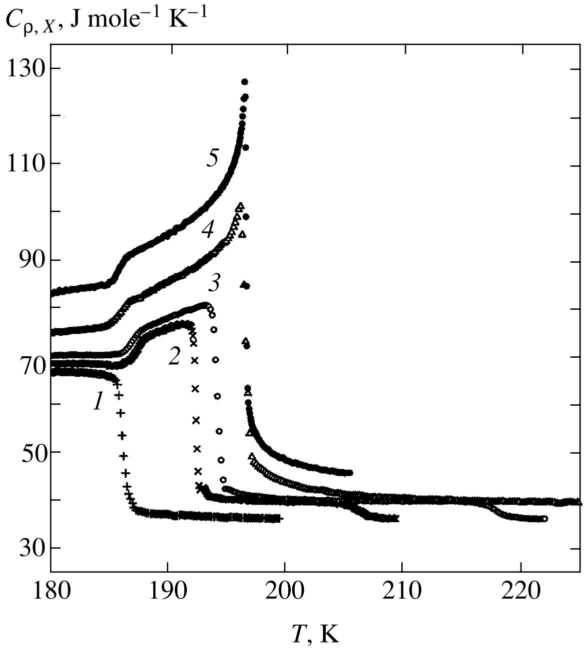


Fig. 3. Temperature dependences of the heat capacity for a mixture of methane + 0.034 of a molar fraction of (pentane + 0.43 of a molar fraction of heptane) for various mean densities of the mixture in the cell: ρ = 0.328 (*1*), 0.300 (*2*), 0.284 (*3*), 0.238 (*4*), and 0.2074 g/cm^3 ≈ ρ_c (*5*).

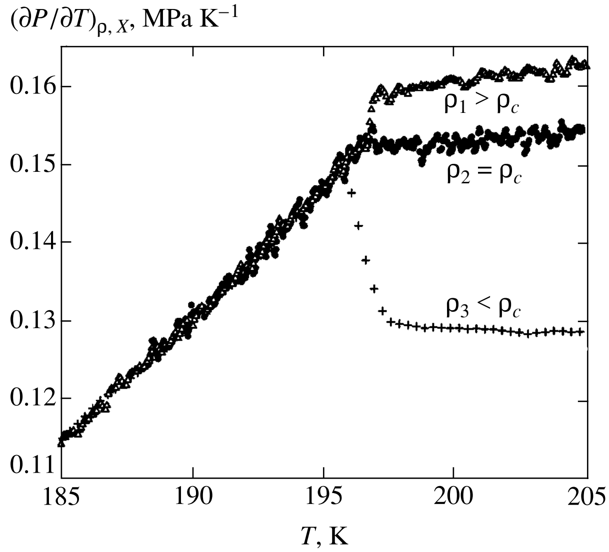


Fig. 4. Temperature dependence of the derivative $(\partial P/\partial T)_{\rho, X}$ on isochores close to the critical isochore for a mixture of methane + 0.034 of a molar fraction of (pentane + 0.43 of a molar fraction of heptane): ρ_1 = 0.2155, ρ_2 = 0.2074 ≈ ρ_c, and ρ_3 = 0.1893 g/cm^3.

of the evolution of this critical anomaly in two cases indicated at the end of Section 1 constitute the main goal of this study.

An important aspect in the experiment was the determination of critical density for each of the mixtures in question. In our opinion, the most precise method of determining the critical density is that based on measuring the jumps on the derivative $(\partial P/\partial T)_{\rho, X}$ on the liquid–vapor transition line for various isochores. The main point of the method is that this jump vanishes on the critical isochore and has different signs for $\rho > \rho_c$ and $\rho < \rho_c$. Consequently, the critical density ρ_c can be determined to a high degree of accuracy from the extrapolation to zero value of the jumps on the derivative dependence on the density. Figure 4 shows the temperature dependences of derivatives $(\partial P/\partial T)_{\rho, X}$ for a mixture of the same composition as in Figs. 2 and 3 on three different isochores. The vanishing of the derivative jump on the middle curve implies that the corresponding isochore is critical.

3. EXPERIMENTAL DATA PROCESSING

Figures 5 and 6 show the result of heat capacity measurements on critical isochores for cases (i) and (ii) (see the Introduction). It can be seen that the behavior of heat capacity is different. In case (i), the amplitude of the heat capacity anomaly decreases with Z (Fig. 5) and vanishes completely for Z = 0.019 of a molar fraction of heptane. In case (ii), a decrease in the total concentration X of heavy components does not affect the amplitude of the heat capacity anomaly (Fig. 6).

Quantitative analysis was carried out using the following procedure. Considering that enthalpy H is the quantity measured directly in experiment, we com-

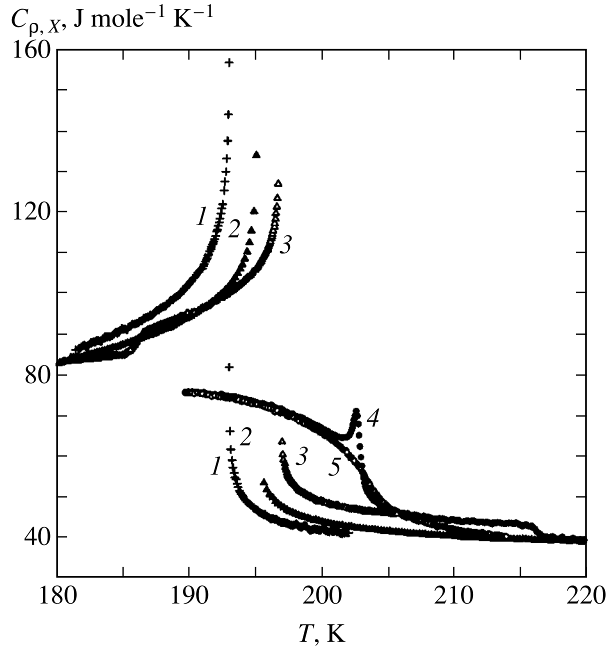


Fig. 5. Curves describing the temperature dependence of heat capacity on critical isochores for a mixture of methane + 0.034 of a molar fraction of (pentane + Z molar fractions of heptane) for Z = 1.0 (*1*), 0.65 (*2*), 0.43 (*3*), 0.05 (*4*), and 0.019 of a molar fraction of heptane (*5*).

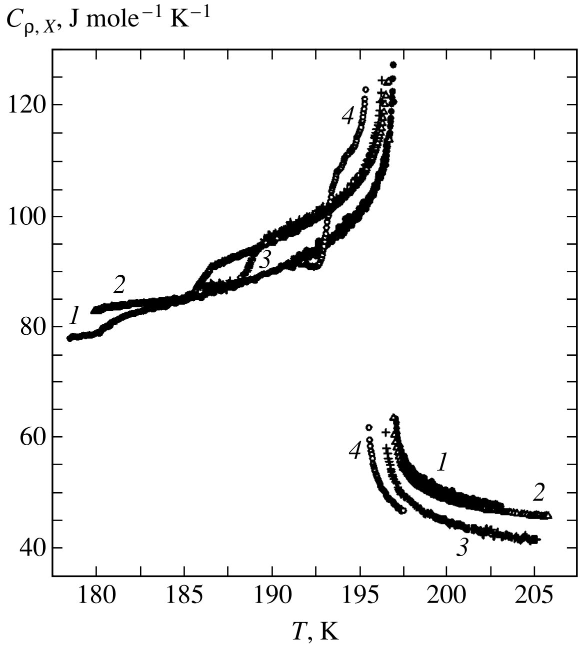


Fig. 6. Curves describing the temperature dependence of heat capacity on critical isochores for a mixture of methane + X molar fractions of (pentane + 0.43 of a molar fraction of heptane) for X = 0.056 (*1*), 0.034 (*2*), 0.0195 (*3*), and 0.0098 of a molar fraction of the solution (*4*).

pared the actually measure enthalpy with the expression derived from its definition by the formula

$$H = \int_{\tau_0}^{\tau} C_{\rho,X} d\tau, \tag{1}$$

where the integration limits $\tau_0 = (T_0 - T_c)/T_c$ and $\tau = (T - T_c)/T_c$ correspond to the temperature interval of experimental data processing and the isochoric heat capacity $C_{\rho,X}$ has the form

$$\begin{aligned} C_{\rho,X} &= A^{\pm}|\tau|^{-\alpha} + B^{\pm}|\tau|^{-\alpha+\Delta} \\ &\quad + C_0^{\pm} + C_1\tau + C_2\tau^2. \end{aligned} \tag{2}$$

Superscripts + and – in this expression indicate the heat capacity branches for $T > T_c$ and $T < T_c$, respectively. The second term is a nonasymptotic Wegner correction with a fixed value of the critical index $\Delta = 0.5$. The last three terms describe the regular part of the heat capacity with identical coefficients for $T > T_c$ and $T < T_c$ in the linear and quadratic terms. The values of the critical index α determined by the least-squares method proved to be close to the theoretical value $\alpha = 0.11$ in all cases studied. In the subsequent comparison of the critical heat capacity amplitudes $A^{\pm}$ in mixtures with different compositions, the value of α was fixed and set equal to 0.11. Tables 1 and 2 show the values of coefficients in Eq. (2) for heat capacity, which were obtained from the experimental data processing using model (1) for both cases studied here.

In case (i), the temperature interval of the observed three-phase region decreases with heptane concentration Z in the solution; for $Z = Z^* \approx 0.019$ of a molar fraction ($T^* = 204.46$ K), the three-phase equilibrium region vanishes. Strictly speaking, point Z^* corresponds to the upper critical end point for the methane + 0.0345 (pentane + 0.019 heptane) mixture. It can be shown, however, that this critical point is close to the tricritical point. Figure 7 shows the experimental dependence of the three-phase equilibrium region width ΔT_{3ph} on Z. The solid curve in this figure corresponds to the theoretical dependence of values of ΔT_{UL} on the closeness to the tricritical point [8]:

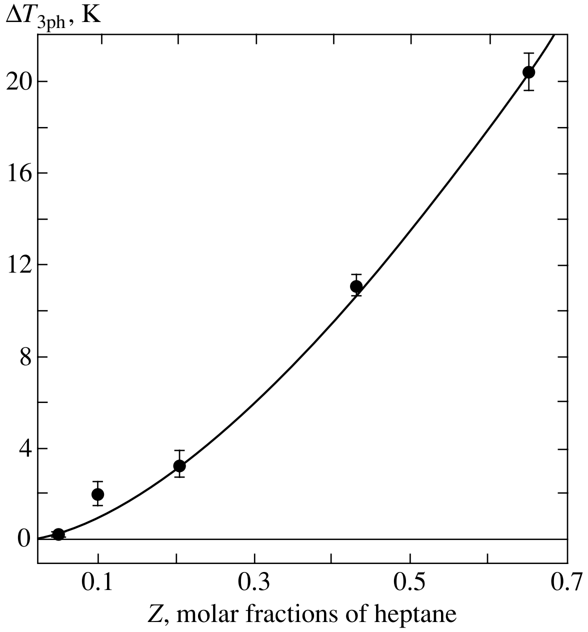


Fig. 7. Dependence of the three-phase equilibrium region width ΔT_{3ph} on concentration Z of heptane in the mixture methane + 0.034 of a molar fraction of (pentane + Z molar fractions of heptane). The solid curve corresponds to the dependence $\Delta T_{3ph} = 41(Z - 0.019)^{3/2}$.

$$\Delta T_{UL} \propto (Z - Z_{TCP})^{3/2}.$$

The closeness of our experimental results to this theoretical dependence indicates that the upper critical end

Table 1. Values of critical temperature and heat capacity coefficients on critical isochores in Eq. (2) for a ternary mixture of methane + 0.034 of a molar fraction of (pentane + Z molar fractions of heptane) for various Z

A^-	A^+	T_c	B^-	B^+	C_0^-	C_0^+	C_1	C_2	A^-/A^+
$Z_1 = 1$									
56.3	28.0	192.932	–6.07	0	15.5	0	47.8	–269	2.01
±0.4	±0.1	±0.001	±1.2		±0.9		±0.9	±16	±0.1
$Z_2 = 0.65$									
39.1	19.7	195.026	–8	–25	37.3	14.7	68.7	–200	1.98
±6.7	±3.7	±0.028	±25	±15	±14	±8	±3.8	±120	±0.5
$Z_3 = 0.43$									
29.6	14.3	196.802	–21.5	–31.4	63	31.0	85.2	–224	2.07
±1.1	±0.8	±0.007	±4	±3.5	±3	±2	±8	±28	±0.13

Note: The heat capacity of the mixture for $Z = 0.056$ was not considered since the critical anomaly region was found to be narrow.

Table 2. Values of critical temperature and heat capacity coefficients on critical isochores in Eq. (2) for a ternary mixture of methane + X molar fraction of (pentane + 0.43 of a molar fraction of heptane) for various X

A^-	A^+	T_c	B^-	B^+	C_0^-	C_0^+	C_1	C_2	A^-/A^+
				$X_1 = 0.056$					
26.5	12.2	196.991	–32.5	–27	60.3	33.8	43.5	88	2.17
±0.7	±0.5	±0.005	±2.2	±2	±1.5	±1.2	±1.5	±12	±0.1
				$X_2 = 0.034$					
29.6	14.3	196.802	–21.5	–31.4	63	31.0	85.2	–224	2.07
±1.1	±0.8	±0.007	±4	±3.5	±3	±2	±8	±28	±0.13
				$X_3 = 0.195$					
29.3	14.7	196.385	0	–39	56.8	27.5	126.0	–228	1.99
±1.5	±1.5	±0.005		±8	±3.5	±3.5	±22	±150	±0.14
				$X_4 = 0.098$					
23.4	12.4	195.467	24.4	–63	67.5	34.8	130.0	0	1.89
±2.2	±1.3	±0.007	±10	±5	±5	±5			±0.26

point in this case is indeed close to the tricritical point. This also means that the values of $\Delta T_{3\text{ph}}$ and ΔT_{UL} are approximately proportional. Consequently, the fact that the heat capacity anomaly amplitudes $A^{\pm}$ decrease as the system approaches the tricritical point (see Table 1) can be regarded as established.

In case (ii), the temperature interval of the observed three-phase region tends to zero upon a decrease in the total concentration X of heavy components. The mixture approaches the upper critical end point.

The extrapolation of experimental data to zero temperature interval of the three-phase region gives the concentration value $X = \tilde{X} \approx 0.0075$ of a molar fraction ($T_c = \tilde{T} \approx 194.87$ K); in this case, the critical heat capacity amplitudes remain practically unchanged (see Table 2 and Fig. 6). It should be noted that the effective molecular mass of the mixture pentane + 0.43 of a molar fraction of heptane is equal approximately to 86, i.e., close to the molecular mass of hexane. On the other hand, it is known that the temperature corresponding to the upper critical end point of the methane + hexane mixture is 195.9 K [3]. The closeness of the temperature corresponding to the upper critical end point determined by us for a ternary mixture and the corresponding temperature in the methane + hexane system indicates that the position of these points is mainly determined by the effective molecular mass of the impurity components.

Tables 1 and 2 show that the amplitude ratio of the heat capacity anomaly below and above the critical point (which is a universal quantity) is close to 2. The theoretical values of this quantity obtained by different authors are slightly different and vary from 1.92 (numerical count based on the three-dimensional Ising model) [9] to 2.08 (ε expansion) [10].

4. DISCUSSION

In the field variables, the role of the thermodynamic potential density of the system is played by pressure. The expression for the pressure differential for a ternary mixture has the form

$$dP = sdT + \rho d\mu_1 + \rho x_2 d\tilde{\mu}_2 + \rho x_3 \tilde{\mu}_3. \qquad (3)$$

Here, s is the entropy per unit volume of the system; ρ is the molar density; x_2 and x_3 are the molar concentrations of pentane and heptane, respectively; and $\tilde{\mu}_2 = \mu_2 - \mu_1$, $\tilde{\mu}_3 = \mu_3 - \mu_1$, where μ_1, μ_2, and μ_3 are the chemical potentials of methane, pentane, and heptane, respectively. Passing from variables x_2 and x_3 to X and Z, we obtain

$$dP = sdT + \rho d\mu_1 + \rho X d\nu_2 + \rho XZ d\nu_3, \qquad (4)$$

where $\nu_2 = \tilde{\mu}_2$ and $\nu_3 = \mu_3 - \mu_2$. We define the variables as

$$\begin{aligned} \tau(\nu_2, \nu_3) &= \frac{T - T_c(\nu_2, \nu_3)}{T_c(\nu_2, \nu_3)}, \\ \tau_c &= \frac{T_c(\nu_2, \nu_3) - T_{TCP}}{T_{TCP}}. \end{aligned} \qquad (5)$$

The first of these variables determines the closeness of the solution to the upper critical point T_c, while the second variable determines the closeness of the upper critical point to the tricritical point T_{TCP}.

In the framework of the scaling hypothesis, the singular part of the thermodynamic potential density in the vicinity of the tricritical point can be written in the form [11]

$$P_{\text{sing}} = |\tau(\nu_2, \nu_3)|^{2-\tilde{\alpha}} f(Y)\ldots, \qquad Y = \frac{|\tau_c|}{|\tau(\nu_2, \nu_3)|^{\varphi}}\ldots, \tag{6}$$

where $f(Y)$ is a universal function of the dimensionless argument Y and φ is the so-called crossover index. It should be noted that the singular part of the thermodynamic potential in relation (6) is written in variables of temperature and reduced chemical potentials of the impurity components. In this sense, P_{sing} is the isomorphic thermodynamic potential [12]. In accordance with the tricritical point theory and the isomorphism hypothesis, we could expect that the critical index $\tilde{\alpha}$ is close to 0.5. It will be shown below that, even qualitatively, such a value of critical index $\tilde{\alpha}$ contradicts the observed vanishing of the amplitudes of heat capacity anomaly as the upper critical point approaches the tricritical point. For this reason, we assume that the critical index $\tilde{\alpha}$ is as yet unknown and we will try to estimate its value from our experimental data.

In the vicinity of the upper critical point, τ_c = const and $\tau(\nu_2, \nu_3)$ tends to zero. The dimensionless parameter Y in this case is much larger than unity. It was noted above that we measured the heat capacity in the vicinity of the upper critical point in the presence of a noncritical phase, which ensures the constancy of the chemical potentials of impurity components in critical phases. The quantity being measured in such experiments is the isomorphic heat capacity C_{ρ, ν_2, ν_3}; in accordance with the isomorphism hypothesis, the type of heat capacity anomaly is the same as for one-component liquids. For this reason, the singular part of the thermodynamic potential density (pressure in our case) in the vicinity of the critical point must have the form

$$P_{\text{sing}} \propto |\tau(\nu_2, \nu_3)|^{2-\alpha}, \tag{7}$$

where $\alpha = 0.11$. It can easily be seen from Eqs. (6) and (7) that, in the limit $Y \longrightarrow \infty$, function $f(Y)$ is given by

$$\lim_{Y \to \infty} f(Y) = \text{const}\, Y^{(-\tilde{\alpha}+\alpha)/\varphi}. \tag{8}$$

This leads to the following expression for the singular part of pressure:

$$P_{\text{sing}} = \text{const}\, |\tau_c|^{(-\tilde{\alpha}+\alpha)/\varphi} |\tau(\nu_2, \nu_3)|^{2-\alpha}. \tag{9}$$

It follows hence that the heat capacity is given by

$$C \propto |\tau_c|^{(-\tilde{\alpha}+\alpha)/\varphi} |\tau(\nu_2, \nu_3)|^{-\alpha} \tag{10}$$

and the critical amplitude has the form

$$A \propto |\tau_c|^{-(\tilde{\alpha}+\alpha)/\varphi}. \tag{11}$$

Table 3. Values of the critical index $\tilde{\alpha} = \alpha - \varepsilon\varphi$ in Eq. (6), calculated from the ratio of the critical heat capacity amplitudes for a mixture of methane + 0.034 of a molar fraction of (pentane + Z molar fractions of heptane) for various Z ($\alpha = 0.11$, $\varphi = 0.5$)

$\tau_{c1} = 0.05642$	$\tau_{c2} = 0.04614$	$\tau_{c3} = 0.03745$
$A_1^-/A_2^- = 1.45$	$A_2^-/A_3^- = 1.32$	$A_1^-/A_3^- = 1.90$
$A_1^+/A_2^+ = 1.42$	$A_2^+/A_3^+ = 1.38$	$A_1^+/A_3^+ = 1.96$
$\tau_{c1}/\tau_{c2} = 1.223$	$\tau_{c2}/\tau_{c3} = 1.232$	$\tau_{c1}/\tau_{c3} = 1.506$
$e_{12}^- = 1.84$	$e_{23}^- = 1.33$	$e_{13}^- = 1.57$
$e_{12}^+ = 1.74$	$e_{23}^+ = 1.54$	$e_{13}^+ = 1.64$
$\tilde{\alpha}_{12}^- = -0.81$;	$\tilde{\alpha}_{23}^- = -0.56$;	$\tilde{\alpha}_{13}^- = -0.67$;
$\tilde{\alpha}_{12}^+ = -0.76$	$\tilde{\alpha}_{23}^+ = -0.66$	$\tilde{\alpha}_{13}^+ = -0.71$

When the critical point approaches the tricritical point, i.e., for $\tau_c \longrightarrow 0$, the experimentally observed decrease in the amplitude of the heat capacity anomaly indicates that $(-\tilde{\alpha} + \alpha)/\varphi > 0$.

As the critical point approaches the upper critical end point, i.e., for $\tau_c \longrightarrow (T_U - T_{TCP})/T_{TCP}$ = const, we have the following expression for the heat capacity amplitude:

$$A \propto \left(\frac{T_U - T_{TCP}}{T_{TCP}}\right)^{(-\tilde{\alpha}+\alpha)/\varphi} = \text{const}. \tag{12}$$

Let us estimate the value of the critical index $\tilde{\alpha}$ in expressions (6)–(12). Knowing the values of heat capacity amplitudes $A_i^{\pm}$ for mixtures with a fixed value of concentration X and various values of $Z = Z_i$ (i = 1, 2, 3 in Table 1) and the corresponding values of

$$\tau_{ci} = \frac{T_{ci} - T_{TCP}}{T_{TCP}},$$

we can determine the exponent $\varepsilon = (-\tilde{\alpha} + \alpha)/\varphi$ in expressions (8)–(12):

$$\varepsilon_{ij}^{\pm} = \frac{\ln(A_i^{\pm}/A_j^{\pm})}{\ln(\tau_{ci}/\tau_{cj})}. \tag{13}$$

Table 3 contains the values of the quantities appearing in formula (13) and the critical index ε determined from this expression. In determining the values of τ_{ci}, we adopted the critical temperature $T^* = 204.46$ K corresponding to $Z^* = 0.019$ as the tricritical temperature T_{TCP}. It is for this value of Z^* that the region of three-phase equilibrium disappears. Setting $\varphi = 0.5$ in accordance with the classical tricritical point theory, we find that $\tilde{\alpha} = \alpha - \varepsilon\varphi$ lies in the interval $-0.81 \le \tilde{\alpha} \le -0.56$. Another value of $\tilde{\alpha}$ can be obtained if we assume that the crossover index φ differs from 0.5. The values of critical indices ε and $\tilde{\alpha}$ given in Table 3 are only estimates of these quantities. It is obvious, however, that the critical index $\tilde{\alpha}$ is rather close to the heat capacity index $\tilde{\alpha} = -1$ completely renormalized in the vicinity of the tricritical point. This value of $\tilde{\alpha}$ can be obtained by setting $\varphi \approx 0.65$. It should be emphasized that the critical index $\tilde{\alpha}$ cannot in principle be equal to 0.5 in accordance with our experimental data.

Note that the form of renormalization of the critical heat capacity anomaly observed in [5] for ^{3}He–^{4}He mixtures also indicates that the amplitude of the isomorphic heat capacity anomaly on the λ line decreases as the system approaches the tricritical point; accordingly, the heat capacity at a constant chemical potential of ^{3}He at the tricritical point proper is found to be finite.

If we assume that the isomorphism hypothesis holds at the tricritical point, the above analysis of the experimental data shows that the fixation of the chemical potential of impurity components is insufficient for the behavior of the heat capacity at the tricritical point to be isomorphic to the behavior of the heat capacity of one-component systems in the vicinity of their tricritical points. The choice of the field variables in the vicinity of the tricritical point, for which isomorphism takes place, remains dubious.

ACKNOWLEDGMENTS

This study was supported by the international oil-and-gas company Schlumberger Oilfield Services.

REFERENCES

1. V. P. Voronov and E. E. Gorodetskii, Pis'ma Zh. Éksp. Teor. Fiz. **72**, 740 (2000) [JETP Lett. **72**, 516 (2000)].
2. E. Fernandez-Fassnacht, A. G. Williamson, A. Sivaraman, *et al.*, J. Chem. Phys. **66**, 4133 (1987).
3. Y.-N. Lin, R. J. J. Chappeler, and R. Kobayashi, J. Chem. Eng. Data **22**, 402 (1973).
4. V. P. Voronov, M. Yu. Belyakov, E. E. Gorodetskii, *et al.*, Transp. Porous Media **52**, 123 (2003).
5. T. A. Alvesalo, P. M. Bergland, S. T. Islander, *et al.*, Phys. Rev. A **4**, 2354 (1971).
6. L. V. Entov, V. A. Levchenko, and V. P. Voronov, Int. J. Thermophys. **14**, 221 (1993).
7. V. P. Voronov, Zh. Éksp. Teor. Fiz. **118**, 163 (2000) [JETP **91**, 144 (2000)].
8. L. Greek, C. M. Knobler, and R. L. Scott, J. Chem. Phys. **74**, 3489 (1981).
9. A. J. Liu and M. E. Fisher, Physica A (Amsterdam) **156**, 35 (1989).
10. C. Berviller, Phys. Rev. B **14**, 4964 (1976).
11. R. B. Griffiths, Phys. Rev. Lett. **24**, 715 (1970); Phys. Rev. B **7**, 545 (1973).
12. M. E. Fisher, Phys. Rev. **176**, 257 (1968).

Translated by N. Wadhwa