Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 197–201.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 227–232.

ATOMS, MOLECULES, OPTICS

Condensate of Excited States in Magnesium

A. V. Popov
Altai State University, Barnaul, 656099 Russia
e-mail: andrey_popov@barnaul.ru
Received October 13, 2004

Abstract—The results of theoretical study of the condensate of excited states in magnesium are reported. Excitations are described in the framework of the Hartree–Fock method taking into account energy level widths. It is shown that conditions for the emergence of condensed excited states are created in the optical range of atomic excitation energies. In the Mg_2 system, such conditions are created for any (indefinitely small) excitation fields in the optical energy range. The weaker the external field, the longer the lifetime and the shallower the potential well for the condensate of excitations. Most stable excitations in Mg_2 were detected at atomic spacings on the order of 9 Bohr radii. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

Excitation condensate (EC) was mentioned for the first time in [1]. The general theory of EC in atoms, molecules, and impurity centers of solids was described in [2]. In [3], experimental and theoretical results on condensed states in a system of excited cesium atoms were described. The existence of an isolated region of a metastable nonideal plasma [4] is not an exceptional property inherent in cesium alone [5]. It was noted in [6] that alkali metal atoms attain a size of 10^{-6}–10^{-5} cm for excitation energies close to the ionization energy and may form a metastable condensate for an atomic density of $\sim 10^{17}$–10^{18} cm^{-3}. EC clusters of hydrocarbons and EC hydrogen layers on the surface were mentioned in [7]. In all probability, the EC can also be formed in a natural way. For example, globe lightning as an EC was described in [8].

In our opinion, the most important experiments were performed with excited cesium (see [3, 5–7, 9] and the literature cited therein). The experiments of Swedish researchers with a thermionic converter of thermal energy into electric energy in the form of a cesium diode, whose emitter and collector are maintained at different temperatures, are of prime importance. Cesium vapor was supplied via a collector grid to the electrode gap and then condensed on the walls of a bell jar in which the thermionic converter was placed. The main evidence of the emergence of an EC was the peculiar shape of the current-voltage characteristics of the thermionic converter with a dynamic cesium vapor supply to the electrode gap (namely, a decrease in the collector work function to a value below 0.7 eV) and a considerable increase in the output voltage. Mass spectroscopy studies revealed the presence of clusters in the electrode gap, which contained up to 1000 atoms. Russian researchers [6] performed experiments analogous to those described above using improved devices and methods. However, they failed to completely reproduce the EC characteristics. For example, when the equilibrium regime of cesium vapor supply was changed to a dynamic mode, the work function of the collector decreased only to 1 eV. According to estimates, the clusters formed in the EC contain approximately 100 atoms with an EC decay time exceeding the decay time of a cluster-free plasma by one or two orders of magnitude.

However, we share the conclusions drawn in [3] that the idea of condensation of excitations has been employed insufficiently by theorists as well as experimenters in spite of its simplicity and has been considered only in a limited number of cases in recent years.

Here, we report on the results of theoretical analysis of EC in magnesium. Excitations are described in the framework of the Hartree–Fock taking into account the widths of atomic levels. The idea that atomic level width should be taken into account was employed for the first time in [10], where excited hydrogen wavefunctions were calculated. We use this idea for calculating the wavefunctions of excited states of multielectron atoms [11]. The basic method in our study was the Hartree–Fock method. In our opinion, the one-electron Hartree–Fock method is the most consistent approach. All approximations are well-defined in it and it is clear (at least, in principle) what should be done to leave this approximation and to obtain more and more exact results. The main point is that this method makes it possible to satisfactorily describe the ground state as well as excitations in simple multielectron systems.

2. COMPUTATIONAL TECHNIQUE

Let us first consider the standard solution to the spectral problem of the ground state of a multielectron

atom in the Hartree–Fock approximation:

$$F\psi = E\psi. \quad (1)$$

The central-field approximation makes it possible to separate radial and angular variables and to seek the eigenfunctions of the Fock operator F in the form of the product of the radial function $R(r)$ and the angular function $Y(\theta, \varphi)$,

$$\psi = R(r)Y(\theta, \varphi), \quad (2)$$

where

$$-\Delta Y = l(l+1)Y, \quad -i\frac{\partial Y}{\partial \varphi} = mY. \quad (3)$$

Here, Δ is the spherical Laplace operator. If we require that solutions to Eqs. (3) be regular on a sphere for $0 \le \theta \le \pi$, $0 \le \varphi \le 2\pi$, continuous for $\theta = 0$ and $\theta = \pi$, and satisfy the condition $Y(\theta, \varphi + 2\pi) = Y(\theta, \varphi)$, we arrive at the eigenvalue problem permitting solutions for Eqs. (3) only for integral values of $l = 0, 1, 2, \ldots$ and $m = 0, \pm 1, \ldots, \pm l$.

Let us now consider the excited states of an atom in equilibrium field. Other particles surrounding the atom under investigation will be included in the concept of a thermostat. Such excitations may decay spontaneously over a finite time τ and possess, in accordance with the indeterminacy relation, a finite width $\Gamma \sim \hbar/\tau$ of the energy level characterizing a quasi-stationary system. The wavefunction describing this system must contain an exponential factor such that the probabilities defined by the squared modulus of the wavefunction decay according to the law $\exp(-\Gamma t/\hbar)$:

$$\Psi(\mathbf{r}, t) = \psi(\mathbf{r})\exp\left\{-i\left(E - \frac{i\Gamma}{2}\right)\frac{t}{\hbar}\right\}. \quad (4)$$

We assume that the central field approximation is justified in describing excitations to the same extent as in describing unfilled shells of the ground state. In this case, it is convenient to seek the function $\psi(\mathbf{r})$ appearing on the right-hand side of Eq. (4) in form (2). Moreover, let us suppose that angular function $Y(\theta, \varphi)$ satisfies the same Eqs. (3), but is regular only in φ. The arbitrary behavior of this function on the sphere in variable θ removes the limitation on integral values of parameter l. We assume that this parameter is complex-valued in the general case ($L = l + x + iy$, where as before, l assumes integral values). The range of the complex correction $x + iy$ can be limited in this case: $|x| < 0.5$ and $|y| < 0.5$. Thus, the problem of determining the spectrum of orbital excitations of an atom can be reduced to the eigenvalue problem for the time-independent Schrödinger equation:

$$\left(F + \frac{\hbar^2}{2m}\frac{u + iv}{r^2}\right)\psi = \varepsilon\psi. \quad (5)$$

Here, F is the Fock operator, $u = x(x + 2l + 1) - y^2$, and $v = y(2x + 2l + 1)$. It should be noted that problem (5) for $y \neq 0$ is essentially non-Hermitian with complex values of energy $\varepsilon = E - i\Gamma/2$. Trying all possible values of $|x| < 0.5$ and $|y| < 0.5$ in the search for self-consistent solutions of Eq. (5), we can trace the spectral characteristics of an excited atom from the minimum of its total energy.

In the subsequent analysis, we will use the atomic system of units. To pass to this system of units, it is sufficient to assume that in all relations used here, the Planck constant $\hbar = 1$, the squared electron charge $e^2 = 2$, and the electron mass $m = 1/2$. Then the energy will be measured in Rydbergs and the distance in Bohr radii.

3. RESULTS OF COMPUTATION AND DISCUSSION

We solved problem (5) using the Routan method in the basis of the Gaussian functions. For the object of investigation, we initially chose beryllium, which is one of the simplest elements in the Periodic Table. As expected, no EC is formed in beryllium in the X-ray excitation energy range [11]. It was found that conditions for the EC formation are created in the case of excitation of atoms in the optical energy range. However, the fields converting the ground state of atoms into the system of excited states whose decay may lead to EC formation are comparable to the fields in the atom itself. To verify the assumption that conditions for EC formation can be created in weaker external fields also, we placed the atom under study in the field of another beryllium atom. Calculations show that short-lived excitations with high-intensity electron transitions to $2p$-symmetry states are formed in Be_2. The Be_2 system becomes more stable for atomic spacings on the order of four Bohr radii. The results obtained by us suggest that the EC that can be formed when the atom in question is placed in the field of two, three, or a larger number of atoms is even more stable in energy and has a longer life. Obviously, the combination of two or three atoms is not a condensate. Here and below, we treat EC as a system of atoms, molecules, or impurity centers in various media, in which a new phase can be formed owing to the interaction between excited electrons. Such an EC phase can also be obtained as a decay product of a strongly Coulomb-bound low-temperature plasma.

Using beryllium as an example, we proved that the engagement of all possible orbitals, including those with decay, makes it possible to describe the result of interaction of excited electrons with the fields. The origin of the fields causing an excitation is immaterial. In particular, these can be the fields produced by a system of excited atoms or molecules surrounding the excitation center under study. It is important that we could discover, along with the ground state, the states with local minima of total energy characterized by a long decay time.

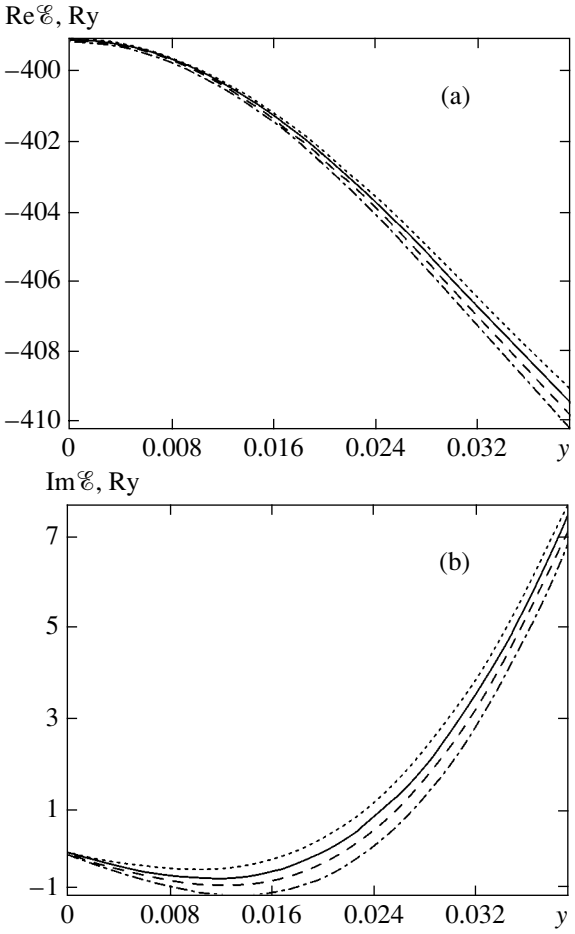

Fig. 1. Dependence of the real (a) and imaginary (b) parts of the total energy of a magnesium atom on parameter y for $x = 0$, 0.001, 0.002, and 0.003. The lower the curve, the higher the value of parameter x corresponding to it.

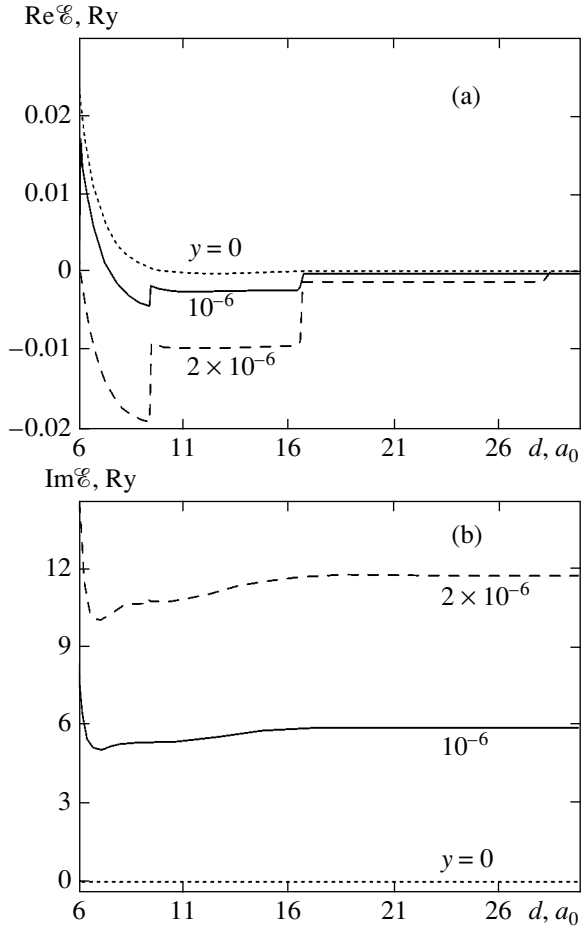

Fig. 2. Dependence of the real (a) and imaginary (b) parts of the total energy of electrons of two magnesium atoms on the distance d between them for three values of parameter y and $x = 0$ in the optical excitation energy range; d is measured in Bohr radii a_0.

In our opinion, it is more promising to work with heavier atoms. To verify the assumption that the heavier a dimer, the longer its excitation time, we chose magnesium, viz., an element also belonging to the second group in the Periodic system. Calculations were made in the framework of the above scheme. We solved problem (5) numerically in a basis of Gaussian-type functions using 12 functions in the expansion in $l = 0$ and 8 functions in the expansion in $l = 1$. Our estimates proved that a basis of such a length is quite suitable for solving Eqs. (5) with the help of the Routan method for magnesium atom if $|x| < 0.03$ and $|y| < 0.03$ since an increase in the length of this basis does not change the required accuracy of all values presented below for discussion.

The results of calculation of the total energy of an excited magnesium atom as a function of parameter y for $x = 0$, 0.001, 0.002, and 0.003 are presented in Fig. 1. The real part Re$\mathscr{E}$ of the total energy shown in Fig. 1a decreases monotonically with increasing x and y. This points towards the existence of more advantageous excited states (from the energy point of view) as compared to the ground state. However, the lifetime of such states, estimated using the uncertainty relation from the imaginary part Im$\mathscr{E}$ of the total energy, is finite everywhere except in the states for which Im$\mathscr{E} = 0$. The presence of points at which Im$\mathscr{E} = 0$ on the curves describing the dependence of Im$\mathscr{E}$ on parameter y (Fig. 1b) indicates that long-lived excitations can exist in principle in magnesium. The value of Im$\mathscr{E} = 0$, which is closest to the principal value, is attained for $x = 0$ and $y = 0.017$. However, it is rather difficult to exert such a strong action on the atom, corresponding to $y = 0.017$, in the optical energy range. Nevertheless, a long-lived excitation can be obtained if, for example, we place such an atom in the field produced by another atom. To verify this assumption, we calculated the total energy of two magnesium atoms as a function of the distance between them for small values of parameter y. The results of calculations presented in Fig. 2a show that Re$\mathscr{E}$ (the real part of the total energy of the electrons of two atoms) in the excited state is everywhere

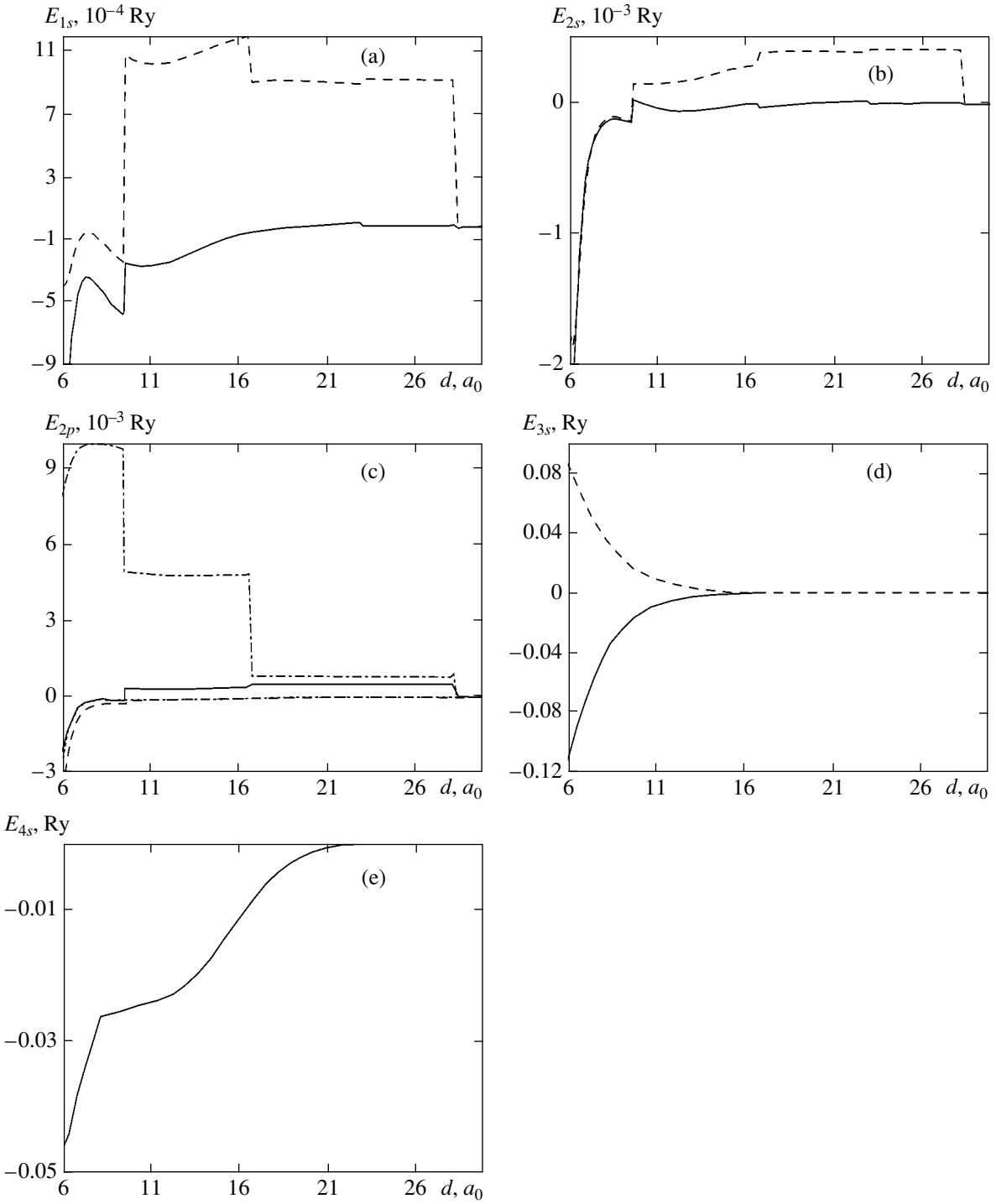

Fig. 3. Curves describing the dependence of the energy of 1*s* (a), 2*s* (b), 2*p* (c), 3*s* (d), and 4*s* (e) electrons of two magnesium atoms on the distance d between them for $x = 0$ and $y = 2 \times 10^{-6}$ in the optical excitation energy range. The energy is measured relative to the energy of 1*s* (a), 2*s* (b), 2*p* (c), 3*s* (d), and 4*s* (e) electrons of a magnesium atom; d is measured in Bohr radii a_0.

smaller than the total energy of the ground state for $y = 0$. Moreover, the value of $\mathrm{Re}\,\mathscr{E}$ is the smaller, the larger the value of y; however, the lifetime of such excitations decreases with increasing y. For instance, it is on the order of 80 fs for $y = 10^{-6}$ and nearly half this value for $y = 2 \times 10^{-6}$. It should be noted that these lifetimes estimated with the help of the uncertainty relation from the imaginary part of the total energy barely change upon an increase in the atomic spacing up to seven Bohr radii for each fixed value of y (see Fig. 2b) in contrast to

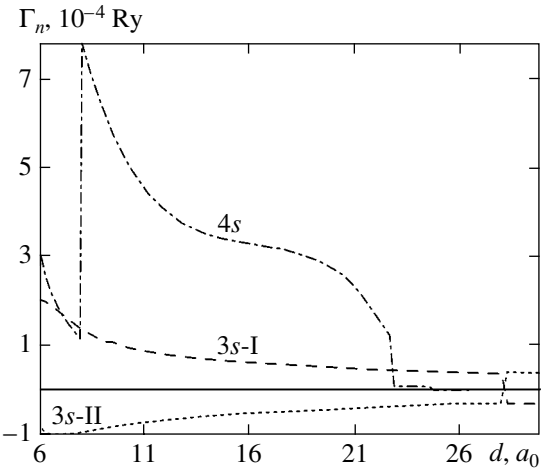

Fig. 4. Dependence of the imaginary part of the energy of s electrons of two magnesium atoms on the distance d between them for $x = 0$ and $y = 2 \times 10^{-6}$ in the optical excitation energy range; d is measured in Bohr radii a_0.

jump-like behavior of $\mathrm{Re}\mathscr{E}$ as a function of distance (Fig. 2a).

We attribute such a jump-like behavior of the real part of the total energy of two magnesium atoms as a function of the distance between these atoms to the rearrangement of the electron structure shown in Fig. 3. The stepwise change in the spectral structure of the core states of the 1s symmetry (Fig. 3a) is substantially smaller than stepwise variations of the spectra of states with 2s symmetry (Fig. 3b) and 2p symmetry (Fig. 3c). We believe that the change in the type of bond in 2p-symmetry states with increasing distance led to a stepwise change in the spectrum in levels with lower energies also. The 3s (Fig. 3d) and 4s-symmetry (Fig. 3e) states with higher energies do not experience jumps, although absolute variations in the spectrum of these states are the strongest.

The lifetime of electrons in the s-states can be estimated from the value of Γ_n (the imaginary part of the electron energy in these states) as a function of the distance between magnesium atoms, which is shown in Fig. 4 for $x = 0$ and $y = 2 \times 10^{-6}$, using the uncertainty relation. Excitations of electrons in the 3s-I states belonging to the first atom and in the 3s-II states belonging to the second atom decay over virtually the same time on the order of 500 fs. Excitations of electrons in the 4s state turn out to be long-lived for large distances between the atoms. For smaller distances beginning from 23 Bohr radii, the lifetime of these excitations rapidly decreases to values on the order of 200 fs, attains its minimum value ~60 fs when the atomic spacing decreases to 8 Bohr radii, followed by a sharp increase to 500 fs.

4. CONCLUSIONS

We proposed here a method for calculating the spectral characteristics and the decay time for excited states. Analysis of the results of calculations leads to the conclusion that excitations of magnesium atoms in the optical energy range create conditions for the emergence of EC. It has been shown that a condensate of excited states can be formed in Mg_2 for any indefinitely small external excitation fields in the optical energy range. The lower the intensity, the longer the lifetime and the shallower the potential well for EC in Mg_2. A competition exists between the depth of the potential well for EC and its lifetime. The most stable excitations exist in Mg_2 for atomic spacing on the order of 9 Bohr radii and in Be_2 for atomic spacings of 4 Bohr radii. Excitation condensate in Mg_2 is more stable than in Be_2 as regards lifetimes as well as the depth of potential wells.

ACKNOWLEDGMENTS

This study was supported by the foundation of noncommercial programs Dynasty and the International Center of Fundamental Physics in Moscow.

REFERENCES

1. É. A. Manykin, M. I. Ozhovan, and P. P. Poluéktov, Dokl. Akad. Nauk SSSR **260**, 1096 (1981) [Sov. Phys. Dokl. **26**, 974 (1981)].
2. É. A. Manykin, M. I. Ozhovan, and P. P. Poluéktov, Zh. Éksp. Teor. Fiz. **84**, 442 (1983) [Sov. Phys. JETP **57**, 256 (1983)].
3. É. A. Manykin, M. I. Ozhovan, and P. P. Poluéktov, Khim. Fiz. **18**, 87 (1999).
4. L. Holmlid, Phys. Rev. A **63**, 13817 (2001).
5. G. É. Norman, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 13 (2001) [JETP Lett. **73**, 10 (2001)].
6. V. I. Yarygin, V. N. Sidel'nikov, I. I. Kasikov, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **77**, 330 (2003) [JETP Lett. **77**, 280 (2003)].
7. L. Holmlid and É. A. Manykin, Zh. Éksp. Teor. Fiz. **111**, 1601 (1997) [JETP **84**, 875 (1997)].
8. G. É. Norman, Khim. Fiz. **18**, 78 (1999).
9. R. Svensson and L. Holmlid, Phys. Rev. Lett. **83**, 1739 (1999).
10. A. Yanavichus and V. Shuchurov, Litov. Fiz. Sb. **8**, 47 (1968).
11. A. V. Popov, Opt. Spektrosk. **93**, 5 (2002) [Opt. Spectrosc. **93**, 1 (2002)].

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 202–211.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 233–242.

ATOMS, MOLECULES, OPTICS

Optical Nonreciprocity in a Magnetically Active Isotropic Medium

V. I. Denisov[a], I. P. Denisova[b], and V. G. Zhotikov[c]

[a]*Moscow State University, Leninskie gory, Moscow, 119992 Russia*
[b]*Tsiolkovskiĭ State Technological University, Moscow, 121552 Russia*
[c]*Moscow Institute of Physics and Technology (Technical University), Dolgoprudnyi, Moscow oblast, 141700 Russia*
e-mail: Denisov@srd.sinp.msu.ru; Zhotikov@complat.ru

Received December 27, 2004

Abstract—The complex permittivity tensor of a magnetically active isotropic medium is expanded, to third-order accuracy, in terms of a small parameter equal to the ratio of the mean distance between the atoms of the medium to the wavelength of an electromagnetic wave. A dispersion equation is constructed. On the basis of this equation, the refractive indices of the medium for normal waves are obtained when a longitudinal magnetic field is applied to the medium. It is shown that calculations to third-order accuracy yield different values for the velocities of all four normal waves that propagate in the medium in forward and backward directions. Calculations are carried out for the experiment conducted with the use of a ring laser for measuring the expansion coefficients of the complex permittivity tensor that are responsible for the small difference between the velocities of the normal waves propagating in forward and backward directions. It is shown that, in the case of an isotropic optically nonactive medium, the third-order expansion coefficients can be measured by means of a ring laser with an absolute accuracy on the order of 10^{-14}. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

By optical nonreciprocity is usually meant the property of a material medium to provide different conditions for the propagation of light in forward and backward directions. Nonreciprocity may manifest itself in the phase (propagation velocity), amplitude, or the polarization of an electromagnetic wave.

A large number of works have been devoted to the theoretical and experimental study of optical nonreciprocity (see, for example, the survey [1] and the references therein). In particular, the contributions of the Faraday effect and the natural optical activity to nonreciprocal phenomena in anisotropic media were studied in [2]. The phase nonreciprocity in optically active (gyrotropic) crystals that results from the above-mentioned physical factors, however small, can be measured in experiments with ring lasers. However, the case considered in [2] does not exhaust all possible mechanisms of the onset of optical nonreciprocity in condensed media because the expansions of the inverse tensor of the complex permittivity used in [2] do not contain third-order terms (with respect to the ratio of the mean interatomic distance to the wavelength of the incident electromagnetic wave), which can be measured by modern experimental techniques.

Let us show by a simple example of a magnetically active isotropic medium that the third-order coefficients must be taken into account, because it is these coefficients that are responsible for the optical nonreciprocity even in the absence of the first-order spatial dispersion.

Consider a nonmagnetic isotropic medium in an external magnetic field **H**. In the general case [3–7], the relation between the flux densities and the fields is expressed as

$$D_j = \varepsilon_{jm}(\omega, \mathbf{k})E_m + \alpha_{jm}(\omega, \mathbf{k})H_m,$$

$$B_j = \beta_{jm}(\omega, \mathbf{k})E_m + \mu_{jm}(\omega, \mathbf{k})H_m.$$

However, taking into account the spatial dispersion, we may assume that [3]

$$\beta_{jm}(\omega, \mathbf{k}) = \alpha_{jm}(\omega, \mathbf{k}) = 0,$$
$$\mu_{jm}(\omega, \mathbf{k}) = \delta_{jm},$$

for nonmagnetic media, where δ_{jm} is the Kronecker delta.

The complex permittivity tensor ε_{mn} of such a medium [3–7] is a function of frequency ω, wavevector **k**, and the external magnetic field **H**. In the transmission band, this tensor must satisfy the following conditions [3]:

$$\begin{aligned}\varepsilon^*_{mn}(\omega, \mathbf{k}, \mathbf{H}) &= \varepsilon_{nm}(\omega, \mathbf{k}, \mathbf{H}),\\ \varepsilon^*_{mn}(\omega, \mathbf{k}, \mathbf{H}) &= \varepsilon_{mn}(-\omega, -\mathbf{k}, \mathbf{H}),\\ \varepsilon_{mn}(\omega, \mathbf{k}, \mathbf{H}) &= \varepsilon_{nm}(\omega, -\mathbf{k}, -\mathbf{H}).\end{aligned} \quad (1)$$

The tensor ε_{mn} as a function of $\mathbf{k}$ is characterized by a small parameter $\delta = a/\lambda$, where a is a characteristic length of the medium (the size of molecules, the lattice constant, the mean distance between molecules, etc.) and λ is the wavelength in the medium. For condensed media in the optical region of the spectrum, the parameter δ is small: $\delta \sim 10^{-3}$.

The measurements [6] of the magnetic rotation angles of the polarization plane in magnetically active media show that, for $H \sim 10^3$ G, the dependence of the tensor ε_{mn} on the external magnetic field can be characterized by the square of the parameter δ.

Thus, the tensor ε_{mn} can be expanded in series in the small parameter $\delta \sim 10^{-3}$. In the simplest case when the optical properties of the medium differ from isotropic ones only in the directions determined by the vectors $\mathbf{k}$ and $\mathbf{H}$, a phenomenological expansion of the tensor ε_{mn} satisfying conditions (1) takes the following form in Cartesian coordinates in the three-dimensional Euclidean space:

$$\varepsilon_{mn}(\omega, \mathbf{k}, \mathbf{H}) = \left[\varepsilon + \alpha_2 k^2 + \frac{\omega}{c}\chi\alpha_4 \mathbf{k}\cdot\mathbf{H}\right]\delta_{nm}$$

$$+ i\chi[\alpha_1 + \alpha_3 k^2]e_{nmj}k_j + \alpha_5 k_n k_m \qquad (2)$$

$$+ i\frac{\omega}{c}\alpha_6 e_{nmj}H_j + \frac{\omega}{c}\chi\alpha_7[k_n H_m + k_m H_n] + O_{nm}(\delta^4),$$

where χ is an axial scalar (pseudoscalar) that is equal to +1 in right-hand coordinates and –1 in left-hand coordinates and e_{nmj} is the Levi-Civita symbol.

To facilitate further calculations, let us introduce the following notation:

$$\mathbf{k} = \frac{\omega}{c}\tilde{n}\mathbf{S},\quad f_1 = \frac{\omega}{c}\alpha_1,\quad \varepsilon_2 = \frac{\omega^2}{c^2}\alpha_2,\quad f_2 = \frac{\omega^2}{c^2}\alpha^5,$$

$$h_2 = \frac{\omega}{c}\alpha_6,\quad \beta_3 = \frac{\omega}{c}\alpha_7,\quad h_3 = \frac{\omega^2}{c^2}\alpha_4,\quad f_3 = \frac{\omega^3}{c^3}\alpha_3,$$

where $\mathbf{S}$ is a unit vector and $\tilde{n}$ is the refractive index of the medium.

Then, expression (2) is rewritten as

$$\varepsilon_{mn}(\omega, \mathbf{k}, \mathbf{H}) = [\varepsilon + \varepsilon_2\tilde{n}^2 + \chi h_3\tilde{n}\mathbf{S}\cdot\mathbf{H}]\delta_{nm}$$

$$+ i\chi[f_1\tilde{n} + f_3\tilde{n}^3]e_{nmj}S_j + f_2\tilde{n}^2 S_n S_m + ih_2 e_{nmj}H_j \qquad (3)$$

$$+ \chi\beta_3\tilde{n}[S_n H_m + S_m H_n] + O_{nm}(\delta^4),$$

where $\varepsilon = \varepsilon(\omega)$ is a zeroth-order term in the small parameter δ; f_1 is a first-order term; ε_2, f_2, and h_2 are second-order terms; f_3, h_3, and β_3 are third-order terms; and $O_{nm}(\delta^4)$ stands for small terms of the fourth and higher orders in δ.

The physical meaning of the expansion coefficients of the tensor ε_{mn} in (3) is obvious: the coefficients $f_1 \sim \delta$ and $f_3 \sim \delta^3$ characterize the optical activity of the substance in the first- and third-order terms, respectively; $f_2 \sim \delta^2$ and $\varepsilon \sim \delta^2$ describe the effect of spatial dispersion on the tensor ε_{mn} in the second order in δ; h_2 characterizes the magnetic activity of an isotropic substance; and the coefficients h_3 and β_3 describe the combined effect of the spatial dispersion and the magnetic activity on the tensor ε_{mn} in the third order in δ.

In view of the second relation in (1), the scalar coefficients ε, ε_2, f_2, h_3, and β_3 in (3) are even functions of frequency ω, while the coefficients f_1, f_3, and h_2 are odd functions.

2. DISPERSION EQUATION AND ITS SOLUTION

Consider a plane electromagnetic wave with the vectors $\mathbf{E}_w$, $\mathbf{D}_w$, and $\mathbf{B}_w = \mathbf{H}_w$ proportional to $\exp\{-i[\omega t - \mathbf{k}\cdot\mathbf{r}]\}$. In this case, the equations of macroscopic electrodynamics are represented as

$$\mathbf{k}\times\mathbf{H}_w = \frac{\omega}{c}\mathbf{D}_w,\quad \mathbf{k}\times\mathbf{E}_w = -\frac{\omega}{c}\mathbf{H}_w. \qquad (4)$$

Eliminating the vector $\mathbf{H}_w$ from the first equation of system (4) and using the second equation, we obtain the system of algebraic equations $\Pi_{nm}E_m = 0$, where

$$\Pi_{nm} = \frac{\omega^2}{c^2}\{[\varepsilon + \varepsilon_2\tilde{n}^2 + \chi h_3\tilde{n}\mathbf{S}\cdot\mathbf{H}]\delta_{nm}$$

$$+ i\chi\tilde{n}[f_1 + f_3\tilde{n}^2]e_{nmj}S_j + f_2\tilde{n}^2 S_n S_m + ih_2 e_{nmj}H_j \qquad (5)$$

$$+ \chi\beta_3\tilde{n}[S_n H_m + S_m H_n] + O_{nm}(\delta^4)\} + [k_n k_m - k^2\delta_{nm}].$$

As is known, for this system of equations to have a nontrivial solution, it is necessary that $\det\|\Pi_{nm}\| = 0$.

It is convenient to perform further calculations with the use of the tensor relations obtained in [8, 9].

Let us introduce certain definitions. Consider an arbitrary second-rank tensor A_{nm} in the three-dimensional Euclidean space whose metric tensor coincides with the Kronecker tensor δ_{nm} in Cartesian coordinates. We define the Nth power of this tensor, where N is a nonnegative integer, as a tensor $A_{nm}^{(N)}$ given by the prod-

uct of N tensors A_{nm} whose indices are contracted by the rule

$$A_{nm}^{(N)} = \underbrace{A_{nj_1}A_{j_1j_2}\ldots A_{j_Nm}}_{N}.$$

Contracting the remaining indices in this expression, we obtain an invariant of the Nth power of this tensor:

$$A_{(N)} \equiv A^{(N)} = A_{kk}^{(N)}.$$

According to this definition, for $N = 0$, we set $A_{ik}^{(0)} = \delta_{ik}$; as a result, the zero-degree invariant of any second-rank tensor in the three-dimensional space is equal to three:

$$A_{(0)} = 3.$$

According to [9], the third power of the tensor A_{nm} in the three-dimensional Euclidean space can be expressed in terms of lower powers of this tensor and its invariants:

$$A_{nm}^{(3)} = A_{nm}^{(2)}A_{(1)} + \frac{1}{2}A_{nm}[A_{(2)} - A_{(1)}^2] + \frac{1}{6}\delta_{nm}[2A_{(3)} - 3A_{(1)}A_{(2)} + A_{(1)}^3].$$

In this notation, the determinant of the third-order matrix whose elements are given by the components of the tensor A_{nm} takes the form

$$\det\|A_{nm}\| = \frac{1}{6}[2A_{(3)} - 3A_{(2)}A_{(1)} + A_{(1)}^3]. \quad (6)$$

If $\det\|A_{nm}\| \neq 0$ and $\det\|A_{nm}\| \neq \pm\infty$, then one can construct the inverse tensor $A_{nm}^{(-1)}$, which satisfies the relation

$$A_{nm}^{(-1)}A_{mp} = \delta_{np}.$$

This tensor is expressed as

$$A_{nm}^{(-1)} = \frac{6A_{nm}^{(2)} - 6A_{nm}A_{(1)} + 3\delta_{nm}[A_{(1)}^2 - A_{(2)}]}{[2A_{(3)} - 3A_{(1)}A_{(2)} + A_{(1)}^3]}. \quad (7)$$

Using formulas (5) and (6), we obtain the following dispersion equation from the condition $\det\|\Pi_{nm}\| = 0$:

$$\varepsilon[\varepsilon - \tilde{n}^2]^2 + \chi\varepsilon\tilde{n}[3\varepsilon h_3 + 2\varepsilon\beta_3 - 2f_1h_2]\mathbf{S}\cdot\mathbf{H} + \varepsilon\tilde{n}^2[3\varepsilon\varepsilon_2 + \varepsilon f_2 - f_1^2] + \tilde{n}^2h_2^2[H^2 - (\mathbf{S}\cdot\mathbf{H})^2] - 4\chi\varepsilon\tilde{n}^3[h_3 + \beta_3]\mathbf{S}\cdot\mathbf{H} - 2\varepsilon\tilde{n}^4[2\varepsilon_2 + f_2] + \chi\tilde{n}^5[h_3 + 2\beta_3]\mathbf{S}\cdot\mathbf{H} + \tilde{n}^6[\varepsilon_2 + f_2] + \tilde{n}^4[3\varepsilon\varepsilon_2^2 + 2\varepsilon\varepsilon_2f_2 - 2\varepsilon f_1f_3 - \varepsilon_2f_1^2 - f_1^2f_2] - \varepsilon h_2^2H^2 - 2\tilde{n}^6\varepsilon_2[\varepsilon_2 + f_2] + [\varepsilon - \tilde{n}^2][\varepsilon O_{mm}(\delta^4) - \tilde{n}^2S_nS_mO_{nm}(\delta^4)] = 0. \quad (8)$$

We will seek a solution to this equation in the form

$$\tilde{n} = \sqrt{\varepsilon + F}, \quad (9)$$

where F is an unknown function that contains small quantities.

Substituting (9) into Eq. (8) and solving it, we find the function F:

$$F = \frac{1}{8\varepsilon}\{\alpha\sqrt{\varepsilon}[8\varepsilon f_1 + 8\chi\sqrt{\varepsilon}h_2\mathbf{S}\cdot\mathbf{H} + [f_1^3 + 12\varepsilon\varepsilon_2f_1 + 8\varepsilon^2f_3]] + 4\varepsilon[f_1^2 + 2\varepsilon\varepsilon_2] + 4\chi\sqrt{\varepsilon}[2\varepsilon h_3 + f_1h_2]\mathbf{S}\cdot\mathbf{H}\} + O(\delta^4), \quad (10)$$

where $\alpha = \pm 1$ (the plus sign corresponds to the first normal wave, and the minus sign, to the second normal wave).

Thus, in a medium with the complex permittivity tensor (3), two normal waves with the phase velocities $V_{\text{ph}} = \omega/k$ may propagate in each direction:

$$V_{\text{ph}}^{1,2} = \frac{c}{\sqrt{\varepsilon}}\left\{1 - \frac{\alpha}{2\sqrt{\varepsilon}}f_1 + \frac{1}{8\varepsilon}[(f_1^2 - 4\varepsilon\varepsilon_2) - 4\alpha\chi h_2\mathbf{S}\cdot\mathbf{H}] - \frac{1}{2\sqrt{\varepsilon^3}}[\alpha\varepsilon^2f_3 + \chi(\varepsilon h_3 - f_1h_2)\mathbf{S}\cdot\mathbf{H}]\right\}.$$

The refractive indices $\tilde{n} = ck/\omega$ of the medium for these waves are given by

$$\tilde{n}(\omega) = \sqrt{\varepsilon} + \frac{\alpha}{2}f_1 + \frac{1}{8\sqrt{\varepsilon}}[f_1^2 + 4\varepsilon\varepsilon_2 + 4\alpha\chi h_2\mathbf{S}\cdot\mathbf{H}] + \frac{1}{2}[\alpha(\varepsilon f_3 + f_1\varepsilon_2) + \chi h_3\mathbf{S}\cdot\mathbf{H}]. \quad (11)$$

Note that, by virtue of relations (9) and (10), the formulas for the refractive indices of the medium represent

expansions in the small parameter δ up to the cubic terms inclusive.

It follows from (9) and (10) that the phase velocities of the normal waves and the refractive indices of an optically active isotropic medium in an external magnetic field depend on the sign of $\mathbf{S}\cdot\mathbf{H}$. This means that the medium is optically nonreciprocal, because the phase velocity of each normal wave depends on which direction, along or opposite the external magnetic field $\mathbf{H}$, a wave propagates.

3. POLARIZATION STATES OF NORMAL WAVES

The solution of dispersion equation (8) for a nonabsorbing isotropic medium with the complex permittivity tensor (3) has shown that, in this medium, two normal waves with different phase velocities may propagate in each direction.

Now, we consider the polarization states of these normal waves. In the optics of anisotropic and gyrotropic media, the polarization of an electromagnetic wave is usually defined by the vector $\mathbf{D}$, which, according to the equation $\operatorname{div}\mathbf{D} = 0$, is perpendicular to the vector $\mathbf{k}$. Therefore, to study the polarizations of normal waves, one should rewrite the constitutive equations as

$$E_n = \varepsilon^{(-1)}_{nm} D_m,$$

where $\varepsilon^{(-1)}_{nm}$ is the inverse of the tensor ε_{mp}.

Using formulas (3) and (7), one can easily find an expansion of the tensor $\varepsilon^{(-1)}_{nm}$ in series in the dimensionless parameter δ up to the cubic terms inclusive:

$$\begin{aligned}\varepsilon^{(-1)}_{nm} &= \frac{1}{\varepsilon^4}\{\varepsilon[\varepsilon^2+\tilde n^2(f_1^2-\varepsilon\varepsilon_2)\\ &+\chi\tilde n(2f_1h_2-h_3\varepsilon)\mathbf{S}\cdot\mathbf{H}]\delta_{nm}\\ &-i[\chi\tilde n[\tilde n^2(\varepsilon^2 f_3-2f_1\varepsilon\varepsilon_2+f_1^3)\\ &+f_1\varepsilon^2]e_{nmj}S_j+\varepsilon^2h_2e_{nmj}H_j]\\ &-\varepsilon\tilde n^2[\varepsilon f_2+f_1^2]S_nS_m-\chi\varepsilon\tilde n[f_1h_2+\varepsilon\beta_3]\\ &\times(H_nS_m+H_mS_n)\}+O_{nm}(\delta^4).\end{aligned} \tag{12}$$

The equations of macroscopic electrodynamics, expressed in terms of the vector $\mathbf{D}$, take the form

$$\mathbf{k}\times\mathbf{H}_{\mathrm w} = \frac{\omega}{c}\mathbf{D}_{\mathrm w},\quad e_{ijn}k_j\varepsilon^{(-1)}_{nm}D^{\mathrm w}_m = -\frac{\omega}{c}H^{\mathrm w}_i.$$

Eliminating $\mathbf{H}_{\mathrm w}$ from these equations, we obtain

$$\Gamma_{nm}D^{\mathrm w}_m = 0,$$

where the tensor Γ_{nm} has the form

$$\Gamma_{nm} = \frac{c}{\omega}\left\{k_nk_p\varepsilon^{(-1)}_{pm}-k^2\varepsilon^{(-1)}_{nm}+\frac{\omega^2}{c^2}\delta_{nm}\right\}.$$

Substituting relation (12) into this expression, we obtain

$$\begin{aligned}\Gamma_{nm} &= \frac{\omega}{\varepsilon^4 c}\{[\varepsilon^3[\varepsilon-\tilde n^2]+\varepsilon\tilde n^4[\varepsilon\varepsilon_2-f_1^2]\\ &+\chi\varepsilon\tilde n^3[\varepsilon h_3-2f_1h_2]\mathbf{S}\cdot\mathbf{H}]\delta_{nm}\\ &+S_nS_m[\chi\varepsilon\tilde n[f_1h_2-\varepsilon(h_3+\beta_3)]\mathbf{S}\cdot\mathbf{H}\\ &-\varepsilon^2\varepsilon_2\tilde n^2+\varepsilon f_1^2\tilde n^2+\varepsilon^3]\\ &+i\chi e_{nmj}S_j[\tilde n^4[\varepsilon^2f_3-2\varepsilon\varepsilon_2f_1+f_1^3]+\tilde n^2\varepsilon^2f_1]\\ &+ie_{nmj}H_j\varepsilon^2\tilde n^2h_2+iS_ne_{mpj}S_pH_j\varepsilon^2h_2\tilde n^2\\ &+H_nS_m\chi\varepsilon\tilde n^2[f_1h_2+\varepsilon\beta_3]\}+\frac{\omega}{c}O_{nm}(\delta^4).\end{aligned} \tag{13}$$

As expected, the equation $\det\|\Gamma_{nm}\| = 0$ has the same roots as the equation $\det\|\Pi_{nm}\| = 0$.

Let us orient the axes of the right-hand ($\chi = +1$) coordinate system so that the z axis is parallel to the wavevector $\mathbf{k}$. Then, $\mathbf{k} = \{0, 0, \xi k\}$, where $\xi = 1$ for the wave propagating in the positive direction of the axis z and $\xi = -1$ for the wave propagating in the opposite direction.

It follows from (13) that, in this system, the coordinates of the components Γ_{31} and Γ_{32} vanish and $\Gamma_{33} = \omega/c$. Therefore, $D_z = 0$ in this system of coordinates.

The remaining two equations of the system $\Gamma_{\alpha\beta}D_\beta = 0$ have a zero determinant. Hence, one of these equations is independent:

$$\Gamma_{11}D_x + \Gamma_{12}D_y = 0.$$

Substituting the coefficients Γ_{11} and Γ_{12} calculated in the chosen system of coordinates to third-order accuracy in the small parameter δ into the above relation and taking into account formulas (11) and (13), we obtain

$$\Gamma_{11}\{D_x - i\alpha\xi D_y\} = 0, \tag{14}$$

where

$$\begin{aligned}\Gamma_{11} &= -\frac{\omega}{8c\sqrt{\varepsilon^3}}\{8\alpha\varepsilon f_1+4\sqrt{\varepsilon}[3\varepsilon f_1^2+2\alpha\xi h_2H_z]\\ &+[20\xi f_1h_2H_z+\alpha(17f_1^3-4\varepsilon f_1\varepsilon_2+8\varepsilon^2f_3)]\}.\end{aligned}$$

Equation (14) yields

$$D_x = i\alpha\xi D_y.$$

This means that both normal waves ($\alpha = \pm 1$) in this case are circularly polarized, but the signs of rotation of the vector **D** in these waves are opposite.

4. TRANSMISSION OF ELECTROMAGNETIC WAVES THROUGH A PLANE-PARALLEL PLATE AT NORMAL INCIDENCE

Consider a plane-parallel layer of a substance with the tensor ε_{nm} given by (3). Without loss of generality, we will assume that the substance layer is enclosed between the planes $z = 0$ and $z = L$.

Suppose that a monochromatic electromagnetic wave of frequency ω, polarized along the x axis, is incident on this layer from vacuum and that the vector **k** of this wave is parallel to the z axis. In the steady-state mode, there exist incident and reflected waves in the region $z < 0$, two pairs of normal waves that propagate in opposite directions in the region $0 < z < L$, and a transmitted electromagnetic wave in the region $z > L$. Let us find the amplitudes and phases of the reflected and transmitted waves.

Away from the absorption bands, the refractive indices of the medium for the normal waves propagating in the positive direction of the axis z are represented as

$$\begin{aligned} n_1 &= \sqrt{\varepsilon} + \frac{1}{2}f_1 + \frac{1}{8\sqrt{\varepsilon}}[(f_1^2 + 4\varepsilon\varepsilon_2) + 4h_2H_z] \\ &\quad + \frac{1}{2}[h_3H_z + (\varepsilon f_3 + f_1\varepsilon_2)], \\ n_2 &= \sqrt{\varepsilon} - \frac{1}{2}f_1 + \frac{1}{8\sqrt{\varepsilon}}[(f_1^2 + 4\varepsilon\varepsilon_2) - 4h_2H_z] \\ &\quad + \frac{1}{2}[h_3H_z - (\varepsilon f_3 + f_1\varepsilon_2)]. \end{aligned} \tag{15}$$

For the waves that propagate in the negative direction of the axis z, we have

$$\begin{aligned} n_3 &= \sqrt{\varepsilon} + \frac{1}{2}f_1 + \frac{1}{8\sqrt{\varepsilon}}[(f_1^2 + 4\varepsilon\varepsilon_2) - 4h_2H_z] \\ &\quad - \frac{1}{2}[h_3H_z - (f_3 + f_1\varepsilon_2)], \\ n_4 &= \sqrt{\varepsilon} - \frac{1}{2}f_1 + \frac{1}{8\sqrt{\varepsilon}}[(f_1^2 + 4\varepsilon\varepsilon_2) + 4h_2H_z] \\ &\quad - \frac{1}{2}[h_3H_z + (\varepsilon f_3 + f_1\varepsilon_2)]. \end{aligned} \tag{16}$$

In the region $z < 0$, the vector **E** can be represented as

$$\begin{aligned} \mathbf{E} &= \{E_{0x}\mathbf{e}_x + iE_{0y}\mathbf{e}_y\}\exp\left[-i\omega\left(t - \frac{z}{c}\right)\right] \\ &\quad + \{E_{1x}\mathbf{e}_x + E_{1y}\mathbf{e}_y\}\exp\left[-i\omega\left(t + \frac{z}{c}\right)\right]. \end{aligned}$$

In the region $0 < z < L$, it is convenient to represent the vector **E** in terms of the components of the vector **D**:

$$\begin{aligned} E_m &= \{\varepsilon_{m2}^{(-1)}(n_1) + i\varepsilon_{m1}^{(-1)}(n_1)\}D_{1y} \\ &\quad \times \exp\left[-i\omega\left(t - \frac{zn_1}{c}\right)\right] + \{\varepsilon_{m2}^{(-1)}(n_2) - i\varepsilon_{m1}^{(-1)}(n_2)\}D_{2y} \\ &\quad \times \exp\left[-i\omega\left(t - \frac{zn_2}{c}\right)\right] + \{\varepsilon_{m2}^{(-1)}(n_3) - i\varepsilon_{m1}^{(-1)}(n_3)\}D_{3y} \\ &\quad \times \exp\left[-i\omega\left(t + \frac{zn_3}{c}\right)\right] + \{\varepsilon_{m2}^{(-1)}(n_4) + i\varepsilon_{m1}^{(-1)}(n_4)\}D_{4y} \\ &\quad \times \exp\left[-i\omega\left(t + \frac{zn_4}{c}\right)\right]. \end{aligned} \tag{17}$$

Finally, in the region $z > L$, we have

$$\mathbf{E} = \{E_{3x}\mathbf{e}_x + E_{3y}\mathbf{e}_y\}\exp\left[-i\omega\left(t - \frac{z}{c}\right)\right].$$

The unknown components E_{1x}, E_{1y}, D_{1y}, D_{2y}, D_{3y}, D_{4y}, E_{3x}, and E_{3y} are determined from the standard conditions that the tangential components of the vectors **E** and $\mathbf{H} = c\mathbf{k} \times \mathbf{E}/\omega$ are continuous across the surfaces $z = 0$ and $z = L$. Taking into account that the relations

$$\varepsilon_{11}^{(-1)}(n) = \varepsilon_{22}^{(-1)}(n), \quad \varepsilon_{12}^{(-1)}(n) = -\varepsilon_{21}^{(-1)}(n)$$

hold in the chosen system of coordinates, after simple but tedious calculations, we obtain

$$\begin{aligned} E_{1x} &= -E_{0x} + Y_1 + Y_2 + Y_3 + Y_4, \\ E_{1y} &= -i[E_{0y} + Y_1 - Y_2 - Y_3 + Y_4], \\ E_{3x} &= \left[Y_1\exp\left[i\frac{\omega L}{c}n_1\right] + Y_2\exp\left[i\frac{\omega L}{c}n_2\right]\right. \\ &\quad \left. + Y_3\exp\left[-i\frac{\omega L}{c}n_3\right] + Y_4\exp\left[-i\frac{\omega L}{c}n_4\right]\right] \\ &\quad \times \exp\left[-i\frac{\omega L}{c}\right], \end{aligned}$$

$$E_{3y} = -i\left[Y_1\exp\left[i\frac{\omega L}{c}n_1\right] - Y_2\exp\left[i\frac{\omega L}{c}n_2\right] - Y_3\exp\left[-i\frac{\omega L}{c}n_3\right] + Y_4\exp\left[-i\frac{\omega L}{c}n_4\right]\right] \times \exp\left[-i\frac{\omega L}{c}\right], \quad (18)$$

$$D_{1y} = -\frac{iY_1}{[\varepsilon_{22}^{(-1)}(n_1) + i\varepsilon_{21}^{(-1)}(n_1)]},$$

$$D_{2y} = -\frac{iY_2}{[\varepsilon_{22}^{(-1)}(n_2) - i\varepsilon_{21}^{(-1)}(n_2)]},$$

$$D_{3y} = -\frac{iY_3}{[\varepsilon_{22}^{(-1)}(n_3) - i\varepsilon_{21}^{(-1)}(n_3)]},$$

$$D_{4y} = -\frac{iY_4}{[\varepsilon_{22}^{(-1)}(n_4) + i\varepsilon_{21}^{(-1)}(n_4)]},$$

where

$$Y_1 = \frac{(1+r_1)E_{0x}}{2\left[1 - r_1r_4\exp\left[i\frac{\omega}{c}(n_1+n_4)L\right]\right]},$$

$$Y_2 = \frac{(1+r_2)E_{0x}}{2\left[1 - r_2r_3\exp\left[i\frac{\omega}{c}(n_2+n_3)L\right]\right]},$$

$$Y_3 = -\frac{(1+r_3)r_2E_{0x}\exp\left[i\frac{\omega}{c}(n_2+n_3)L\right]}{2\left[1 - r_2r_3\exp\left[i\frac{\omega}{c}(n_2+n_3)L\right]\right]},$$

$$Y_4 = -\frac{(1+r_4)r_1E_{0x}\exp\left[i\frac{\omega}{c}(n_1+n_4)L\right]}{2\left[1 - r_1r_4\exp\left[i\frac{\omega}{c}(n_1+n_4)L\right]\right]},$$

$$r_\alpha = \frac{1-n_\alpha}{1+n_\alpha}, \quad \alpha = 1, 2, 3, 4.$$

Neglecting the spatial dispersion and the magnetic activity, we obtain the well-known [3] relations for the complex amplitudes from (18) in the case when an electromagnetic wave is transmitted through a plane-parallel layer of an isotropic substance.

If a linearly polarized monochromatic electromagnetic wave propagates in the negative direction of the axis z, then there are incident and reflected waves in the region $z > L$:

$$\mathbf{E} = E_{0x}\mathbf{e}_x\exp\left[-i\omega\left(t + \frac{z}{c}\right)\right] + \{E_{rx}\mathbf{e}_x + E_{ry}\mathbf{e}_y\}\exp\left[-i\omega\left(t - \frac{z}{c}\right)\right].$$

In the region $0 < z < L$, the vector $\mathbf{E}$ is defined by an expression similar to (17), and, in the region $z < 0$, there is a transmitted wave given by

$$\mathbf{E} = \{E_{px}\mathbf{e}_x + E_{py}\mathbf{e}_y\}\exp\left[-i\omega\left(t + \frac{z}{c}\right)\right].$$

Then, by virtue of the boundary conditions at $z = 0$ and $z = L$, we have

$$E_{px} = X_1 - X_2 - X_3 + X_4,$$

$$E_{py} = -i[X_1 + X_2 + X_3 + X_4],$$

$$E_{rx} = \left\{-E_{0x}\exp\left[-i\frac{\omega L}{c}\right] + X_1\exp\left[i\frac{\omega L}{c}n_1\right] - X_2\exp\left[i\frac{\omega L}{c}n_2\right] - X_3\exp\left[-i\frac{\omega L}{c}n_3\right] + X_4\exp\left[-i\frac{\omega L}{c}n_4\right]\right\}\exp\left[-i\frac{\omega L}{c}\right],$$

$$E_{ry} = -i\left[X_1\exp\left[i\frac{\omega L}{c}n_1\right] + X_2\exp\left[i\frac{\omega L}{c}n_2\right] + X_3\exp\left[-i\frac{\omega L}{c}n_3\right] + X_4\exp\left[-i\frac{\omega L}{c}n_4\right]\right]\exp\left[-i\frac{\omega L}{c}\right], \quad (19)$$

$$D_{1y} = -\frac{iX_1}{[\varepsilon_{22}^{(-1)}(n_1) + i\varepsilon_{21}^{(-1)}(n_1)]},$$

$$D_{2y} = -\frac{iX_2}{[\varepsilon_{22}^{(-1)}(n_2) - i\varepsilon_{21}^{(-1)}(n_2)]},$$

$$D_{3y} = -\frac{iX_3}{[\varepsilon_{22}^{(-1)}(n_3) - i\varepsilon_{21}^{(-1)}(n_3)]},$$

$$D_{4y} = -\frac{iX_4}{[\varepsilon_{22}^{(-1)}(n_4) + i\varepsilon_{21}^{(-1)}(n_4)]},$$

where

$$X_1 = -\frac{(1+r_1)r_4E_{0x}\exp\left[i\frac{\omega}{c}(n_4-1)L\right]}{2\left[1-r_1r_4\exp\left[i\frac{\omega}{c}(n_1+n_4)L\right]\right]},$$

$$X_2 = \frac{(1+r_2)r_3E_{0x}\exp\left[i\frac{\omega}{c}(n_3-1)L\right]}{2\left[1-r_2r_3\exp\left[i\frac{\omega}{c}(n_2+n_3)L\right]\right]},$$

$$X_3 = -\frac{(1+r_3)E_{0x}\exp\left[i\frac{\omega}{c}(n_3-1)L\right]}{2\left[1-r_2r_3\exp\left[i\frac{\omega}{c}(n_2+n_3)L\right]\right]},$$

$$X_4 = \frac{(1+r_4)E_{0x}\exp\left[i\frac{\omega}{c}(n_4-1)L\right]}{2\left[1-r_1r_4\exp\left[i\frac{\omega}{c}(n_1+n_4)L\right]\right]}.$$

5. ANALYSIS OF THE RESULTS

It follows from formulas (18) and (19) that the reflected and transmitted waves are elliptically polarized; the major axes of these ellipses prove to be rotated through different angles with respect to the polarization plane of the incident electromagnetic wave.

Since the refractive indices (15) and (16) of the medium for all four normal waves in the plane-parallel layer are different, an isotropic medium placed in an external magnetic field will be nonreciprocal in phase to a certain degree of accuracy. This effect should most clearly manifest itself in experiments with a ring laser.

In our view, one prospective method for determining the optical nonreciprocity in material media consists in measuring the difference between the phase velocities of electromagnetic wave propagating in opposite directions in this medium.

As is known [10], the relative measurement accuracy of the absolute value of the velocity of light is small, $\delta c/c \sim 10^{-9}$ at present. However, in searching for the optical nonreciprocity in material media, one should measure the difference between the velocities of counter propagating waves, rather than the absolute velocity of light, with relative accuracy on the order of 10^{-15}–10^{-16}.

For this purpose, it is convenient to use a ring laser. The ring laser [10, 11] represents a triangular, rectangular, or other closed-loop high-Q optical resonator a part of whose loop contains an active medium that provides generation of electromagnetic waves propagating along the loop in opposite directions.

When the identity of conditions for counter propagating electromagnetic waves is violated (for example, due to the Fizeau effect caused by a gas pumped through a part of the loop of the ring laser), the generation frequencies of counter propagating electromagnetic waves become different. Therefore, if one inserts a plane-parallel layer of the substance under consideration into the path of the waves in the ring laser and applies a magnetic field parallel to the propagation direction of the waves to this layer, then the conditions for the waves that propagate in this part of the loop in different directions become different; as a result, the frequencies of these waves will be different. Since the modern state of the art in experimental technology allows one to measure frequency differences of about $\delta\nu \sim 10^{-3}$ Hz in ring lasers [12], it is clear that the technique described is one of the most promising methods for studying the optical nonreciprocity in isotropic media and for measuring the coefficients h_2 and h_3 in expansion (3) of the complex permittivity tensor.

Let us calculate the accuracy to which these coefficients can be measured in experiments with a ring laser.

In a ring laser, the active medium amplifies only those electromagnetic waves whose phase is changed by $\delta\Psi = 2\pi N$, where N is an integer, when these waves make one rotation along the loop.

This condition allows one to determine the generation frequencies of electromagnetic waves that propagate in opposite directions in the ring laser in the experimental scheme suggested.

Denote by P the perimeter of the ring laser, and by L, the thickness of the plate inserted into the path of the waves. Assume that the forward wave ($\xi = 1$) propagates clockwise along the loop of the ring laser.

The frequency ω of the generated electromagnetic wave that propagates clockwise in the ring laser can be determined from the equation

$$\frac{\omega}{c}P + \Psi_+ = 2\pi N_+,$$

where N_+ is an integer and Ψ_+ is the phase increment due to the layer of the substance under investigation in a dc magnetic field on the path of the forward wave.

The frequency $\tilde{\omega}$ of the generated backward wave ($\xi = -1$) can be determined analogously:

$$\frac{\tilde{\omega}}{c}P + \Psi_- = 2\pi N_-.$$

Now, if we extract these waves from the ring laser and direct them to a detector, then we can obtain the following expression for the frequency $\Delta\omega$ of the output sig-

nal, which is equal to the difference of the frequencies of the mixed electromagnetic waves:

$$\Delta\omega = \tilde{\omega} - \omega = \frac{c[2\pi(N_- - N_+) + \Delta\Psi]}{P}. \quad (20)$$

Now, we find the phase difference

$$\Delta\Psi = \Psi_+ - \Psi_-.$$

Since one component of the vector **E** of electromagnetic waves in ring lasers is intentionally suppressed and only the component orthogonal to it is amplified, we will determine the phase difference $\Delta\Psi$ for the component E_x.

According to (18), the component E_{3x} of the wave transmitted through a plane-parallel layer of the substance under investigation in the forward direction can be represented as

$$E_{3x} = \rho_+ e^{i\Psi_+}. \quad (21)$$

Introducing

$$k_1 = \omega/c, \quad n_1 = n_1(\omega), \quad n_2 = n_2(\omega),$$
$$n_3 = n_3(\omega), \quad n_4 = n_4(\omega),$$

we obtain

$$z_1 = 1 + r_1^2 r_4^2 - 2r_1 r_4 \cos[k_1 L(n_1 + n_4)],$$

$$z_2 = 1 + r_2^2 r_3^2 - 2r_2 r_3 \cos[k_1 L(n_2 + n_3)],$$

$$\rho_+^2 = \frac{(1 - r_1 r_4)^2}{z_1} + \frac{(1 - r_2 r_3)^2}{z_2}$$

$$+ \frac{2(1 - r_1 r_4)(1 - r_2 r_3)}{z_1 z_2}\{\cos[k_1 L(n_1 - n_2)] \quad (22)$$

$$- r_1 r_4 \cos[k_1 L(n_2 + n_4)]$$

$$- r_2 r_3 \cos[k_1 L(n_1 + n_3)]$$

$$+ r_1 r_2 r_3 r_4 \cos[k_1 L(n_4 - n_3)]\}.$$

Analogously, introducing

$$k_2 = \tilde{\omega}/c, \quad \tilde{n}_1 = n_1(\tilde{\omega}), \quad \tilde{n}_2 = n_2(\tilde{\omega}),$$
$$\tilde{n}_3 = n_3(\tilde{\omega}), \quad \tilde{n}_4 = n_4(\tilde{\omega}),$$

from expression (19) for the counterpropagating wave, we obtain

$$E_{px}^* = \rho_- e^{-i\Psi_-}, \quad (23)$$

where

$$\tilde{z}_1 = 1 + \tilde{r}_1^2 \tilde{r}_4^2 - 2\tilde{r}_1 \tilde{r}_4 \cos[k_2 L(\tilde{n}_1 + \tilde{n}_4)],$$

$$\tilde{z}_2 = 1 + \tilde{r}_2^2 \tilde{r}_3^2 - 2\tilde{r}_2 \tilde{r}_3 \cos[k_2 L(\tilde{n}_2 + \tilde{n}_3)],$$

$$\rho_-^2 = \frac{(1 - \tilde{r}_1 \tilde{r}_4)^2}{\tilde{z}_1} + \frac{(1 - \tilde{r}_2 \tilde{r}_3)^2}{\tilde{z}_2}$$

$$+ \frac{2(1 - \tilde{r}_1 \tilde{r}_4)(1 - \tilde{r}_2 \tilde{r}_3)}{\tilde{z}_1 \tilde{z}_2}\{\cos[k_2 L(\tilde{n}_4 - \tilde{n}_3)] \quad (24)$$

$$- \tilde{r}_1 \tilde{r}_4 \cos[k_2 L(\tilde{n}_1 + \tilde{n}_3)]$$

$$- \tilde{r}_2 \tilde{r}_3 \cos[k_2 L(\tilde{n}_2 + \tilde{n}_4)]$$

$$+ \tilde{r}_1 \tilde{r}_2 \tilde{r}_3 \tilde{r}_4 \cos[k_2 L(\tilde{n}_1 - \tilde{n}_2)]\}.$$

Formulas (21) and (23) imply

$$\sin\Delta\Psi = \mathrm{Im}\left(\frac{E_{3x} E_{px}^*}{\rho_+ \rho_-}\right).$$

Using (22) and (24), we obtain

$$\sin\Delta\Psi = \frac{(1 - r_1 r_4)(1 - \tilde{r}_1 \tilde{r}_4)}{z_1 \tilde{z}_1 \rho_+ \rho_-}$$

$$\times \{\sin[(k_1 n_1 - k_2 \tilde{n}_4 + k_2 - k_1)L]$$

$$+ r_1 r_4 \sin[(k_1 n_4 + k_2 \tilde{n}_4 + k_1 - k_2)L]$$

$$- \tilde{r}_1 \tilde{r}_4 \sin[(k_1 n_1 + k_2 \tilde{n}_1 + k_2 - k_1)L]$$

$$+ r_1 r_4 \tilde{r}_1 \tilde{r}_4 \sin[(k_2 \tilde{n}_1 - k_1 n_4 + k_2 - k_1)L]\}$$

$$+ \frac{(1 - r_2 r_3)(1 - \tilde{r}_2 \tilde{r}_3)}{z_2 \tilde{z}_2 \rho_+ \rho_-}$$

$$\times \{\sin[(k_1 n_2 - k_2 \tilde{n}_3 + k_2 - k_1)L]$$

$$+ r_2 r_3 \sin[(k_1 n_3 + k_2 \tilde{n}_3 + k_1 - k_2)L]$$

$$- \tilde{r}_2 \tilde{r}_3 \sin[(k_1 n_2 + k_2 \tilde{n}_2 + k_2 - k_1)L]$$

$$+ r_2 r_3 \tilde{r}_2 \tilde{r}_3 \sin[(k_2 \tilde{n}_2 - k_1 n_3 + k_2 - k_1)L]\}$$

$$+ \frac{(1 - r_2 r_3)(1 - \tilde{r}_1 \tilde{r}_4)}{\tilde{z}_1 z_2 \rho_+ \rho_-} \quad (25)$$

$$\times \{\sin[(k_1 n_2 - k_2 \tilde{n}_4 + k_2 - k_1)L]$$

$$+ r_2 r_3 \sin[(k_1 n_3 + k_2 \tilde{n}_4 + k_1 - k_2)L]$$

$$- \tilde{r}_1 \tilde{r}_4 \sin[(k_1 n_2 + k_2 \tilde{n}_1 + k_2 - k_1)L]$$

$$+ r_2 r_3 \tilde{r}_1 \tilde{r}_4 \sin[(k_2 \tilde{n}_1 - k_1 n_3 + k_2 - k_1)L]\}$$

$$+ \frac{(1 - r_1 r_4)(1 - \tilde{r}_2 \tilde{r}_3)}{z_1 \tilde{z}_2 \rho_+ \rho_-}$$

$$\times \{\sin[(k_1 n_1 - k_2 \tilde{n}_3 + k_2 - k_1)L]$$

$$+ r_1 r_4 \sin[(k_1 n_4 + k_2 \tilde{n}_3 + k_1 - k_2)L]$$

$$- \tilde{r}_2 \tilde{r}_3 \sin[(k_1 n_1 + k_2 \tilde{n}_2 + k_2 - k_1)L]$$

$$+ r_1 r_4 \tilde{r}_2 \tilde{r}_3 \sin[(k_2 \tilde{n}_2 - k_1 n_4 + k_2 - k_1)L]\}.$$

Depending on whether the coefficient f_1 is equal to or different from zero, this formula yields somewhat different results.

When the isotropic medium is gyrotropic in the first order ($f_1 \neq 0$), even a plane-parallel plate of thickness $L = 0.1$ cm can rotate the vector **E** of the transmitted wave by an angle of several radians. Therefore, even such a thin layer of substance makes it possible to rotate the vector **E** by an angle of 90° and thus may suppress the E_x component in the transmitted wave. To avoid such a situation, we will assume that the thickness L of the layer of the substance under investigation is chosen so that

$$\frac{\omega L f_1}{c} = 2m\pi,$$

where m is an integer.

Expanding expression (25) under these assumptions and retaining only the asymptotically leading terms, we obtain

$$\Delta\Psi = \frac{(n_0^4 - 1)h_2 H_z \sin^2(\omega L n_0/c)}{2n_0^2[4n_0^2 + (n_0^2 - 1)^2 \sin^2(\omega L n_0/c)]}, \qquad (26)$$

where $n_0 = \sqrt{\varepsilon}$.

This formula describes the optical nonreciprocity in phase, which results from the combined effect of the Faraday rotation and the first-order spatial dispersion. A similar expression for this phenomenon in gyrotropic crystals was obtained in [2].

Another interesting particular case of optical nonreciprocity, which was not considered in [2], occurs in a weakly optically active medium.

Let $f_1 = 0$; i.e., let an isotropic medium be not gyrotropic in the first order in the small parameter δ. Then, substituting (15) and (16) into formula (25), expanding the latter formula in series up to the terms of order δ^3 inclusive, and retaining only the asymptotically leading terms, we obtain

$$\Delta\Psi = \frac{\omega h_3 H_z L}{c}. \qquad (27)$$

In this case, the difference (20) of the frequencies of the generated waves that propagate in opposite directions in a ring laser is expressed as

$$\Delta\omega = \tilde{\omega} - \omega = \frac{2\pi(N_- - N_+)c}{P} + \frac{\omega h_3 H_z L}{P}.$$

Thus, it is the third-order coefficient h_3 in the expansion of the tensor ε_{mn} that guarantees the phase nonreciprocity in an isotropic optically nonactive ($f_1 = 0$) medium. Since h_3 ranges from 10^{-12} to 10^{-13} G^{-1} in many isotropic optically nonactive media, this phenomenon can reliably be measured even for a longitudinal magnetic field of $H \sim 10^2$ G.

Now, let us estimate the minimal values of the coefficients h_2 and h_3 that can be measured by means of a ring laser.

Suppose that experiments are carried out on a KM-type ring laser [12] produced by FGUP NII Polyus, Moscow, with a perimeter of 170 cm and the operating frequency $\nu \sim 10^{14}$ Hz and that the thickness of the substance under investigation is 0.17 cm. Since this commercial ring laser allows one to measure values of $\Delta\nu$ as small as 10^{-3} Hz, it follows from formulas (20), (26), and (27) that, in the single-mode regime ($N_- = N_+$), the coefficients $h_2 H_z$ and $h_3 H_z$ can be measured with an absolute accuracy on the order of 10^{-14}.

Note that we have only demonstrated that, in principle, ring lasers enable one to observe the phase nonreciprocity and measure the coefficient h_2 in optically active media and the coefficient h_3 in optically nonactive media that enter the expansion (3) of the tensor ε_{nm}. To conduct a real experiment, one has to solve a number of technical problems. One of such problems concerns the quenching of electromagnetic waves reflected from a plane-parallel layer, and, first of all, the quenching of the E_x component of the field. This can be done in several ways: either by choosing the thickness L of the substance layer or by placing a plane-parallel layer at a Brewster angle with respect to the incident wave.

Besides, we did not take into account the frequency pedestal in calculations, which is specially introduced to avoid the locking of the frequency of counter propagating waves in a ring laser.

However, the consideration of all technical features of the experiments with ring lasers makes the calculation of this phenomenon more complicated, although does not rule out the possibility of observing weak phase nonreciprocity in an isotropic optically nonactive ($f_1 = 0$) substance in a longitudinal (with respect to the propagation direction of the electromagnetic wave) magnetic field and measuring the coefficient $h_3 H \sim \delta^3$ in expansion (3) of the complex permittivity tensor.

The experimental verification of the results obtained may serve as a basis for designing a magnetic-field-controlled nonreciprocal element for various optical devices.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 04-02-16604.

REFERENCES

1. N. V. Kravtsov and N. N. Kravtsov, Kvantovaya Élektron. (Moscow) **27**, 98 (1999).
2. M. A. Novikov, Kristallografiya **34**, 1354 (1989) [Sov. Phys. Crystallogr. **34**, 816 (1989)].
3. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 8: *Electrodynamics of Continuous Media*, 2nd ed. (Nauka, Moscow, 1982; Pergamon, Oxford, 1984).
4. A. F. Andreev and V. I. Marchenko, Usp. Fiz. Nauk **130**, 39 (1980) [Sov. Phys. Usp. **23**, 21 (1980)].
5. F. I. Fedorov, *Theory of Gyrotropy* (Nauka i Tekhnika, Minsk, 1976) [in Russian].
6. *The Physics of Magnetic Dielectrics*, Ed. by G. A. Smolenskiĭ (Nauka, Leningrad, 1974) [in Russian].
7. G. A. Smolenskiĭ, R. V. Pisarev, and I. G. Siniĭ, Usp. Fiz. Nauk **116**, 231 (1975) [Sov. Phys. Usp. **18**, 410 (1975)].
8. I. P. Denisova and B. V. Mehta, Gen. Relativ. Gravit. **29**, 583 (1997).
9. I. P. Denisova, *Introduction to the Tensor Calculus and Its Applications* (UNTs DO, Moscow, 2003) [in Russian].
10. *Physical Encyclopedia* (Bol'shaya Rossiĭskaya Éntsiklopediya, Moscow, 1994), Vol. 4, p. 548 [in Russian].
11. V. V. Ragul'skiĭ, Usp. Fiz. Nauk **167**, 1022 (1997) [Phys. Usp. **40**, 972 (1997)].
12. V. V. Grishachev, V. I. Denisov, V. G. Zhotikov, *et al.*, Opt. Spektrosk. **98**, 51 (2005) [Opt. Spectrosc. **98**, 47 (2005)].

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 212–223.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 243–256.

ATOMS, MOLECULES, OPTICS

Kinetic Description of the Coulomb Explosion of a Spherically Symmetric Cluster

V. F. Kovalev[a] and V. Yu. Bychenkov[b]

[a]*Institute for Mathematical Modeling, Russian Academy of Sciences, Moscow, 125047 Russia*
[b]*Lebedev Physical Institute, Russian Academy of Sciences, Moscow, 119991 Russia*
e-mail: kovalev@imamod.ru

Received January 27, 2005

Abstract—The particle distribution function is calculated for the Coulomb explosion of a spherically symmetric charged cluster formed through the interaction of intense ultrashort laser pulses with a cluster gas. The particle density and mean velocity distributions as well as the energy spectra of the accelerated particles are obtained. These characteristics are analyzed in detail for a cold cluster plasma, where the kinetic effects determine the physics of multiple flows emerging after the turnover of the cluster particle velocity profile. We find the boundaries of the multiple-flow regions and study the characteristics of an exploding cluster as a function of its initial density profile. The energy spectra of the accelerated ions are obtained for a cluster plasma with a specified cluster size distribution. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

The explosion dynamics of a charged cluster is of interest in connection with the discussed practical applications of the interaction effects between intense ultrashort laser pulses and a cluster gas [1, 2] or solid nanoobjects [3]. It is well known that the explosion of clusters accelerates ions to high energies [4–6], and, for example, in the case of clusters of deuterons, their collisions provide thermonuclear neutrons [7], which opens up prospects for producing subnanosecond neutron sources to be used in material science, nondestructive testing, etc.

A relatively simple model for describing the explosion dynamics of a cluster can be formulated when it is exposed to such an intense laser field that the electron oscillation amplitude, $\max\{1, a\}c/\omega_0$, is much larger than the characteristic cluster radius, r_c, and the kinetic energy of the electrons, $mc^2a^2/2$, is much higher than the energy of their Coulomb interaction with the cluster, ZNe^2/r_c. In this case, the laser is intense enough to detach the electrons from the field-ionized cluster almost instantaneously, and the subsequent Coulomb explosion will proceed as the explosion of a charged cluster into a vacuum if the pressure of the surrounding electron gas, $n_emc^2a^2/2$, is much lower than the characteristic pressure of the exploding cluster, e^2Z^2Nn/r_c. Here, m and e are the electron mass and charge, c is the speed of light, $a = eE_0/m\omega_0c$ is the dimensionless vector potential of the electromagnetic field with electric field strength E_0 and frequency ω_0, Z is the ion charge number, N is the number of atoms in the cluster, n_e is the electron density in the intercluster space, and n is the atomic density in the cluster. Such conditions are realized for fairly small clusters with a relatively small number of atoms in the cluster when exposed to laser radiation of relativistic intensity [5, 7–10]. For example, for a deuterium cluster with a mean atomic density in the cluster $n \sim 3 \times 10^{22}$ cm^{-3}, this corresponds to the laser fluxes $\gtrsim 10^{19}$ W cm^{-2} at a wavelength of 1 μm and $r_c \lesssim 100$ Å.

For larger clusters or at moderate laser intensities, the electrons produced through photoionization remain within the cluster [8]. In this situation, the absorption of laser radiation energy and the electron heating are determined by the collision processes and plasma effects. In an electron–ion plasma cluster in which the bulk of the energy is concentrated in hot electrons, the ion acceleration during its explosion is attributable precisely to the electron component, which gives rise to a charge-separating field. Under this field, the ions follow the expanding electron halo with the formation of a quasi-neutral expanding plasma on long time scales. This scenario was discussed starting from [11] in both the hydrodynamic [12] and kinetic [13] models of spherical plasma expansion. In this paper, we do not consider this ion acceleration scenario, but study the Coulomb explosion of a cluster by which we mean both the cluster proper and any spherically symmetric nanostructure.

During the Coulomb explosion of a cluster, the effective acceleration of its ions is attributable to a strong spatial charge separation [10, 14, 15]. This regime was investigated using numerical simulations [16–19]. However, the effects that arise during Coulomb explosions cannot be fully understood without developing the corresponding theoretical models, which are clearly lacking. The simplest approach corre-

sponds to the model of an ideal cluster, an exploding homogeneous spherical bunch of a given radius [6, 16, 20]. This model allows the ion velocity, density, electric field, and energy distribution of the accelerated ions to be easily calculated at an arbitrary time. The ion energy and charge spectra as a function of the cluster distribution in radius and the parameters of the irradiating cluster ultrashort laser pulse beam were analyzed in [21] in terms of the model of an ideal cluster. However, as was shown in [14, 15], for an initially nonuniform ion density profile in a spherically symmetric cluster, in which the density decreases monotonically to zero with increasing distance from the center, the solution constructed using the hydrodynamic approximation exists only for a certain limited time, $t < t_s$. According to [14, 15], a singularity that corresponds to the ion density becoming infinite arises in the hydrodynamic solution at $t = t_s$, and a multiple flow emerges on longer time scales. This suggests that the well-known bounded solution for an ideal cluster is actually unstable due to the large, but finite density gradient at the cluster–vacuum boundary. Our results prove that this assumption is valid.

A consistent allowance for the spatial inhomogeneity of a charged cluster is the principal element of our theory that determines the formation of a Coulomb explosion "shock wave" [14, 15, 19], which does not emerge in the popular model of an ideal cluster [6, 16, 20]. Allowance for the nonuniformity of the cluster density profile is necessitated by a number of practically important problems. First of all, note that the currently prepared clusters can be initially spatially inhomogeneous due to artificial layering using coverings of different densities [22]. The cluster ion density distribution can also be significantly nonuniform due to a prepulse. Analysis of currently available experiments shows [23] that the propagation of ultrashort intense laser pulses is always accompanied by a prepulse with a typical duration from tens of picoseconds to nanoseconds. In this case, the maximum contrast of the laser radiation does not exceed 10^8–10^9 even under extreme conditions; at laser fluxes higher than 10^{19} W cm^{-2}, this inevitably results in a smearing of the cluster ion density profile with the formation of a transition region near the outer boundary. Numerical simulations of the explosion dynamics of a deuterium cluster using the particle-in-cell code also suggest the formation of a transition region near the cluster boundary [19]; as a result, the formation of a singularity in the ion density distribution attributable to the density nonuniformity is observed. The discussed [14] and realized [3] interaction of short laser pulses with nanoobjects gives another example of the possible practical application of target inhomogeneities, especially in connection with the well-developed nanosphere [24] and nanotube [25] production technology.

The analytical description of a spherically symmetric cluster explosion [14, 15] is consistent with the general proposition of [26] that the hydrodynamic problem of a Coulomb explosion (without including the thermal motion of the cluster particles) that uses the equations of motion and continuity for the cluster ions and the Poisson equation has an exact analytical solution. At the same time, the hydrodynamic description of a Coulomb explosion beyond the singularity, $t > t_s$, is not possible, and the kinetic approach realized in this paper should be used.

The emergence of a multiple flow in the explosion dynamics of a charged cluster is similar to the turnover of the nonuniform velocity profile of disturbances propagating in a gas of noninteracting particles [27, 28]. Similar phenomena were also discussed in astrophysical applications [29, 30]. The particle flows emerge and merge on singular surfaces, caustics. In the dynamics of a cold self-gravitating gas, the caustics multiply and crowd over the course of time, which is accompanied by an increase in the number of regions of multiple flows, and a structure that was called in [30] a nondissipative gravitational singularity is formed. A characteristic feature of this structure is the formation of an infinitely deep potential well after the emergence of the singularity followed by the capture of particles flows in this well. The description of this process using the adiabatic capture theory [29] is also confirmed by numerical simulations [31].

When passing from gravitational problems to the problem of the expansion of a charged plasma, replacing the forces of mutual attraction between particles with the forces of repulsion leads to qualitatively different particle dynamics beyond the turnover point. Thus, for example, in a plane geometry, no singularity emerges in the hydrodynamic approximation when the gas of charged particles expands with a zero initial velocity. For a spherically symmetric explosion of a charged cluster, the potential and its derivative at the flow turnover point are finite; i.e., the singularity is kinematic in nature, and the particle flow is not captured. For a fairly smooth, monotonically decreasing initial profile, the number of singularities (caustics) after the turnover does not increase. After the emergence of a singularity, two caustics [14] that separate the regions of triple and double flows exist on the density profile, although the region occupied by the triple flow expands with time.

This paper is devoted to the kinetic description of a spherically symmetric cluster explosion in which the cluster ion distribution function can be found by solving the Vlasov equation in a self-consistent electrostatic field for a given initial ion distribution in radial velocity and radius. We obtain a general solution of the kinetic equation that includes both the transverse thermal motion of the cluster ions and their thermal radial velocity spread. This solution is analyzed in detail in the limit of a negligible thermal cluster ion velocity (a cold cluster). The kinetic effects here manifest themselves in the emergence of regions of multiple flows.

We determine the laws of motion for the boundaries of these regions in the presence of double and triple flows and study the spatial ion density and mean velocity distributions for an exploding cluster and the particle energy spectrum for several typical initial particle distributions in radius. By disregarding the thermal particle motion, we prove that the result of the hydrodynamic theory for the ion spectrum is also valid in the case where the hydrodynamic validity conditions are formally violated; i.e., the emergence of regions of multiple flows at $t > t_s$ does not change the shape of the accelerated ion spectrum.

2. SOLUTION OF THE CAUCHY PROBLEM FOR A COULOMB EXPLOSION: THE KINETIC APPROACH

The dynamics of charged plasma particles (ions) in a cluster is described by the Vlasov kinetic equation for the ion distribution function f and by the Poisson equation for the self-consistent electric field $\mathbf{E}$:

$$\begin{aligned} f_t + \mathbf{v} f_{\mathbf{r}} + (Ze/M)\mathbf{E} f_{\mathbf{v}} &= 0, \\ \operatorname{div}\mathbf{E} &= 4\pi Ze\int d\mathbf{v} f, \end{aligned} \tag{1}$$

where M is the ion mass. When the sign in front of the "field" term on the left-hand side of the Poisson equation is changed, the system of equations (1) transforms to the dynamic equations for a gas of neutral gravitating particles (see, e.g., [26, Section 81] and [29]). For this reason, there is a close relationship between the dynamics of a self-gravitating gas and the explosion dynamics of a charged cluster.

The kinetic equation (1) admits of spherically symmetric solutions (a spherical cluster). In this case, the ion distribution function may be assumed to depend only on the radius r, the radial velocity v, and the square of the velocity component orthogonal to the radius vector at a given point, $f = f(t, r, v, \mathbf{v}_\perp^2)$. For definiteness, we assume that the distribution function f is factorized as follows:

$$f = f^r(t, r, v) f^\perp(\mathbf{v}_\perp^2);$$

i.e., the transverse velocity distribution is assumed to be stationary and uniform, for example, Maxwellian, with temperature $T_\perp$. The distribution function integrated over the transverse velocities, $F(t, r, v) = \int f\, d\mathbf{v}_\perp$, is then defined by the system of equations

$$\begin{aligned} F_t + vF_r + \frac{2v}{r}F + \frac{1}{M}\left(ZeE + \frac{2T_\perp}{r}\right)F_v &= 0, \\ (r^2E)_r - 4\pi Zer^2\int_{-\infty}^{\infty} dvF &= 0, \end{aligned} \tag{2}$$

where

$$T_\perp = \frac{\int f^\perp d\mathbf{v}_\perp \mathbf{v}_\perp^2}{2\int f^\perp d\mathbf{v}_\perp}$$

is the transverse cluster ion temperature.

The initial conditions for Eqs. (2) correspond to specifying the ion coordinate and velocity distribution function, $f_0(v, r)$, and the radial electric field that satisfies the Poisson equation (the second equation in (2)) at $t = 0$:

$$\begin{aligned} F|_{t=0} &= f_0(v, r), \\ E|_{t=0} &= \frac{4\pi Ze}{r^2}\int_0^r r'^2 dr' \int_{-\infty}^{\infty} dv f_0(v, r'); \end{aligned} \tag{3}$$

since the problem is symmetric and the charge is localized, the latter becomes zero at $r = 0$ and decreases at infinity as $r \longrightarrow \infty$. For uniform initial velocity distributions, the function f_0 is represented by the product $f_0(v, r) = f_c(v)n_c(r)$. In particular, for the Maxwellian initial ion velocity distribution, this corresponds to a spatially uniform initial temperature (T = const):

$$f_c(v) = \sqrt{\frac{M}{2\pi T}}\exp\left(-\frac{Mv^2}{2T}\right).$$

When the Lagrangian variables (velocity and coordinates) are used, the solution of the Cauchy problem (2) and (3) can be represented as

$$\begin{aligned} F(t, r, v) = r^{-2}\int_{-\infty}^{\infty} dv f_c(v)\int_0^{\infty} dh h^2 n_c(h) \\ \times \delta(r - R(t, h, v))\delta(v - U(t, h, v)), \end{aligned} \tag{4}$$

where the functions R and U are the solutions of the following initial-value problem:

$$R_t = U, \quad U_t = \frac{2v_{T\perp}^2}{R} + \frac{w}{R^2}, \quad R|_{t=0} = h,$$

$$U|_{t=0} = v, \quad w(h) = \int_0^h dr r^2 \omega_L^2(r)\int_{-\infty}^{\infty} dv f_c(v), \tag{5}$$

$$v_{T\perp} = \sqrt{\frac{T_\perp}{M}}, \quad \omega_L = \sqrt{\frac{4\pi Z^2 e^2 n_c}{M}}.$$

We write the solution of the latter separately for posi-

tive and negative initial velocities ν:

$$\nu \geq 0,$$

$$U = \sqrt{\frac{2w}{h}}\left[\frac{h}{2w}\nu^2 + 1 - \frac{h}{R} - \frac{2v_{T\perp}^2 h}{w}\ln\frac{h}{R}\right]^{1/2},$$

$$(t+t_0)\sqrt{\frac{2w}{h^3}} = \int_{h/R}^{h/R_0}\frac{dp}{p^2}\left(1 + \frac{\nu^2 h}{2w} - p - \frac{2v_{T\perp}^2 h}{w}\ln p\right)^{-1/2},$$

$$\nu < 0, \qquad (6)$$

$$U = \operatorname{sgn}(t-t_0)\sqrt{\frac{2w}{h}} \times\left[\frac{h}{2w}\nu^2 + 1 - \frac{h}{R} - \frac{2v_{T\perp}^2 h}{w}\ln\frac{h}{R}\right]^{1/2},$$

$$|t_0 - t|\sqrt{\frac{2w}{h^3}} = \int_{h/R}^{h/R_0}\frac{dp}{p^2}\left(1 + \frac{\nu^2 h}{2w} - p - \frac{2v_{T\perp}^2 h}{w}\ln p\right)^{-1/2}.$$

Here, the functions t_0 and R_0 are defined by

$$t_0 = \sqrt{\frac{h^3}{2w}} \times\int_1^{h/R_0}\frac{dp}{p^2\sqrt{1 + \nu^2 h/2w - p - (2v_{T\perp}^2 h/w)\ln p}}, \qquad (7)$$

$$1 + \frac{\nu^2 h}{2w} - \frac{h}{R_0} - \frac{2v_{T\perp}^2 h}{w}\ln\frac{h}{R_0} = 0.$$

Accordingly, the electric field distribution is given by

$$E(t,r) = r^{-2}\int_{-\infty}^{\infty} d\nu\, f_c(\nu) \times\int_0^{\infty} dh h^2 n_c(h)\theta(r - R(t,h,\nu)). \qquad (8)$$

Dimensionless variables are used in Eqs. (6)–(8) and below: we normalize the time to the inverse ion Langmuir frequency $\omega_L(0)$ calculated from the initial density at the cluster center $n_c(0)$, the coordinates r, h, and R to r_c, $\omega_L(r)$ to $\omega_L(0)$, the densities n and n_c to $n_c(0)$, the velocities v, ν, $v_{T\perp}$, and U to $\omega_L(0)r_c$, the variable w to $\omega_L^2(0)r_c^3$, the distribution functions F and f to $n_c(0)/\omega_L(0)r_c$ and $1/\omega_L(0)r_c$, respectively, and the electric field E to $4\pi Zen_c(0)r_c$. We associate r_c with the characteristic initial cluster radius determined by the density profile. The presence of a unit step function (the θ function) in Eq. (8) for the field indicates that the contribution only from the particle for which the condition $R(t, h, \nu) \leq r$ is satisfied should be taken into account when calculating this field.

At a zero transverse cluster ion temperature ($T_\perp = 0$), the integrals in Eqs. (6) can be calculated in explicit form to yield the algebraic relations

$$\nu \geq 0,$$

$$(t+t_0)\sqrt{\frac{2w}{R_0^3}} = \frac{\sqrt{1 - R_0/R}}{R_0/R} + \frac{1}{2}\ln\frac{1 + \sqrt{1 - R_0/R}}{1 - \sqrt{1 - R_0/R}},$$

$$U = \sqrt{\frac{2w}{h}\left(\frac{h}{R_0} - \frac{h}{R}\right)}, \quad T_\perp = 0, \qquad (9)$$

$$\nu < 0,$$

$$|t_0 - t|\sqrt{\frac{2w}{R_0^3}} = \frac{\sqrt{1 - R_0/R}}{R_0/R} + \frac{1}{2}\ln\frac{1 + \sqrt{1 - R_0/R}}{1 - \sqrt{1 - R_0/R}},$$

$$U = \operatorname{sgn}(t-t_0)\sqrt{\frac{2w}{h}\left(\frac{h}{R_0} - \frac{h}{R}\right)}, \quad T_\perp = 0.$$

The functions t_0 and R_0 at $T_\perp = 0$ are defined by

$$t_0 = \sqrt{\frac{R_0^3}{2w(h)}}\left[\frac{\sqrt{1 - R_0/h}}{R_0/h} + \frac{1}{2}\ln\frac{1 + \sqrt{1 - R_0/h}}{1 - \sqrt{1 - R_0/h}}\right], \qquad (10)$$

$$R_0 = \frac{h}{1 + (h/2w)\nu^2}.$$

Similar solutions can be obtained in one-dimensional and two-dimensional geometries, i.e., for the plane (the

explosion of a charged plasma layer) and cylindrical (the explosion of a charged plasma cylinder) cases.

Below, we also give the formulas that describe the explosion of an ion cluster at a zero radial cluster ion temperature when $T = 0$ and $f_c(v) = \delta(v)$. The suggested kinetic approach then corresponds to the model of multiple-flow hydrodynamics [31] arising here from the first principles. In this case, as follows from (7), we have $R_0 = h$, $t_0 = 0$, and Eqs. (6) take the form

$$U = \sqrt{\frac{2w}{h}}\left(1 - \frac{h}{R} - \frac{2v_{T\perp}^2 h}{w}\ln\frac{h}{R}\right)^{1/2},$$
$$t\sqrt{\frac{2w}{h^3}} = \int_{h/R}^{1} \frac{dp}{p^2\sqrt{1 - p - (2v_{T\perp}^2 h/w)\ln p}}. \quad (11)$$

To conclude this section, we give the relations that define such global characteristics of the cluster explosion into a vacuum as the particle density and mean velocity as well as the ion spectral energy distribution. As follows from many experiments, the latter is the main characteristic for describing the Coulomb explosions of clusters. Whether the model used to calculate the energy of the accelerated ions is adequate can be judged from the dependence of the shape of the ion spectrum on macroscopic cluster ion characteristics (such as the thermal ion velocity, the spatial cluster ion density profile, the cluster size distribution, etc.). A significant dependence of the spectrum of the accelerated particles in an exploding cluster on the cluster size distribution was pointed out in several papers [20, 21].

The cluster ion density $n(t, r)$ and mean radial velocity $u(t, r)$ are defined in a standard way via the zeroth and first moments of the particle distribution function $F(t, r, v)$, while the ion energy distribution function (ion spectrum),

$$\frac{dN}{d\varepsilon} = 4\pi r^2 n(t, r)/\frac{d\varepsilon}{dr},$$

is introduced in such a way that, being integrated over all energies, $\varepsilon = Mv^2/2$, it yields the total number of particles in the cluster. Assuming that the functions $R(t, h, \nu)$ and $U(t, h, \nu)$ are uniquely defined at any t for all h and ν, let us write the following relations for $n(t, r)$, $u(t, r)$, and $dN/d\varepsilon$ using solution (4):

$$n(t, r) = \int_{-\infty}^{\infty} d\nu\, f_c(\nu) \times \sum_s \frac{h_s^2 n_c(h_s)/R^2(t, h_s, \nu)}{|\partial R(t, h, \nu)/\partial h|_{h = h_s}},$$

$$u(t, r) = \frac{1}{n(t, r)}\int_{-\infty}^{\infty} d\nu\, f_c(\nu) \times \sum_s U(t, h_s, \nu)\frac{h_s^2 n_c(h_s)/R^2(t, h_s, \nu)}{|\partial R(t, h, \nu)/\partial h|_{h = h_s}}, \quad (12)$$

$$\frac{dN}{d\varepsilon} = \frac{2\pi}{\sqrt{\varepsilon}}\int_{-\infty}^{\infty} d\nu\, f_c(\nu) \times \sum_l \frac{h_l^2 n_c(h_l)}{|\partial U(t, h, \nu)/\partial h|_{h = h_l,\, U = \pm\sqrt{\varepsilon}}}.$$

Here, we normalize N to $n_c(0)r_c^3$ and ε to $M\omega_L^2(0)r_c^2/2$. In the formulas for the density and the mean velocity, the summation is over all possible $s = 1, 2, \ldots, S$ roots of the equation $r = R(t, h_s, \nu)$ defining the function $h_s(t, r, \nu)$. Similarly, in the formula for the ion spectrum, the summation is over all possible $l = 1, 2, \ldots, L$ roots of the equations $U(t, h_l, \nu) = \pm\sqrt{\varepsilon}$.

The formulas presented in this section give a complete kinetic description of the Coulomb explosion of a plasma cluster including the thermal motion of its particles (ions). However, even when the thermal motion of the cluster particles is disregarded, the analytical results obtained are of fundamental importance, since they allow the characteristics of the cluster explosion to be described after the emergence of a singularity at the hydrodynamic explosion stage, when the hydrodynamic theory is inapplicable [15]. We discuss this problem in the next section.

3. THE KINETIC DESCRIPTION OF THE EXPLOSION OF A COLD CLUSTER

The choice of an initial distribution function in the form $f_c = \delta(\nu)$ and $T_\perp = 0$ corresponds to the hydrodynamic limit of a cold plasma, which is characterized by the emergence of a singularity in the spatial density distribution at a certain time t_s. The cold-ion approximation yields a result that is identical to that obtained previously using the hydrodynamic equations [15] before the emergence of a singularity. However, it has the undeniable advantage that it allows one to extend this solution beyond the singularity, i.e., to describe the multiple-flow regime of cluster explosion. There is no analytical description of this regime for a Coulomb explosion as yet. The extension of the "cold" solution [15] to times $t > t_s$ arises as a natural calculation of the integral with the δ function in (4) and requires no ad hoc assumptions [14]. Indeed, by performing integration over the coordinate h and the velocity ν, we can represent the cold solution (4) in the following form (conve-

nient for calculating the moments of the velocity distribution function):

$$F(t, r, v) = \sum_{k=1}^{K} \frac{n_c(h_k)}{|\partial R(t, h, 0)/\partial h|_{h=h_k}} \times (1 - q^2(t, h_k))^2 \delta(v - U(t, h_k, 0)),$$
$$R(t, h_k, 0) = \frac{h_k}{1 - q^2}, \quad U(t, h_k, 0) = \sqrt{\frac{2w}{h_k}} q, \tag{13}$$
$$t\sqrt{\frac{2w(h_k)}{h_k^3}} = \frac{q}{1 - q^2} + \frac{1}{2}\ln\frac{1+q}{1-q}.$$

Here, the summation is over all possible $k = 1, 2, \ldots, K$ roots of the equation

$$r = \frac{h_k}{1 - q^2(h_k, t)}$$

that defines the function $h_k(t, r)$. When the dependence of r on h is single-valued, i.e., $K = 1$, we obtain a hydrodynamic solution [15] that is valid only until $t = t_s$. The value of t_s for a smooth initial density profile whose curvature changes sign with increasing r can be found from the system of equations that consists of the last equation in (13) and the following equations (see [15]):

$$1 - \frac{3t_s}{2} q(t_s, h)(1 - q^2(t_s, h))\sqrt{\frac{2w(h)}{h^3}} \times \left(1 - \frac{h^3\omega_L^2(h)}{3w(h)}\right) = 0,$$
$$3\left(1 - \frac{h^3\omega_L^2(h)}{3w(h)}\right) - q^2\left[\frac{h^3\omega_L^2(h)}{w(h)} + \frac{h^6\omega_L^4(h)}{w^2(h)} - 4\frac{h^3\omega_L^2(h)}{w(h)} - 2\frac{h^4\partial_h(\omega_L^2(h))}{w(h)}\right] = 0. \tag{14}$$

Beyond the singularity, $t > t_s$, the dependence of r on h is multivalued, $K > 1$, which necessitates including the contributions from various branches of $h_k(t, r)$ to the distribution function F at a given point r.

The phenomenon of multiple flow at $t > t_s$ also manifests itself in the integral characteristics of the distribution function. Thus, for example, the cluster ion density n and mean velocity u beyond the turnover point can be calculated using the formulas

$$n(t, r) = \sum_k \frac{n_c(h_k)}{|\Delta(t, h_k)|}(1 - q^2(t, h_k))^3,$$
$$u(t, r) = \frac{1}{n(t, r)}\sum_k U(t, h_k)\frac{n_c(h_k)}{|\Delta(t, h_k)|} \times (1 - q^2(t, h_k))^3, \tag{15}$$
$$\Delta(t, h) = 1 - \frac{3t}{2} q(1 - q^2)\sqrt{\frac{2w(h)}{h^3}} \times \left(1 - \frac{h^3\omega_L^2(h)}{3w(h)}\right).$$

For definiteness, let us discuss the explosion of a cluster with an initial density profile $n_c(r)$ that monotonically decreases to zero with increasing radius.

Since the density distribution (15) decreases monotonically at $t = 0$, it flattens out with time in the inner regions of the cluster, while the density increases at its periphery. This stems from the fact that the initially nonuniform electric field with its maximum inside the cluster accelerates the cluster ions differently: the ions at larger radii move more slowly than those between the center and the region of field maximum. Catching up with the peripheral ions, the latter at $t = t_s$ give rise to a density singularity called a Coulomb explosion shock wave in [14] (though this singularity differs from the shock waves observed in a cluster plasma [32]). This corresponds to the turnover of the mean velocity profile in the hydrodynamic model [14, 15] and to the intersection of particle trajectories. The kinetic approach naturally describes the passage of various particle groups through one another, which encounters difficulties with the description in the hydrodynamic model, whose extension beyond the point $t > t_s$ requires at least a modification of the single-flow model [29]. At $t > t_s$, the formed singularity breaks up into two singularities (called leading and trailing shocks in [14]), which can be associated with the groups of central and peripheral ions, respectively, described above.

In this paper, we will adhere to the terminology traditional for the physics of gravitating systems,[1] where the singularities of this type are called caustics (see, e.g., [31, 34]). For a monotonically decreasing density profile with an inflection point at $t > t_s$, there exist two caustic surfaces, the singularities r_{c1} and r_{c2}, that sepa-

[1] It is also well known that discontinuous time functions can be used in the solutions of the dynamic equations for a one-dimensional gravitating medium, for example, to describe the sticking of particles [33].

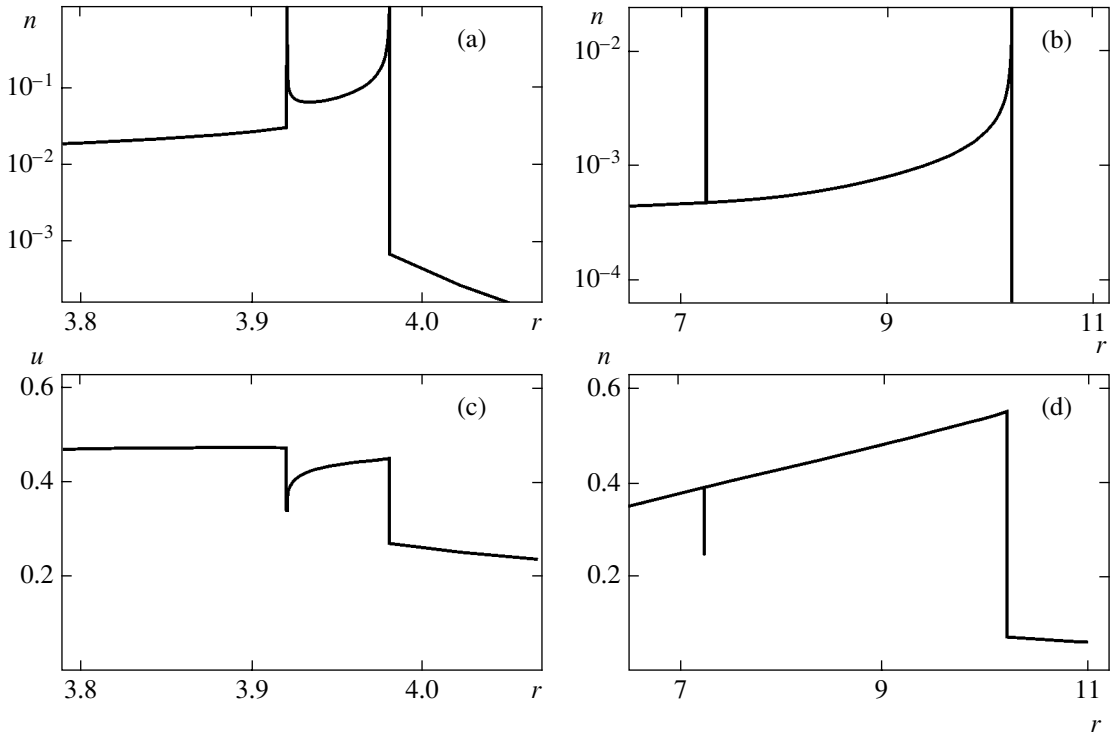

Fig. 1. Spatial distributions of density n and mean velocity u for a cluster with an exponential initial density profile at times $\omega_{L0}t$ = 8 (a) and (20) (b).

rate the regions of existence of single ($0 \le r \le r_{c1}$, $r > r_{c2}$) and triple ($r_{c1} < r < r_{c2}$) flows. The positions of the caustic surfaces in space at an arbitrary time, $t > t_s$, are defined by a pair of equations:

$$1 - \frac{3t}{2}q(1-q^2)\sqrt{\frac{2w(h)}{h^3}}\left(1 - \frac{h^3\omega_L^2(h)}{3w(h)}\right) = 0,$$
$$t\sqrt{\frac{2w(h)}{h^3}} = \frac{q}{1-q^2} + \frac{1}{2}\ln\frac{1+q}{1-q}. \quad (16)$$

At long times, $t \gg t_s$, the radii of the caustic surfaces, r_{c1} and r_{c2}, are given by the asymptotic relations

$$\frac{3}{2} - \frac{t^2 w(h_{c1})}{h_{c1}^3} = 0, \quad r_{c1} \approx 1.635 h_{c1}, \quad (17)$$

$$1 - \frac{h_{c2}^3\omega_L^2(h_{c2})}{w(h_{c2})} = -2\delta\left(\ln\frac{2}{\delta} - 2\right),$$
$$r_{c2} \approx t\sqrt{\frac{2w(h_{c2})}{h_{c2}}}\left[1 + \delta\left(\ln\frac{2}{\delta} - 1\right)\right]^{-1},$$
$$\delta = \frac{2}{bt}\left[1 + \frac{1}{bt}(-1 + \ln(b^2t^2))\right], \quad (18)$$
$$b = \sqrt{\frac{32w(h_{c2})}{h_{c2}^3}}.$$

The pair of relations (17) is obtained if we discard the last term in the first equation in (16) at $h \gg 1$. In this case, the combination $t\sqrt{2w/h^3}$ has a stationary (t-independent) value. Relations (18) follow from Eqs. (16) if we assume that $q \longrightarrow 1$ in them, i.e., $1 - q = \delta \ll 1$. The contributions proportional to δ in the first two relations of (18) determine the slow time dependence of h_{c2}. For $t \longrightarrow \infty$, h_{c2} tends to the stationary value of h_{cs} that is specified by the equation $\omega_L^2(h_{cs}) = w(h_{cs})/h_{cs}^3$ derived from (18) at $\delta = 0$.

It follows from Eqs. (17) and (18) that at fairly long t, h_{c1} and, hence, r_{c1} increase as $t^{2/3}$, while r_{c2} increases as t. This implies that the separation between the caustics increases as t, i.e., linearly with time. Thus, the asymptotically outer caustic virtually coincides with the current radius of the exploding cluster (the front of accelerated ions), whose size increases linearly with time.

As an illustration, Fig. 1 shows the density and mean velocity distributions (15) after the turnover, $r > t_s \approx 7.1974$, for a cluster with an exponentially decreasing density profile, $n_c(r) = (4/3\sqrt{\pi})\exp(-r^2)$, for t = 8 and 20. We clearly see that after the formation of a density singularity, two infinite peaks exist at any time [14]; the separation between them increases with time according to the law of motion of the caustics established above. The transition from a single flow to a triple flow is

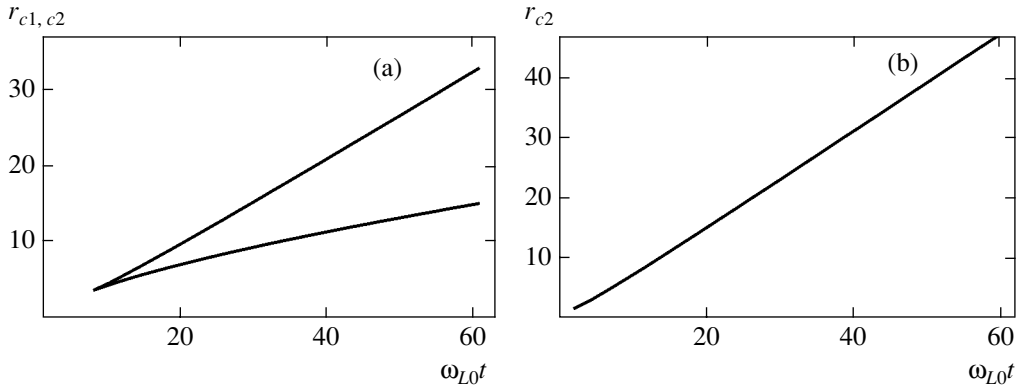

Fig. 2. Time evolution of the caustic surfaces for (a) a cluster with an exponential initial density profile and (b) a cluster with a nearly ideal initial density profile.

accompanied by a discontinuity in density: it changes abruptly. In addition to the density discontinuity, the transition from a single flow to a triple flow is also accompanied by a discontinuity in mean velocity, but the mean velocity remains finite. Such a pattern was also observed in a gas of noninteracting particles [28], suggesting that the singularity is kinematic in nature. Note that on long time scales, there is a significant accumulation of particles near the boundary of the exploding cluster (the upper panel in Fig. 1b) on a shell with a thickness of the order of the initial gradient length (r_c). This may prove to be important for the effects attributable to collisions between clusters, for example, fusion reactions. Figure 2a shows the time evolution of the caustics (the boundaries of the region of multiple flows), r_{c1} (inner) and r_{c2} (outer). With the exception of the times near t_s, the evolution of the caustics is described well by the asymptotic analytical formulas (17) and (18), which essentially yield a result based on the exact formulas (16).

For a cluster in which the density distribution has no inflection point but decreases monotonically to zero at $r = 1$, i.e., $\omega_L^2(1) = 0$, for example, for a cluster with a linear or parabolic density profile, the time at which the solution becomes multivalued is defined by the root of the algebraic equation for $q_s \equiv q(t_s, 1)$:

$$t_s = \frac{2}{3\sqrt{2w(1)}}\frac{1}{q_s(1-q_s^2)},$$
$$\frac{1}{2}\ln\frac{1+q_s}{1-q_s} + \frac{1}{1-q_s^2}\left(q_s - \frac{2}{3q_s}\right) = 0. \quad (19)$$

As was shown by using the hydrodynamic approach [15], the solution of this equation, $q_s \approx 0.6232$, yields the following "universal" formula for the time at which the solution becomes multivalued: $t_s \approx 1.237/\sqrt{w(1)}$. For all such clusters that have the same total number of ions and differ only by the pattern of monotonic decrease in density with increasing radius, the value of $w(1)$ is the same, $w(1) = 1/3$, which yields the following singularity time: $t_s \approx 2.142$. Accordingly, after the singularity time, $t > t_s$, the multiple-flow region (to be more precise, the double-flow region, since only one caustic r_{c2} exists) extends from the radius $r_b(t) \equiv R(t, 1, 0)$, which is implicitly specified by the relations

$$r_b = \frac{1}{1-q^2}, \quad t\sqrt{2w(1)} = \frac{q}{1-q^2} + \frac{1}{2}\ln\frac{1+q}{1-q} \quad (20)$$

to the radius r_{c2}, which coincides with the outer boundary of the cluster.

Equation (20) for $r_b(t)$ at $w(1) = 1/3$ is identical to the equation that describes the front radius $r_f(t)$ for an ideal cluster [15]; i.e., the inner boundary of the region of double flow for a cluster with a density profile monotonically decreasing with increasing radius formally coincides with the outer boundary of an ideal cluster with the same total number of particles. To illustrate this assertion, Fig. 3 shows the density and mean velocity distributions for a cluster in which the initial density is almost constant up to the outer boundary and decreases linearly to zero on the ξ scale further out,

$$n_c^{\mathrm{tri}}(r) = [1 - \theta(-1 + r + \xi) + \frac{1-r}{\xi}$$
$$\times(\theta(-1 + r + \xi) - \theta(r-1))] \quad (21)$$
$$\times\left(1 + \xi^2 - \frac{3\xi}{2} - \frac{\xi^3}{4}\right)^{-1}.$$

The parameter ξ for Fig. 3 was chosen to be $\xi = 0.1$. In contrast to the case with a monotonically decreasing density profile having an inflection point, a density singularity emerges here and maintains its position at the cluster–vacuum boundary over the course of time. The mean velocity also remains continuous and finite up to this boundary. The time evolution of the caustic r_{c2} (the

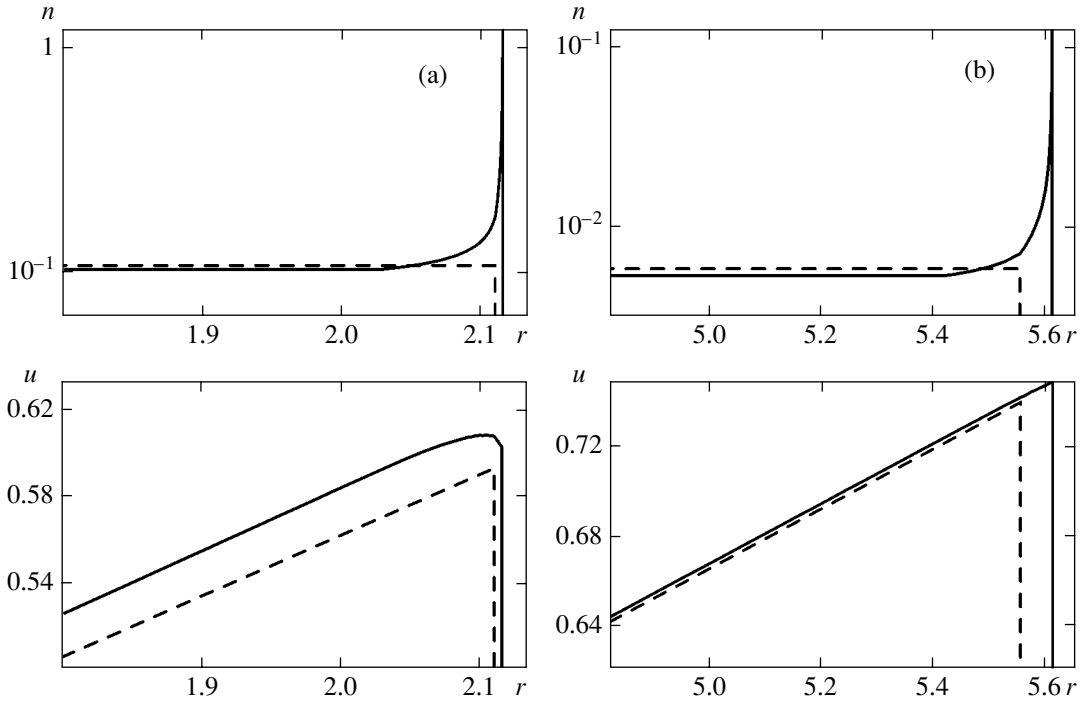

Fig. 3. Spatial distributions of density n and mean velocity u for a cluster with a nearly ideal initial density profile ($\xi = 0.1$) at times $\omega_{L0}t = 3$ (a) and 8 (b). The dashed line corresponds to an ideal cluster.

outer boundary of the region of double flow) is shown in Fig. 2b for $\xi = 0.1$.

For comparison, Fig. 3 shows the spatial density and mean velocity distributions for an ideal cluster, a cluster with a sharp density cutoff, $n_c^{\mathrm{id}}(r) = \theta(1 - r)$. We see that allowance for the small transition region with a decreasing density profile at the edge of a homogeneous cluster gives rise to an infinite density peak at the time specified by condition (19). This suggests that the explosion of an ideal cluster is unstable. As follows from the shape of the density distribution, it differs significantly from that for an ideal cluster near the singularity with a size of the order of

$$t\sqrt{\frac{2}{3}}\left(\sqrt{\frac{3w(h_{c2})}{h_{c2}}} - 1\right).$$

The total number of ions in the region of double flow is determined by the spatial inhomogeneity scale ξ and for $t \gg t_s$ is of the order of

$$\frac{1}{3}\left(1 - \sqrt{1 + \xi^2 - \frac{3\xi}{2} - \frac{\xi^3}{4}}\right),$$

i.e., $\sim\xi/4$ for $\xi \ll 1$.

Note that a significant change in density on a small ξ scale is required for a thin spherical high-density edge to be formed at the cluster–vacuum boundary in a finite time; i.e., no singularity emerges for a small jump in density near the boundary when, in contrast to distribution (21), the density decreases only slightly near the cluster edge. An arbitrary small periodic (with the characteristic wavelength smaller than the cluster radius) or nonperiodic initial perturbation in the constant density of an ideal cluster does not give rise to a singularity either. This can be easily understood from Eqs. (16), which define the conditions for the emergence of a singularity. The product $tq(1 - q^2)\sqrt{2w(h)/h^3}$ in the first equation in (16) is limited above by a value close to unity; hence, the factor $1 - h^3\omega_L^2(h)/3w(h)$ must also be of the order of unity for a singularity to emerge. However, for a nearly ideal cluster for which the density $n_c(h)$ and $w(h)$ differ from 1 and $h^3/3$, respectively, by small corrections proportional to a parameter $\mu \ll 1$, the factor $1 - h^3\omega_L^2(h)/3w(h)$ is also $\sim\mu$; i.e., the condition for the emergence of a singularity is not satisfied. In this way, the stability of an ideal cluster against small density perturbations is established.

4. THE SPECTRUM OF ACCELERATED IONS

Let us now discuss the ion spectral distribution under multiple-flow cluster explosion conditions. Using an initial distribution function of the form $f_c =$

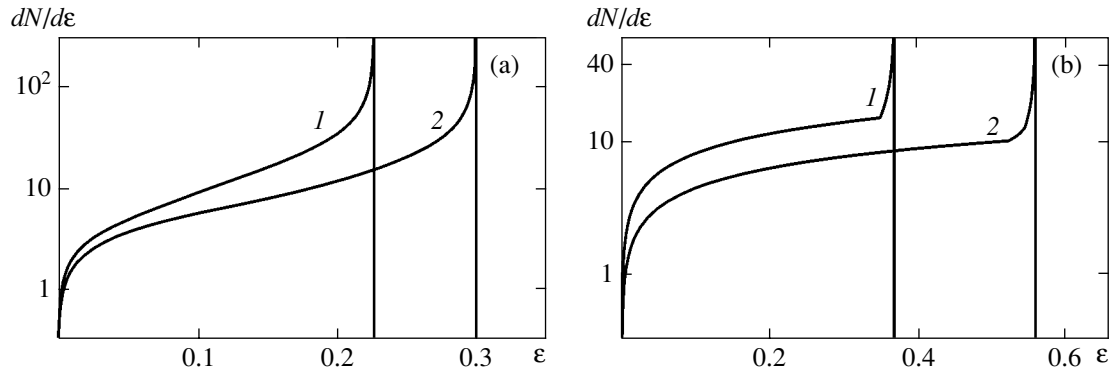

Fig. 4. Ion spectral energy distributions for clusters with the initial density profiles and at the times corresponding to (a) Fig. 1a (curve 1), Fig. 1b (curve 2) and (b) Fig. 3a (curve 1), Fig. 3b (curve 2).

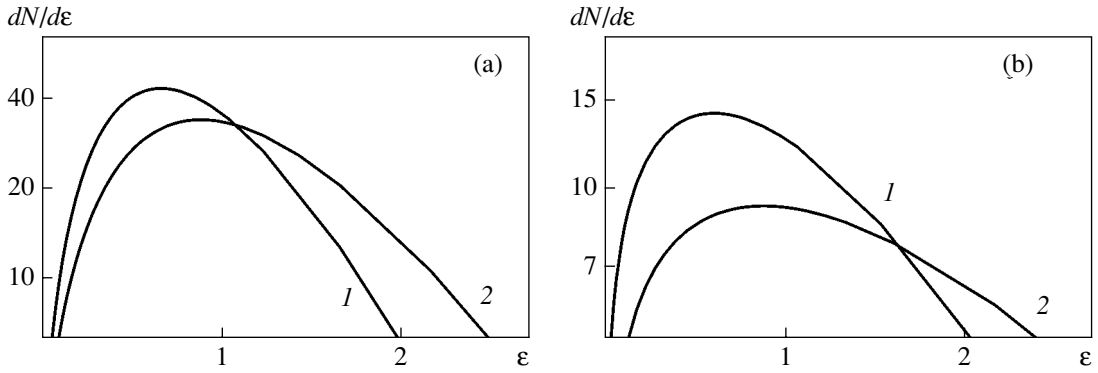

Fig. 5. Averaged ion spectral energy distributions for clusters with the same parameters as those in Fig. 4.

$\delta(\nu)$ in Eq. (12) for $dN/d\varepsilon$ and Eqs. (9), let us write the following formula for the ion spectral distribution:

$$\frac{dN}{d\varepsilon} = 2\pi\sum_l \frac{h_l^4 n_c(h_l)}{qw(h_l)}\left|q\left(1-\frac{h_l^3\omega_L^2(h_l)}{w(h_l)}\right)\right.$$
$$\left. + \frac{3t}{2}(1-q^2)^2\sqrt{\frac{2w(h_l)}{h_l^3}}\left(1-\frac{h_l^3\omega_L^2(h_l)}{3w(h_l)}\right)\right|^{-1}, \quad (22)$$
$$\varepsilon = \frac{2w(h_l)}{h_l}q(t,h_l)^2,$$
$$t\sqrt{\frac{2w(h_l)}{h_l^3}} = \frac{q}{1-q^2}+\frac{1}{2}\ln\frac{1+q}{1-q}.$$

Here, the summation is over all possible $l = 1, 2, \ldots, L$ roots of the equation $\varepsilon = (2w/h_l)q^2$ defining the function $h_l(t, \varepsilon)$. In [15], spectrum (22) with allowance made for the single-valued dependence of the velocity on the Lagrangian coordinate was represented as the sum of two terms, $(dN/d\varepsilon)^{\pm}$, that include the contributions to $dN/d\varepsilon$ from the particles for which the energy increases (+) or decreases (–) with increasing h. Formally, this breakdown is also possible in Eq. (22). It should be noted that there can be several regions in which the velocity increases (or decreases) with increasing h. However, when passing from a single flow to a multiple flow, the dependence of $U(t, h, 0)$ on the Lagrangian coordinate h remains single-valued at any time t. It thus follows that the formal shape of the cluster ion spectral distribution remains unchanged irrespective of the single- or multiple-flow regime. Thus, when passing to a multiple flow, the formulas describing the cluster ion spectral distribution remain the same as those for a cold cluster in the hydrodynamic model; the validity condition for this model, $t < t_s$, does not appear any longer. In this way, we prove that the formula of the hydrodynamic approximation for the ion spectral distribution is valid outside the range of its formal validity.

To illustrate the characteristic shapes of the ion spectra, Fig. 4 shows the ion spectral distributions after the emergence of a density singularity for the same initial density profiles and the same times as those in Fig. 1. As was noted in [15], the ion spectra observed in an actual experiment with a cluster plasma are smoother than the theoretical spectra for an individual cluster, $dN/d\varepsilon$. This may be because there is a spread in characteristic cluster radii in the cluster plasma. This

spread is very large for the cluster medium formed when a gas jet cools down [20], and the cluster size distribution is nearly Gaussian,

$$G(r_c) \propto r_c \exp\left[-\frac{(r_c - r_0)^2}{d^2}\right],$$

with $d \sim r_0$. Accordingly, the averaged ion energy distribution is determined by the convolution

$$\left\langle \frac{dN}{d\varepsilon} \right\rangle = \int_0^\infty \frac{dN}{d\varepsilon} G(r_c) dr_c, \quad \int_0^\infty G(r_c) dr_c = 1. \qquad (23)$$

Figure 5 shows the averaged ion energy spectra (23) at $d = r_0$ for the same initial density profiles as those in Fig. 4. We see that the shape of the averaged spectral distribution for the initial density profiles in the cluster under consideration has a smoother energy dependence and is largely determined by the cluster size distribution. At the same time, it follows from Fig. 5 that the height and width of the spectral peak as well as its position at a fixed time depend on the atomic density distribution in the cluster. These spectral characteristics could be the subject of experimental testing.

5. CONCLUSIONS

We have analytically solved the problem of the particle distribution function in an exploding charged cluster for the first time. We determined the global characteristics of this explosion: the spatial–temporal particle density and mean velocity distributions and the accelerated-particle spectra. These characteristics were analyzed in detail for a cold cluster, which is typical of the problem of the Coulomb explosion of a cluster ionized by a strong laser field. We showed that the kinetic effects determine the existence of multiple-flow regions after the turnover of the cluster particle velocity profile and found the boundaries of these regions. Passing to a multiple flow significantly modifies the cluster particle density and mean velocity distributions: they are described by functions with discontinuities instead of smooth functions. At the same time, it was proven that the cluster particle spectral energy distribution does not change qualitatively compared to that characteristic of the hydrodynamic approach if its inapplicability after the time corresponding to the emergence of a multiple flow is formally ignored. In this case, the accelerated-ion spectrum depends on the initial density profile in clusters, and its detailed experimental study would allow us to judge whether the theory is adequate.

Note the recently published results of particle-in-cell numerical simulations of the explosion of a spherical charged cluster exposed to an intense laser pulse [19]. The formation of singularities, density "shock waves," was observed during the cluster explosion, and there is a characteristic feature in the spectra of accelerated cluster ions similar in shape to that in Fig. 4. In this case, the initial ion distribution was assumed to be either nearly uniform with a small transition region near the outer boundary or smoothed through the successive irradiation by two laser pulses. The authors point out that one density shock in the former case and two shocks in the latter case were formed in the course of time. Thus, the results of [19] are consistent with our theory.

The general formulas given in this paper for the cluster particle distribution function allow the effect of the thermal ion motion on the global characteristics of an exploding cluster to be analyzed. This is a problem for further studies. Its significant complication is expressed, for example, in the fact that, when the nonzero thermal cluster particle velocity is taken into account, the regime of multiple flow is realized from the very beginning of the cluster explosion. However, the derived formulas allow certain assumptions about the pattern of changes in the main characteristics of the cluster plasma with the thermal particle motion to be made even now. Thus, for example, at a nonzero longitudinal temperature ($T \neq 0$), the cluster particle density and mean velocity are found as the result of kinetic "stirring" of the partial cluster particle densities and velocities. Consequently, instead of the discontinuities in the density and mean velocity profiles, one may expect their smoothing. How intense the thermal motion of the cluster particle must be to noticeably suppress the density singularity and to broaden its peaks during kinetic stirring should be judged from the parameter $T/M\omega_L^2 r_c^2$.

Kinetic stirring has the following meaning. In contrast to the "cold" solution for the dynamic equations of a cluster existing at a zero thermal velocity, a continuum spectrum of the regions with multiple flows emerges here when the nonzero radial thermal particle velocity is taken into account. This assertion is obvious, since particles moving toward and away from the cluster center exist when the thermal motion is taken into account. The partial velocities and coordinates for an ensemble of cluster ions with a given initial velocity profile change with time in accordance with Eqs. (9). The total cluster particle density at a given point is determined by the partial densities of flows with different velocities; for flows with nonzero initial velocities, new singularities emerge in the partial density distributions that are absent in the cold solution. After the summation of the partial densities of flows with different initial velocities, the singularities in the particle density distribution disappear; kinetic stirring takes place with the formation of smoother density distributions. Allowance for the transverse thermal particle motion also leads to a similar effect.

Finally, note that, apart from clusters, our results can be used to study the impact of intense laser pulses on various nanostructures, not only spherically symmetric,

but also axially symmetric, and on standard film targets of submicron thickness. This is of interest, for example, in connection with the advanced production technology of such nanoobjects as nanoballs [35], nanospheres [24], nanowires [36], and nanotubes [25].

ACKNOWLEDGMENTS

We are grateful to D.V. Shirkov and A.P. Chupakhin for interest in this paper and a discussion of the results obtained.

This work was supported by the Russian Foundation for Basic Research (project nos. 03-02-16428 and 05-01-00631), INTAS (grant nos. 01-0233 and 01-0572), ISTC (grant no. 2289), and the State Program for Support of Leading Science Schools (grant no. NSh-2339.2003.2).

REFERENCES

1. O. Echt, in *Elemental and Molecular Clusters*, Ed. by G. Benedek, T. P. Marten, and G. Pacchioni (Springer, Berlin, 1988).
2. *Large Finite Systems*, Ed. by J. Jortner, A. Pullman, and B. Pullman (Reidel, Dordrecht, 1987).
3. G. Kulcsar, D. AlMawlawi, F. W. Budnik, *et al.*, Phys. Rev. Lett. **84**, 5149 (2000).
4. T. Ditmire, J. Zweiback, V. P. Yanovsky, *et al.*, Nature **398**, 489 (1999).
5. M. Lezius, S. Dobosz, D. Normand, and M. Schmidt, Phys. Rev. Lett. **80**, 261 (1998).
6. V. P. Krainov and M. B. Smirnov, Phys. Rep. **370**, 237 (2002).
7. J. Zweiback, R. A. Smith, T. E. Cowan, *et al.*, Phys. Rev. Lett. **84**, 2634 (2000).
8. T. Ditmire, J. W. G. Tisch, E. Springate, *et al.*, Phys. Rev. Lett. **78**, 2732 (1997).
9. J. Purnell, E. M. Snyder, S. Wei, and A. W. Castleman, Jr., Chem. Phys. Lett. **229**, 333 (1994).
10. S. V. Bulanov, T. Zh. Esirkepov, J. Koga, *et al.*, Fiz. Plazmy (Moscow) **30**, 21 (2004) [Plasma Phys. Rep. **30**, 18 (2004)].
11. A. V. Gurevich, L. V. Pariĭskaya, and L. P. Pitaevskiĭ, Zh. Éksp. Teor. Fiz. **49**, 647 (1965) [Sov. Phys. JETP **22**, 449 (1966)].
12. Yu. I. Chutov and A. Yu. Kravchenko, Fiz. Plazmy (Moscow) **6**, 272 (1980) [Sov. J. Plasma Phys. **6**, 151 (1980)].
13. V. F. Kovalev and V. Yu. Bychenkov, Phys. Rev. Lett. **90**, 185004 (2003).
14. A. E. Kaplan, B. Y. Dubetsky, and P. L. Shkolnikov, Phys. Rev. Lett. **91**, 143401 (2003).
15. V. Yu. Bychenkov and V. F. Kovalev, Fiz. Plazmy (Moscow) **31**, 203 (2005) [Plasma Phys. Rep. **31**, 178 (2005)].
16. K. Nishihara, H. Amitani, M. Murakami, *et al.*, Nucl. Instrum. Methods Phys. Res. A **464**, 98 (2001).
17. Y. Kishimoto, T. Masaki, and T. Tajima, Phys. Plasmas **9**, 589 (2002).
18. http://www.billingpreis.mpg.de/hbp99/pu4hbp99.pdf.
19. F. Peano, R. A. Fonseca, and L. O. Silva, Phys. Rev. Lett. **94**, 033401 (2005).
20. J. Zweiback, T. E. Cowan, J. H. Hartley, *et al.*, Phys. Plasmas **9**, 3108 (2002).
21. A. A. Katasonov and M. B. Smirnov, Zh. Éksp. Teor. Fiz. **126**, 566 (2004) [JETP **99**, 494 (2004)].
22. J. Mansikka-aho, M. Manninen, and H. Nishioka, Phys. Rev. B **48**, 1837 (1993).
23. A. Maksimchuk, K. Flippo, H. Krause, *et al.*, Fiz. Plazmy (Moscow) **30**, 514 (2004) [Plasma Phys. Rep. **30**, 473 (2004)].
24. J. Cao and T. Matsoukas, J. Nanopart. Res. **6**, 447 (2004).
25. http://www.pa.msu.edu/cmp/csc/nanotube.html.
26. K. P. Stanyukovich, *Unsteady Motions of Continuous Medium* (Gostekhizdat, Moscow, 1955) [in Russian].
27. T. Ditmire, K. Shigemori, B. A. Remington, *et al.*, Astrophys. J., Suppl. Ser. **127**, 299 (2000).
28. Ya. B. Zel'dovich and A. D. Myshkis, *Elements of Mathematical Physics* (Nauka, Moscow, 1973) [in Russian].
29. I. S. Veselovskiĭ, Zh. Éksp. Teor. Fiz. **77**, 1352 (1979) [Sov. Phys. JETP **50**, 681 (1979)].
30. A. V. Gurevich and K. P. Zybin, Zh. Éksp. Teor. Fiz. **94**, 3 (1988) [Sov. Phys. JETP **67**, 1 (1988)].
31. Ya. B. Zel'dovich, Astrofizika **5**, 3 (1970).
32. A. V. Gurevich and K. P. Zybin, Usp. Fiz. Nauk **165**, 723 (1995) [Phys. Usp. **38**, 687 (1995)].
33. E. Weinan, Yu. G. Rykov, and Ya. G. Sinaĭ, Usp. Mat. Nauk **50**, 193 (1995).
34. V. I. Arnol'd, Tr. Seminara im. I. G. Petrovskogo, No. 8, 21 (1982).
35. http://www.microspheres-nanospheres.com.
36. J. Cao, Zh. Xu, X. Wei, *et al.*, Chem. Commun. **6**, 543 (2001).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 224–229.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 257–262.

ATOMS, MOLECULES, OPTICS

Superlight Source of Radiation with a Waveguide Used for Bunching an Electron Beam

A. V. Serov

Lebedev Institute of Physics, Russian Academy of Sciences, Leninskii pr. 53, Moscow, 119991 Russia
e-mail: serov@x4u.lebedev.ru

Received February 10, 2005

Abstract—A scheme of a generator of electromagnetic-wave radiation is proposed in which a radiating region moves along a radiator with a velocity greater than the velocity of light in vacuum. The superlight motion of the generating region leads to the situation in which the resulting radiation has the properties of Vavilov–Cherenkov radiation. The electron beam of a superlight source is formed while the particles travel across a waveguide along which an electromagnetic wave propagates. The construction of the generator makes it possible to vary the velocity of the radiating region, the radiation pattern, and the radiation beamwidth. Calculations are performed that allow one to evaluate the parameters of the generator and the characteristics of radiation. *© 2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

By a superlight source in vacuum is meant a source of coherent radiation from a large number of particles rather from a single charged particle [1]. The source of radiation is a certain domain that moves with a velocity greater than the velocity of light, each pulse of radiation being generated by new particles. Due to the motion of the radiating domain with a velocity greater than the velocity of light, the intensity and the spatial distribution of coherent radiation are characterized by the same features as Vavilov–Cherenkov radiation. Various approaches to the design of superlight sources in vacuum were considered for the first time in [2]. In particular, the radiation due to a wave impulse formed by plane electromagnetic waves that is obliquely incident on an interface, as well as the radiation emitted when a charged thread falls onto a conducting surface, were considered in that paper. The mechanism of radiation of different particles may be different; however, a characteristic feature of superlight sources is the interference of waves emitted along the trajectory of the radiating domain [1]. The electrodynamics of superlight charges was studied in survey [3].

In recent years, there have been publications in which the authors describe theoretical and experimental investigations of superlight sources of radiation. In [4], the authors considered electromagnetic radiation emitted when a pulse of X-ray radiation with a plane front is obliquely incident on a conducting surface. The X-ray radiation induces a photoelectron emission from the conducting surface, and the emission front propagates along the surface with a velocity greater than the velocity of light. In [5, 6], certain schemes of superlight sources were studied theoretically and the intensities of the radiation induced by such photoelectron emission were evaluated. Later, at the Russian Research Institute of Experimental Physics, Russian Federal Nuclear Center, equipment was designed that employs X-ray pulses, and the time and angular characteristics of the generated radiation were measured [7, 8]. The X-ray radiation is generated by a plasma that is formed when a subnanosecond laser pulse is focused on an aluminum target. According to the authors of [7, 8], the generation of electromagnetic radiation with intensities of practical importance requires electron energies on the other of hundreds of kiloelectron-volts. Therefore, the photoelectrons emitted under the irradiation of a plate by X-ray pulses are further accelerated in the space between the cathode (the emitting plate) and the grid anode.

Another type of a superlight source of radiation is observed when measuring coherent transient radiation of a bunch of particles accelerated in a microtron [9]. Upon leaving the microtron, the bunch of particles passes through a metal foil. In the experiment, the horizontal size of the bunch was much greater than its vertical size and the size in the direction of motion; therefore, the bunch had the shape of a length of charged thread that moves at an angle to the surface of the foil. In addition to the maxima that are typical of the transient radiation, the measured angular distribution of radiation contains peaks whose asymmetry and directivity are characteristic of Vavilov–Cherenkov radiation.

In [10, 11], an equipment with a superlight source is described and experimental results obtained on this equipment are presented. A characteristic feature of this equipment is that radiation in it is generated by polarization currents induced in dielectric plates.

In the present paper, we propose a design for a superlight source in which the domain that emits radia-

tion and moves with a superlight velocity is formed by a continuous electron beam that travels through the waveguide and reaches the space of radiation.

2. DESIGN OF THE SOURCE

The schematic diagram of the source is shown in Fig. 1. The basic element of the source is a rectangular waveguide W. The cross section of the waveguide has the dimensions a along the x axis and b along the y axis. An H_{10} electromagnetic wave propagates along the waveguide in the positive direction of the z axis. The electric and magnetic fields of this wave are described by the formulas

$$E_y = -E_0 \sin\left(\frac{\pi x}{a}\right)\sin\varphi,$$

$$H_x = E_0 \eta \sin\left(\frac{\pi x}{a}\right)\sin\varphi,$$

$$H_z = E_0 \frac{\lambda}{2a}\cos\left(\frac{\pi x}{a}\right)\cos\varphi, \quad (1)$$

$$E_x = E_z = H_z = 0,$$

where E_0 is the electric-field strength of the wave, $\varphi = \omega t - k_z z$, $k_z = k\sqrt{1-(\lambda/2a)^2}$, $k = 2\pi/\lambda$, and λ is the free-space wavelength. The H_{10} wave contains a single electric component, E_y. The phase velocity v_f of the wave depends only on the transverse dimension a of the waveguide,

$$v_f = \frac{c}{\sqrt{1-(\lambda/2a)^2}}, \quad (2)$$

and is always greater than the velocity of light in vacuum c.

There are narrow slots of length d along the z axis in the upper and lower walls of the waveguide. These slots are drawn by heavy dashed lines in the figure. If the width of the slots is much less than the transverse dimensions of the waveguide, the slots virtually do not perturb the electromagnetic field in the waveguide.

A source of electrons of small size Δx in the x direction and of length d in the z direction is situated under the waveguide. This source generates a continuous electron beam, which is injected into the waveguide through the lower slot. The velocity v_i of the injected particles is directed along the y axis. Consider a beam of particles that are injected at the point $(x_0, 0, z_0)$. Depending on the moment of injection, the particles get either into a decelerating or an accelerating electric field of the wave in the waveguide. If the initial velocity v_i of the particles is nonrelativistic and the height b of the waveguide and the electric-field strength E_0 are sufficiently large, then the bunching of particles occurs in the waveguide. Some particles of the continuous beam of electrons are decelerated, change the direction of their motion, and leave the waveguide through the lower slot. Other particles are accelerated, pass through the waveguide, and leave it through the upper slot. Thus, a continuous beam of electrons injected at the point $(x_0, 0, z_0)$ is transformed into a train of bunches at the output of the waveguide (at the point (x_0, b, z_0)). The length l_b of the bunches is determined by the initial velocity v_i of the particles, the transverse dimensions a and b of the waveguide, and the electric-field strength E_0. The distance between the bunches is $L = \lambda v_k/c$, where v_k is the velocity of particles at the output of the waveguide.

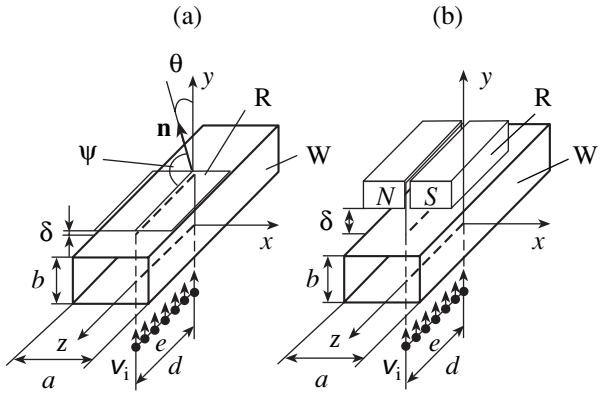

Fig. 1. Scheme of a superlight source that employs (a) transient and (b) synchrotron radiation; W is a waveguide, and R is a radiator.

At the point $(x_0, b, z_0 + \delta z)$, which is shifted by a distance δz along the z axis, the time dependence of the current through the waveguide is the same as that at the point (x_0, b, z_0) but the time delay $\delta t = \delta z/v_f$. If the output time of the first particle of a bunch at the point (x_0, b, z_0) is t_0, then the output time of the first particle of the bunch that leaves the waveguide at the point $(x_0, b, z_0 + \delta z)$ is $t_0 + \delta z/v_f$. Thus, the point at which the first particles of bunches leave the waveguide moves along the z axis with the velocity v_f, which is greater than the velocity of light. Since an electron beam, rather than a single particle, travels through the waveguide, a certain extended region in which particles travel across the waveguide moves along the waveguide. This region moves along the z axis, while the particles themselves travel in the transverse direction, along the y axis. While the beam injected into the waveguide is continuous, the output beam represents a train of charged strips. The length of a strip (the size along the z axis) is equal to the length d of the slot, the size of a strip along y is equal to the length l_b of a bunch, and the size along x is equal to the width Δx of the slot in the waveguide.

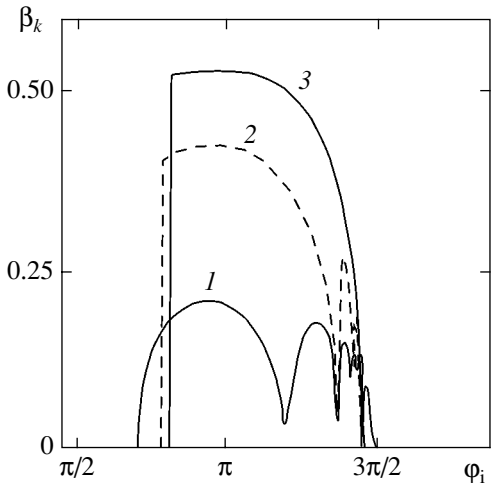

Fig. 2. Velocity β_k of electrons at the output of the waveguide as a function of the phase φ_i of injection into the waveguide for the electric-field strength E_0 in the waveguide equal to (*1*) 30, (*2*) 60, and (*3*) 90 kV/cm.

The strip moves at angle ψ to the z axis such that $\tan\psi = v_k/v_f$.

If the field strength in the waveguide is small, then the velocity of the continuous electron beam is modulated while it travels across the waveguide. The velocity modulation of the beam results in the bunching of particles while they move inside the waveguide and in the free space upon leaving the waveguide.

There is a radiating element R at a certain distance δ from the waveguide (see Fig. 1). It was stressed in [2] that the particles in a superlight source may have different mechanisms of radiation. In particular, the radiating element of the superlight source shown in Fig. 1a is a thin foil (a metal grid), and the particles generate transient radiation while passing through this foil. In the source shown in Fig. 1b, the particles generate synchrotron radiation while moving in an external magnetic field. The drift space δ is chosen so that, at any point of the radiating element, the length of the current pulse is minimal and the amplitude of the current is maximal.

3. NUMERICAL ANALYSIS OF A SUPERLIGHT SOURCE

We applied numerical methods to the analysis of the operation features of a superlight source. As a model, we used the device shown in Fig. 1a. The width and the height of the waveguide are a = 60 mm and b = 10 mm, respectively. An H_{10} wave with a wavelength of λ = 10 cm propagates along the waveguide in the positive direction of the z axis. For such a wavelength and the dimensions of the waveguide, the propagation velocity is equal to $v_f = 1.81c$. The radiating element is a metal foil fixed to the upper wall of the waveguide. In our calculations, we assumed that the drift space δ is zero.

First, we consider the dynamics of particles in the waveguide. To this end, we solve a relativistic equation of motion of a particle in a given electromagnetic field [12],

$$\frac{d\boldsymbol{\beta}}{dt} = \frac{e}{mc\gamma}\{\mathbf{E} + \boldsymbol{\beta}\times\mathbf{H} - \boldsymbol{\beta}\cdot(\boldsymbol{\beta}\cdot\mathbf{E})\}, \qquad (3)$$

where $\boldsymbol{\beta} = \mathbf{v}/c$ and $\gamma = 1/\sqrt{1-\beta^2}$ are the relative velocity and the energy of the particle, respectively.

The field of the electromagnetic wave is described by formulas (1). The particles are injected at the point $x = y = z = 0$ on the lower wall of the waveguide and have an initial velocity of $v_i = 0.025c$ directed along the y axis. We consider the motion of two hundred particles that are injected successively over intervals of $dt = 0.005T$ ($T = \lambda/c$ is the period of the wave). Such a calculation simulates the motion of a bunch of length $l_b = v_i T$ injected into a waveguide within a time period equal to one period of the wave.

Next, we calculate the positions and the velocities of particles while they travel in the wave and at the output of the waveguide. We also calculate the moment when the particles leave the waveguide. The calculations are performed for various values of the electric-field strength E_0 of the wave. Figure 2 shows the particle velocity β_k at the waveguide output as a function of the phase φ_i of particles at the moment of injection. One can see that not all the electrons that are injected into the waveguide during the accelerating half-period pass through the waveguide. Therefore, the bunch length l_b (the distance between the first and the last particles that travel through the waveguide at a given point within a time period of T) is less than $L/2$, and the phase length of the bunch is less than π. As the field strength increases, the bunch length decreases and the length of the flat region of the impulse increases, $\beta_k = f(\varphi_i)$; i.e., the number of particles whose velocity is close to the maximal velocity increases. Figure 3 shows the positions of particles on the plane (β_k, φ_k), where φ_k is the particle phase at the output of the waveguide. One can see that the bunching of particles occurs: the electron beam, which is continuous while entering the waveguide, turns into a bunch of electrons concentrated in a rather narrow phase interval at the output of the waveguide. Moreover, even in this interval, the particles are distributed nonuniformly. This fact is illustrated by the diagram of the particle density ρ as a function of phase φ_k shown in Fig. 3. The calculations show that about 75% of the particles that pass through the waveguide are concentrated within the phase interval of 0.2π; i.e., at the waveguide output, the phase length of a bunch is about an order of magnitude less than 2π. If the slot size along the z axis is much greater than its size along the x axis and the length of the bunch along the y

axis (in our case, $l_b \approx 5$ mm), then we can assume that a charged thread is incident on the radiator.

Using the data obtained while calculating the motion of particles across the waveguide, we calculated the angular distribution of the field strength E^r of the transient radiation in the plane *yz*. Here, we assumed that the radiation field at a given frequency generated by an individual particle that passes through the plane metal boundary and flies out to vacuum is described by the formula [13]

$$E_\omega^r = \frac{e}{\pi c R}\left(\frac{\beta_k|\varepsilon-1|\sin\theta\cos\theta}{\varepsilon\cos\theta+\sqrt{\varepsilon-\sin^2\theta}}\right)\exp\left(i\frac{\omega}{c}R-i\omega t_k\right), \quad (4)$$

where *e* is the electron charge, *R* is the distance form the point where the particles pass through the radiator to the observation point, ε is the permittivity of the foil material, θ is the angle between the particle velocity and the wavevector **n**, and t_k is the moment at which the particle leaves the waveguide.

In the centimeter- and millimeter-wave bands, one may assume that the spectrum of the transient radiation of a particle is uniform. For most metals, the angular distribution of the radiation of a single charged particle has a qualitatively similar character. In the case under consideration, the transient radiation is generated by weakly relativistic particles. Therefore, the distribution has a maximum between 60° and 70°.

The radiation field of the bunch as a whole is equal to the sum of the fields generated by individual particles. The contribution of each particle is determined by the distance *R* to the observation point and the moment t_k at which the particle leaves the waveguide. Figure 4 (curve *1*) shows the angular distribution of the intensity $I \sim (E^r)^2$ of radiation at a wavelength of λ = 3 cm generated by particles that leave the waveguide at the point (0, *b*, 0). One can see that the distribution is symmetric about the direction of motion of the particles. Such an angular distribution is characteristic of the radiation due to an electron beam with small transverse dimensions. Further, we analyze the effect of the slot length *d* on the distribution of radiation.

Using formula (4), we can obtain an expression for the angular distribution of the field of transient radiation generated by a piece of a charged thread that passes through a metal foil. We will neglect the difference between the velocities of electrons at the waveguide output and assume that all particles have the same velocity β_b. In addition, we will assume that the coordinates *x* and *z* of a particle at the waveguide output are the same as those at the injection point. Under such an assumption, the oscillations and the drift of particles along the longitudinal *z* axis are not taken into account.

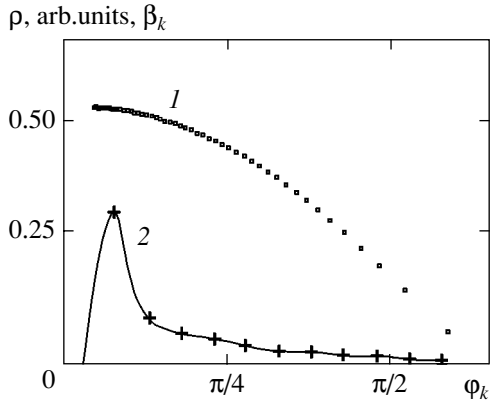

Fig. 3. (*1*) Velocity β_k and (*2*) density ρ of electrons as a function of the phase φ_k at the output of the waveguide for an electric-field strength in the waveguide of E_0 = 90 kV/cm.

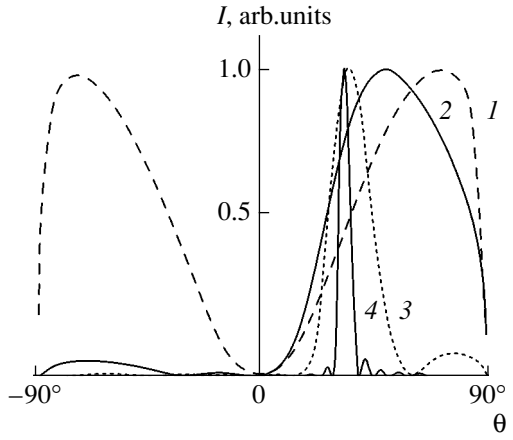

Fig. 4. Angular distribution of the radiation intensity at a wavelength of λ = 3 cm. The radiator is an electron beam with small transverse dimensions (*1*) or an electron source of length *d* = 3 (*2*), 10 (*3*), and 40 cm (*4*).

Consider the fields at a large distance from the radiation source, when $R \gg d$. In this case, we have

$$R = R_0 - z\sin\theta, \quad (5)$$

where R_0 is the distance from the observation point to the origin and *z* is the coordinate of the point at which the particle leaves the waveguide.

Assume that the length of the thread is much greater than its transverse dimensions (dimensions in the *x* and *y* directions). The moment t_k at which the particle that passes through the waveguide at the point with coordinate *z* leaves the waveguide can be expressed as

$$t_k = t_0 + \frac{z}{v_f}, \quad (6)$$

where t_0 is the moment at which the particle injected at the origin leaves the waveguide.

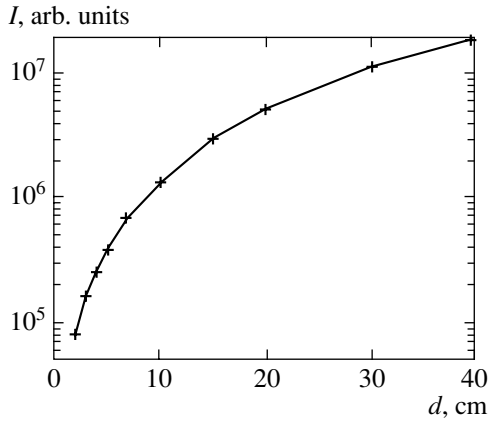

Fig. 5. Maximal intensity of radiation as a function of the length of the electron source.

Substituting (5) and (6) into (4), we obtain

$$E_\omega^r = \frac{q}{\pi c R}\left(\frac{\beta_b|\varepsilon-1|\sin\theta\cos\theta}{\varepsilon\cos\theta+\sqrt{\varepsilon-\sin^2\theta}}\right) \times \exp\left(i\frac{\omega}{c}R_0 - i\omega t_0\right)\int_0^d e^{i\omega\alpha z}dz, \quad (7)$$

where q is the charge of a unit length of the thread,

$$\alpha = \frac{\sin\theta}{c} - \frac{1}{v_f}. \quad (8)$$

After integrating with respect to z, we obtain the following expression for the strength of the field of the transient radiation generated by a charged thread:

$$E_\omega^r = \frac{q}{\pi c R}\left(\frac{\beta_b|\varepsilon-1|\sin\theta\cos\theta}{\varepsilon\cos\theta+\sqrt{\varepsilon-\sin^2\theta}}\right) \times \exp\left[i\frac{\omega}{c}R_0 - i\omega t_0 + \frac{\alpha}{2}d\right]\frac{\sin\eta}{\eta}d, \quad (9)$$

where $\eta = \alpha d/2$.

It follows from (9) that the angular distribution of the field strength of the transient radiation generated by a thread is determined by the field distribution of an isolated charge (given by the expression in parentheses), the function $\sin\eta/\eta$, and the length d of the source of electrons. The function $\sin\eta/\eta$ attains its principal maximum at $\eta = 0$ and oscillates while decreasing as $\eta \longrightarrow \pm\infty$. The condition $\eta = 0$ for the principal maximum is fulfilled when $\alpha = 0$. It follows from (8) that this condition corresponds to the radiation angle θ_{max} given by

$$\sin\theta_{max} = \frac{c}{v_f}. \quad (10)$$

Substituting v_f from (2) into (10) and taking into account the wavelength λ and the waveguide size a, we obtain $\theta_{max} = 56°$.

Denote by ψ the angle between the z axis and the wavevector **n** (see Fig. 1a). Since $\psi = \pi/2 - \theta$, we have the following equation for this angle:

$$\cos\psi_{max} = \frac{c}{v_f}. \quad (11)$$

Formula (11) defines the angle ψ_{max} of Vavilov–Cherenkov radiation from a source that moves along the z axis with a velocity v_f greater than the velocity of light. Note that each particle of a bunch travels along a normal to the foil and produces only the transient radiation while passing through the foil. However, the radiation source itself moves along the foil at a superlight velocity. The interference of the waves emitted by individual particles results in a directional radiation that is characteristic of the Vavilov–Cherenkov radiation [14].

We calculated the angular distribution of the intensity of radiation generated by an electron beam for different values of the length d of the electron source. The results of these calculations are shown in Fig. 4 (curves *2–4*). One can see that, as d increases, the angular distribution becomes asymmetric, and the width of the angular distribution at the maximum intensity decreases. Moreover, the angle at which the radiation intensity is maximal varies from $\theta_{max} = 70°$, which is typical for the transient radiation of weakly relativistic particles, to $\theta_{max} = 34°$. Recall that $\psi_{max} = \pi - \theta_{max} = 56°$ is the angle of the Vavilov–Cherenkov radiation.

All the curves in Fig. 4 are normalized by the maximal intensity. The radiation intensity as a function of d is shown in Fig. 5. It follows from the calculations that the radiation intensity increases with the source length d; for $d > 5\lambda$, the function of the intensity of radiation emitted at the angle ψ_{max} versus the source length becomes quadratic, $I \propto d^2$.

4. CONCLUSIONS

We have considered a variant of a superlight source of radiation that employs a waveguide for bunching an electron beam. We have calculated the motion of particles through the waveguide and the angular distribution of the intensity of transient radiation generated by an electron beam. In the millimeter- and centimeter-wave bands, the spectrum of the transient radiation can be considered to be uniform. Therefore, such a design of the source can produce intense radiation over a wide range of wavelengths. The radiation direction and the width of the angular distribution can be varied by varying the shape of the radiating surface and its length.

If a source employs synchrotron radiation (see Fig. 1b), then the wavelength at which the radiation

power attains its maximum can be controlled by varying the strength of the magnetic field in the radiator.

The scheme of the radiation source proposed allows one to realize a radiation mode when the velocity of radiation region varies along its trajectory. This is achieved by varying the transverse dimensions of the waveguide along its longitudinal axis.

ACKNOWLEDGMENTS

I am grateful to B.M. Bolotovskii for the discussion of the results.

This work was supported by the Russian Foundation for Basic Research (project no. 04-02-16376) and by the Ministry of Education and Science (project no. RSh-2060.2003.2).

REFERENCES

1. V. L. Ginzburg, *Theoretical Physics and Astrophysics* (Nauka, Moscow, 1975; Pergamon, Oxford, 1979).
2. B. M. Bolotovskiĭ and V. L. Ginzburg, Usp. Fiz. Nauk **106**, 577 (1972) [Sov. Phys. Usp. **15**, 184 (1972)].
3. B. M. Bolotovskiĭ and V. P. Bykov, Usp. Fiz. Nauk **160**, 141 (1990) [Sov. Phys. Usp. **33**, 477 (1990)].
4. N. J. Carron and C. L. Longmire, IEEE Trans. Nucl. Sci. **23**, 1897 (1976).
5. Yu. N. Lazarev and P. V. Petrov, Pis'ma Zh. Éksp. Teor. Fiz. **60**, 625 (1994) [JETP Lett. **60**, 634 (1994)].
6. Yu. N. Lazarev and P. V. Petrov, Zh. Éksp. Teor. Fiz. **115**, 1689 (1999) [JETP **88**, 926 (1999)].
7. A. V. Bessarab, A. V. Kunin, S. P. Martynenko, *et al.*, Tr. Ross. Fed. Yad. Tsentra VNIIÉF, No. 1, 518 (2001).
8. A. V. Bessarab, A. A. Gorbunov, S. P. Martynenko, and N. A. Prudkoy, IEEE Trans. Plasma Sci. **32** (3) (2004).
9. A. V. Serov, S. V. Levonyan, and B. M. Bolotovskiĭ, Zh. Éksp. Teor. Fiz. **120**, 1346 (2001) [JETP **93**, 1163 (2001)].
10. H. Ardavan, A. Ardavan, and J. Singleton, J. Opt. Soc. Am. A **20**, 2137 (2003).
11. H. Ardavan, A. Ardavan, and J. Singleton, J. Opt. Soc. Am. A **21**, 858 (2004).
12. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 2: *The Classical Theory of Fields*, 7th ed. (Nauka, Moscow, 1984; Pergamon, Oxford, 1975).
13. V. L. Ginzburg and V. N. Tsytovich, *Transition Radiation and Transition Scattering* (Nauka, Moscow, 1984) [in Russian].
14. V. M. Bolotovskiĭ and A. V. Serov, Zh. Tekh. Fiz. **72** (1), 3 (2002) [Tech. Phys. **47**, 1 (2002)].

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 230–252.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 263–291.

ATOMS, MOLECULES, OPTICS

Practical Error-Correction Procedures in Quantum Cryptography

A. P. Makkaveev[b,c], S. N. Molotkov[a,b], D. I. Pomozov[b,c], and A. V. Timofeev[b]

[a]*Institute of Solid-State Physics, Russian Academy of Sciences, Chernogolovka, Moscow oblast, 142432 Russia*
[b]*Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, 119992 Russia*
[c]*Institute of Physics and Technology, Russian Academy of Sciences, Moscow, 117428 Russia*
e-mail: molotkov@issp.ac.ru
Received February 25, 2005

Abstract—Quantum cryptography (secure key distribution) systems must include procedures for correcting errors in the raw key transmitted over a quantum communication channel. Several reconciliation protocols are discussed and compared in terms of efficiency. *© 2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

Quantum cryptography (secure key distribution) [1, 2] makes it possible, in principle, to implement absolutely secure "one-time pad" cryptography systems (undecipherable by an eavesdropper even theoretically) [3–5]. The security of quantum key distribution is based on the fundamental principles of quantum mechanics [6]. More precisely, Heisenberg's uncertainty principle forbids simultaneous measurement of observables represented by noncommuting operators. In quantum cryptography, these observables are represented by the density matrices of carrier states corresponding to the classical bits, 0 and 1. For pure states, the impossibility of simultaneous measurement (complete indistinguishability) of density matrices is equivalent to nonorthogonality of the carrier states [7]. In other words, it is impossible to distinguish nonorthogonal states with zero probability of error and leave the system in a post-measurement state is identical to its initial state. Thus, any eavesdropping attempt (i.e., a perturbation of the carrier state) must change the measurement statistics at the receiver end as compared to the statistics of measurement on the input states. Transformation of quantum states in a noisy quantum channel also leads to changes in the measurement statistics. In quantum cryptography, any change in measurement statistics should be attributed to the eavesdropper's intervention, because it is impossible in principle to distinguish between changes in statistics induced by channel noise and the eavesdropper.

If the laws of quantum mechanics were such that only the fact that a carrier state had been perturbed could be revealed, then this possibility would be of no use for secure key distribution. However, not only perturbation can be detected, but the change in measurement statistics can also be related to the amount of information that can be extracted by the eavesdropper.

In quantum cryptography, not only a quantum communication channel (fiber or free space) is required to transmit quantum states, but also a public classical communication channel is necessary for the legitimate partners to check for changes in the measurement statistics and correct errors in transmissions over the quantum communication channel. The only requirement to be satisfied by the classical channel is that the published information cannot be changed by the eavesdropper's intervention; i.e., the channel must be unjammable. A public classical channel of this kind is a mathematical idealization, because it cannot physically exist. The integrity of the announced data can be guaranteed only by using authentication and integrity control procedures, which require the use of a special secret key. If the Internet is used as a public classical channel, then authentication keys can be generated by means of the Diffie–Hellman algorithm [8]. However, the Diffie–Hellman scheme cannot be used to generate authentication keys if the same fiber-optic link is employed as both a public classical and a quantum channel for the fundamental reason that a "man-in-the-middle" attack then obviously becomes possible. To deal with this situation, one needs a short initial key to be used only in the first session and then discarded. In each subsequent session, the authentication and integrity of the message transmitted over the classical channel are guaranteed by using part of the key generated the preceding communication. The remaining (longer) part of the key received over the quantum channel is used for encoding. If the authentication and encryption procedures are based on a Russian state standard (GOST), then the initial key must be 256 bits long. However, a much longer new secret key can be transmitted in a few seconds by communicating over the quantum channel.

The initial key could obviously be used for encoding a new key to be sent to the other legitimate partner. However, the absolute security of the new key would be

guaranteed only if its length did not exceed that of the key used to encode it. In quantum cryptography, the initial key is never used directly for transmitting a newly generated key over the quantum channel. It is shown below that the amount of information transmitted over the public classical channel per bit of a new secret key is less than unity; therefore, key expansion can be implemented.

The use of a short initial key is more attractive, because the number of rounds of public communication required to complete key reconciliation and privacy amplification can thus be minimized.

In this approach, the main theoretical problem is the evaluation of the length of a secret key that can be received under the observed change in measurement statistics at the receiver end as compared to the statistics of measurement on the unperturbed states. Generally, the change in measurement statistics is quantified in terms of the bit error rate, i.e., the probability that a transmitted 0 is received as a 1 or vice versa. (Note, however, that alternative measures of change in measurement statistics can be used as well.) The error rate is evaluated by publicly comparing part of the bit sequence transmitted over the quantum channel, and then the disclosed bits are discarded.

At the next stage, Alice and Bob correct errors in the rest of the bit sequence, communicating over the public channel. (Alice and Bob are the conventional names of the sender and receiver, respectively; *A* and *B* denote their respective locations; and Eve is the conventional name of the eavesdropper.) As a result, Alice and Bob have shorter strings that are identical with probability arbitrarily close to one, e.g., $1–2^{-200} \approx 1–10^{-70}$ (recall that the estimated number of atoms in the Universe is 10^{77}).

After the key has been reconciled, Eve has a string of bits, or quantum states stored in a quantum memory, or both. At the last stage, a final secret key is generated by compressing the bit string shared by Alice and Bob (actually, by using a random hash function). The resulting compressed string becomes a secret key shared by the legitimate partners, and it is guaranteed that the amount of private information known to Eve is exponentially small as a function of a certain security parameter defined by Alice and Bob.

It is natural to require that error-correction and privacy-amplification procedures leave as many bits as possible in the final key. Another requirement is the minimization of the number of rounds of public communication per bit in the final secret key.

The legitimate partners must not only correct the errors in the raw key, but also calculate an upper bound for the amount of information about the remnant key that can be extracted by Eve from public transmissions. Error correction can be performed by using various procedures, including any classical error-correcting code. Note that it is all but obvious which one will prove the most efficient by the criteria mentioned above.

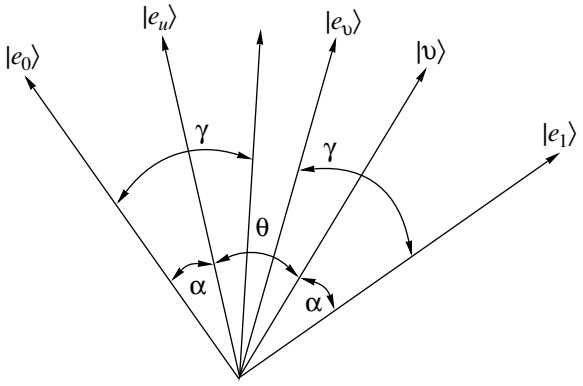

Fig. 1.

In what follows, we discuss error correction by bisective search, the CASCADE algorithm, and classical Bose–Chaudhuri–Hocquenghem and Hamming codes. It is important that the efficiency of any error-correction procedure cannot be evaluated without taking into account the quantum part of the key generation protocol.

Irrespective of the error-correction procedure employed, the analysis presented below shows that the upper bound for the amount of information about the final key that can be extracted by Eve corresponds to the case when she can perform collective measurements on several quantum states stored in a quantum memory. Even though measurement of many entangled states is beyond the scope of present-day technology, it is not forbidden by any fundamental principle of quantum mechanics. This problem is eliminated if Eve's instrumentation can be used to perform only individual measurements.

2. THE QUANTUM PART OF A PROTOCOL FOR RAW KEY GENERATION

Conceptually, the simplest quantum key distribution protocol, known as the B92 protocol, makes use of two nonorthogonal states. Even though attenuation in a fiber-optic cable is the only factor limiting its implementation, this protocol works only if the quantum channel length does not exceed 20 km. Its discussion in the context of the present study is motivated by the simplicity of its analysis and the fact that it contains all elements common to other, more practicable, key distribution protocols.

The quantum part of the protocol has a standard form [9, 10]. Two nonorthogonal states of the information carrier corresponding to classical bits are prepared by Alice: $0 \longleftrightarrow |u_0\rangle$, $1 \longleftrightarrow |u_1\rangle$. The overlap between them is conveniently parameterized by the corresponding angle (see Fig. 1):

$$\langle u_0|u_1\rangle = \sin 2\alpha. \quad (1)$$

The orthogonal basis vectors in the span of $|u_0\rangle$ and $|u_1\rangle$ are denoted by $|0\rangle$ and $|1\rangle$.

Bob performs individual measurements described by the following partition of the identity operator:

$$I = A_0 + A_1 + A_?, \quad A_0 = \frac{(I - |u_1\rangle\langle u_1|)}{1 + \langle u_1|u_0\rangle},$$
$$A_1 = \frac{(I - |u_0\rangle\langle u_0|)}{1 + \langle u_1|u_0\rangle}, \quad A_? = I - A_0 - A_1. \tag{2}$$

The outcome space Ω consists of three outcomes: $\Omega = \{0, 1, ?\}$. The operator A_0 represents measurements whose outcomes are interpreted as 0. In the absence of Eve's intervention, such a measurement occurs only if the state $|u_0\rangle$ has been sent. If the state $|u_1\rangle$ has been sent, then the measurement represented by A_1 (but not A_0) can occur, and its outcome is interpreted as 1. The operator $A_?$ represents measurements with inconclusive outcomes, which may correspond to both $|u_0\rangle$ and $|u_1\rangle$.

The conditional probability $\Pr(i|j)$ of an outcome i ($i = 0, 1, ?$) when a state $|u_j\rangle$ ($j = 0, 1$) has been sent is expressed as follows:

$$\Pr(0|0) = \Pr(1|1) = \mathrm{Tr}\{|u_0\rangle\langle u_0|A_0\}$$
$$= \mathrm{Tr}\{|u_1\rangle\langle u_1|A_1\} = 1 - \cos\theta,$$
$$\Pr(1|0) = \Pr(0|1) = \mathrm{Tr}\{|u_0\rangle\langle u_0|A_1\}$$
$$= \mathrm{Tr}\{|u_1\rangle\langle u_1|A_0\} = 0. \tag{3}$$

$$\Pr(?|0) = \Pr(?|1) = \mathrm{Tr}\{|u_0\rangle\langle u_0|A_?\}$$
$$= \mathrm{Tr}\{|u_1\rangle\langle u_1|A_?\} = \cos\theta. \tag{4}$$

All possible eavesdropping strategies can be tentatively classified as follows.

Opaque eavesdropping. (This strategy is also frequently called *intercept–resend attack.*) Eve intercepts Alice's transmission, performs a direct measurement on it, and resends a new state depending on the measurement outcome.

Entangled translucent eavesdropping. Eve uses an auxiliary system (ancilla) interacting with the information carrier to create an entangled state of the carrier and the ancilla and performs a measurement on the ancilla state while the carrier state is forwarded to Bob.

Collective attack. This strategy is analogous to translucent eavesdropping, with the exception that Eve stores each ancilla state in a quantum memory and performs a collective measurement on the stored ancilla states only after Alice and Bob have completed their error-correction and privacy-amplification procedures by communication over a public channel. Collective attacks are more effective than individual measurements.

Joint attack. This strategy is the most general and, probably, the most effective. It is analogous to a collective attack, except that Eve performs a measurement on a single state from a Hilbert space of large dimension interacting with the carrier state.

Formally, entangled translucent eavesdropping is described by a unitary transformation of the ancilla state $|e\rangle$ interacting with the carrier state prepared by Alice:

$$|u_0\rangle \otimes |e\rangle \longrightarrow U(|u_0\rangle \otimes |e\rangle)$$
$$= a|u_0\rangle \otimes |e_0\rangle + b|u_1\rangle \otimes |e_1\rangle = |\phi_0\rangle, \tag{5}$$

$$|u_1\rangle \otimes |e\rangle \longrightarrow U(|u_1\rangle \otimes |e\rangle)$$
$$= b|u_0\rangle \otimes |e_0\rangle + a|u_1\rangle \otimes |e_1\rangle = |\phi_1\rangle, \tag{6}$$

where $|e_0\rangle$ and $|e_1\rangle$ are (nonorthogonal) ancilla states, and a and b are real coefficients chosen by Eve to specify the unitary evolution.

While the ancilla and carrier are in an entangled state after the evolution, Eve and Bob can measure only the states of the ancilla and carrier, respectively.

In the case of a collective attack, the states measured by Bob before Eve performs any operation on the ancilla state are obtained by tracing out the ancilla subspace:

$$\rho(|u_0\rangle) = \mathrm{Tr}_E\{|\phi_0\rangle\langle\phi_0|\}$$
$$= a^2|u_0\rangle\langle u_0| + ab\sin 2\gamma(|u_0\rangle\langle u_1| + |u_1\rangle\langle u_0|)$$
$$+ b^2|u_1\rangle\langle u_1|,$$
$$\rho(|u_1\rangle) = \mathrm{Tr}_E\{|\phi_1\rangle\langle\phi_1|\} \tag{7}$$
$$= b^2|u_0\rangle\langle u_0| + ab\sin 2\gamma(|u_1\rangle\langle u_0| + |u_0\rangle\langle u_1|)$$
$$+ a^2|u_1\rangle\langle u_1|.$$

After Bob has performed the measurements described by (2) on the states represented by (7), conditional probabilities (3) and (4) become

$$\Pr(0|0) = \Pr(1|1) = \mathrm{Tr}\{\rho(|u_0\rangle)A_0\}$$
$$= \mathrm{Tr}\{\rho(|u_1\rangle)A_1\} = a^2(1 - \sin 2\gamma),$$
$$\Pr(1|0) = \Pr(0|1) = \mathrm{Tr}\{\rho(|u_0\rangle)A_1\} \tag{8}$$
$$= \mathrm{Tr}\{\rho(|u_1\rangle)A_0\} = b^2(1 - \sin 2\gamma),$$

$$\Pr(?|0) = \Pr(?|1) = \mathrm{Tr}\{\rho(|u_0\rangle)A_?\}$$
$$= \mathrm{Tr}\{\rho(|u_1\rangle)A_?\} = 1 - (a^2 + b^2)(1 - \sin 2\gamma). \tag{9}$$

The deviation of Bob's measurement statistics from Alice's is quantified by an error rate. In particular, the

probability that a 0 is transmitted and a 1 is received or vice versa corresponds to the error rate

$$Q = \frac{\Pr(1|0) + \Pr(0|1)}{\Pr(0|0) + \Pr(1|1)} = \frac{b^2}{a^2 + b^2}. \quad (10)$$

The inconclusive outcomes are rejected by communicating over a public channel. To evaluate Q, Alice and Bob compare by public discussion the bits in a randomly sampled subsequence (with a length of about half that of the transmitted sequence) and then reject the published part of the sequence.

The ancilla states stored by Eve and the bits received by Bob are in one-to-one correspondence. In particular, if Alice has sent $|u_0\rangle$ and Bob has received a 0, then the unnormalized ancilla state is

$$\begin{aligned}\rho_{0\to 0}^{\text{Eve}}(|u_0\rangle) &= \mathrm{Tr}\{\sqrt{A_0}|\phi_0\rangle\langle\phi_0|\sqrt{A_0}\} \\ &= (1-Q)|e_0\rangle\langle e_0|.\end{aligned} \quad (11)$$

If $|u_1\rangle$ sent by Alice has been received by Bob as a 1, then Eve has

$$\begin{aligned}\rho_{1\to 1}^{\text{Eve}}(|u_1\rangle) &= \mathrm{Tr}\{\sqrt{A_1}|\phi_1\rangle\langle\phi_1|\sqrt{A_1}\} \\ &= (1-Q)|e_1\rangle\langle e_1|.\end{aligned} \quad (12)$$

Furthermore, if Bob has received a 1, whereas $|u_0\rangle$ has been sent by Alice, then the ancilla state will also be "incorrect:"

$$\begin{aligned}\rho_{0\to 1}^{\text{Eve}}(|u_0\rangle) &= \mathrm{Tr}\{\sqrt{A_1}|\phi_1\rangle\langle\phi_1|\sqrt{A_1}\} \\ &= Q|e_1\rangle\langle e_1|,\end{aligned} \quad (13)$$

$$\begin{aligned}\rho_{1\to 0}^{\text{Eve}}(|u_0\rangle) &= \mathrm{Tr}\{\sqrt{A_0}|\phi_1\rangle\langle\phi_1|\sqrt{A_1}\} \\ &= Q|e_0\rangle\langle e_0|.\end{aligned} \quad (14)$$

Table 1 shows the combinations of the bits sent and received by Alice and Bob, respectively, and the ancilla states stored by Eve that are possible after the communication over the quantum channel has been completed and the inconclusive outcomes have been rejected by public discussion.

Thus, Q is the fraction of bit errors in the total number of received bits, and the ancilla states kept by Eve in a quantum memory are in one-to-one correspondence with the classical bits received by Bob: $0 \longleftrightarrow |e_0\rangle$ and $1 \longleftrightarrow |e_1\rangle$.

3. SECRET KEY LENGTH IN THE SHANNON LIMIT

After the inconclusive outcomes have been rejected, Alice and Bob formally have a binary symmetric channel characterized by the conditional probabilities

Table 1. Correspondence between Alice's and Bob's bit strings and the quantum states stored by Eve

Alice	0	1	…	0	0	1
Bob	0	1	…	1	0	0
Eve	$\|e_0\rangle$	$\|e_1\rangle$	…	$\|e_1\rangle$	$\|e_0\rangle$	$\|e_0\rangle$
Error position	…	…	…	x	…	x

$$\begin{aligned}\Pr(0|0) &= \Pr(1|1) = 1-Q, \\ \Pr(1|0) &= \Pr(0|1) = Q\end{aligned} \quad (15)$$

and equal a priori probabilities of input states

$$\Pr(0) = \Pr(1) = \frac{1}{2}.$$

By virtue of the one-to-one correspondence between the bits received by Bob and the ancilla states stored by Eve, the Alice–Eve channel is a quantum binary symmetric one with input states

$$\rho_0 = |u_0\rangle\langle u_0|, \quad \rho_1 = |u_1\rangle\langle u_1|$$

and Eve's output states

$$|e_0\rangle\langle e_0|, \quad |e_1\rangle\langle e_1|.$$

The channel can be described by a superoperator $\mathcal{T}_{AE}[\ldots]$ as a linear, trace-preserving or trace-reducing, completely positive mapping of Alice's input density matrix to Eve's output density matrix. The unitary representation of the superoperator is

$$\begin{aligned}\mathcal{T}_{AE}[\ldots] &= \mathrm{Tr}_B\{\sqrt{A_0}U([\ldots]\otimes|e\rangle\langle e|)U^{-1}\sqrt{A_0}\} \\ &+ \mathrm{Tr}_B\{\sqrt{A_1}U([\ldots]\otimes|e\rangle\langle e|)U^{-1}\sqrt{A_1}\}.\end{aligned} \quad (16)$$

Furthermore,

$$\begin{aligned}\mathcal{T}_{AE}[|u_0\rangle\langle u_0|] &= (1-Q)|e_0\rangle\langle e_0| + Q|e_1\rangle\langle e_1|, \\ \mathcal{T}_{AE}[|u_1\rangle\langle u_1|] &= (1-Q)|e_1\rangle\langle e_1| + Q|e_0\rangle\langle e_0|.\end{aligned} \quad (17)$$

Now, let us find an upper bound for the error rate compatible with secure key distribution between Alice and Bob and evaluate the corresponding key length in the limit of $n \longrightarrow \infty$ by means of random coding in the spirit of Shannon's mathematics. (We note, however, that this approach is obviously impracticable, since it will require exhaustive search over an exponentially large number of code words.) Let the first code word be the bit string sent by Alice. Then, $M-1$ n-bit code words are randomly generated. All of the M code words

are sent to Bob over a public classical channel. We denote them by

$$w^{(1)}, w^{(2)}, \ldots, w^{(M)},$$

where

$$w^{(j)} = (j_1, j_2, \ldots, j_n), \quad j_k = 0, 1.$$

(They are obviously known to Eve.)

We note that the total number of code words must be sufficiently redundant for Bob to be able to identify the first code word (the bit string sent by Alice over a quantum channel) with probability one. By the Shannon coding theorem, if the number of code words is

$$\begin{aligned} M &< 2^{n[C_{AB}(Q)-\delta]}, \quad \delta \longrightarrow 0, \\ C_{AB}(Q) &= 1 + Q\log Q + (1-Q)\log(1-Q), \end{aligned} \tag{18}$$

where $C_{AB}(Q)$ is the capacity of a binary symmetric classical channel, then the probability of decoding averaged over the ensemble of code words tends to unity [11, 12]. Comparing the received bit string with all of the code words, Bob finds one that is Hamming closest to the string (i.e., differing from the output string by the minimal number of bits) and identifies the bit string sent by Alice with probability one. Then, he corrects errors in the received string to obtain a raw key. The probability of decoding error averaged over the ensemble of code words conditioned on (18) is

$$\begin{aligned} P_e(n, M) &= \frac{1}{M}\sum_{j=1}^{M}[1 - \Pr(w_B^j | w^j)] \\ &\le \varepsilon + (M-1)2^{-n[C_{AB}(Q)-\delta]} < \varepsilon(n, M) \longrightarrow 0, \end{aligned} \tag{19}$$

where $\Pr(w_B^j|w^j)$ is the conditional probability of decoding the code word w^j. In other words, for any $\varepsilon(n, M)$, there exists an arbitrarily small δ such that (19) holds for $n < N$.

The use of M code words randomly sampled from the 2^n-dimensional space of n-bit strings means that the Hamming distance between the nearest code words is slightly larger than the number Qn of bit errors, which makes it possible for Bob to decode the received string into the input code word with probability one.

Eve can perform either an individual measurement on each ancilla state stored in a quantum memory or a collective measurement on all (mutually nonorthogonal) stored ancilla states (see Table 1). In the case of entangled translucent eavesdropping, the error of distinction between two nonorthogonal states is minimized by performing an optimal individual measurement represented as the partition of the identity operator [13]

$$I = \mathcal{M}_0 + \mathcal{M}_1, \quad \mathcal{M}_{0,1} = |m_{0,1}\rangle\langle m_{0,1}|, \tag{20}$$

where $|m_{0,1}\rangle$ denotes the eigenvectors of the operator $|e_0\rangle\langle e_0| - |e_1\rangle\langle e_1|$. The corresponding probability of error is

$$\varepsilon(Q) = \frac{1}{2}[1 - \sqrt{1 - |\langle e_0|e_1\rangle|^2}] = \frac{1}{2}(1 - \cos 2\gamma), \tag{21}$$

$$\sin 2\gamma = \frac{\sqrt{1-(1-2Q)^2} - \sin 2\alpha}{\sin 2\alpha[\sqrt{1-(1-2Q)^2}\sin 2\alpha - 1]}.$$

These results are obtained by using the unitarity and normalization conditions for the functions $|\phi_{0,1}\rangle$ (defined by (5) and (6)), which lead to the relations

$$\begin{aligned} \sin 2\alpha &= 2ab + (a^2 + b^2)\sin 2\alpha \sin 2\gamma, \\ a^2 + b^2 + 2ab\sin 2\alpha \sin 2\gamma &= 1. \end{aligned} \tag{22}$$

Note that a similar error corresponds to the case when Eve performs measurement (20) on each transmission without using quantum memory. However, further analysis is more transparent when quantum memory is employed.

Before Bob corrects errors after Eve's individual measurements have been completed, the error rates in the Alice–Bob, Bob–Eve, and Alice–Eve channels are, respectively,

$$\Pr\{b_A(i) = b_B(i)\} = 1 - Q, \tag{23}$$

$$\Pr\{b_B(i) = b_E(i)\} = 1 - \varepsilon(Q), \tag{24}$$

$$\Pr\{b_A(i) = b_E(i)\} = (1-Q)(1-\varepsilon) + Q\varepsilon. \tag{25}$$

Therefore, Eve's error rate is

$$E(Q) = 1 - (1-Q)(1-\varepsilon(Q)) + Q\varepsilon(Q). \tag{26}$$

Using (21), we rewrite (26) as

$$E(Q) = Q\cos^2\gamma + (1-Q)\sin^2\gamma.$$

If $E \le Q$, then Eve can decode the bit string sent by Alice with probability one. The equation

$$Q = E(Q)$$

determines the highest error rate compatible with secure key distribution between Alice and Bob.

If the error rate in the Alice–Eve channel is higher than that in the Alice–Bob channel, i.e.,

$$C_{AB}(Q) > C_{AE}(E(Q)),$$

then the length of a secret key that can be shared by Alice and Bob in the limit of $n \longrightarrow \infty$ is

$$n_{\text{secret}} \longrightarrow n. \tag{27}$$

The probability of Eve's error averaged over all code words is

$$\begin{aligned} P_e(n, M) &= \frac{1}{M}\sum_{j=1}^{M}[1 - \Pr(w_E^j | w^j)] \\ &\geq 1 - O(1)2^{-n[C_{AB}(Q) - C_{AE}(E(Q))]} \longrightarrow 1, \\ &n \longrightarrow \infty, \\ C_{AE}(E(Q)) &= 1 + E(Q)\log E(Q) \\ &+ (1 - E(Q))\log(1 - E(Q)). \end{aligned} \tag{28}$$

This expression is actually a corollary to a theorem on error rate for transmission at a data rate higher than the capacity of a memoryless classical communication channel [11, 12, 14]. In other words, relation (28) follows from the "strong inverse" to the Shannon coding theorem, where "strong inversion" means that the error rate tends to unity when the data rate exceeds the channel capacity. The case when a nonzero upper bound for error rate does not approach unity with increasing n corresponds to the "weak inverse" to the coding theorem, as in collective measurements.

When Eve undertakes a collective attack, the quantum Alice–Eve channel is formally described by the superoperator $\mathcal{T}_{AE}[\ldots]$ that maps Alice's input density matrix to Eve's output density matrix as in (16) and (17). Eve's measurements are represented by outcome operators making up a partition of the identity operator in the span of ancilla states (see [15–18] for details):

$$I = \sum_{k=1}^{M} \chi_{w^k}, \tag{29}$$

$$\begin{aligned} \chi_{w^k} &= \left(\sum_{l=1}^{M} \mathcal{P}\mathcal{P}_{w^l}\right)^{-1/2} \\ &\times \mathcal{P}\mathcal{P}_{w^k}\mathcal{P}\left(\sum_{l=1}^{M} \mathcal{P}\mathcal{P}_{w^l}\mathcal{P}\right)^{-1/2}, \end{aligned} \tag{30}$$

where $\mathcal{P}$ is the projector onto the typical subspace of the density matrix

$$\begin{aligned} &\left(\frac{1}{2}\mathcal{T}_{AE}[|u_0\rangle\langle u_0|] + \frac{1}{2}\mathcal{T}_{AE}[|u_1\rangle\langle u_1|]\right)^{\otimes n} \\ &= \left(\frac{1}{2}(|e_0\rangle\langle e_0| + |e_1\rangle\langle e_1|)\right)^{\otimes n}, \end{aligned} \tag{31}$$

and $\mathcal{P}_{w^k}$ is the projector onto the typical subspace of the density matrix corresponding to the code word w^k,

$$\begin{aligned} \rho_{w^k} &= \rho_{i_1} \otimes \rho_{i_2} \otimes \ldots \rho_{i_n}, \\ i_k &= 0, 1, \quad k = 1, 2, \ldots, M, \end{aligned} \tag{32}$$

$$\rho_{i_k} = \begin{cases} (1 - Q)|e_0\rangle\langle e_0| + Q|e_1\rangle\langle e_1|, & i_k = 0, \\ (1 - Q)|e_1\rangle\langle e_1| + Q|e_0\rangle\langle e_0|, & i_k = 1. \end{cases} \tag{33}$$

The projector $\mathcal{P}_{w^k}$ is constructed as the spectral projector corresponding to the product $\Lambda_I = \lambda_{i_1} \cdot \lambda_{i_2} \ldots \lambda_{i_n}$ of eigenvalues of ρ_{w^k} such that

$$e^{-n[\bar{H}(\rho) + \delta]} < \Lambda_I < e^{-n[\bar{H}(\rho) - \delta]}, \tag{34}$$

where

$$\begin{aligned} \bar{H}(\rho) &= \frac{1}{2}[S((1 - Q)|e_0\rangle\langle e_0| + Q|e_1\rangle\langle e_1|) \\ &+ S((1 - Q)|e_1\rangle\langle e_1| + Q|e_0\rangle\langle e_0|)], \end{aligned} \tag{35}$$

with

$$S(\rho) = -\mathrm{Tr}\{\rho\log(\rho)\} \tag{36}$$

being the von Neumann entropy.

According to [15],

$$\begin{aligned} \mathcal{P}_{w^k} &\leq \rho_{w^k} e^{n[\bar{H}(\rho) + \delta]}, \\ \mathbf{E}\{\mathrm{Tr}\{\rho_{w^k}(I - \mathcal{P}_{w^k})\}\} &\leq \varepsilon, \end{aligned} \tag{37}$$

where $\mathbf{E}$ denotes an average over all random codes. The probability of error of decoding by means of Eve's outcome operators (30) is

$$\begin{aligned} P_e(n, M) &= \frac{1}{M}\sum_{k=1}^{M}[1 - \Pr(w^k = w_E^k)], \\ \Pr(w^k = w_E^k) &= \mathrm{Tr}\{\rho_{w^k}, \chi_{w^k}\}. \end{aligned} \tag{38}$$

It was shown in [15, 18] that if

$$M \leq 2^{n[\bar{C}_{\mathcal{T}_{AE}} - \delta]}, \tag{39}$$

then

$$P_e(n, M) < 8\varepsilon + (M - 1)2^{-n[\bar{C}_{\mathcal{T}_{AE}} - 3\delta]} \longrightarrow 0. \tag{40}$$

The exponent in (39) contains the classical information capacity of a noisy quantum channel [15, 18] described

by superoperator (16), (17):

$$\begin{aligned}\bar{C}_{\mathcal{T}_{AE}} &= S\left(\frac{1}{2}\mathcal{T}_{AE}[|u_0\rangle\langle u_0|] + \frac{1}{2}\mathcal{T}_{AE}[|u_1\rangle\langle u_1|]\right)\\ &- \frac{1}{2}S(\mathcal{T}_{AE}[|u_0\rangle\langle u_0|]) - \frac{1}{2}S(\mathcal{T}_{AE}[|u_1\rangle\langle u_1|])\\ &= -\frac{1-\sin 2\gamma}{2}\log\frac{1-\sin 2\gamma}{2} - \frac{1+\sin 2\gamma}{2}\\ &\times \log\frac{1+\sin 2\gamma}{2} + \Lambda_-\log\Lambda_- + \Lambda_+\log\Lambda_+,\end{aligned} \tag{41}$$

where

$$\Lambda_\pm = \frac{1}{2}(1 \pm \sqrt{1-4\mathcal{L}}), \tag{42}$$

$$\begin{aligned}\mathcal{L} &= [(1-Q)\cos^2\gamma + Q\sin^2\gamma]\\ &\times [(1-Q)\sin^2\gamma + Q\cos^2\gamma] - \cos^2\gamma\sin^2\gamma.\end{aligned} \tag{43}$$

Secure key distribution implies that Eve cannot decode the transmitted code word. If the number M of code words used in transmissions over the Alice–Bob channel is smaller than that in the Alice–Eve channel, then secure key distribution is impossible a priori; i.e., Eve decodes any bit sequence with probability one. If Eve's instrumentation can be used to perform only individual measurements, while Alice and Bob use a random code to correct output errors, then secure key distribution is possible if

$$C_{AB}(Q) > C_{AE}(E(Q)) \tag{44}$$

combined with (28) holds (see also [19]). This means that Bob can correct all errors in the received bit string with probability one, while Eve's error rate tends to unity (in accordance with the "strong inverse" to the Shannon coding theorem), and the entire n-bit string can be used by Alice and Bob as a secret key.

Using instrumentation for collective measurements described by (29) and (30), Eve can decode the bit string sent by Alice with probability one if

$$M < 2^{n[\bar{C}_{\mathcal{T}_{AE}} - \delta]}.$$

Accordingly, the critical condition for secure key distribution is

$$C_{AB}(Q) = \bar{C}_{\mathcal{T}_{AE}}, \tag{45}$$

if Eve can perform collective measurements.

When $M > 2^{n\bar{C}_{\mathcal{T}_{AE}}}$, Eve cannot decode the bit string sent by Alice even by means of collective measurements, whereas Bob can correct all errors in an n-bit string with probability one if

$$M < 2^{n[C_{AB}(Q) - \delta]}.$$

The corresponding probability of Eve's decoding error averaged over the ensemble of code words is

$$P_e(n, M = 2^{nC_{AB}(Q)}) > C_{AB}(Q) - \bar{C}_{\mathcal{T}_{AE}}; \tag{46}$$

i.e., the "weak inverse" to the coding theorem applies. In contrast to the case of classical or individual quantum measurement, Eve's error rate tends to unity slower than any exponential of

$$n[C_{AB}(Q) - C_{AE}(E(Q))].$$

When Eve performs individual (essentially, classical) measurements, if

$$C_{AB}(Q) > C_{AE}(E(Q)),$$

then she fails to decode any n-bit string with probability one in the limit of large n, and the entire string can be used by Alice and Bob as a secret key.

In the case of a collective attack (an essentially quantum measurement), the probability of Eve's decoding error per code word is not lower than $C_{AB}(Q) - \bar{C}_{\mathcal{T}_{AE}}$, and the corresponding probability of decoding is

$$1 - [C_{AB}(Q) - \bar{C}_{\mathcal{T}_{AE}}].$$

A conservative estimate (even slightly biased in favor of Eve) for the number of code words known to Eve is $2^{n\bar{C}_{\mathcal{T}_{AE}}}$ if their total number is M. Since Bob can correct all errors in a received n-bit string if

$$M < 2^{nC_{AB}(Q)},$$

we can use the space dimension 2^n as an estimate for the number of code words, and then the fraction of all bits in the entire space known to Eve is $2^{n\bar{C}_{\mathcal{T}_{AE}}}/2^n$. Since the probability distribution of randomly sampled code words is uniform, the fraction of bits in each particular string known to Eve does not exceed $n\bar{C}_{\mathcal{T}_{AE}}$, and the remaining $n[1 - \bar{C}_{\mathcal{T}_{AE}}]$ bits can be used as secret ones by the legitimate partners.

Figure 2 illustrates the variation of Bob's error rate Q for several values of the angle of overlap between $|u_0\rangle$ and $|u_1\rangle$.

Note that the fact that Eve's error rate tends to unity cannot guarantee secure key distribution. Indeed, (27) and (28) are asymptotic expressions valid in the limit of

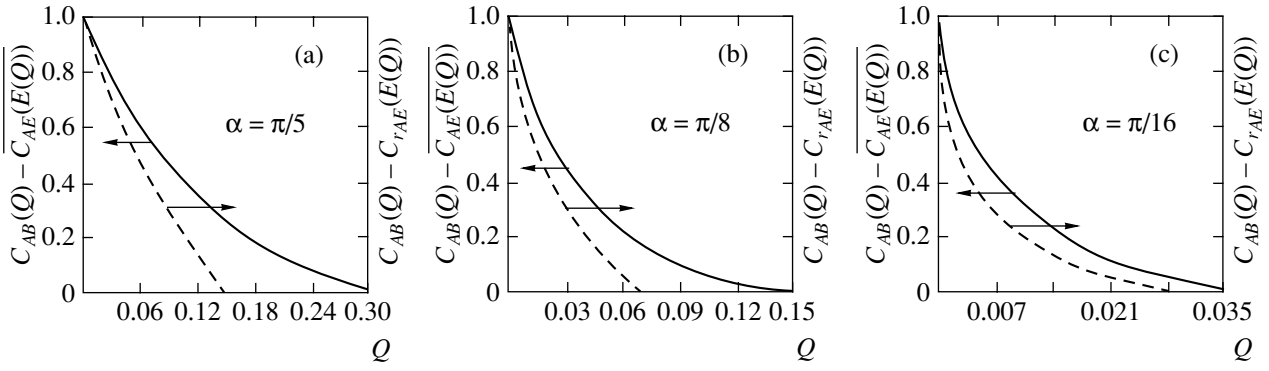

Fig. 2.

$n \longrightarrow \infty$, whereas any actual bit string has a finite length. Secure key distribution requires that the amount of private information contained in an n-bit string w_E singled out by Eve from the ensemble of all possible strings W_E relative to the set of bit strings shared by Alice and Bob must be an exponentially decreasing function of the security parameter s,

$$I(W;\, W_E = w_E) < \frac{2^{-s}}{\ln 2}. \tag{47}$$

In terms of conditional entropy, this requirement is expressed as

$$H(W|W_E = w_e) > n - \frac{2^{-s}}{\ln 2}. \tag{48}$$

Thus, the n-bit reconciled key shared by Alice and Bob still cannot be used as a secret key, because the probability given by (28) for a finite n may not be exponentially small as a function of s. Recall that

$$I(W;\, W_E = w_E) = 0$$

is the probability of a correct guess.

At the last stage, a final key is generated by compressing the n-bit strings shared by Alice and Bob into a string of length r, and it is guaranteed that conditions (47) and (48) are satisfied. The hashing procedure (privacy amplification) makes use of a theorem [20] that relies on the properties of the second-order Renyi entropy and universal hash functions [21].

A 2-universal hash function is a function

$$g(x) : \{0,1\}^n \longrightarrow \{0,1\}^r (X \longrightarrow Y)$$

such that, for arbitrary

$$x_1, x_2 \in X, \quad x_1 \neq x_2,$$

the probability that $y_1 = y_2$ ($y_1 = g(x_1)$ and $y_2 = g(x_2)$) is not higher than

$$\frac{1}{|Y|} = \frac{1}{2^r}$$

($|Y|$ is the volume of the space of r-bit strings). The set of random functions $g \in G$ is a set of 2-universal hash functions if a random sample from a uniform distribution on G contains no more than $|G|/|Y|$ functions for which collision of their values corresponding to distinct values of the argument is possible. In other words, if a hash function is randomly sampled from a uniform distribution, then the probability of its equal values corresponding to two distinct n-bit strings does not exceed 2^{-r}.

With regard to cryptography (including quantum cryptography), this means that the r-bit string obtained by compressing Eve's n-bit string different from that shared by the legitimate partners with a randomly sampled hash function is identical to the r-bit string shared by Alice and Bob with a probability not higher than 2^{-r}.

By the privacy amplification theorem, the parameters n, r, and s can be related to the properties of universal hash functions by means of the second-order Renyi entropy defined in terms of the conditional probabilities $\Pr(W|w_E)$. To do this, we need the following definitions. Suppose that a random variable $x \in X$ is characterized by the probability distribution $P_X(x)$ on X. The collision probability is defined as the probability that equal values of x are obtained in two independent trials:

$$P_c(X) = \sum_{x \in X} P_X^2(x). \tag{49}$$

The second-order Renyi entropy is defined as

$$R(X) = -\log P_c(X). \tag{50}$$

Conditional probability distributions are defined similarly:

$$P_c(X|Y = y) = \sum_{y \in Y} P_{X|Y=y}^2(y), \tag{51}$$

$$R(X|Y = y) = -\log P_c(X|Y = y). \tag{52}$$

The mean value of the Renyi entropy is

$$R(X|Y) = \sum_{y \in Y} P_Y(y) R(X|Y = y). \quad (53)$$

To evaluate the amount of private information known to Eve, the following relations are required:

$$R(X) \le H(X),$$
$$H(X) = -\sum_{x \in X} P_X(x) \log P_X(x). \quad (54)$$

For any joint distribution, it holds that

$$R(X|Y) \le H(X|Y). \quad (55)$$

The following theorem plays a fundamental role in cryptography.

Privacy amplification theorem [20]. *If $x \in X$ is a random variable characterized by the probability distribution $P_X(x)$, $R(X)$ is second-order Renyi entropy, $g \in G$ is a random variable with uniform distribution on the set G of 2-universal hash functions $g : X \longrightarrow \{0, 1\}^r$, and $K = G(X)$, then*

$$H(K|G) \ge R(K|G) \ge r - \log(1 + 2^{r - R(X)}) \ge r - \frac{2^{r - R(X)}}{\ln 2}, \quad (56)$$

where $H(K|G) = H(G(X)|G)$ is the mean conditional Shannon entropy.

With regard to quantum cryptography problems, we have the following important corollary to the theorem. Suppose that the joint probability distribution P_{WW_E} is not known in the general case; the Renyi entropy is

$$R(W|W_E = w_E) = c.$$

If Alice and Bob use a randomly sampled hash function from $\{0, 1\}^n \longrightarrow \{0, 1\}^r$ in G with uniform probability distribution to generate a secret key by hashing their shared strings from the set $K = G(X)$, then

$$H(K|G, W_E = w_E) \ge r - \log(1 + 2^{r - c}) \ge z - \frac{2^{r - c}}{\ln 2}. \quad (57)$$

The conditional probability distribution $P_{W|W_E = w_E}(w_E)$ for individual measurement is

$$P_{W|W_E = w_E}(w_E) = (1 - \varepsilon)^{n - d(w_E, W)} \varepsilon^{d(w_E, W)}, \quad (58)$$

where $d(w_E, W)$ is the Hamming distance between the strings W and $W_E = w_E$. The corresponding Renyi entropy is

$$R(W|W_E = w_E) = -n \log[\varepsilon^2 + (1 - \varepsilon)^2], \quad \varepsilon = \frac{1}{2}(1 - \cos 2\gamma). \quad (59)$$

The mutual information known to Eve is

$$I(K; GW_E) = H(K) - H(K|GW_E) \le \frac{2^{-s}}{\ln 2}, \quad (60)$$

where

$$r = c - s = -n \log[\varepsilon^2 + (1 - \varepsilon)^2] - s, \quad (61)$$

and s is a security parameter.

It is essential for further analysis that compression ratio depends on the specific procedure used by the legitimate partners to correct errors in the raw key. The probability $P_{W|W_E = w_E}(w_E)$ introduced above is conditioned on the key reconciliation performed by the legitimate partners. Initially, Alice and Bob have different strings:

$$P_{W_A|W_B = w_B}(w_B) \ne 1.$$

Accordingly, for Eve's initial string,

$$P_{W_A|W_E = w_E}(w_E) \ne 1.$$

After the error correction by public discussion has been completed, $w = w_A = w_B$ and $w_E \ne w$. Moreover, Eve can also use these public transmissions to partially correct errors in her bit string, and the conditional probability corresponding to the resulting string will be different in the general case:

$$P_{W|W_E = w_E}(w_E) \ne P_{W_A|W_E = w_E}(w_E).$$

Therefore, the legitimate partners must not only correct errors, but also evaluate the change in the conditional probability for Eve's knowledge of their private information.

In what follows, we discuss several procedures for correcting errors in the raw key. Eve's conditional information may either change or remain invariant, depending on the procedure employed. Note that the efficiency of a particular procedure depends on the final key length, as well as by the length of the reconciled key. In particular, when the reconciled key obtained by using one error-correction procedure is longer than that obtained by another, the conditional probability for Eve is higher in the former case and a higher compression ratio is required (for a given security parameter), which leads to a shorter final key. It is shown below that this is

the case when the raw keys are reconciled by using classical codes.

4. ERROR CORRECTION IN THE RAW KEY BY BISECTIVE SEARCH (BINARY)

Consider bisective search with subsequent removal of errors [22], which makes it easy to monitor changes in conditional probabilities.

After the quantum part of the protocol has been executed, Alice and Bob have distinct N-bit strings in the general case. The states stored in Eve's quantum memory are in one-to-one correspondence with the classical bits received by Bob. The exact number of single-bit errors in the string received by Bob is not known in advance. Alice or Bob randomly samples a substring of length $N/2$ and announces the bits contained in the sampled positions. Comparing these bits, Alice and Bob estimate the error rate Q. Then, they execute the following procedure for locating bit errors.

1. Using the estimated error rate, Bob calculates the average size L of a block containing one single-bit error.

2. Alice partitions her input string into blocks of size L, calculates the parity bits for each block, and sends the results to Bob.

3. Bob compares the received parity bits with the ones corresponding to each respective block in his output string. If the parity bits disagree (the block contains an odd number of bit errors), then the bisective search described below is performed.

4. The disclosure of a parity bit for each block leaks exactly one bit of secret information. Alice and Bob take out one bit from each block to recover from the leakage.

5. Alice performs a random permutation of bits in the remaining string and informs Bob about it. This operation does not change the amount of mutual information known to Eve. The process is then repeated, starting from step 1.

6. If no error is revealed after repeating the above steps 20 to 30 times, then Alice's and Bob's keys are likely to be identical.

7. Alice sends a randomly generated string to Bob. Having compared the parity bits for the sent and received strings, Alice and Bob exchange the results and simultaneously take out one bit. If the parity bits coincide after repeating this step M times, then Alice's and Bob's keys are identical with probability $1 - 2^{-M}$. The choice of the parameter M is dictated by technical requirements.

Since the average number of single bit errors per block must not exceed one, the average number of single-bit errors in a string of length $N/2$ characterized by an error rate Q is $N_{\text{err}} \approx QN/2$, and the block size is $L = p(N/N_{\text{err}})$, where $0 < p < 1$. According to numerical experiment, $p = 0.5$ is a good estimate; i.e., there is one single-bit error per two blocks.

The following procedure of bisective search is executed to locate errors in the blocks for which the parity bits calculated by Alice and Bob disagree.

1. Bob bisects each block to be searched (increasing the resulting noninteger subblocks) and sends the parity bits corresponding to the first subblock to Alice.

2. If the parity bits disagree, then step 1 is executed.

3. If the parity bits corresponding to the first subblock agree, then Bob executes step 1 for the second subblock. All disclosed bits are removed.

4. If the subblock size is less than 4 bits, then the subblock is removed and the search is terminated. The revealed errors are also removed.

The results of calculations are illustrated by Figs. 3–5.

5. RELATIONSHIP BETWEEN THE FINAL KEY LENGTH AND THE QUANTUM PART OF THE PROTOCOL

After the error correction has been completed, Alice and Bob share identical strings. Using the information extracted from public transmissions, Eve removes the states corresponding to the bits removed by Alice and Bob from the ancilla, but does not perform any measurement on the ancilla states. Since the bits in Bob's

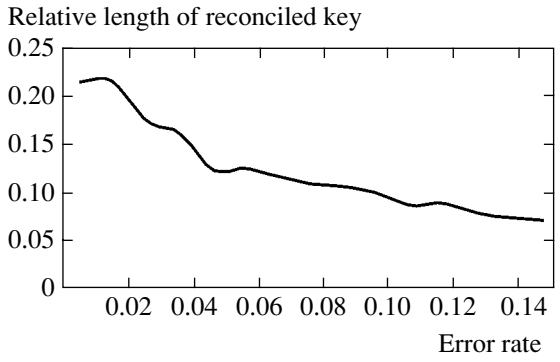

Fig. 3.

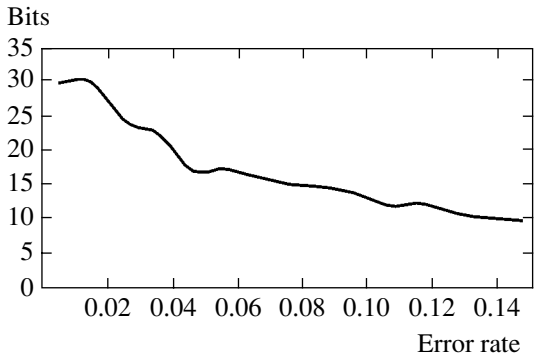

Fig. 4.

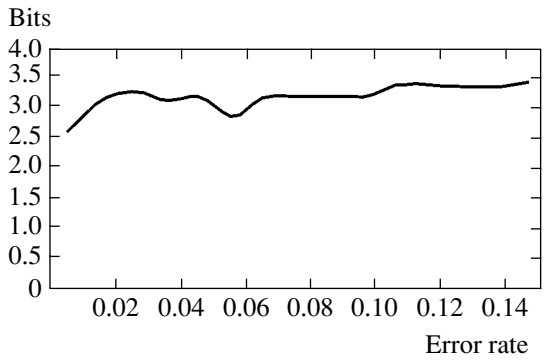

Fig. 5.

string are in one-to-one correspondence with the ancilla states before the error correction, the correspondence after the error correction will be as shown in Table 2.

The one-to-one correspondence holds, but since Bob's string now contains no errors, there are no incorrect quantum states in Eve's quantum memory; i.e., the probability of any *i*th bit shared by Alice and Bob conditioned on the corresponding ancilla state is

$$\Pr(i||e_i\rangle) = 1 \quad (i = 0, 1).$$

After Eve has obtained a certain bit string by performing optimal measurement (20), (21) on the ancilla states, the conditional probability is as given by (58), where n should be interpreted as the length of the reconciled key shared by Alice and Bob. The compression ratio for the final secret key is given by (61), whereas the relative amount of private information known to Eve does not exceed that given by (60).

Collective measurement on all states stored in Eve's quantum memory is not forbidden by any fundamental principle of quantum mechanics. The resulting amount of mutual information gained by Eve can be greater than in the case of individual measurement.

After the keys have been reconciled, the legitimate partners have identical bit strings, and Eve can formally treat the Alice–Eve and Bob–Eve channels as quantum binary ones carrying classical information. Indeed, since

$$0 \longleftrightarrow |e_0\rangle, \quad 1 \longleftrightarrow |e_1\rangle,$$

Eve can interpret every quantum state $|e_0\rangle$ and $|e_1\rangle$ as corresponding to (encoding) a 0 or 1 randomly generated by Alice or Bob with probability 1/2 and sent to her over a noise-free quantum channel. Eve wants to decode every transmission from Alice or Bob that she has intercepts by quantum measurement with minimum error, not knowing the classical bit string actually transmitted.

Table 2. Correspondence between Alice's and Bob's bit strings and the quantum states stored by Eve after error correction

Alice	0	1	…	0
Bob	0	1	…	0
Eve	$\|e_0\rangle$	$\|e_1\rangle$	…	$\|e_0\rangle$

Each code word in the set $\{w^1, w^2, \ldots, w^M\}$ encoded into quantum states by Alice and Bob is an n-bit string $w^i = \{i_1, i_2, \ldots, i_n\}$ ($i_k = 0, 1$), where n is the length of the reconciled key shared by Alice and Bob and $M = 2^n$. However, we tentatively treat the number M as arbitrary.

If the quantum states were orthogonal, Eve would be able to distinguish between them, i.e., record a bit string identical to that received by Bob. However, Eve's measurement error is higher than Bob's because of the indistinguishability of nonorthogonal states. Whereas the outcome of any individual quantum measurement performed by Eve can be interpreted as a 0 or 1, Eve's error rate will be lower if she uses the operators described below to perform collective measurements [15].

Since each particular code word is generated randomly and independently, the corresponding probability is (see details in [15, 18])

$$\Pr\{w = (i_1, i_2, \ldots, i_n)\} = p_{i_1}p_{i_2}\ldots p_{i_n} = \frac{1}{2^n}, \qquad p_{i_k} = \frac{1}{2}, \tag{62}$$

and the expected value of $\rho_{w^i} = |\psi_{w^i}\rangle\langle\psi_{w^i}|$ is

$$\mathbf{E}(\rho_{w^i}) = \sum_{i_1, i_2, \ldots, i_n} p_{i_1}p_{i_2}\ldots p_{i_n}|e_{i_1}\rangle\langle e_{i_1}| \otimes \ldots \otimes |e_{i_n}\rangle\langle e_{i_n}| = \rho_E^{\otimes n}, \tag{63}$$

where

$$\rho_E = \frac{1}{2}|e_0\rangle\langle e_0| + \frac{1}{2}|e_1\rangle\langle e_1| \tag{64}$$

is the density matrix describing the ancilla state.

To decode a code word (i.e., map quantum states to a bit string), Eve makes use of the collective outcome operators

$$\chi_k = |\tilde{\psi}_{w^k}\rangle\langle\tilde{\psi}_{w^k}|, \quad |\tilde{\psi}_{w^k}\rangle = \mathcal{P}|\psi_{w^k}\rangle, \tag{65}$$

with $\mathcal{P}$ denoting the projector onto the typical subspace

of the density matrix $\rho_E^{\otimes n}$:

$$\mathcal{P} = \sum_{J \in B} |\lambda_J\rangle\langle\lambda_J|, \qquad |\lambda_J\rangle = |\lambda_{j_1}\rangle \otimes |\lambda_{j_2}\rangle \otimes \ldots \otimes |\lambda_{j_n}\rangle, \tag{66}$$

where $|\lambda_{j_k}\rangle$ and $\lambda_J = \lambda_{j_1}\lambda_{j_2}\ldots\lambda_{j_n}$ denote the eigenstates and eigenvalues of ρ_E, respectively. The typical subspace is defined as

$$B = \{J : 2^{-n[H(\rho_E)+\delta]} < \lambda_J < 2^{-n[H(\rho_E)-\delta]}\}, \tag{67}$$

and it holds that (see [15])

$$\left\|\rho_E^{\otimes n}\mathcal{P}\right\| < 2^{-n[H(\rho_E)-\delta]}, \quad \mathrm{Tr}\{\rho_E^{\otimes n}(1-\mathcal{P})\} < \varepsilon. \tag{68}$$

Here, the von Neumann entropy is equal to the classical capacity of a quantum binary channel [15]:

$$\begin{aligned}\bar{C}(\varepsilon(Q)) &= H(\rho_E) \\ &= -\mathrm{Tr}\{\rho_E \log \rho_E\} = -\lambda_1\log\lambda_1 - \lambda_2\log\lambda_2 \\ &= -\frac{1-\sin 2\gamma}{2}\log\frac{1-\sin 2\gamma}{2} - \frac{1+\sin 2\gamma}{2} \\ &\quad \times \log\frac{1+\sin 2\gamma}{2},\end{aligned} \tag{69}$$

where

$$\lambda_{1,2} = \frac{1 \pm \sin 2\gamma}{2}$$

are the eigenvalues of ρ_E.

Eve's attempt to decode M code words generated by Alice is characterized by the error [15]

$$p_E(n, M) = \frac{1}{M}\sum_{k=1}^{M}[1 - \Pr(w^k = w_E^k)], \qquad \Pr(w^k = w_E^k) = \langle\psi_{w^k}|\chi_k|\psi_{w^k}\rangle. \tag{70}$$

Therefore (see details in [15]),

$$\begin{aligned}p_E(n, M) &\le 2\mathrm{Tr}\{\rho^{\otimes n}(1-\mathcal{P})\} \\ &\quad + (M-1)\mathrm{Tr}\{(\rho^{\otimes n}\mathcal{P})^2\} \\ &\le 2\varepsilon + (M-1)2^{-n[H(\rho_E)-\delta]}.\end{aligned} \tag{71}$$

If Eve performed optimal individual measurements as defined above, then $\bar{C}(\varepsilon(Q))$ would be equal to the one-shot classical capacity of a quantum binary channel [15],

$$\begin{aligned}C(\varepsilon(Q)) &= \frac{1}{2}\Big[(1+\sqrt{1-\sin^2 2\gamma}) \\ &\quad \times \log\left(1+\sqrt{1-\sin^2 2\gamma}\right) \\ &\quad + \left(1-\sqrt{1-\sin^2 2\gamma}\right)\log\left(1-\sqrt{1-\sin^2 2\gamma}\right)\Big].\end{aligned} \tag{72}$$

Note that $C(\varepsilon(Q))$ cannot exceed $\bar{C}(\varepsilon(Q))$ and is actually equal to the capacity $C_{AE}(E(Q))$ given by (28).

These results are interpreted as follows. Suppose that the number of randomly generated n-bit strings is

$$M < 2^{n[\bar{C}((Q))-\delta]}, \quad \delta \longrightarrow 0,$$

and one of these is the string encoded by Alice or Bob into quantum states $|e_{0,1}\rangle$ by using a published substitution table. Then, Eve can construct outcome rules for collective measurements and identify the encoded string among these M strings with probability one.

However, Eve cannot identify the string encoded by Alice or Bob if

$$M < 2^{n[\bar{C}(\varepsilon(Q))-\delta]},$$

$$M = 2^{nR} > 2^{n[\bar{C}(\varepsilon(Q))-\delta]}, \quad 0 < R < 1.$$

After the error correction has been completed, the number of code words known to Eve is the total number of n-bit strings, $M = 2^n$. Therefore, the error rate associated with her measurement is

$$P_e(n, M) > R - \bar{C}(\varepsilon(Q)), \quad M > 2^{n[\bar{C}(\varepsilon(Q))-\delta]}. \tag{73}$$

This result is the quantum analog of the "weak inverse" to the coding theorem.

The error rate associated with classical (as well as individual quantum) measurements tends to unity exponentially in $n[C_{AB}(Q) - C_{AE}(E(Q))]$ (see (28)), in accordance with the "strong inverse" mentioned above.

In other words, if

$$C_{AB}(Q) > C_{AE}(E(Q))$$

and the number of code words used and announced by Alice is

$$M < 2^{n[C_{AB}(Q)-\delta]},$$

where n is sufficiently large, then Bob will decode n-bit string encoded by Alice with probability one, whereas Eve will fail to decode it with probability one by performing individual measurements (by virtue of (23) and (26) under condition (27)); i.e., the n-bit string can be used by Alice and Bob as a secret key.

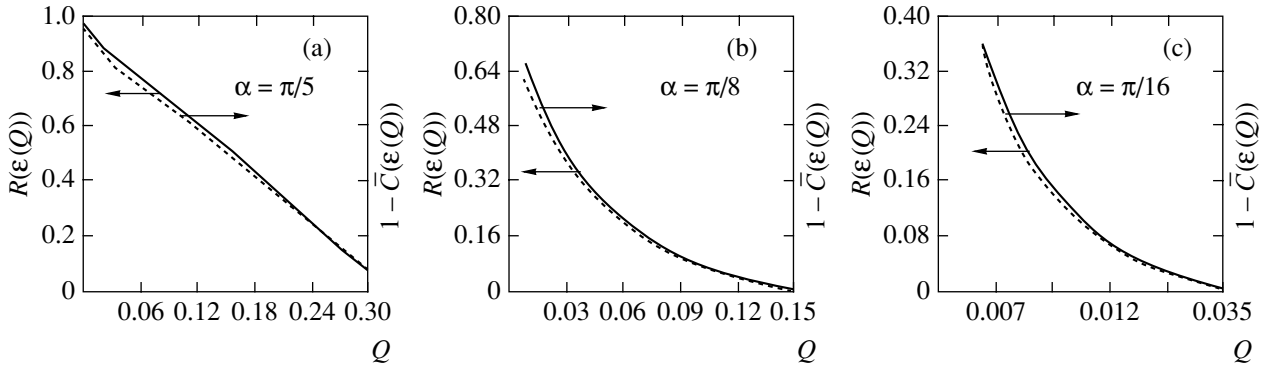

Fig. 6.

Whereas Eve can identify the string encoded by Alice among M ones by collective measurement if

$$M < 2^{n[\bar{C}(\varepsilon(Q)) - \delta]},$$

the probability of decoding in the case of

$$M = 2^n > 2^{n[\bar{C}(\varepsilon(Q)) - \delta]}$$

will be

$$1 - P_e(n, M) < 1 - [R - \bar{C}(\varepsilon(Q))].$$

If $R = 1$, then a conservative estimate (biased in favor of Eve) for the number of bits in the string known to Eve cannot exceed

$$n\bar{C}(\varepsilon(Q)); \tag{74}$$

i.e., she will fail to identify

$$n[1 - \bar{C}(\varepsilon(Q))] \tag{75}$$

bits in the key. Therefore, Alice and Bob can generate a secret key of length

$$r = n[1 - \bar{C}(\varepsilon(Q))] - s \tag{76}$$

using a universal hash function

$$g : \{0, 1\}^n \longrightarrow \{0, 1\}^r,$$

where s is a security parameter preset by Alice and Bob and n is the length of the key reconciled by using the BINARY protocol. In this case, the relative amount of information contained in the final secret key known to Eve does not exceed the quantity given by (60).

Collective measurements are unfeasible to this day, but we can rely on the conservative estimate above in evaluating the length of the final secret key. If Eve's instrumentation can be used to perform only individual measurements, then the final length of the secret reconciled key is

$$\begin{aligned} r &= nR(\varepsilon(Q)) - s, \\ R(\varepsilon(Q)) &= -\log[\varepsilon^2(Q) + (1 - \varepsilon(Q))^2], \end{aligned} \tag{77}$$

where $R(\varepsilon(Q))$ is the Renyi entropy (cf. (59)).

Figure 6 shows the number of secret-key bits per bit of an n-bit reconciled key if Eve performs either individual or collective measurements and s is set to zero.

The surprising fact that the Renyi entropy calculated for optimal individual measurements is virtually equal to the classical capacity of a quantum communication channel (attainable only in the case of collective measurements) may reflect a hidden fundamental relationship between these quantities.

For example, if $s = 200$ and Bob's error rate is 6%, then a 256-bit final secret key can be generated only if the raw key contains at least 17812 bits. About half of these (8900) are removed after the error rate is evaluated. Error correction leaves 16% of the raw key (about 2850 bits, see Fig. 3). Further compression (by means of a universal hash function) depends on the overlap of nonorthogonal states of the information carrier prepared by Alice. In particular, 16% of the reconciled key is left when $\alpha = \pi/8$ and Bob's error rate is 6% (see Fig. 6b). Thus, the 256-bit string generated by hashing the 2850-bit reconciled key when $s = 200$ is the final secret key, and the relative amount of private information known to Eve does not exceed 2^{-200} according to (60).

Another important parameter of an error-correction procedure is the number of bits transmitted over the public channel per bit of the final secret key, because the authenticity and integrity of data transmitted over a public channel must be guaranteed.

We used two approaches to evaluate this parameter. If a random number generator is available only to Alice, then the information about the bits randomly sampled for estimating the error rate, the parity bits calculated for random strings, etc., must be transmitted between Alice and Bob over a public channel. Figure 4 shows the number of transmitted bits per bit of the final secret key plotted versus the error rate Q estimated by the legitimate partners. In particular, about 20 bits must be transmitted per bit of the final secret key if $Q = 6\%$.

The other approach is applied when Alice and Bob have a short initial key to be used only in the first ses-

sion. In this case, the necessary amount of information to be transmitted over a public classical communication channel can be reduced if identical pseudo-random number generators can be used by both Alice and Bob. Specifically, we used a pseudo-random number generator based on a cipher standardized in GOST 28147-89 to reduce the required number of transmitted bits to two to three per bit of the final secret key (see Fig. 5).

6. COMBINED PROCEDURE OF ERROR CORRECTION IN THE RAW KEY (CASCADE)

The CASCADE protocol is an iterative procedure including BINARY as a subroutine and using previously gained information at each new step [23].

Again, Alice and Bob disclose part of the raw bit sequence to estimate the error rate Q for calculating the block size to be used in the first cycle. The subsequent steps are as follows.

1. A hash function is randomly generated in each cycle. If the size of the current block used in the ith cycle is k_i and the raw key length is N, then

$$F_i(m) : \{1, 2, \ldots, N\} \longrightarrow \left\{1, 2, \ldots, \left[\frac{N}{k_i}\right]\right\}. \quad (78)$$

The jth block obtained by partitioning contains the bits whose positions in the source string,

$$K_j^i = \{m | F_i(m) = j\},$$

correspond to collisions in the hash function. In essence, each particular block is a string of k_i randomly picked bits, which are uniformly distributed when a hash function is used.

2. Alice calculates the parity bit

$$b_j^{\text{parity}} = \bigoplus_{l \in K_j^i} b_l$$

for each block and announces the results. Bob compares the received parity bits with the ones corresponding to each respective block in his output string. If the parity bits disagree, then bisective search is performed. Each located error is labeled with the corresponding index j, but not removed, and the parity bits announced while BINARY is being executed are added up.

3. The set $\mathcal{H}$ is defined as the totality of previously constructed jth blocks, each containing an odd number of single bit errors after the jth one has been labeled.

4. BINARY is executed to find an error in the smallest block in this set, and the located error is labeled by l and corrected, but not removed. The set $\mathcal{B}$ is defined as the totality of previously constructed lth blocks.

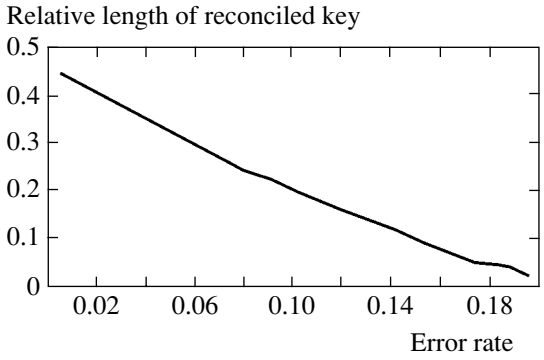

Fig. 7.

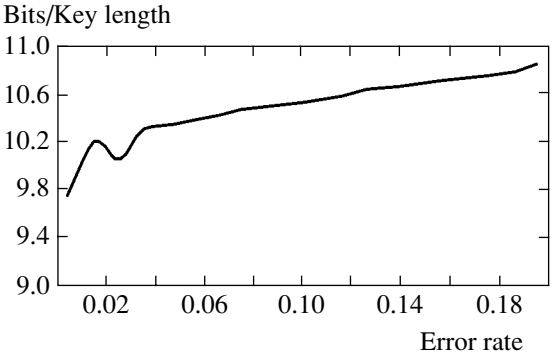

Fig. 8.

5. If the set

$$\mathcal{H}' = (\mathcal{B} \cup \mathcal{H})/(\mathcal{B} \cap \mathcal{H})$$

of blocks containing either an odd number of single-bit errors or none is not empty, then step 1 is executed.

A cycle is terminated when the parity bits of all blocks are equal. After both single-bit errors and announced parity bits have been removed, the amount of information known to Eve is similar to that gained after the BINARY protocol has been executed. Therefore, the compression ratio for random hashing is as in BINARY, but the reconciled key that remains after executing CASCADE is longer. Note that all corrected bits and parity bits can be retained, but then the ensuing analysis of conditional probabilities becomes much more complicated.

The efficiency of CASCADE is twice as high as that of BINARY. For example, the raw key required to generate a 256-bit final key contains 9500 bits if $s = 200$ and the error rate is 6%, which is half the size required in BINARY (see Fig. 7).

If a random number generator can be used only by Alice, then the required number of bits to be transmitted per bit of the final secret key is half that required in BINARY (Fig. 8). When identical pseudo-random number generators can be used by Alice and Bob, this

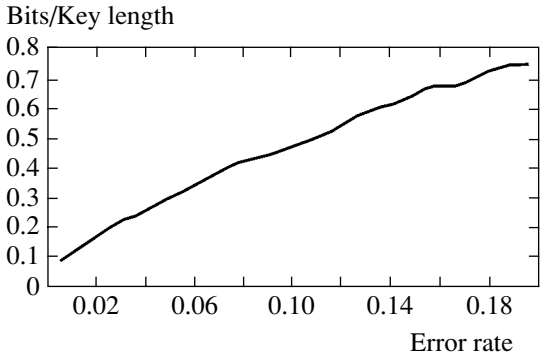

Fig. 9.

number is reduced to a fraction of a bit, i.e., by an order of magnitude as compared to BINARY (see Fig. 9).

7. ERROR CORRECTION IN THE RAW KEY BY MEANS OF BOSE–CHAUDHURI–HOCQUENGHEM CODES

One natural approach to correcting errors in the raw key relies on the use of classical error-correcting codes. We reiterate here that the efficiency of such a code cannot be evaluated without taking into account the quantum part of the protocol, because a classical code applied in quantum cryptography not only corrects errors, but also substantially changes the conditional probabilities for Eve's measurements on the raw key.

In what follows, we discuss key reconciliation by means of Bose–Chaudhuri–Hocquenghem (BCH) codes [24, 25], which constitute a broad class. These codes can be used to correct several single-bit errors per block (code word), which makes it possible to dynamically change a code depending on the estimated error rate Q.

First, we provide the definitions minimally required to analyze t-error correcting BCH codes.

The Galois field $GF(2^n)$ is the vector space of binary words of length n with arithmetic operations modulo 2. A linear code $\mathscr{C}$ spans a linear subspace in $GF(2^n)$. A code is called *cyclic* if the fact that $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1})$ $(c_i = 0, 1)$ is a code word implies that $\mathbf{c}' = (c_1, \ldots, c_{n-1}, c_0)$ is also a code word. A vector from $GF(2^n)$ is conveniently represented by the polynomial of x of degree not higher than $n-1$ whose coefficients are the components of the vector:

$$c(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_{n-1} x^{n-1}. \qquad (79)$$

A cyclic code is defined by a corresponding generator polynomial $g(x)$ of degree $n-k$ (k is the number of information bits), which is a divisor of the polynomial $x^{n-1}-1$:

$$x^{n-1} - 1 = g(x)h(x),$$

where $h(x)$ is a parity check polynomial.

A t-error correcting BCH code is constructed as follows. The primitive element α is defined for the field $GF(2^n)$. (By definition, each element in the field is a power of the primitive element, and $\alpha = \alpha^2$ in $GF(2)$.) The generator polynomial $g(x)$ whose roots make up the set

$$[\alpha, \alpha^2, \ldots, \alpha^3, \ldots, \alpha^{2t}] \qquad (80)$$

is constructed for the preset t. The code-word length n is defined as $n = 2^m - 1$ by choosing a number m. The polynomials $f_j(x)$ of minimal degree whose roots are α^j $(j = 1, 2, \ldots, 2t)$ are found in $GF(2) - GF(2^n)$. The generator polynomial $g(x)$ for a code of length n is calculated as the least common multiple:

$$g(x) = \mathrm{LCM}[f_1(x), f_2(x), \ldots, f_{2t}(x)]. \qquad (81)$$

The Peterson–Gorenstein–Zierler algorithm is used for decoding [24, 25]. Suppose that the input code word is $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1})$, the output

$$\mathbf{y} = \mathbf{c} + \mathbf{e} \qquad (82)$$

contains an error vector $\mathbf{e} = (e_{i_1}, \ldots, e_{i_\nu})$ represented as $e(x) = e_{i_1} x^{i_1} + \ldots + e_{i_\nu} x^{i_\nu}$ $(e_{i_l} = 0, 1)$, and the number ν of single-bit errors is not known in the general case.

A decoding cycle consists of three stages: the calculation of an error syndrome $S(x)$; the construction of an error locator polynomial $\Lambda(x)$; and the calculation of the roots of $\Lambda(x)$, i.e., the determination of e_j and error correction.

Suppose that the code vector is $C(\alpha^j) = 0$ for several α^j:

$$y(\alpha^j) = e(\alpha^j). \qquad (83)$$

After introducing

$$Y_l = e_{i_l}, \quad X_l = \alpha^{i_l}$$

for convenience, an error syndrome is defined componentwise:

$$\begin{aligned} S_1 &= Y_1 X_1 + \ldots + Y_\nu X_\nu, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ S_{2t} &= Y_1 X_1^{2t} + \ldots + Y_\nu X_\nu^{2t}. \end{aligned} \qquad (84)$$

By the definition of the error syndrome, there exists a unique solution to this system of equations (nonlinear in the general case). The error locator polynomial is defined as

$$\Lambda(x) = \Lambda_\nu x^\nu + \ldots + \Lambda_x x + 1, \qquad (85)$$

with roots X_l^{-1} ($l = 1, 2, \ldots, \nu$), i.e.,

$$\Lambda(x) = (1 - xX_1)\ldots(1 - xX_\nu). \quad (86)$$

Its roots can be found if the coefficients Λ_l are known. If S_l are known, then Λ_l can be found by solving a linear system of equations as follows. Since

$$Y_l(X_l^{j+\nu} + \Lambda_l X_l^{j+\nu-1} + \ldots + \Lambda_\nu X_l^j) = 0, \quad (87)$$

the summation of (87) over all l yields

$$\sum_{l=1}^{\nu} Y_l(X_l^{j+\nu} + \Lambda_l X_l^{j+\nu-1} + \ldots + \Lambda_\nu X_l^j) = 0, \quad (88)$$

which can be rewritten as a linear system of equations for Λ_l by setting $j = 1, \ldots, \nu$ and using (84):

$$\Lambda_1 S_{j-\nu-1} + \Lambda_2 S_{j+\nu-2} + \ldots + \Lambda_\nu S_j = -S_{j+\nu}, \quad (89)$$

$$\begin{pmatrix} S_1 & \ldots & S_\nu \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ S_\nu & \ldots & S_{2\nu-1} \end{pmatrix} \begin{pmatrix} -S_{\nu+1} \\ \cdot \\ \cdot \\ -S_{2\nu} \end{pmatrix} = \begin{pmatrix} \Lambda_\nu \\ \cdot \\ \cdot \\ \Lambda_1 \end{pmatrix}. \quad (90)$$

The system of equations (90) is nondegenerate if ν is equal to the (yet unknown) actual number of single-bit errors and degenerate otherwise. Accordingly, if the determinant calculated for $\nu = t$ is zero, then ν is reduced by unity and the calculation is repeated until a nonzero determinant is obtained for ν equal to the number of single-bit errors. Then, system (90) is solved to find a unique set of coefficients Λ_l, and the roots of polynomial (85) are determined by successively calculating its value for each element of the field. Since

$$X_l^{-1} = \alpha^{-i_l},$$

the position i_l of the single-bit error e_{i_l} to be corrected is found as the exponent in

$$X_l \cdot \alpha^{i_l} = 1.$$

The procedure of error correction by means of BCH codes consists of the following steps.

1. An estimate for the error rate Q is obtained by public comparison of about half the transmitted sequence. For the preset code-word length $n = 2^m - 1$, the average number of single-bit errors per code word is estimated as $t = [Q \cdot n]$. The sequence is partitioned into blocks of size n. (We examined the BCH codes corresponding to $n = 2^m - 1 = 15$, 31, 63, 127, and 255.) The generator polynomial $g(x)$ corresponding to n and t is constructed.

2. Alice generates code words. In each block, the first k bits carry information, and the remaining $n - k$ are used as control bits. The inversion of a necessary number of bits is announced, which does not change the private information known to Eve since Alice generates code words by calculating the control bits as functions of the information ones. For the code (n, k) with generator polynomial $g(x)$ in the field $GF(2^m)$, the polynomial representing $n - 2^m - 1$ encoding bits is the remainder of the division of $i(x) \cdot x^{n-k}$ modulo

$$g(x) - c(x) = -R_{g(x)}[x^{n-k} i(x)],$$

where

$$i(x) = i_0 + i_1 x + \ldots + i_{k-1} x^{k-1}, \quad i_l = 0, 1$$

is the information polynomial. The bits i_l ($l = 0, \ldots, k - 1$) are the first bits of the blocks in the bit string sent by Alice.

3. Bob decodes the received code words, and then Alice and Bob remove the control bits in each code word.

Table 3 shows the parameters of only those BCH codes examined in this study for which t lie in the interval corresponding to error rates not higher than $Q = 0.15$ estimated with step $\Delta Q = 0.01$. The extreme right column lists the highest error rate for each particular code.

The polynomials $g(x)$ and $h(x)$ written out below for the BCH codes with 15-, 31-, 63-, 127-, and 255-bit code words correspond to an estimated error rate of about 10%; i.e., these are 2-, 4-, 7-, 13-, and 26-error correcting codes, respectively. For the 2-error correcting BCH code [15, 7],

$$g(x) = x^8 + x^7 + x^6 + x^4 + 1,$$
$$h(x) = x^7 + x^6 + x^4 + 1.$$

For the 4-error correcting BCH code [31, 11],

$$g(x) = x^{20} + x^{18} + x^{17} + x^{13} + x^{10} + x^9 + x^7 + x^6 + x^4 + x^2 + 1,$$
$$h(x) = x^{11} + x^9 + x^8 + x^7 + x^2 + 1.$$

For the 7-error correcting BCH code [63, 24],

$$g(x) = x^{39} + x^{38} + x^{37} + x^{36} + x^{34} + x^{33} + x^{31} + x^{28} + x^{27} + x^{25} + x^{23} + x^{22} + x^{17} + x^{11} + x^8 + x^5 + 1,$$
$$h(x) = x^{24} + x^{23} + x^{20} + x^{18} + x^{17} + x^{16} + x^{15} + x^{11} + x^{10} + x^8 + x^5 + 1.$$

Table 3. Parameters of BCH codes used in the case of different error rates of the legitimate partners

Code-word length	Number of carrier bits	Number of errors per word	Critical error rate
$15 = 2^4–1$	11	1	0.066
	7	2	0.132
$31 = 2^5–1$	26	1	0.032
	21	2	0.065
	16	3	0.097
	11	5	0.161
$63 = 2^6–1$	57	1	0.016
	51	2	0.032
	45	3	0.048
	39	4	0.063
	36	5	0.079
	30	6	0.095
	24	7	0.111
	18	10	0.159
$127 = 2^7–1$	120	1	0.008
	113	2	0.016
	106	3	0.024
	99	4	0.031
	92	5	0.039
	85	6	0.047
	78	7	0.055
	71	9	0.070
	64	10	0.079
	57	11	0.087
	50	13	0.102
	43	14	0.110
	36	15	0.118
	29	21	0.165
$255 = 2^8–1$	239	2	0.011
	231	3	0.012
	223	4	0.016
	215	5	0.023
	207	6	0.024
	199	7	0.027
	191	8	0.031
	187	9	0.035
	179	10	0.039
	171	11	0.043
	163	12	0.047
	155	13	0.051
	147	14	0.055
	139	15	0.059
	131	18	0.071
	123	19	0.075
	115	21	0.082
	107	22	0.086
	99	23	0.090
	91	25	0.098
	87	26	0.102
	79	27	0.106
	71	29	0.114
	63	30	0.118
	55	31	0.122
	47	42	0.165

For the 13-error correcting BCH code [127, 50],

$$g(x) = x^{77} + x^{75} + x^{74} + x^{71} + x^{68} + x^{65} + x^{64} + x^{62} + x^{60} + x^{57} + x^{55} + x^{53} + x^{51} + x^{49} + x^{46} + x^{45} + x^{43} + x^{42} + x^{39} + x^{38} + x^{30} + x^{28} + x^{26} + x^{22} + x^{18} + x^{17} + x^{15} + x^{9} + x^{8} + x^{4} + 1,$$

$$h(x) = x^{50} + x^{48} + x^{47} + x^{46} + x^{44} + x^{43} + x^{39} + x^{38} + x^{36} + x^{35} + x^{30} + x^{27} + x^{26} + x^{25} + x^{24} + x^{23} + x^{16} + x^{15} + x^{12} + x^{9} + x^{4} + 1.$$

For the 26-error correcting BCH code [255, 87],

$$g(x) = x^{168} + x^{165} + x^{159} + x^{157} + x^{156} + x^{155} + x^{154} + x^{152} + x^{151} + x^{150} + x^{149} + x^{148} + x^{145} + x^{144} + x^{143} + x^{138} + x^{137} + x^{134} + x^{133} + x^{132} + x^{131} + x^{127} + x^{126} + x^{124} + x^{121} + x^{120} + x^{119} + x^{118} + x^{116} + x^{112} + x^{111} + x^{110} + x^{108} + x^{106} + x^{103} + x^{102} + x^{99} + x^{98} + x^{97} + x^{94} + x^{93} + x^{92} + x^{88} + x^{87} + x^{83} + x^{82} + x^{81} + x^{78} + x^{77} + x^{76} + x^{75} + x^{73} + x^{68} + x^{65} + x^{64} + x^{61} + x^{56} + x^{55} + x^{53} + x^{52} + x^{51} + x^{49} + x^{46} + x^{44} + x^{42} + x^{41} + x^{38} + x^{36} + x^{34} + x^{32} + x^{31} + x^{30} + x^{28} + x^{27} + x^{25} + x^{24} + x^{21} + x^{18} + x^{17} + x^{16} + x^{15} + x^{13} + x^{9} + x^{7} + x^{6} + x^{3} + x^{2} + x + 1,$$

$$h(x) = x^{87} + x^{84} + x^{81} + x^{76} + x^{74} + x^{73} + x^{71} + x^{66} + x^{64} + x^{63} + x^{61} + x^{58} + x^{57} + x^{56} + x^{54} + x^{52} + x^{50} + x^{48} + x^{46} + x^{45} + x^{42} + x^{40} + x^{38} + x^{36} + x^{35} + x^{34} + x^{33} + x^{32} + x^{31} + x^{29} + x^{27} + x^{23} + x^{22} + x^{21} + x^{20} + x^{17} + x^{16} + x^{15} + x^{14} + x^{13} + x^{11} + x^{9} + x^{7} + x^{6} + x^{5} + x^{4} + x + 1.$$

We used these BCH codes to correct errors in the raw key by executing one cycle. Figure 10 shows the corresponding reconciled key lengths plotted as functions of estimated error rates Q. Note that the error rate remains finite after only one cycle is executed, because new errors are generated (incorrect words are obtained by decoding) if the number of single-bit errors in a code word exceeds the number of errors that can be corrected by a particular code. However, the execution of another cycle is not justified, because the string length is multiplied by k/n after each cycle and the resulting final key will be too short.

When the code-word length is 127 or 255 and Q = 6%, no errors are left after a single error-correction cycle has been executed (see Fig. 11). The reconciled key length is approximately 20% of the raw key length, i.e., similar to that obtained by executing CASCADE for Q = 6%. When the code word is 15, 31, or 63 bits long, the reconciled key generated in a single cycle contains a significant number of single-bit errors, even though its relative length can amount to 35%, but the efficiency attained by executing another cycle is lower than in the procedures discussed above.

The efficiency of "one-cycle" error correction executed for 127- or 255-bit code words is comparable to that of CASCADE. However, it is important that the conditional probability for Eve's measurement on the bit strings shared by Alice and Bob remains invariant after the error correction in CASCADE with error removal, being determined by the error $\varepsilon(Q)$ of optimal measurement on nonorthogonal states $|e_0\rangle$ and $|e_1\rangle$ (see (20) and (21)), whereas it changes substantially after a BCH code is applied. The conditional probability for Eve's measurement on the reconciled key determines the compression ratio for hashing into a final secret key. Therefore, the fact that the length of the reconciled key is similar to that in CASCADE does not necessarily imply that a final key of the same length will be obtained.

Figure 12 shows the number of bits transmitted over the public classical channel per bit of the reconciled key plotted versus Bob's error rate Q. It is clear that this number is a fraction of a bit, as in CASCADE.

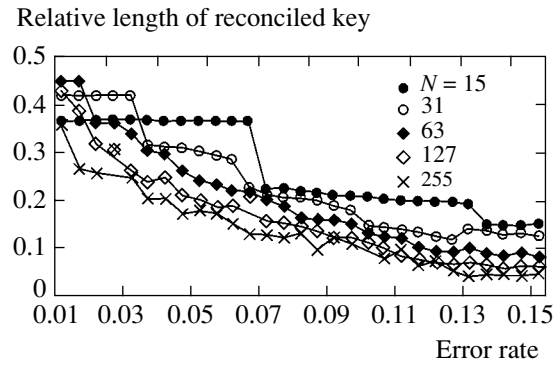

Fig. 10.

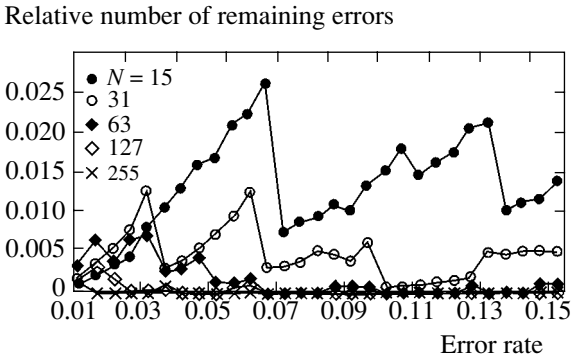

Fig. 11.

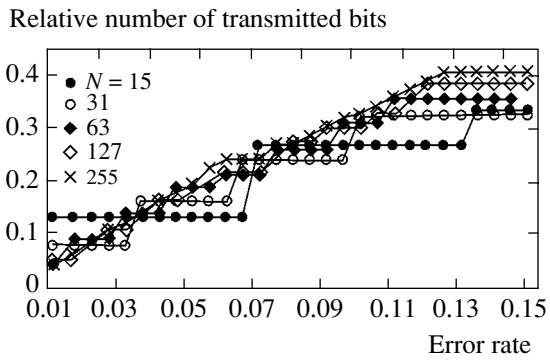

Fig. 12.

8. CALCULATION OF EVE'S ERROR RATE AFTER ERROR CORRECTION BY MEANS OF BOSE–CHAUDHURI–HOCQUENGHEM CODES

First, we evaluate the conditional probability for Eve's individual measurements. Before error correction, Eve's error rate is

$$E(Q) = 1 - (1 - Q)(1 - \varepsilon(Q)) + Q\varepsilon(Q)$$

(see (26)), and Bob's error rate is Q. While Bob's knowledge of the reconciled key is virtually perfect, Eve's error rate is determined as follows.

Since the parameters of the error-correcting code actually employed (code-word length, the number of information and control bits, partition into blocks, etc.) are supposed to be known, Eve can try to correct errors in her output string by using decoding rules. It is well known that the probability of decoding error is

$$P_e = \frac{1}{M}\sum_{i=1}^{M} \Pr(w_E^i \neq w^i), \qquad (91)$$

where w_E^i is the code word generated by Eve from the code word w^i sent by Alice [24, 25]. For the code $[n, k]$,

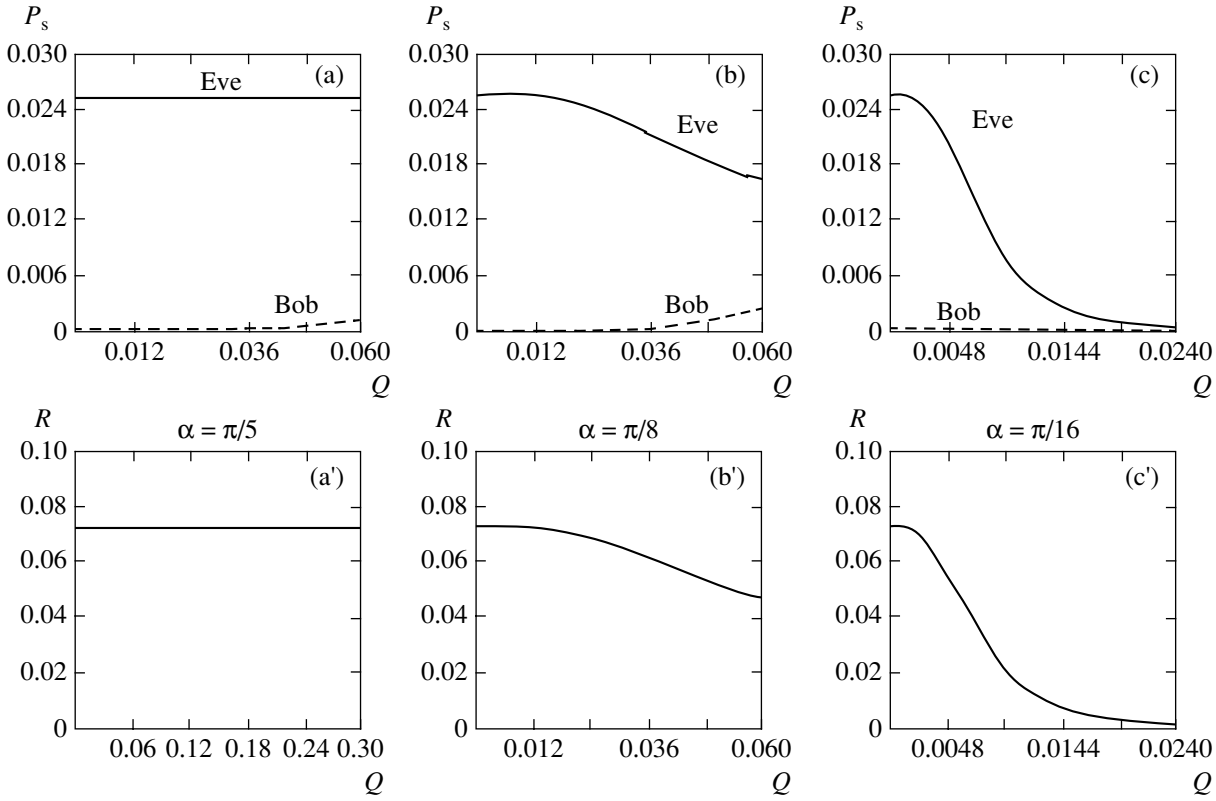

Fig. 13.

the error probability is

$$P_e(p) = 1 - \sum_{j=0}^{n} \alpha_j p^j (1-p)^{n-j}, \tag{92}$$

where α_j is the number of coset leaders with Hamming weight j and p is the initial error rate per bit. We seek the error rate per bit $P_{symb}(p)$, which is known to be bounded from below (underestimated in favor of Eve):

$$P_{symb}(p) \geq \frac{P_e(p)}{k}. \tag{93}$$

Straightforward calculation of $P_{symb}(p)$ requires exhaustive search, which is particularly difficult to perform for codes with long code words. Therefore, use should be made of lower estimates. For any code $[n, k]$,

$$P_e(p) \geq [C_n^{t+1} - \alpha_{t+1}] p^{t+1} (1-p)^{n-t-1} + \sum_{i=t+2}^{n} C_n^i p^i (1-p)^{n-i}, \tag{94}$$

where

$$\alpha_{t+1} = 2^{n-k} - 1 - \sum_{i=1}^{t} C_n^i \geq 0, \tag{95}$$

and t is the largest integer for which (95) holds.

Eve's error rate per bit of the reconciled key shared by Alice and Bob is given by (93) and (94) with

$$p \longrightarrow E(Q) = 1 - (1-Q)(1-\varepsilon(Q)) + Q\varepsilon(Q).$$

Bob's error rate after one error-correction cycle is given by (93) and (94) with p replaced by Q.

Eve's and Bob's error rates after one cycle of error correction by means of the BCH code [63, 39] "tuned" to $Q = 6\%$ are shown in Fig. 13 for several angles of overlap between the carrier states prepared by Alice.

Eve's entropy $R(P_e(E(Q)))$ shown in Figs. 13a'–13c' determines the relative length of the final key. According to Fig. 13, the final key length does not exceed 8% of the length of the reconciled key. When $\alpha = \pi/16$ (the prepared carrier states are almost orthogonal), the Renyi entropy approaches zero even for $Q \approx 1\%$. More precisely, $R(P_e(E(Q))) \approx 0.02$ (see Fig. 13c'); i.e., the final key length is not greater than 2% of the reconciled key length.

Figure 14 shows analogous curves of Eve's and Bob's error rates evaluated for several angles of overlap between the carrier states after one cycle of error correction by means of the BCH code [127, 78]. In the case of a significant overlap ($\alpha = \pi/5$), the error rates are almost constant after error correction. When the prepared carrier states are almost orthogonal, Eve's error

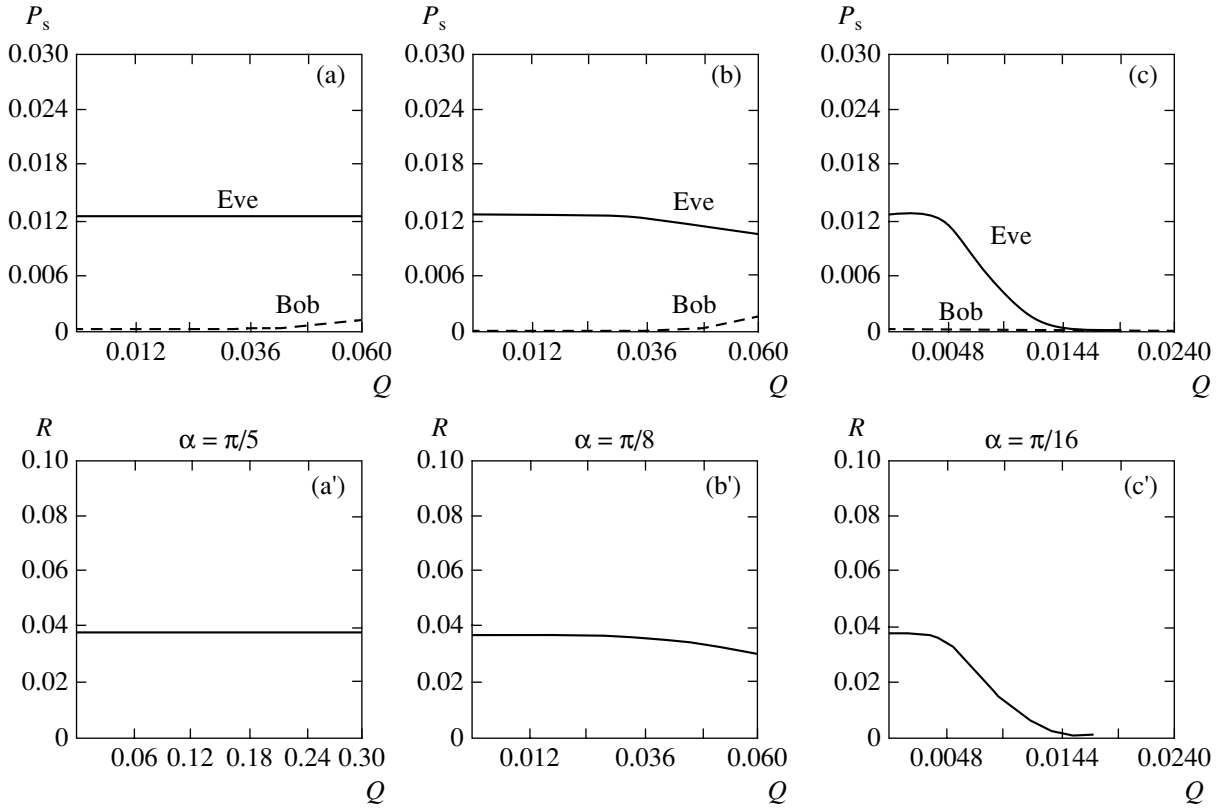

Fig. 14.

rate sharply decreases with increasing Q. Eve's entropy $R(P_e(E(Q)))$ shown in Figs. 14a'–14c' demonstrates that the length of the final key does not exceed 4% of the reconciled key length. Note that error correction by means of the BCH code [63, 39] leaves 8% even though the relative length of the reconciled key is larger.

These results illustrate the fact that an efficient error-correcting code can be used by the eavesdropper as well as by the legitimate partners: the conditional probability for Eve's measurement on the reconciled key increases with the error-correction efficiency, and the length of the final secret key decreases accordingly.

Thus, the error-correction efficiency of a code evaluated without taking into account the quantum part of the protocol cannot be used to quantify its efficiency with regard to the final secret key length.

9. ERROR CORRECTION IN THE RAW KEY BY MEANS OF HAMMING CODES

For comparison, we perform error correction in the raw key by means of Hamming codes, which correct one single-bit error per code word and are the easiest to decode [24, 25]. (A discussion of these codes as applied in quantum cryptography can be found in [26].)

The Hamming code with primitive length $n = 2^m - 1$ can be defined by choosing a parity check matrix of the form

$$\mathbf{H} = [\alpha^0, \alpha^1, \ldots, \alpha^{n-1}], \tag{96}$$

where α^0 is the primitive element of the field $GF(2^m)$. Each code vector $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1})$ satisfies the relations

$$\mathbf{c} \cdot \mathbf{H}^{\mathrm{T}} = 0, \quad \sum_{i=0}^{n-1} c_i \alpha^i = 0, \tag{97}$$

or, in the polynomial representation,

$$c(x) = \sum_{i=0}^{n-1} c_i x^i, \tag{98}$$

where α^i solves the equation $c(\alpha) = 0$. Since $\alpha^{2^m-1} = 1$ (i.e., α is a root of the polynomial $x^{2^m-1} - 1$), the minimal polynomial $f_j(x)$ whose roots are α^i in the field $GF(2^m)$ is a divisor of $x^{2^m-1} - 1$ and can therefore be used as the generator polynomial of a code. The code is parameterized by the length $n = 2^m - 1$ of a code word, which consists of $2^m - 1 - m$ information bits and m control bits.

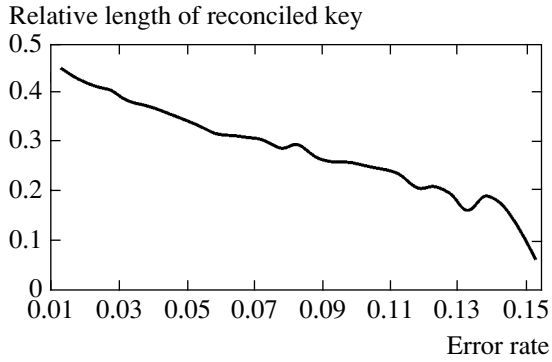

Fig. 15.

If the input code word is $c(x)$ and the output one containing one single-bit error is

$$y(x) = c(x) + e(x),$$

then

$$y(\alpha) = c(\alpha) + e(\alpha) = e(\alpha) = \alpha^i,$$

where i is the position of the error. The (single-) error polynomial has the form

$$e(x) = \sum_{i=0}^{n-1} e_i x^i$$

with only one $e_i = 1$.

The error correction in the raw key is performed as follows.

1. Bob uses the error rate to estimate the number m of control bits per code word and sends the result to Alice. Since one single-bit error per code word can be corrected by means of a Hamming code, the code-word length is set to minimize the probability of more than one error. The number of control bits is calculated as

$$m = \max\left(3, \log\frac{1}{Q}\right).$$

2. Alice partitions the key she has prepared into 2^m-bit blocks (with locations in each block labeled by indices starting from 0) and sends the corresponding parity bits to Bob.

3. When a parity bit calculated by Bob differs from the corresponding one sent by Alice, the Hamming code is applied to the corresponding block (with the exception of the zeroth bit); i.e., the control bits are placed at the positions labeled by 1, 2, …, 2^i, …, 2^{m-1} to make up an m-bit binary word (with positions in the block labeled by indices starting from 1). The XOR operation is executed bitwise on the remaining 1's, and the results are appended as control bits.

4. Bob performs decoding by executing the XOR operation bitwise on the indices labeling the nonzero bits, including the control ones. If the resulting number (syndrome) is one, then it is the index of an error; otherwise, there are no errors. After the decoding, the control bits are removed to preclude transfer of any additional private information to Eve. The decoding procedure applied to a block containing more than one single-bit error either adds another single-bit error or corrects the error.

5. Alice and Bob remove the zeroth bit from each block to preclude transfer of any information with the parity bit, which is known to Eve since it has been announced by the legitimate partners.

6. Alice and Bob use the previously estimated error rate to evaluate a new one (e.g., see [26] for details).

7. Alice performs a random permutation of bits in the remaining string and informs Bob about the new sequence, and the process is repeated by starting from step 1.

The algorithm is executed until a certain number of cycles are wasted (with parity bits equal for all blocks).

Figure 15 shows the length of reconciled key generated by means of Hamming error-correcting codes. In particular, Fig. 15 demonstrates that about 30% of the raw key is left if $Q = 6\%$, which is comparable to the efficiency of CASCADE and twice higher than in BINARY. Moreover, key reconciliation by means of Hamming codes is more efficient as compared to error correction by means of BCH codes, because each block is checked for parity before it is encoded by Alice and no encoding is performed if the parity bit of the block is zero (i.e., the block contains an even number of single-bit errors, and this number can only be increased by encoding or decoding the block). If a block contains an odd number of single-bit errors, then the corresponding parity bits calculated by Alice and Bob are different. Since the block size is preset to minimize the probability of three single-bit errors per block, almost every block contains one single-bit error and its correction by means of a Hamming code is guaranteed. Then, the remaining bits are permuted, and the procedure is repeated.

In terms of the number of bits transmitted over a public classical channel per bit of the reconciled key, the Hamming codes are comparable to CASCADE and BCH codes and are twice as efficient as BINARY. Figures 16 and 17 show the number of bits transmitted over the public channel per bit of the raw key evaluated in the cases when a random number generator can be used only by Alice and by both Alice and Bob, respectively.

Recall that the length of the reconciled key is an intermediate result. The overall efficiency must be estimated by using the length of the final secret key, which requires knowledge of the conditional probability for Eve's measurement on the reconciled key. We have not found any conclusive answer to this day, because the conditional probability remains invariant for blocks

that have not been encoded after being checked for parity and changes for the reconciled positions in encoded blocks. Since the bits corresponding to different conditional probabilities associated with Eve's measurement are randomly distributed among the new blocks obtained by random permutation, it is hardly possible to find an analytical expression for the new conditional probability (and Renyi entropy), which determines the final key length. According to our preliminary estimates, the advantage of Hamming codes over CASCADE at an intermediate stage is eliminated in the final key.

Error correction in the case of a nonuniform distribution of conditional probabilities over the key is a separate problem of special interest.

10. OPEN QUESTIONS CONCERNING COLLECTIVE MEASUREMENTS IN QUANTUM CRYPTOGRAPHY

Even though the B92 protocol is a simplified model, it contains all elements common to other, more practicable key distribution protocols, such as BB84.

In the case of entangled translucent eavesdropping, the basic elements can be outlined as follows. After a measurement has been performed at the receiver end according to any protocol, the bits measured by Bob are in one-to-one correspondence with the ancilla states, $i_k \longleftrightarrow \rho_{i_k}$, where the state ρ_{i_k} stored by Eve in a quantum memory is either pure or mixed. This correspondence holds for all bit positions, no matter whether the corresponding bits are true or wrong as compared to their counterparts in the string sent by Alice. The ancilla states stored by Eve depend on the error rate Q for Bob's measurements.

Further results depend on the error-correction procedure used by the legitimate partners. If the procedure relies on the use of a random code table (which may be exponentially large), then Eve can extract the secret key by performing individual measurements when the capacity of the Alice–Bob channel is higher than that of the Alice–Eve channel. When n is large (formally, as $n \longrightarrow \infty$), the final key length is equal to the length n of the input string. Since the "strong inverse" to the Shannon coding theorem is applicable, the amount of information extracted by Eve from the key tends to zero with increasing n (see (28)). If collective measurements can be performed on all states stored in Eve's quantum memory (which is not forbidden by any fundamental principle), then Bob's reconciled n-bit string cannot be used as a secret key before it is compressed by using a universal hash function. The compression ratio (relative length of the secret key) depends on the conditional probabilities for Eve's collective measurements,

$$P_{W|W_E = w_E^k}(w_E^k) = \mathrm{Tr}\{\chi_{w_E^k}\hat{\rho}_W\},$$
$$\hat{\rho}_W = \rho_{i_1} \otimes \ldots \otimes \rho_{i_n}, \tag{99}$$

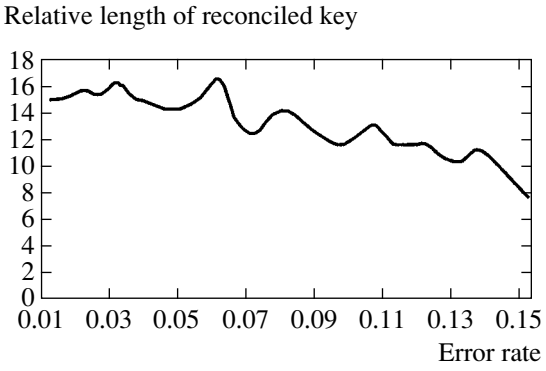

Fig. 16.

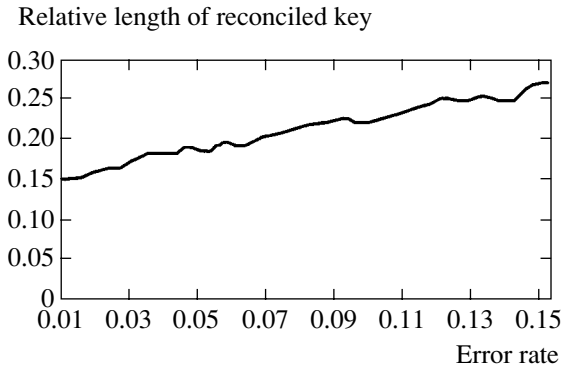

Fig. 17.

where $\chi_{w_E^k}$ is a collective outcome operator analogous to (30) and (65). A measurement outcome $\chi_{w_E^k}$ is interpreted by Eve as a binary string $w_E^k = (i_1^E, i_2^E, \ldots, i_n^E)$ if the true string is W. The compression ratio depends on the Renyi entropy (52) calculated for the conditional probabilities given by (99). To the best of our knowledge, no exact inequalities have yet been found for the Renyi entropy associated with collective measurements.

Error correction performed by legitimate partners using an exponentially large random code table is not a constructive procedure. It can be used only to find a theoretical upper bound for the error rate corresponding to secure key distribution. However, questions concerning collective measurements arise with regard to constructive practical error-correction procedures as well.

The execution of the BINARY or CASCADE protocol leads to a situation analogous to that outlined above, with the only exception that Bob's string does not contain any single-bit errors. Accordingly, the ancilla states stored by Eve do not include any of those corresponding to single-bit errors. The reconciled key cannot be used as a secret one before the hashing (determined by conditional probabilities (99)) is performed.

Collective measurements can also be performed by Eve when a classical error-correcting code is employed. After the legitimate partners have estimated the error rate Q, chosen an appropriate classical code, and partitioned the string into code words, Eve can perform collective quantum measurements on code words. In this case, the outcome operator defined by (30), (65), or (99) defines a partition of the identity operator in the subspace spanned by all code-word vectors, which further complicates analysis of compression ratio.

As of today, collective measurements pose a rather theoretical threat. However, real experiments on quantum memory (e.g., see [27]) may lead to their implementation in the nearest future.

Thus, the procedure used to correct errors in the raw key has a drastic effect on the length of the final secret key. Moreover, since the length of the key depends not only on the error-correction procedure, but also on the quantum part of the protocol, knowledge of the efficiency of the procedure is not sufficient to characterize the overall efficiency including the quantum part of the protocol.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 05-02-17387 and INTAS, grant no. 04-77-7284.

REFERENCES

1. S. Wiesner, SIGACT News **15**, 78 (1983).
2. C. H. Bennett and G. Brassard, in *Proceedings of IEEE International Conference on Computer Systems and Signal Processes* (Bangalore, India, 1984), p. 175.
3. G. S. Vernam, J. Am. Inst. Electr. Eng. **55**, 109 (1926).
4. V. A. Kotel'nikov, A Report (Moscow, 1941).
5. C. E. Shannon, Bell Syst. Tech. J. **28**, 658 (1949).
6. W. K. Wootters and W. H. Zurek, Nature **299**, 802 (1982).
7. C. H. Bennett, Phys. Rev. Lett. **68**, 3121 (1992); C. H. Bennett, G. Brassard, and N. D. Mermin, Phys. Rev. Lett. **68**, 557 (1992).
8. W. Diffie and M. E. Hellman, IEEE Trans. Inf. Theory **22**, 644 (1976).
9. A. K. Ekert, B. Huttner, G. M. Palma, and A. Peres, Phys. Rev. A **50**, 1047 (1994).
10. H. E. Brandt, J. M. Myer, and S. J. Lomonaco, Jr., Phys. Rev. A **56**, 4456 (1997).
11. C. E. Shannon, Bell Syst. Tech. J. **27**, 397 (1948); **27**, 623 (1948).
12. R. Gallagher, *Information Theory and Reliable Communication* (Wiley, New York, 1968; Sovetskoe Radio, Moscow, 1974).
13. C. W. Helstrom, *Quantum Detection and Estimation Theory* (Academic, New York, 1976).
14. J. Wolfowitz, Ill. J. Math. **1**, 591 (1957).
15. A. S. Kholevo, *Introduction to the Quantum Theory of Information* (MTsNMO, Moscow, 2002), Sovrem. Mat. Fiz., No. 5; Probl. Peredachi Inf. **8**, 63 (1972); **15**, 3 (1979); Usp. Mat. Nauk **53**, 193 (1998).
16. R. Jozsa and B. Schumacher, J. Mod. Opt. **41**, 2343 (1994).
17. P. Hausladen, R. Jozsa, B. Schumacher, *et al.*, Phys. Rev. A **54**, 1869 (1996).
18. B. Schumacher and M. D. Westmoreland, Phys. Rev. A **56**, 131 (1997).
19. D. Maurer, IEEE Trans. Inf. Theory **39**, 733 (1993).
20. C. H. Bennett, G. Brassard, C. Crépeau, and U. Maurer, IEEE Trans. Inf. Theory **41**, 1915 (1995).
21. J. L. Carter and M. N. Wegman, J. Comput. Syst. Sci. **18**, 143 (1979).
22. C. H. Bennett, F. Bessette, G. Brassard, *et al.*, J. Cryptology **5**, 3 (1992).
23. G. Brassard and L. Salvail, Lect. Notes Comput. Sci. **765**, 410 (1994).
24. E. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes* (North-Holland, Amsterdam, 1977).
25. W. W. Peterson and E. J. Weldon, *Error-Correcting Codes* (MIT Press, Cambridge, 1972).
26. W. T. Buttler, S. K. Lamoreaux, J. R. Torgerson, *et al.*, quant-ph/0203096.
27. B. Julsgaard, J. Sherson, J. I. Cirac, *et al.*, Nature **432**, 482 (2004).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 253–258.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 292–299.

ATOMS, MOLECULES, OPTICS

Diffusion-Limited Aggregation: A Continuum Mean Field Model

A. B. Ryabov[a], E. B. Postnikov[b], and A. Yu. Loskutov[a]

[a]*Moscow State University, Moscow, 119992 Russia*
[b]*Kursk State University, Kursk, 305000 Russia*
e-mail: loskutov@chaos.phys.msu.ru

Received March 15, 2005

Abstract—Mean field theory is used as a basis for a new approach to analyzing fractal pattern formation by diffusion-limited aggregation. A coarse time scale is introduced to take into account the discrete nature of DLA clusters. A system of equations is derived and solved numerically to determine the fractal dimension and density of a cluster as a function of distance from its center. The results obtained are in good agreement with direct numerical simulations. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

Diffusion-limited aggregation (DLA) is the generic name for a class of models describing the formation of a cluster by addition of randomly walking particles sticking to it. The original model introduced in [1] has drawn continuing attention (e.g., see reviews in [2, 3]), because it applies to a variety of physical processes, including dielectric breakdown, solute aggregation, and growth of bacterial colonies. Moreover, the DLA cluster is of mathematical interest as an object with fractal dimension.

To date, detailed numerical simulations of the growth and scaling properties of DLA clusters have been performed. In particular, the aggregate density field was found to exhibit fractal behavior: the aggregate mass is a power of its characteristic radius, with exponent $d = 1.71$ and 2.50 in the two- and three-dimensional continuum models, respectively [4].

Furthermore, the statistical properties of an individual branch of the DLA cluster have been analyzed by examining quasi-one-dimensional random walk as a model of fractal aggregate growth in a domain of length much greater than its width. A summary of early studies in this area was presented in [5]. It was found that $d = 0.66$ for the direction of power-law density profile. This exponent was corrected in extensive numerical experiments: $d = 0.72$ [6]. Currently, diffusion-limited aggregation in the absence of external forces or interaction between particles is a well-studied process that may serve as a benchmark test for new numerical methods [7].

In theory, the most substantial progress has been achieved by applying renormalization-group methods, but only in discrete DLA of two types. In one approach, a parameter is sought that leaves invariant the mean occupancy distribution in a lattice DLA model [8]. In the other [9], the cluster is generated by using a conformal map of the unit circle such that the random walker is mapped to a randomly chosen point on the growing cluster perimeter at each step of an iterative process. A modification of the latter approach was used in [6] to simulate one-dimensional diffusion-limited aggregation in channel geometry.

However, it still remains unclear if DLA is tractable as a continuum model formulated in terms of differential equations. In the original model proposed in [10], a mean field theory that relates the continuous aggregate density distribution to the probability distribution of a random walker coming from infinity was developed by using a power series expansion in cluster density. A refined mean field theory was proposed in [11]. These models provided a qualitative explanation of the process, but the predicted dimension was substantially lower than the measured one. For this reason, various modified mean field approaches were proposed in subsequent studies.

In [12], the growth rate was assumed to be proportional to the gradient of the distribution of the randomly walking particle, rather than to the distribution itself. In [13], the growth rate was represented as a power of the field variable with a phenomenological exponent greater than unity to cut off growth at small density. As a consequence, density buildup was observed in the cluster front zone, and a higher fractal dimension is obtained. In [14], the latter model was substantiated by showing that the growth rate is proportional to a quadratic combination of density and its derivatives if the probabilities of attachment and interaction between random walkers are equal. This result was obtained by replacing the Boltzmann collision integral with a system of differential equations.

Thus, even though the understanding of aggregation kinetics has improved owing to progress in the framework of mean field theory, the "final solution" has never been found. In this paper, we show how the mean field

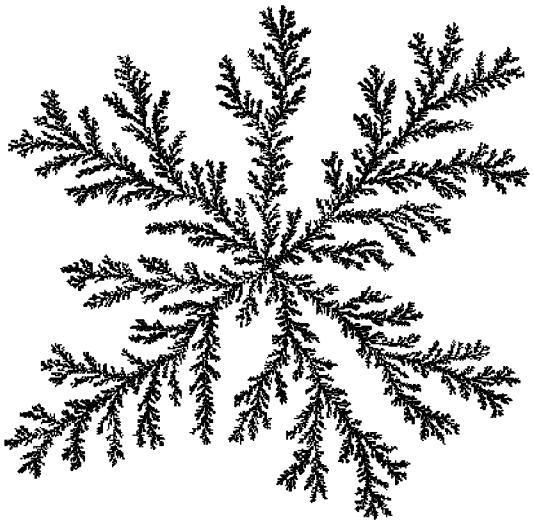

Fig. 1. Typical DLA cluster.

theory can be modified to provide adequate description of the scaling behavior of the system.

2. MODEL

In this section, we formulate a mean field DLA model based on the equations proposed in [11]. We present the complete derivation of a system of equations and demonstrate how a coarse time scale can be introduced.

2.1. Preliminary Analysis

Generally, two-dimensional aggregation kinetics are modeled as follows. While an immobile seed of diameter a is located at the origin, a particle born at a distance R from the seed executes a random walk inside the circle of radius R until it either escapes from the circle or comes in contact with the seed. The particle crossing the circle reappears at a random point on it. The particle that contacts the seed sticks to it, and then a new particle is launched from a random point on the circle. As the process repeats, each new particle can stick to any particle on the cluster boundary. To minimize the effects due to the finite size of the system, the birth radius R is assumed to be much larger than the characteristic radius of the cluster.

A typical DLA cluster has a highly ramified fractal structure (see Fig. 1). New particles are captured in its front zone. In mean field theory, the structure is smeared, and the average perimeter of a cluster is difficult to define.

To describe the growth of a cluster, we introduce the characteristic function $\hat{\rho}(r, \phi)$, equal to unity at the locations occupied by the aggregate and zero elsewhere. The characteristic function $\hat{u}(r, \phi)$ is defined analogously for the random walker. In mean field theory, these functions are ensemble averaged by assuming that the particle distribution is isotropic. As a result, we obtain probability density profiles $\rho(r)$ and $u(r)$ for the cluster and the random walker, respectively.

The entire two-dimensional domain is partitioned into square cells of area a^2, where a is the particle diameter. For each cell containing a particle with probability one, we set $\rho = 1/a^2$, so that its integral over the cell is unity. To facilitate analysis, we perform the change $\rho \longrightarrow a^2\rho$, $u \longrightarrow a^2u$ so that ρ or u is unity in each cell occupied by a particle. Then, the number of particles in the cluster is expressed as

$$N = \frac{1}{a^2}\int_D \rho\, ds, \tag{1}$$

where ds is an area element and D is the region occupied by the cluster.

In the model proposed in [11], the cluster growth is modeled by the system of equations

$$\begin{aligned} \frac{\partial \rho}{\partial t} &= u(\rho + a^2\Delta\rho), \\ \frac{\partial u}{\partial t} &= a^2\Delta u - u(\rho + a^2\Delta\rho), \end{aligned} \tag{2}$$

supplemented with the boundary condition $u(R, t) = c$, which represents a source of random walkers located at the birth radius. The kinetic equation for the random-walker distribution describes diffusion and adsorption onto the aggregate. In the kinetic equation for the cluster density distribution, the terms $u\rho$ and $a^2\Delta\rho$ represent the contributions of random-walker–cluster contact and the lateral growth due to nonlocal interaction, respectively.

Since random walk is much faster than cluster growth, we set $u_t = 0$:

$$\begin{aligned} \frac{\partial \rho}{\partial t} &= u(\rho + a^2\Delta\rho), \\ 0 &= a^2\Delta u - u(\rho + a^2\Delta\rho). \end{aligned}$$

As mentioned in the Introduction, this model contains the basic features of the model, but fails to predict the correct fractal dimension, because it ignores the discrete nature of the DLA cluster.

This explanation is corroborated by numerical analyses of the following model [15, 16]. The region occupied by the cluster is divided into annuli of width a. The annulus of radius r_n is characterized by the largest number M_n of particles that can be placed inside it (estimated as $M_n \sim 2\pi r_n a/a^2$) and the actual number N_n of particles contained in it. The probability of adsorption of a random walker onto the annulus is N_n/M_n. A ran-

dom walker contained in the nth annulus can be adsorbed on the $(n-1)$th, $(n+1)$th, or nth annulus. If $N_n/M_n \ll 1$, then the total adsorption probability can be approximated as

$$\sum_{k=-1}^{1} \frac{N_{n+k}}{M_{n+k}}.$$

When a random walker is adsorbed onto the nth annulus, the value of N_n increases by unity. Otherwise, it drifts into an adjoining annulus with a probability of $1/2 \pm a/r_n$, where the plus and minus signs in $\pm a/r_n$ correspond to the outer and inner annuli, respectively.

This model is simpler than direct numerical simulation, because it makes use of the axial symmetry of the cluster and retains information about the number of particles in each annulus. Nevertheless, the predicted fractal dimensions agree with measured ones, 1.65 [15] and 1.72 [16].

Returning to the mean field theory and taking into account the discrete nature of the cluster, we define $u(r, \tau)$ as the probability distribution for a particle to be located at a distance r from the origin at an instant τ. Furthermore, we introduce an initial distribution $u(r, 0)$ and impose the impermeability condition $u_r|_{r=0,R} = 0$. We use a random-walk time variable τ and a discrete cluster-growth time n to allow for the disparity between the corresponding time scales, because it is obvious that the adsorption of a single random walker does not result in any significant change in the overall cluster geometry.

To derive a kinetic equation for $u(r, \tau)$, we use simple partition into cells in the (x, y) plane. We introduce the probability $\rho(x, y)$ of finding a cluster particle in the cell with coordinates x and y and the probability $u(x \pm a, y \pm a, \tau)$ that the random walker occupies a neighboring cell (see Fig. 2). Treating the random walk followed by adsorption onto the cluster and the ensuing evolution of cluster density as independent processes, i.e., assuming that the change in cluster density over the random-walker lifetime is negligible, we write the following difference equation for $u(x, y, \tau)$:

$$u(x, y, \tau+1) = (1-\rho(x, y))[u(x-a, y, \tau) + u(x+a, y, \tau) + u(x, y-a, \tau) + u(x, y+a, \tau)]/4. \quad (3)$$

Then, we introduce a time increment $\delta t \sim a^2$ and use a Taylor series expansion to rewrite (3) as

$$\frac{\partial u}{\partial \tau} = \frac{a^2}{4}\Delta u - \rho\left(u + \frac{a^2}{4}\Delta u\right), \quad (4)$$

where the Laplacian to be calculated in Cartesian coordinates reduces to

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}$$

under the assumption of axial symmetry.

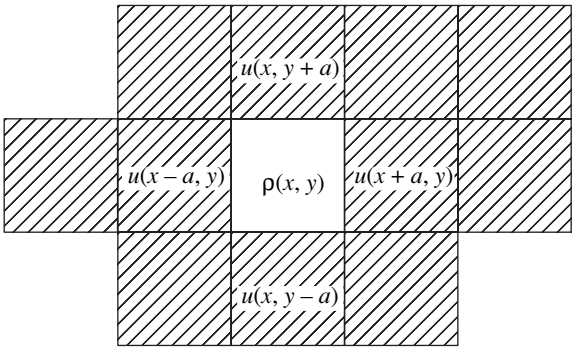

Fig. 2. Lattice model.

The resulting equation is similar to the second one in (2) up to second-order terms, but it allows for nonlocal adsorption, vanishing diffusion through fully occupied regions (where $\rho_n(r) = 1$), and free diffusion into unoccupied regions (where $\rho_n = 0$). We believe that this equation provides a more accurate description.

2.2. Refined Model

Since the asymptotic solution to the diffusion–absorption equation derived above is such that

$$u(\tau) \le \exp(-\rho\tau),$$

there exists a bounded function

$$U(r) = \frac{1}{T^*}\int_0^{T^*} u(r, \tau)d\tau$$

describing, up to normalization, the time-averaged random-walker distribution in space. The integral of Eq. (4) with respect to τ from zero to the random-walker lifetime T^* combined with the condition $u(T^*) = 0$ yields

$$-u(r, 0) = \frac{a^2}{4}\Delta U(r) - \rho\left(U(r) + \frac{a^2}{4}\Delta U(r)\right).$$

This equation is supplemented with the impermeability condition

$$\frac{\partial}{\partial r}U(r)|_{r=0,R} = 0.$$

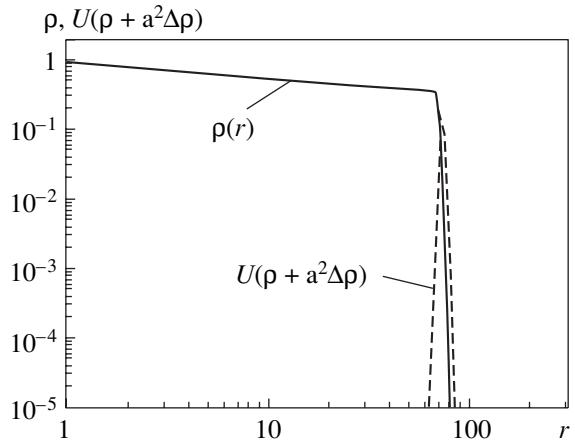

Fig. 3. Cluster density ρ and density of a newly adsorbed single-particle layer.

The equation for the cluster density ρ corresponding to the first equation in (2) can be written as

$$\frac{\partial\rho}{\partial n} = CU(\rho + \Delta\rho),$$

where C is a normalization factor and n is the number of an iteration step. In the conventional mean field theory, the random walker hitting the cluster "spreads" over its perimeter, which results in a slight increase in the cluster size. According to (1), the ensuing change in the cluster density is such that

$$\int[\rho(n+1) - \rho(n)]2\pi r dr/a^2 = 1.$$

Hence,

$$C_0 = \frac{a^2}{2\pi\int U(\rho + \Delta\rho)r dr}.$$

A numerical analysis of the corresponding system of equations leads to results analogous to those reported in [11]; in particular, the fractal dimension of the 2D cluster is unity (see Section 3.1).

However, the adsorption of a particle at a point on the real cluster perimeter has a negligible effect on the probability of particle addition at other locations. Therefore, we can introduce a coarse time scale by assuming that the density of a cluster changes only after its entire perimeter is covered by newly adsorbed random walkers. Then, we obtain

$$C = \frac{2\pi R_b}{a}C_0 = \frac{2\pi R_b}{a}\frac{a^2}{2\pi\int U(\rho + \Delta\rho)r dr}, \qquad (5)$$

where R_b is the characteristic distance from the origin to the cluster boundary.

Figure 3 shows a typical density profile and a curve $U(\rho + \Delta\rho)$ characterizing the location of new adsorbed particles. The latter curve has a sharp peak in the neighborhood of the cluster boundary and vanishes outside it. Accordingly, we can approximately replace R_b with r in expression (5) and write the final system of equations as follows:

$$\rho_{n+1} = \rho_n + CU(\rho + \Delta\rho),$$
$$-u(r, 0) = \Delta U(r) - \rho_n(r)\left(U(r) + \frac{a^2}{4}\Delta U(r)\right), \qquad (6)$$

where

$$C = \frac{2\pi r}{a}\frac{a^2}{2\pi\int U(\rho + \Delta\rho)r dr},$$

and n is the number of an iteration step at which the cluster is covered by a new layer of adsorbed random walkers and its boundary shifts by an increment on the order of particle diameter.

3. ANALYSIS IN A PLANAR GEOMETRY

3.1. Numerical Analysis

Using the initial density distribution

$$\rho_0(r) = \exp(-2r/a),$$

we solved Eqs. (6) by successive iteration with steps corresponding to cluster size increments. At each step, the second equation in (6) was computed on a 10^5 point spatial grid by using an explicit finite-difference scheme, the distribution $\rho_n(r)$ obtained at the preceding step, and

$$u_0(r) = \exp(-(r - R)^2/a^2).$$

The resulting solution was substituted into (6) to calculate a normalized increment of the cluster distribution. We set the increment at $r = 0$ to zero, since the region occupied by the seed is impenetrable.

The dashed curve in Fig. 4 is a numerical solution to system (6) for the two-dimensional cluster. Its log–log slope, -0.22 ± 0.02, corresponds to the cluster fractal dimension $d = 1.78 \pm 0.02$. The dotted curve predicted by the mean field theory developed in [11] without introducing any coarse time scale corresponds to a lower dimension ($d = 1$).

3.2. Analytical Solution

Since system (6) is amenable to iterative computation, it can be solved analytically by applying the finite Hankel transform

$$U(p_i, t) = \frac{1}{R^2}\int_0^R U(r, t) J_0\left(p_i \frac{r}{R}\right) r dr,$$

where $J(p_i)$ is the Bessel function. The desired solution can be represented by a series expansion:

$$U(r, t) = 2\sum_i U(p_i, t)\frac{J_0\left(p_i \frac{r}{R}\right)}{J_0^2(p_i)},$$

where p_i denotes the roots of the equation $J_1(p_i) = 0$.

Using the initial condition

$$u(r, 0) = \delta(r - R)/2\pi(r/R)$$

and the boundary conditions

$$\frac{\partial}{\partial r} U(r)|_{r = 0, R} = 0,$$

we obtain an infinite algebraic system of equations for $U(p_i, t)$:

$$U(p_i) + \sum_{p_k} U(p_k) K(p_k, p_i) = \frac{2R^2}{\pi a^2}\frac{J(p_i)}{p_i^2},$$

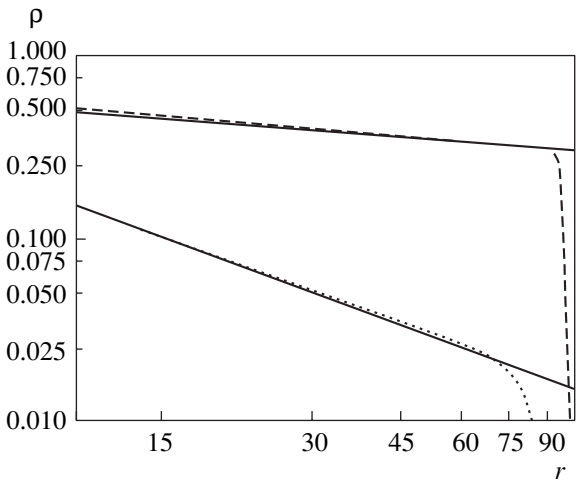

Fig. 4. Cluster density vs. radius predicted in [11] (dotted curve) and by model (6) with coarse time scale (dashed curve). The fractal density is 1.0 and 1.78 in the former and latter cases, respectively.

with

$$K(p_k, p_i) = \frac{8\left(1 - \frac{a^2}{4R^2}p_k^2\right)_R}{a^2 p_i^2 J_0^2(p_k)}\int_0^R \rho(r) J_0\left(p_i \frac{r}{R}\right) J_0\left(p_k \frac{r}{R}\right) r dr.$$

By analogy with Fredholm integral equation of the second kind, an exact solution to this system can be represented in terms of a resolvent kernel:

$$U(p_i) = \frac{2R^2}{\pi a^2}\left[\frac{J(p_i)}{p_i^2} + \frac{1}{D}\sum_k \frac{J(p_k)}{p_k^2} D(p_i, p_k)\right], \quad (7)$$

where

$$D = 1 + \sum_{m=1}^{\infty}\frac{(-1)^m}{m!}\sum_{\alpha_1=1}^{\infty}\cdots\sum_{\alpha_m=1}^{\infty}\begin{vmatrix} K(p_{\alpha_1}, p_{\alpha_1}) & \cdots & K(p_{\alpha_1}, p_{\alpha_m}) \\ \vdots & \ddots & \vdots \\ K(p_{\alpha_m}, \rho_{\alpha_1}) & \cdots & K(p_{\alpha_m}, p_{\alpha_m}) \end{vmatrix},$$

$$D(p_i, p_k) = K(p_i, p_k) + \sum_{m=1}^{\infty}\frac{(-1)^m}{m!}\sum_{\alpha_1=1}^{\infty}\cdots\sum_{\alpha_m=1}^{\infty}\begin{vmatrix} K(p_i, p_k) & K(p_i, p_{\alpha_1}) & \cdots & K(p_i, p_{\alpha_m}) \\ K(p_{\alpha_1}, p_k) & K(p_{\alpha_1}, p_{\alpha_1}) & & \\ \vdots & & \ddots & \vdots \\ K(p_{\alpha_m}, p_k) & K(p_{\alpha_m}, p_{\alpha_1}) & \cdots & K(p_{\alpha_m}, p_{\alpha_m}) \end{vmatrix}.$$

Substituting the derivative of this solution into the system of kinetic equations, we obtain a recursive procedure for calculating cluster growth. However, since it involves calculations of slowly convergent series and integrals containing Bessel functions, this procedure is more difficult to use for evaluating the fractal dimension, as compared to numerical analysis.

4. CONCLUSIONS

A new mean field approach to analyzing fractal pattern formation by diffusion-limited aggregation is proposed. Following previous studies (e.g., see [11, 14, 15]), we assume that the aggregate is axially symmetric. The key distinction of the present approach is the use of a coarse time scale taking into account the discrete nature of DLA clusters. The coarse time scale is introduced by requiring that the addition of a particle to the cluster does not lead to any significant effect on the continuous radial distribution obtained by changing from the discrete to the continuum model and averaging over the angle. To meet this requirement, we assume that the cluster density changes jumpwise only after the entire accessible perimeter of the cluster is covered by a new layer of adsorbed random walkers. As a result, we derive a difference–differential equation with discrete time step. Furthermore, this form of the governing equation entails the use of a normalization factor C proportional to the accessible cluster perimeter.

The value of the normalization factor depends on the dimension of the spatial domain of aggregate growth. For cluster growth on the surface of a cylinder along its element [5, 18], C = const since the corresponding fractal dimension is evaluated by solving a quasi-one-dimensional problem on a line parallel to the direction of aggregate growth. For the two-dimensional cluster, this factor is proportional to the circle perimeter. For the three-dimensional one, it scales with the volume of a spherical layer with thickness equal to the particle diameter, where random walkers are adsorbed. Since the density profile drops near the cluster boundary (see Fig. 3), the function $\rho + \Delta\rho$ has a sharp maximum, which makes it possible to define absorption region and find the normalization factor.

The approach developed in this study is used to obtain a fractal dimension of 1.78 in the two-dimensional DLA model, in good agreement with direct numerical simulations. We have also considered the quasi-one-dimensional and three-dimensional DLA models. In the former case, our solution of the proposed system of equations yields a fractal dimension of 0.80 ± 0.02, in fair agreement with direct numerical simulations [5, 18]. As the space dimension increases to three and higher, the assumption of isotropy progressively becomes less accurate (even if physically plausible) and entails a systematic overestimation of the fractal dimension of the cluster.

Finally, we note that the proposed approach can be naturally generalized to aggregation in a system of particles of several types. In particular, if we assume that particles of the same type stick together, then the normalization factor in each equation corresponding to a particular type of particles will be proportional to the largest volume fraction that can be occupied by particles of this type in the current spherical layer.

ACKNOWLEDGMENTS

We thank N. Brilliantov for fruitful discussions during our study of the problem.

REFERENCES

1. T. A. Witten, Jr. and L. M. Sander, Phys. Rev. Lett. **47**, 1400 (1981).
2. T. Halsey, Phys. Today **53**, 36 (2000).
3. V. I. Roldughin, Russ. Chem. Rev. **72**, 823 (2003).
4. P. Meakin, Physica D (Amsterdam) **86**, 104 (1995).
5. A. Arneodo, J. Elezgaray, M. Tabard, *et al.*, Phys. Rev. E **53**, 6200 (1996).
6. E. Somfai, R. Ball, J. DeVita, *et al.*, Phys. Rev. E **68**, 020401 (2003).
7. D. Tillberg and J. Machta, Phys. Rev. E **69**, 051403 (2004).
8. A. Erzan, L. Pietronero, and A. Vespignani, Rev. Mod. Phys. **67**, 545 (1995).
9. M. Hastings, Phys. Rev. E **55**, 135 (1997).
10. T. A. Witten, Jr. and L. M. Sander, Phys. Rev. B **27**, 5686 (1983).
11. R. Ball, M. Nauenberg, and T. A. Witten, Jr., Phys. Rev. A **29**, 2017 (1984).
12. K. Ohno, K. Kikuchi, and H. Yasunara, Phys. Rev. A **46**, 3400 (1992).
13. H. Levine and Yu. Tu, Phys. Rev. A **45**, 1053 (1992).
14. V. A. Bogoyavlenskiy and N. A. Chernova, Phys. Rev. E **61**, 5422 (2000).
15. H. Sakaguchi, J. Phys. Soc. Jpn. **68**, 61 (1999).
16. A. Loskutov, D. Andrievsky, V. Ivanov, *et al.*, Macromol. Symp. **160**, 239 (2000).
17. E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge Univ. Press, Cambridge, 1996).
18. C. Evertsz, Phys. Rev. A **41**, 1830 (1990).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 259–269.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 300–311.

NUCLEI, PARTICLES, FIELDS, GRAVITATION, AND ASTROPHYSICS

Geometry of a Centrosymmetric Electric Charge

Yu. A. Khlestkov

Moscow Engineering Physics Institute, Moscow, 115409 Russia
e-mail: khlestkov@yandex.ru

Received August 4, 2003; in final form, March 10, 2005

Abstract—The gravitational description given for an electric charge on the basis of exact solution of the Einstein–Maxwell equations eliminates Coulomb divergence. The internal pulsating semiconfined world formed by neutral dust is smoothly joined with parallel Reissner–Nordstrem vacuum worlds via two static bottlenecks. The charge, rest mass, and electric field are expressed in terms of space curvatures. The internal and external parameters of the maximon, electron, and the universe form a power series. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

Elementary particles are treated as point objects in planar spacetime. Their internal structure can be considered in the space of the general theory of relativity (GTR), whose curvature is equivalent to the gravitational field [1].

Initially, Einstein planned, first, to geometrize physical fields and their sources and, second, to represent the discreteness of spacetime as well as parameters of elementary particles, their nonlocalizability and stochastic behavior (quantum effects) as manifestations of properties of a "continuous" gravitational field.

This study provides a simple example illustrating the possibility of implementation of the first part of this program. The internal structure of an electric charge is described on the basis of the exact solution to the Einstein–Maxwell equations for a nonstationary centrosymmetric dustlike matter (dust) and an electric field [2].

It should be noted that the second part of the program has not been realized as yet; i.e., the quantum behavior of elementary particles has neither been described in a nonlinear geometrized theory nor interpreted, and the spin has not been taken into account; the solution obtained for an electric charge cannot be extended to real charged elementary particles (electron, proton, etc.). We will only consider the description of geometry of an abstract electric charge whose charge and rest mass coincide with the corresponding quantum numbers of the real electron, maximon, universe, etc.

This solution clarifies the meaning of two universal constants, viz., electric charge e and rest mass m_0, which are found to be first integrals of the Einstein–Maxwell equations (integrals of motion) and which can be determined from the spacetime curvatures at any point of the spacetime.

The procedure in which a source of gravitational field (energy–momentum tensor of matter appearing on the right-hand side of the equations) must be specified "manually" has been usually regarded as poor since Einstein's time. For this reason, the possibility of complete geometrization of physical fields is considered as dubious.

The solution obtained here shows that this is not quite correct. We specify only the "filling" of the system to be studied (matter, physical fields and interactions, the presence of charges, pressure, rotation, torsion, etc.). Then, in view of the nonlinearity and self-consistency of the Einstein system of equations and equations for potentials of physical fields, all physical parameters of the system (densities, velocities, and field strengths) are expressed in terms of geometry as a result of the solution of such a system of equations.

In contrast to the linear situation, which is responsible in all probability for such an attitude to the Einstein equations with a right-hand side, it is impossible in the general case to characterize the properties of the functions satisfying the nonlinear equations prior to their solution in view of the absence of a superposition principle. A nonlinear equation has no "source" or the "right- and left-hand sides" in the linear sense.

It is also equally important that the spacetime curvature eliminates the main disadvantage of theories in the Minkowski spacetime, viz., Coulomb divergence of the field of a point charge, which generates divergences in the existing quantum theories.

Strictly speaking, a planar space must be empty, since if the Riemann–Christoffel curvature tensor is zero, the conservative Einstein tensor must also be zero, as well as the energy–momentum tensor of matter and, hence, the densities and potentials of physical fields.

In addition, this solution at last clarifies the reason for the prevailing concept of negligibility of gravitational effects over the classical length (due to the extremely small value of the gravitational radius as compared to the classical radius), as well as the idea that gravitational interaction (spacetime curvature) becomes significant either over limiting lengths in the

1063-7761/05/10102-0259$26.00

microworld (such as Planck's length) or over the scale of the universe in the macro- and megaworlds.

The reason appears paradoxical: such a situation in vacuum (vanishing smallness of the "potential gravitational field" as compared to the potential electric field of an isolated stationary point charge, $\sqrt{k}\,m_0/r^2 \ll e/r^2$, where k is the gravitational constant) is explained precisely by the large "gravitational mass defect" due to the strong gravitational field (large spacetime curvature) inside the charge, which reduces the energy of charge in vacuum observed from outside (from the "Newtonian" point of view, the gravitation field possesses a sort of "negative energy").

The gravitational interaction, in accordance with its definition, is universal and is manifested over any length, mapping physical fields on geometry [1].

A brief review of the literature concerning this problem in GTR and the reason for which a rigorous solution of the problem of the source of electric charge and rest mass of a particle could not be obtained earlier are given in the Appendix.

2. FORMULATION OF THE PROBLEM

Let us suppose that the gravitational field in a centrosymmetric orthogonal nonstationary metric [1] (in the τ, r, θ, φ coordinates)

$$ds^2 = e^{\nu}d\tau^2 - e^{\lambda}dr^2 - R^2(\tau, r)d\sigma^2, \qquad d\sigma^2 = d\theta^2 + \sin^2\theta d\varphi^2 \tag{1}$$

in a synchronous comoving frame of reference is defined by the energy–momentum tensor, whose mixed components are given by

$$\mathrm{diag}(\varepsilon_s + \varepsilon_f, \varepsilon_f, -\varepsilon_f, -\varepsilon_f),$$

which corresponds to ideal dustlike generally charged matter with a charge density ρ_f, energy density $\varepsilon_s(\tau, r)$, and electromagnetic field with energy density $\varepsilon_f(\tau, r)$.

The system of the Einstein–Maxwell equations in the given case can be written in the form

$$\Phi = \frac{4\pi}{\kappa}R(1 - e^{-\lambda}R'^2 + e^{-\nu}\dot{R}^2),$$

$$\dot{\Phi} = 4\pi R^2\dot{R}\varepsilon_f,$$

$$\Phi' = 4\pi R^2 R'(\varepsilon_f + \varepsilon_s),$$

$$\dot{\varepsilon}_f + 4\dot{R}\varepsilon_f/R = 0,$$

$$\varepsilon'_f + 4R'\varepsilon_f/R = \sqrt{8\pi\varepsilon_f}\rho_f e^{\lambda/2},$$

$$\dot{\varepsilon}_s + (\dot{\lambda} + 4\dot{R}/R)\varepsilon_s/2 = 0,$$

$$\dot{\rho}_f + (\dot{\lambda} + 4\dot{R}/R)\rho_f/2 = 0$$

(we assume that the cosmological term is zero). Here, $\dot{R} = \partial R/\partial\tau$, $R' = \partial R/\partial r$, and $\kappa = 8\pi k/c^4$ is the Einstein constant.

The general solution to the Cauchy problem depends on three arbitrary functions of r, viz., integrals of motion corresponding to the initial distribution of energy density $\varepsilon_s(0, r)$ of matter, charge density $\rho_f(0, r)$, and radial velocity $\beta_r(0, r)$, which is defined for $\tau = 0$.

For such functions, we can choose [2] the electric charge in the form

$$Q(r) = 4\pi\int\rho_f e^{\lambda/2}R^2 dr + Q_0,$$

the total energy of matter in the form

$$E(r) = 4\pi\int(\varepsilon_s + \varepsilon_f)R^2 R' dr + Q^2/2R + E_0,$$

and the function

$$f(r) = e^{-\lambda/2}R' + qQ/R,$$

where Q_0 and E_0 are constants and $q = \rho_f/\varepsilon_s$.

In terms of these quantities, we can express the energy densities of matter and field [2]:

$$\varepsilon_s = \frac{E'}{4\pi R^2 R'(1 - qQe^{\lambda/2}/RR')}, \quad \varepsilon_f = \frac{Q^2}{8\pi R^4}.$$

The solution to the Einstein–Maxwell equations is obtained in the cases when one of the arbitrary functions of r becomes constant [2]. There exist three types of solutions,

$$Q = Q_0, \quad R_g = R_{g0}, \quad R_f = R_{f0},$$

where

$$R_g = \kappa E/4\pi$$

is the gravitational radius,

$$R_f = Q^2/2E = R_c^2/R_g$$

is the classical (electromagnetic) radius, and

$$R_c = \sqrt{k}Q/c^2$$

is the so-called critical radius.

In the given problem, we are interested in the first type of solutions with a constant electric charge Q_0. In this case, $\rho_f = 0$; i.e., the dust is neutral. An interesting situation arises: charge Q_0 is an integral of motion of a neutral gravitating matter and, in turn, generates an electromagnetic field with energy density ε_f and field

strength $E_r = \pm\sqrt{8\pi\varepsilon_f}$ of the radial electric field in the comoving reference frame (or is generated by this field and the neutral matter).

It would be interesting to discover what the inner and outer (vacuum) worlds of the object called the electric charge are and to express its physical parameters (charge, mass, and radius) in terms of the curvature of the pseudo-Riemannian space of the inner world. Let us prove that even in the simplest centrosymmetric case in the absence of inevitable rotation, the charge is a non-Euclidean topological construction, viz., a semiconfined pulsating world with two static extremal surfaces (bottlenecks) in spacetime.

3. INNER WORLD

We choose an attractive (from the standpoint of physics) solution for $Q = Q_0$, which corresponds to a semiconfined time-periodic world [2]

$$\begin{gathered} e^{\nu} = 1, \quad e^{\lambda} = R'^2/f^2, \\ f^2 < 1, \quad 4R_f(1-f^2)/R_g < 1, \\ R = \frac{R_g}{2(1-f^2)}(1-\delta\cos\eta), \\ \tau - \tau_r = \frac{R_g}{2(1-f^2)^{3/2}}(\eta - \delta\sin\eta), \end{gathered} \tag{2}$$

where

$$\delta = \sqrt{1 - 4R_f(1-f^2)/R_g}$$

and $\tau_r(r)$ is an arbitrary function of r, which is determined by the method for measuring time in the congruence of observers.

For $Q_0 = 0$, this solution is transformed to the well-known nonstationary metrics of the Tolman–Friedmann confined world of neutral dust [1]; in the absence of matter the solution is transformed into the Reissner–Nordstrem static world of a solitary charge, which in turn is transformed to the Schwarzschild world of a point mass.

A remarkable property of metric (2) is the absence of pointlike singularity of the type of an infinite Gaussian curvature of radial spheres: since $\delta < 1$, the radius $R(\tau, r)$ of the internal scalar curvature of the 2-surface (τ = const, r = const) never vanishes anywhere ($R(\tau, r) \neq 0$) if the integral of motion $R_g(r)$ is defined appropriately.

Since $E_r = Q_0/R^2$, gravitation (i.e., the spacetime curvature) removes the Coulomb divergence of the classical point charge field in the Minkowski spacetime.

Let us specify the initial conditions for $\tau = 0$. Let us suppose that the density of matter in the state of maximal expansion of the inner world is constant:

$$\eta = \pi, \quad \dot{R}(0) = 0, \quad \varepsilon_s(0) = \varepsilon_0.$$

Here, zero in the parentheses indicates the dependence of quantities on r for $\tau = 0$. Integrating the relation

$$R_g' = \kappa R^2 R'\varepsilon_s$$

for $\tau = 0$, we obtain the relation between the total energy and the density of matter in the initial state,

$$R_g = \frac{\kappa\varepsilon_0}{3}R^3(0) + C_g, \tag{3}$$

where C_g is the integration constant. All quantities can now be expressed in terms of $R(0)$ and R (we set $C_g = 0$):

$$\begin{gathered} R_g = \frac{\kappa\varepsilon_0}{3}R^3(0), \quad R_f = \frac{3R_c^2}{\kappa\varepsilon_0}\frac{1}{R^3(0)}, \\ f^2 = 1 - \frac{R_g}{R(0)} + \frac{R_c^2}{R^2(0)}, \\ \varepsilon_s = \varepsilon_0\frac{R^2(0)R'(0)}{R^2R'}, \quad \varepsilon_f = \frac{R_c^2}{\kappa R^4}. \end{gathered}$$

4. STATIC SURFACE

Let us define the object, viz., a static 2-surface (h)

$$r = r_h, \quad R = R_h, \quad \dot{R}_h = 0.$$

The static conditions ($\dot{R}_h = 0$) lead to the conditions

$$\delta_h = 0, \quad 1 - f_h^2 = \frac{R_{gh}^2}{4R_c^2};$$

substituting these conditions into Eq. (2), we find that the scalar curvature radius of a static 2-surface is always equal to a doubled classical radius,

$$R_h = 2R_{fh},$$

i.e., to the ratio $R_h = Q_0^2/E_h$ of the squared charge to the total energy of the inner world for $r = r_h$. For $Q_0 \neq 0$, this radius always differs from zero.

All parameters on the given surface can be expressed in terms of critical radius R_c and dimensionless parameter ξ:

$$\begin{gathered} R_h = \xi R_c = \frac{R_{gh}}{2}\xi^2, \quad R_{gh} = \frac{2R_c}{\xi}, \\ 1 - f_h^2 = \frac{1}{\xi^2}. \end{gathered} \tag{4}$$

Since $f_h^2 < 1$, parameter $\xi > 1$; consequently, the radius of curvature of the static sphere is larger than the critical radius ($R_h > R_c$) for any charge.

The extremal object for which $\xi = 1$ is called a maximon. It is the only static 3-object with a constant curvature and a constant density, with a spherical system of coordinates (which is nondegenerate on its surface), with the minimal radius, and with the maximal gravitational radius.

In solving the Cauchy problem, we choose the initial conditions so that $R(0) \geq R_h$ (i.e., the radius of curvature of the static sphere has its minimal value in the initial state of maximal expansion). In this case, for the integrals of motion we have

$$R_g = R_{gh}\frac{R^3(0)}{R_h^3} \geq R_{gh}, \quad R_f = R_{fh}\frac{R_h^3}{R^3(0)} \leq R_{fh};$$

i.e., the total energy of the inner world on the static surface is minimal, while the electromagnetic field energy is maximal.

Let us define the "rest mass" m_0 as the total mass (energy) of the inner world on the static sphere,

$$R_{gh} = \frac{2kE_h}{c^4} = \frac{2km_0}{c^2}, \quad m_0 = \frac{m_c}{\xi},$$

where $m_c = Q_0/\sqrt{k}$ is the maximon mass.

Applying relation (3) to the static sphere and taking into account relation (4) between its radius and the gravitational radius, we find that the dust energy density in the initial state (and, hence, on the given sphere) is finite and unambiguously determined by its parameters,

$$\varepsilon_0 = \varepsilon_c/\xi^4,$$

where $\varepsilon_c = 6/\kappa R_c^2$ is the critical energy density.

5. MAXIMAL EXPANSION STATE

The inner world of a charge in the initial state $\eta = \pi$ can be described by the differential equation

$$e^{-\lambda(0)}(R^2(0)')^2 = 4\left[R^2(0) - \frac{2}{\xi^2 R_h^2}R^4(0) + R_c^2\right], \quad (5)$$

whose solution can be written in the form

$$R^2(0) = 2a_0^2(1 - \delta_0\cos 2\chi), \qquad \int e^{\lambda(0)/2}dr = 2a_0\chi, \quad (6)$$

where the dimensionless coordinate $\chi \in [0, \pi]$ [1],

$$\delta_0 = \sqrt{1 + 8/\xi^4}, \quad 2a_0 = \frac{\xi R_h}{\sqrt{2}} = \sqrt{\frac{3}{\kappa\varepsilon_0}},$$

and $e^{\lambda(0)}$ is a function of r, which depends on the method for measuring radial lengths.

Solution (6) implies that a semiconfined world has two static spheres (geometrical images of charges with opposite polarities, i.e., charged particles–antiparticles), on which the radial electric field has opposite directions. These spheres are located at points χ_h and $\pi - \chi_h$, where

$$\chi_h = \frac{1}{2}\arccos\frac{1 - 4/\xi^2}{\delta_0}.$$

The given solution for $R_c = 0$ is transformed into the well-known solution for the Tolman world [1]. The maximal *radial length* of the inner world in the confined model is given by

$$l_r(\pi) = \int e^{\lambda(0)/2}dr = 2\pi a_0.$$

Thus, the radial length of the inner world in the maximal expansion state is defined, in accordance with relation (6), by the energy density of the matter in the initial state.

For $\eta = \pi$ and $\chi = \pi/2$, quantities $R(0)$ and R_g assume their maximal values, such that $R_{\max} \approx R_{g\max}$:

$$R_{\max} = \frac{\xi R_h}{2}\sqrt{1 + \delta_0} = \sqrt{2}a_0\sqrt{1 + \delta_0},$$

$$R_{g\max} = \frac{\xi R_h}{4}(1 + \delta_0)^{3/2} = \frac{a_0}{\sqrt{2}}(1 + \delta_0)^{3/2}.$$

The maximal value of the gravitational radius (total energy) can be juxtaposed to the total mass M of the inner world in the maximal expansion state:

$$R_{g\max} = 2kM/c^2.$$

6. GEOMETRIZATION OF THE CHARGE

The curvature ${}_\alpha K_\beta^{(a)}$ of the 2-surface $S^{(2)}$ formed by the coordinate lines $\{x^\mu, x^\nu\}$ is perpendicular to coordinates x^α, x^β, $\alpha \neq \beta \neq \mu \neq \nu$, and observed from the measurement space a ($a = 2, 3, 4, \ldots$) can be expressed in terms of the Riemann–Christoffel tensor of the corresponding space and the modulus of the metric on the surface [1]:

$${}_\alpha K_\beta^{(a)} = R^{(a)}_{\mu\nu\mu\nu}/(g_{\mu\mu}g_{\nu\nu} - g_{\mu\nu}^2) \quad (7)$$

(summation over the indices is not envisaged). In metric (1), we obtain from expression (7)

$$
\begin{aligned}
&{}_0K_r^{(2)} = \frac{1}{R^2}, \quad {}_0K_r^{(3)} = \frac{1}{R^2}(1 - e^{-\lambda}R'^2),\\
&{}_0K_r^{(4)} = \frac{1}{R^2}(1 - e^{-\lambda}R'^2 + e^{-\nu}\dot{R}^2),\\
&{}_0K_\varphi^{(2)} = {}_0K_\varphi^{(3)}, \quad {}_0K_\varphi^{(3)} = \frac{1}{R^2}(1 - e^{-\lambda}R'^2)',\\
&{}_0K_\varphi^{(4)} = \frac{1}{R^2}(1 - e^{-\lambda}R'^2 + e^{-\nu}\dot{R}^2)',\\
&{}_0K_\theta^{(a)} = {}_0K_\varphi^{(a)}.
\end{aligned}
\tag{8, 9}
$$

The curvature of the 3-hypersurface $S^{(3)}$ orthogonal to coordinate x^α is equal to the sum of curvatures (7) over index β:

$$
{}_\alpha K^{(a)} = {}_\alpha K_\beta^{(a)} + {}_\alpha K_\gamma^{(a)} + {}_\alpha K_\delta^{(a)}, \quad \alpha \neq \beta \neq \gamma \neq \delta. \tag{10}
$$

The sum of 4-curvatures of all area elements orthogonal to the x^0 axis is equal to the G_0^0 component of the Einstein tensor,

$${}_0K^{(4)} = {}_0K_r^{(4)} + 2{}_0K_\varphi^{(4)} = G_0^0 = \kappa(\varepsilon_f + \varepsilon_s). \tag{11}$$

The sum of 4-curvatures of all area elements orthogonal to the x^1 axis is equal to the G_1^1 component of the Einstein tensor,

$${}_rK^{(4)} = {}_rK_0^{(4)} + 2{}_rK_\varphi^{(4)} = G_1^1 = \kappa(\varepsilon_f - p_s),$$

where p_s is the pressure of the matter (which differs from zero in the general case).

The scalar curvature of the 4-space (Gaussian, or internal, curvature [1]) is equal to the sum of all curvatures (10) orthogonal to the axes x^0, x^1, x^2, and x^3; in the present case, it is given by

$$
\begin{aligned}
K^{(4)} &= {}_0K^{(4)} + {}_rK^{(4)} + 2{}_\varphi K^{(4)}\\
&= G = G_0^0 + G_1^1 + 2G_2^2 = \kappa\varepsilon_s = R_g'/R^2R'.
\end{aligned}
\tag{12}
$$

The Einstein–Maxwell equations lead to the following relation between curvatures (7)–(12) and physical characteristics:

$$
\begin{aligned}
\kappa\varepsilon_s &= K^{(4)},\\
\kappa\varepsilon_f &= {}_0K^{(4)} - K^{(4)},\\
1 - f^2 &= {}_0K_r^{(2)-1}{}_0K_r^{(3)},\\
R_c &= {}_0K_r^{(2)-1}({}_0K^{(4)} - K^{(4)})^{1/2},\\
R_g &= {}_0K_r^{(2)-3/2}({}_0K^{(4)} - K^{(4)} + {}_0K_r^{(4)}).
\end{aligned}
\tag{13}
$$

It was noted above that ${}_0K_r^{(a)} \neq \infty$; i.e., singularity of radial spheres is absent ($\varepsilon_f \neq \infty$).

Relations (13) make it possible to express two fundamental constants (charge e and rest mass m_0), which are the first integrals of the given gravitating system (m_0 is equal to the total mass $M(r)$ of the inner world on the static sphere $r = r_h$) in terms of curvatures and two other constants, viz., c (corresponding to the presence of a light cone for the chosen signature) and k (relating the geometry to the physics):

$$
\begin{aligned}
e &= \frac{c^2}{\sqrt{k}}[{}_0K_r^{(2)-1}({}_0K^{(4)} - K^{(4)})^{1/2}],\\
m_0 &= \frac{c^2}{2k}[{}_0K_r^{(2)-3/2}({}_0K^{(4)} - K^{(4)} + {}_0K_r^{(4)})].
\end{aligned}
\tag{14}
$$

To find constant e, we can determine the curvatures at any point of the inner and outer vacuum world of the electric charge ($K^{(4)} = 0$ in vacuum). Constant m_0 can be sought from the curvatures at any point of the vacuum world and on the static sphere. Inside the charge, the mass will be a function of r.

The electromagnetic field of a charge that is at rest in vacuum, which is represented outside and inside by the radial electric field in a reference frame comoving with the dustlike matter, can also be expressed in the entire space in terms of its curvatures:

$$E_r = \frac{c^2}{\sqrt{k}}({}_0K^{(4)} - K^{(4)})^{1/2}.$$

Thus, an interesting possibility of experimental determination of physical parameters of objects from measuring geometrical quantities; for fundamental constants, the same values will be obtained at any point in space.

7. BOTTLENECK

Let us define a bottleneck in the spacetime as a 2-surface of extremal curvature. In the simplest centrosymmetric case, we can speak of the bottleneck as an extremal surface of radial spheres, which is orthogonal to time and radial coordinates. If the bottleneck is static in the comoving reference frame (i.e., its curvature does not change with time and it does not move along the radial coordinate), it coincides with a static sphere and the conditions for its existence have the form

$${}_0K_{rh}^{(4)} > 0, \quad {}_0K_{rh}^{(4)'} = 0, \quad {}_0\dot{K}_{rh}^{(4)} = 0. \tag{15}$$

The bottleneck will display the maximal curvature of radial spheres ($\eta = \pi$) or inflection for ${}_0K_{rh}^{(4)''} \leq 0$, and the minimal curvature for ${}_0K_{rh}^{(0)''} > 0$ ($\eta = 0$). Substitut-

ing expressions (8) for curvatures into conditions (15), we obtain

$$
\begin{aligned}
{}_0K_r^{(4)'} &= R^{2'}\frac{{}_0K_\varphi^{(4)} - {}_0K_r^{(4)}}{R^2} \\
&= \frac{R'}{R^4}\left(\frac{4R_c^2}{R} - 3R_g + \kappa\varepsilon_s R^3\right) = 0,
\end{aligned}
\tag{16}
$$

$$
{}_0\dot{K}_r^{(4)} = \frac{\dot{R}}{R^4}\left(\frac{4R_c^2}{R} - 3R_g\right) = 0, \tag{17}
$$

which means that a static bottleneck exists for $\dot{R}_h = 0$ and either for

$$
\kappa\varepsilon_0 R_h^3 = 3R_{gh} - 4R_c^2/R_h,
$$

or for $R_h' = 0$. In the former case, the quantity C_g in expression (3) for the gravitational radius cannot be set equal to zero any longer; $R(0)$ as a solution to Eq. (5) will be expressed in terms of elliptical functions. Let us consider the second condition $R_h' = 0$.

Since $e^{\lambda_h} = R'^2/f^2 = 0$ for $\xi \neq 1$, the determinant of the metric tensor vanishes at the bottleneck. Consequently, the spherical system of coordinates degenerates on it. However, all its geometrical parameters (curvatures) and the corresponding physical quantities (mass, dust density, electromagnetic field energy density, and field strength) are finite; i.e., this singularity is of purely coordinate nature. It is not reflected in physics or geometry in any way.

It should also be noted that function $e^{\lambda(0)}$ in solution (6) must vanish at the bottleneck in this case.

8. JOINING WITH VACUUM

Solution (2) covers the entire spacetime and does not require any supplements. Nevertheless, using static bottlenecks, the inner semiconfined world (2) can be smoothly continued to two Reissner–Nordstrem vacuum worlds. In the curvature coordinates, we have

$$
\begin{aligned}
&R_g' = 0, \quad \dot{R} = 0, \quad R' = 1, \quad R = r, \\
&\varepsilon_s = 0, \quad e^\nu = e^{-\lambda} = A(r), \\
&ds^2 = A(r)d\tau^2 - \frac{dr^2}{A(r)} - r^2 d\sigma^2, \\
&A(r) = 1 - \frac{r_g}{r} + \frac{r_c^2}{r^2},
\end{aligned}
\tag{18}
$$

where

$$
r_g = R_{gh} = \frac{2km_0}{c^2}, \quad r_c = R_c = \frac{Q_0\sqrt{k}}{c^2}.
$$

Since joining is carried out over the bottleneck whose curvature is extremal from inside and on which the metric has a coordinate singularity $g_{11} = 0$, metric (18) must be transformed to another radial coordinate $\tilde{r}$ [1] to nullify the metric coefficient $g_{\tilde{1}\tilde{1}}$ on the 2-surface of joining $r = 2r_f$ as in the case of the internal solution. Naturally, the transformation Jacobian $J = r_{,\tilde{r}}$ also vanishes on this surface (e.g., $r = \tilde{r} + 4r_f^2/(2r_f + \tilde{r})$, where $r_f = R_{fh} = Q_0^2/2E_h^2$). Direct substitution of an arbitrary transformation $r = r(\tilde{r})$ into the Einstein equations [1] under the condition $J_h = 0$ readily shows that transformed metric (18), which is independent of the world time, satisfies these equations at the bottleneck.

At the bottleneck, $R_h' = r_{,\tilde{r}} = 0$ and the quantity G_1^1 (the mixed component of the Einstein conservative tensor) also turns out to be continuous, which is physically equivalent to the continuity of the electric field upon a transition to vacuum. Component G_0^0 of this tensor experiences a first-kind discontinuity, which corresponds to a sharp dust–vacuum interface from the standpoint of physics.

The joining procedure satisfies the Likhnerovich conditions: if $f(x^\mu) = 0$ (equation for the joining surface, i.e., the equation $r - r_h = 0$ in the present case), product $G_\mu^\nu f_{,\nu}$ is found to be continuous. In fact, cutting a part $r < 2r_f$ from the Reissner–Nordstrem metric, we discard the singularity $r = 0$ inherent in the vacuum solution and obtain an extended (bulk) material Reissner–Nordstrem field source.

Figure 1 shows qualitatively the hierarchy of spaces for various simple field sources, where the three-dimensional hypersurface (observed physical space) is represented for better visualization by the curve along which coordinate r changes (i.e., the cross section of a 2D surface of a revolution with one of cyclic coordinates, θ or φ, varying along its second direction). The distance from the rotational axis is proportional to radius $R(r)$ of the 2D Gaussian curvature; the convexity or concavity of the surface depends on the signs of curvatures ${}_0K_r^{(3)}$ and ${}_0K_\varphi^{(3)}$. It can be seen from Fig. 1 that the given field had to be slightly deformed by a transformation (dotted curve) for passing from a pointlike to a bulk source in the Reissner–Nordstrem metric and for its smooth joining with the internal solution.

Thus, the electric charge appears geometrically as two parallel vacuum spaces in which tensions are equivalent to an electric field and which are connected via two static bottlenecks with a radius equal to a doubled classical radius. Between these bottlenecks, a nonstationary confined world is situated, in which tensions are equivalent to neutral dustlike matter and an electric field, and which pulsates from the maximal expansion

state to the maximal compression state. This pattern is shown qualitatively in Fig. 2 in the form of a 2-surface of revolution in accordance with the same rules as the curves in Fig. 1. Quantity $R(\eta, \chi)$ does not vanish anywhere. The strength of the radial electric field attains its maximal value at the bottlenecks and decreases in inverse proportion to the square of the Gaussian curvature radius with increasing distance from the bottlenecks to the inner or outer space of the charge.

The only divergence preserved in view of central symmetry and nonstationarity of the metric is the kink on the 2-surface formed by coordinates $\{r, \theta\}$ or $\{r, \varphi\}$ when expansion is replaced by compression. The kink moves according to a definite law $\chi(\eta)$ (singularity of the type $R' = 0$, $R'_g \neq 0$, ${}_0K_\varphi^{(a)} \longrightarrow \infty$) and corresponds to an infinitely large value of the energy of the matter ($\varepsilon_s \longrightarrow \infty$). However, this singularity is immaterial; it appears due to the simplicity of the model and is absent at the bottleneck.

9. EXTERNAL AND INTERNAL PARAMETERS OF THE CHARGE

Let us suppose that the electric charge of the world is equal to the fundamental charge ($Q_0 = e$) and that constants c, k, and e have experimental values. Then the parameters of the world at the bottleneck and inside the charge are determined by the value of ξ. The maximon ($\xi = 1$) is an object with the critical parameters

$$f_h = 0, \quad \delta = 0, \quad \delta_0 = 3, \quad \dot{R} = R' = 0,$$
$$R_h = 2R_{fh} = R_{gh}/2 = R_c = R_{\max} = R_{\min} = 2\sqrt{2}a_0, \tag{19}$$
$$m_0 = M = m_c = \frac{e}{\sqrt{k}}, \quad \varepsilon_0 = \frac{6}{\kappa R_c^2} = \varepsilon_c,$$
$$R_c = \frac{km_c}{c^2}.$$

Here, $R_{\min}$ is the radius of curvature for $\eta = 0$ and $\chi = \pi/2$, i.e., in the state of maximal compression. The parameters of the remaining objects can be expressed in terms of the critical parameters (19):

external,

$$R_h = \xi R_c, \quad m_0 = m_c/\xi, \tag{20a}$$

and internal,

$$R_{\max} = \frac{\xi^2}{2}\sqrt{1+\delta_0}R_c,$$
$$M = \frac{\xi^2}{8}(1+\delta_0)^{3/2}m_c, \tag{20b}$$
$$\varepsilon_0 = \rho_0 c^2 = \varepsilon_c/\xi^4.$$

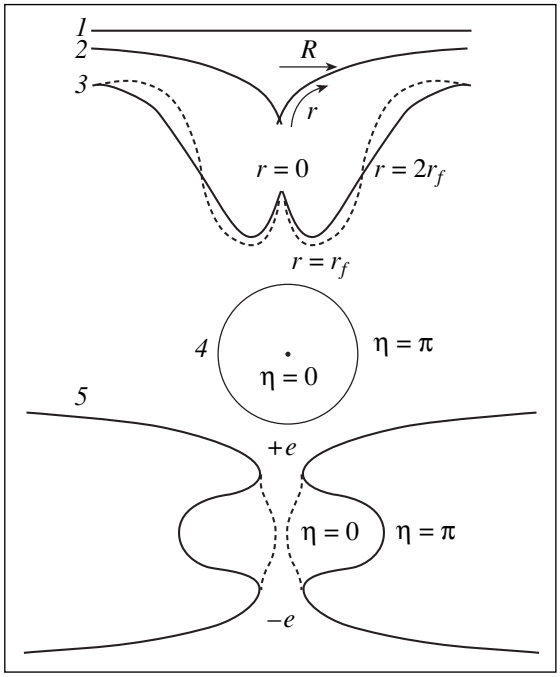

Fig. 1. Geometry of centrosymmetric metrics: planar empty Minkowski spacetime (*1*); Schwarzschild solution for the vacuum field of point mass m_0 (*2*); Reissner–Nordstrem solution for the vacuum field of point charge e having a rest mass of m_0 (dotted curves show the geometry of the transformed metric with a Gaussian curvature extremum for $r = 2r_f$) (*3*); Tolman–Friedmann solution for a confined world of dustlike matter in the maximal expansion state ($\eta = \pi$), having a singularity in the maximal compression state ($\eta = 0$) (*4*); solution for the inner world of an electric charge consisting of neutral dust and a radial electric field pulsating from the maximal expansion state ($\eta = \pi$) joined by two parallel vacuum Reissner–Nordstrem worlds via two static bottlenecks (charges $+e$, $-e$) (*5*).

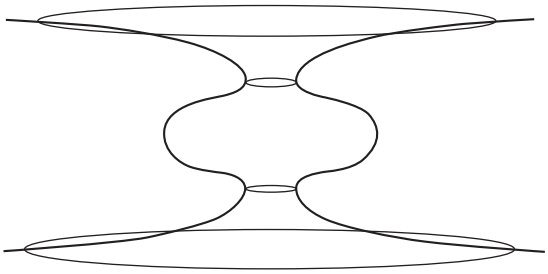

Fig. 2. Geometry of the inner and outer worlds of an electric charge: the physical 3-hypersurface is simplified to a 2-surface of revolution, which is formed by the radial and one of cyclic coordinates. This surface is connected with the parallel vacuum world and an antiworld via two bottlenecks (equivalent to a particle and an antiparticle with opposite charges).

In this model, the value of ξ is not bounded from above. An important factor is that the experimental value of ξ for the electron is known exactly from its external parameters (20a):

$$\xi_e = e/\sqrt{k}m_0 = 2.04 \times 10^{21}$$

Spectrum of electric charge parameters

Parameters			External		Internal		
n	Object	$\xi = q^n$	R_h, cm	m_0, g	R_{max}, cm	M, g	ρ_0, g/cm^3
0	Maximon	1	1.38×10^{-34}	1.86×10^{-6}	1.38×10^{-34}	1.86×10^{-6}	1.69×10^{95}
1	Mifion	4.52×10^{10}	0.62×10^{-23}	4.11×10^{-17}	1.99×10^{-13}	1.34×10^{15}	4.06×10^{52}
2	Electron	2.04×10^{21}	2.82×10^{-13}	0.91×10^{-27}	4.06×10^{8}	2.73×10^{36}	0.98×10^{10}
3	Universe	0.92×10^{32}	1.27×10^{-2}	2.02×10^{-38}	0.83×10^{30}	0.56×10^{58}	2.35×10^{-33}

(it should be recalled that we are speaking here of an abstract object with a charge and a rest mass of a real elementary particle, electron, naturally disregarding rotation in the simplest case of central symmetry). The experimental estimate of the value of ξ for our universe is also known, but now from its internal parameters (20b), viz., the total mass and radius, under the assumption that a closed cosmological model is applicable [1]:

$$\xi_u \approx (2\sqrt{2}M/m_c)^{1/2} \approx (\sqrt{2}R_{max}/R_c)^{1/2}$$
$$(\delta_0 \approx 1 \text{ for the Universe}).$$

We assume that the electric charge of the Universe is also equal to the fundamental charge.

Consequently, the internal parameters of the electron and, accordingly, the external parameter of our universe are known exactly. The result is curious: the electron contains a universe whose mass in the state of maximal expansion is $M = 2.73 \times 10^{36}$ g (i.e., on the order of 10^3 masses of the Sun) and whose radius $R_{max} = 4.06 \times 10^8$ cm (i.e., on the order of the Earth's radius), and our universe appears from outside as a particle (bottleneck, see Fig. 2) having a very small rest mass $m_0 \approx 2 \times 10^{-37}$ g and a curvature radius of $R_h \approx 10^{-3}$ cm.

It is interesting to note that, if we assume that the internal radius and the total mass of the observed universe are two orders of magnitude higher than the value estimated from the Hubble constant and apply relations (20) to charges with parameters of the known objects (maximon, electron, and universe), we will readily find that they form a power series,

$$\xi = q^n,$$

where $q = 4.52 \times 10^{10}$. Then one more "particle" (with $n = 1$), which can be conditionally called the "mifion," can exist between the maximon ($n = 0$) and the electron ($n = 2$). The electron is followed by the universe ($n = 3$) (see table).

10. COSMOLOGICAL CONSEQUENCE

The fact that the internal radius and mass of the universe can be larger than those estimated in astrophysics in the approximation of the invariability of the Hubble constant h and from the linear Doppler effect does not contradict the Einstein–Friedman cosmology [1]: over long distances and time intervals, we must take into account the fact that "radius" $a(\eta)$ does not remain constant during light propagation in an expanding hypersphere, but increases, and that the relationship of the observed frequency of light with the radiation frequency and the velocity of the source is determined not by the linear Lorentz transformations (in which the velocity is limited to the velocity of light), but by general covariant transformations.

The relation between the "recession velocity" $\beta = \chi d\ln a/d\eta$ and the "red shift" in the homogeneous model disregarding the charge (in the Friedmann–Tolman metric) has the form [1]

$$\frac{\omega}{\omega_0} = \left(\sin\frac{\eta - \chi}{2}/\sin\frac{\eta}{2}\right)^2,$$
$$\beta = \frac{hl}{c} = \chi\cot\frac{\eta}{2}, \qquad (21)$$
$$h = \frac{c}{a^2}\frac{da}{d\eta} = \frac{c}{2a_0}\frac{\cos(\eta/2)}{\sin^3(\eta/2)}.$$

These relations show that the Hubble constant h and the recession velocity β turn to infinity at singularity ($\eta = 0$) and are zero at the maximal expansion state ($\eta = \pi$). This does not lead to any contradiction since β in the present case is not a physical velocity: all points of the hypersphere are at rest relative to it, and the space itself expands.

11. EXPERIMENTAL CONFIRMATIONS

The assumption that the charge of the Universe is equal to e does not contradict the astrophysical data indicating the presence of a nonvanishing mean electromagnetic field in it. It is this field, which is strong enough in a state close to maximal compression, that ensures acceleration of relict particles generated ("tempered") at this stage of evolution of the universe to ultrahigh energies (10^{21}–10^{22} eV) [2]. In the present state, which is close to maximal expansion, these

particles become primary sources of experimentally observed extensive air showers (EAS).

It is difficult to explain the presence of such particles in cosmic noise by other mechanisms (stellar or galactic). A particle may accumulate an ultrarelativistic energy of 10^{22} eV on the scale of the Universe. Due to radiation losses during accelerated motion along a curvilinear trajectory, this energy is lower than the critical energy $ec^2/\sqrt{k}$ (the maximon rest energy), which is on the order of 10^{28} eV [2].

Further, a rigorous result of the theory is that a charge with the parameters of the electron contains (in the case of a uniform density of the inner world in the initial state) a universe with a mass on the order of 10^3 masses of the Sun. This can also be verified experimentally if we assume that bursts of supernovas observed in astrophysics can be interpreted not as a result of a catastrophic collapse of burnt-out cooled stars, but as the release of a part of the internal energy by an elementary particle that loses its stability for one reason or other (say, as a result of insufficiently considered experiments on accelerators of civilizations existing on planets of the given galaxies).

12. CONCLUSIONS

First integrals of the Einstein–Maxwell equations for the system under study were obtained by Markov and Frolov in 1972 [3]. In [2], these equations were integrated completely. The exact solution of the Einstein–Maxwell equations for the inner space of a spherically symmetric electric charge [2, 4] implies that

(i) the electric charge is a gravitational object in the GTR; the radius of the Gaussian curvature of the bottleneck connecting the inner nonstationary pulsating semiconfined world of dust and the outer Reissner–Nordstrem vacuum world is equal to a doubled classical radius; (ii) the space curvature (gravitational field) removes the Coulomb divergence of the field produced by a point charge in a planar Minkowski spacetime; the radial electric field in the comoving reference frame attains its maximal value on two parallel bottlenecks and decreases in inverse proportion to the squared radius of the Gaussian curvature with increasing distance from the bottlenecks to the bulk of the charge and to vacuum; (iii) physical constants e and m_0 are first integrals of the Einstein–Maxwell equations; all physical parameters (electric charge, rest mass, radial electric field, and dust density) can be expressed in terms of the curvatures of the 4-space and can be determined by measuring the curvatures at any point of the space; (iv) the rest mass is the total (gravitational and observed) mass of the inner world at the bottleneck; the smallness of the gravitational radius as compared to the classical radius of many "elementary" particles indicates not a negligibly small role of gravitational effects over the classical length, but rather a strong "gravitational mass defect" of the inner world at the bottleneck due to the focusing (attracting) action of the gravitational field, when the large space curvature reduces the observed mass of the object; for a charge with the electron parameters, the value of $\xi \gg 1$: $e \gg \sqrt{k}\, m_0$; i.e., the electric charge is much larger than the "gravitational charge"; (v) an elementary particle (electron, proton, maximon, etc., in the "nonquantum" representation of the GTR) and the Universe are formally a single object considered from outside and inside; i.e., the micro- and macroworlds are identical; for uniform initial conditions, the "electron" contains a universe with a mass on the order of 10^3 masses of the Sun and a maximal radius on the order of the Earth's radius, while our Universe (if its charge is e) appears from outside as a particle whose mass is on the order of 10^{-37}–10^{-38} g and a bottleneck radius on the order of 10^{-3}–10^{-2} cm; (vi) in the given problem, the global spacetime is topologically nontrivial and "layered" (this should not be confused with stratification in the gauge field theory); various objects (electron, universe, etc.) are tunnels connecting these parallel layers; this solution theoretically confirms the correctness of the ideas concerning the neighborhood of the GTP (e.g., Wheeler's mole burrows [5]); if an "electron" emerges in the world (the bottleneck corresponding to a negative charge), a "positron" emerges in the parallel space (the bottleneck with a positive charge), which is an antiworld; thus, a world of particles and an antiworld of charged antiparticles are located on two parallel orientable 3-hypersurfaces; (vii) since a charge particle consists of dust, which in turn is formed by charged particles (burrows between the layers of vacuum spaces), the space as a whole can be supplied with a nontrivial topological structure of a closed set that is not dense anywhere (an everywhere "perforated" Cantor-type set); consequently, the general problem of existence might have a paradoxical solution: there exists something equal to zero (having zero measure).

The problem considered here shows that the Einstein GTR can be geometrized: the electromagnetic field and matter have a gravitational (geometric) mapping. This statement is of methodological importance: the prevailing opinion that gravitation is a field equivalent in properties to other physical fields and having only a geometrical interpretation like other fields is archaic. On the contrary, the gravitational field has a unifying meaning: any physical field possessing an energy–momentum tensor can be mapped on the geometry of space whose curvature is precisely the gravitational field. Matter is equivalent to a gravitational field, which is equivalent to curved space. This is the essence of the Einstein equations proper and of the rigorous GTR principle of equivalence. All other formulations of this principle (equality of the inertial and gravitational masses, local "vanishability" of gravitational field, local "rectifiability" of space, etc.) are of limited (nonrelativistic) nature.

The prevailing idea that the gravitational field is significant either over limiting lengths of the Planck type (which is approximately an order of magnitude larger than the critical radius) in the microworld or on the scale of the Universe in the megaworld is refined in the GTR: gravitational fields are manifested over any length as the maps of physical fields onto spacetime geometry. Thus, a gravitational field cannot be "weaker" than, say, an electromagnetic field since it is precisely this field in the form of tensions of the curved space.

It should be noted that the traditional concept on a gravitational field as a "classical" field can also be revised in the near future. The GTR sets no intrinsic limitations on the values of any parameters. Conversely, not gravitational field should be subjected to quantization, but the origin of discrete quantum effects should be explained with the help of a "continuous" gravitational field (it was mentioned above that this formed the initial Einstein program).

ACKNOWLEDGMENTS

The author is obliged to mention Ya.B. Zeldovich, who once showed his interest in the problem, and to thank N.V. Mitskevich, I.D. Novikov, and V.P. Frolov for their remarks made in the 1970–80s. Thanks are also due to Yu.S. Vladimirov, V.N. Mel'nikov, and participants of their seminars (especially K.A. Bronnikov); participants of the seminar carried out at the Moscow Physical Engineering Institute (especially N.S. Trushkin); and also E.D. Zhizhin and A.V. Berkov for discussions and valuable remarks. The author is grateful to B.Yu. Bogdanovich, B.N. Onykiĭ, and É.Ya. Shkolnikov for their support and to M.Yu. Lukashin, L.A. Sukhanova, and A.Yu. Khlestkov for his help in this research.

APPENDIX

A Brief Review of Publications on Nontrivial Geometrical Structures in the Theory of Gravitation

Obtaining regular solutions to the Einstein equations with a nontrivial topology (black holes, bottlenecks, burrows, tubes, bubbles, etc.) has been underway for a long time and certain advances have been made in this direction (see, for example, [6–15]). However, the general conclusion that has been drawn is distressing [6, 9]: a space with a bottleneck (horn, wormhole) and two asymptotically planar worlds can be constructed only in "pathological" cases from the standpoint of the generally accepted concept of causality (negative energy density of matter generating a gravitational field, violation of the weak energy condition [16], etc.).

Nevertheless, we have obtained a solution that describes the internal structure of an electric charge without any singularities (geometrical or physical) at the bottleneck. This can be explained by the difference in the formulation of the problem. In previous works, static ($\dot{R} = 0$) vacuum ($\varepsilon_s = 0$, $\rho_f = 0$) worlds with a scalar field, as well as worlds with an electric and scalar field with different Lagrangians were considered in the framework of spherical symmetry [6–15]. In most cases, metric (1) of these worlds can be reduced to the form [6] $\nu(r) + \lambda(r) = 0$. For such worlds, the right-hand side of the Einstein equations implies that the difference between two mixed components of the conservative Einstein tensor must be

$$G_0^0 - G_1^1 = \kappa(\varepsilon_s + p_s);$$

on the other hand, it follows from the left-hand side of the Einstein equations that this difference is given by

$$-e^{-\lambda(r)}R''/R.$$

In [6–15], the bottleneck was defined as an infinitely long 3D tube with a finite radius of curvature $R(r)$, which is minimal on the 2-surface $r = r_h$, i.e.,

$$R_h' = 0, \quad R_h'' > 0.$$

It was also assumed that metric coefficient e^{ν_h} on this surface has a finite value. It turns out that these conditions can be satisfied only for a negative sum of the energy density and pressure of any matter.

Worlds whose source was supplemented with the so-called cosmological Λ term were considered as a version. If the latter term is identified as the first type of energy–momentum tensor according to Petrov [1], we can consider matter with an energy density Λ/κ and isotropic pressure $-\Lambda/\kappa$ as exotic (i.e., matter with an ultrarelativistic equation of state for which the sign of energy density is opposite to the sign of pressure). This is usually regarded as pathology both for $\Lambda > 0$ and for $\Lambda < 0$.

Here, we managed to avoid this owing to another formulation of the problem: we considered a class of nonstationary metrics (1), $\dot{R} \neq 0$, generated by dustlike neutral matter and an electromagnetic field, which is represented by a radial electric field in the reference frame comoving with the dust.

Now, the left-hand side of the difference of two Einstein equations appears quite differently. It does not contain the second derivative of radius R of the Gaussian curvature with respect to coordinate r and, by virtue of the solution to the Einstein–Maxwell equations, is identically equal to the positive right-hand side:

$$G_0^0 - G_1^1 = 2(1 - f(r)^2 + \dot{R}^2)'/R^{2'} - 2\ddot{R}/R = \kappa\varepsilon_s > 0.$$

This expression (as well as other physical and geometrical quantities) remains finite at the bottleneck. If we define the bottleneck as a 2-surface on which $R_h' = 0$,

$\dot{R}_h = 0$ (this does not exhaust all possible definitions of the bottleneck), then $e^{\lambda_h} = 0$ for $\xi > 1$; i.e., only a purely coordinate singularity exists at it (the spherical system of coordinates degenerates).

REFERENCES

1. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 2: *The Classical Theory of Fields*, 5th ed. (Nauka, Moscow, 1967; Pergamon, Oxford, 1975), Chaps. 10–12.
2. Yu. A. Khlestkov, Zh. Éksp. Teor. Fiz. **68**, 387 (1975) [Sov. Phys. JETP **41**, 188 (1975)].
3. M. A. Markov and V. P. Frolov, Teor. Mat. Fiz. **13**, 41 (1972).
4. Yu. A. Khlestkov, in *Collection of Scientific Works of Moscow Institute of Engineering Physics* (Moscow, 2003), Vol. 5, p. 177 [in Russian].
5. J. A. Wheeler, Ann. Phys. (N.Y.) **2**, 604 (1957).
6. K. A. Bronnikov, gr-qc/0104092.
7. I. Dimnikova, gr-qc/0010016.
8. K. A. Bronnikov, Phys. Rev. D **63**, 044005 (2001).
9. D. V. Gal'tsov and J. P. S. Lemos, gr-qc/0008076.
10. J. D. Bekenstein, gr-qc/9808028.
11. S. V. Krasnikov, Phys. Rev. D **62**, 084028 (2000); gr-qc/9909016.
12. K. A. Bronnikov and S. Grinyok, Gravit. Cosmol. **7**, 297 (2001); gr-qc/0201083.
13. S. V. Sushkov and S.-W. Kim, Class. Quantum Grav. **19**, 4909 (2002); gr-qc/0208069.
14. S. A. Hayward, gr-qc/0306051.
15. S. Kim and H. Lee, Phys. Rev. D **63**, 064014 (2001); gr-qc/0102077.
16. S. W. Hawking and G. F. R. Ellis, *The Large-Scale Structure of Spacetime* (Cambridge Univ. Press, Cambridge, 1973; Mir, Moscow, 1977).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 270–279.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 312–321.

NUCLEI, PARTICLES, FIELDS, GRAVITATION, AND ASTROPHYSICS

Effective Dynamics of an Electrically Charged String with a Current

P. O. Kazinski
Tomsk State University, Tomsk, 634050 Russia
e-mail: kpo@phys.tsu.ru
Received February 21, 2005

Abstract—Equations of motion for an electrically charged string with a current in an external electromagnetic field with regard to the first correction due to the self-action are derived. It is shown that the reparameterization invariance of the free action of the string imposes constraints on the possible form of the current. The effective equations of motion are obtained for an absolutely elastic charged string in the form of a ring (circle). Equations for the external electromagnetic fields that admit stationary states of such a ring are derived. Solutions to the effective equations of motion of an absolutely elastic charged ring in the absence of external fields as well as in an external uniform magnetic field are obtained. In the latter case, the frequency at which one can observe radiation emitted by the ring is evaluated. A model of an absolutely nonstretchable charged string with a current is proposed. The effective equations of motion are derived within this model, and a class of solutions to these equations is found. *© 2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

The description of the effective dynamics of electrically charged low-dimensional objects, such as particles, strings, and membranes, is one of the traditional problems in classical electrodynamics. The application of such models allows one to considerably simplify the solution of the system of Maxwell–Lorentz integrodifferential equations. For a nonrelativistic charged particle, effective equations of motion were obtained already by Lorentz [1]. The relativistic generalization of the Lorentz equations was obtained by Dirac [2]. At present, the effective equations of motion are known for a point charge in curved-background spacetime [3], for a spinning particle [4, 5], for a massive particle in higher dimensions [6, 7], and for a massless charged particle in four-dimensional spacetime [8]. The general scheme for the description of the self-action of electric currents in the string models is given in [9]. In [10], the general theory of moving, electrically charged relativistic membranes is described.

In the present paper, we present an approximate (neglecting the effect of radiative friction) Poincaré-invariant description of the effective dynamics of a thin electrically charged string with a current. The importance of studying the effective dynamics of such strings is beyond doubt because of the numerous applications, both in practice and theoretical models, of extended charged and/or conducting objects with negligible transverse dimensions. For instance, the effective equations of motion obtained in Section 2 are applied to two specific models of strings in Sections 3 and 4. In Section 3, we consider the effective dynamics of an absolutely elastic ring-shaped charged string.[1] This model describes the dynamics of a high-current beam of charged particles that move along a circle. In Section 4, we study the effective dynamics of an absolutely nonstretchable charged string with a current. In Section 2, we derive the effective equations of motion for a charged string with a current and discuss some of their properties; in particular, in the case of reparameterization-invariant free action of a string, we find the generators of gauge transformations and the constraints on the possible form of the current that flows along the string.

We will describe a charged string within the model of an infinitely thin string. It is well known that the self-action of such a string leads to a diverging expression for the force of the self-action, because infinitely close points of an infinitely thin charged string interact with infinite force. The regularization procedure, whose physical meaning consists in "smearing" a singular source of the electromagnetic field, allows one to represent the self-action force as an asymptotic series in the regularization parameter—the cross-section radius of the string—which contains one logarithmically divergent term. The smaller the cross-section radius of the string, the greater the contribution of this divergent term to the self-action force. For a sufficiently thin

[1] We define an absolutely elastic string as a string that does not significantly resist both external forces and the forces induced by its own fields. For example, an imaginary line with a current may serve as such a string. One should not confuse this concept with the well-known model of the Nambu–Goto string in the limit of zero tension (see, for example, [11]), where the string yet has its own dynamics.

1063-7761/05/10102-0270$26.00

string, one can neglect other terms of the asymptotic series to obtain effective equations of motion for a thin charged string in the form of a system of differential equations rather than integrodifferential equations, as in the case when all terms of the asymptotic series are taken into account. Similar equations are obtained when describing the effective dynamics of cosmic strings (see survey [12]).

2. A CHARGED STRING WITH A CURRENT

In this section, we find the leading contribution of the self-action of an electrically charged string with a current and derive equations of motion for the string in an external field with regard to this correction. We show that the requirement of the reparameterization invariance of the free action of a string imposes constraints on the possible form of the current flowing through the string.

Suppose given a closed string N with coordinates $\{\tau, \sigma\}$, $\sigma \in [0, 2\pi)$, that is embedded, by a smooth mapping $x(\tau, \sigma)$, into the Minkowski space $\mathbb{R}^{3,1}$ with coordinates $\{x^\mu\}$, $\mu = 0, 1, 2, 3$, and the metric $\eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$. Suppose that $e(\tau, \sigma)$ is a vector density, defined on the string N, that characterizes the electric current flowing through the string. Then, from the viewpoint of an ambient space, the current density is given by (see, for example, [10])

$$j^\mu(x) = c\int \delta^4(x - x(\tau, \sigma))e^i(\tau, \sigma) \times \partial_i x^\mu(\tau, \sigma)d\tau d\sigma, \tag{1}$$

where c is the velocity of light; it is obvious that the charge conservation law $\partial_\mu j^\mu = 0$ immediately implies $\partial_i e^i = 0$. Hereupon, the Latin indices run through the values 0 and 1 and correspond to τ and σ, respectively.

Let us introduce a nondegenerate symmetric scalar product in a linear space of n-forms on $\mathbb{R}^{3,1}$ as follows:

$$\langle X, Y\rangle = n!\int\limits_{\mathbb{R}^{3,1}} X \wedge *Y = \int\limits_{\mathbb{R}^{3,1}} X_{\mu_1\ldots\mu_n}Y^{\mu_1\ldots\mu_n}d^4x; \tag{2}$$

here, $*$ is the Hodge operator that sends n-forms to $(4-n)$-forms, and $\wedge$ denotes the exterior product of forms. In these terms, the action of the model in question is expressed as

$$S[A, x] = -\frac{1}{8\pi c}\langle A, \delta dA\rangle - \frac{1}{c^2}\langle j, A\rangle + S_0[x], \tag{3}$$

where d is the exterior differential, $\delta = *d*$, A_μ is the 4-potential of the electromagnetic field, and $S_0[x]$ is the action that describes the free dynamics of the string. The equations of motion for action (3) are given by

$$\delta F = -\frac{4\pi}{c}j, \quad \frac{\delta S_0[x]}{\delta x^\mu} = \frac{1}{c^2}\left\langle \frac{\delta j[x]}{\delta x^\mu}, A\right\rangle = \frac{1}{c}F_{\mu\nu}e^i\partial_i x^\nu, \tag{4}$$

where $F = dA$ is the strength tensor of the electromagnetic field.

To obtain effective equations of motion of a string, we should solve the Maxwell equations for an arbitrary configuration of the string and substitute the solutions of these equations into the expression for the Lorentz force. This yields an ill-defined (divergent) expression for the self-action force of the string:

$$F^{\mathrm{rr}}_\mu[x] = -\frac{4\pi}{c^3}\left\langle \frac{\delta j[x]}{\delta x^\mu}, Gj[x]\right\rangle, \tag{5}$$

where G is an operator whose kernel is a retarded Green function. Applying a regularization procedure [13] to this expression, we obtain an asymptotic series in the regularization parameter that contains one logarithmically divergent term. The regularization parameter makes the sense of the cross-section radius of the string; when this radius tends to zero, the radiation-reaction force diverges. If the cross-section radius is small but finite, then this divergent term makes the most essential contribution to the self-action force; moreover, the smaller the cross-section radius, the larger this contribution.

Using the formalism developed in [13], we can easily show that the logarithmically divergent term that arises in the expression for the self-action force can be obtained by varying the action with the Lagrangian[2]

$$L^{\mathrm{sing}} = -\frac{1}{c}\frac{e^i\partial_i x_\mu e^j\partial_j x^\mu}{\sqrt{|h|}} \times 2\ln\frac{\Lambda}{\varepsilon}\bigg|_{\varepsilon\to 0} = -\frac{1}{c}\frac{e^2}{\sqrt{|h|}}2\ln\frac{\Lambda}{\varepsilon}\bigg|_{\varepsilon\to 0}, \tag{6}$$

where $e^2 = e^i e^j h_{ij}$, $h_{ij} = \partial_i x_\mu \partial_j x^\mu$ is the induced metric on the string, $h = \det h_{ij}$, the parameter Λ characterizes the cut-off of the integral at the upper limit (its magnitude is on the order of the string length), and ε is the cutoff parameter of the integral at the lower limit (its magnitude is on the order of the cross-section radius of the string).

[2] This result can even be obtained without invoking the general covariant procedure, proposed in [13], for regularizing the radiation reaction in theories with singular sources. The leading divergent term is uniquely determined by the Poincaré invariance and the reparameterization invariance, as well as by the expression multiplying the δ function in formula (1). These arguments are frequently used for deriving leading divergent terms [14, 9, 15].

Let us introduce a vector field $V^i = e^i/\sqrt{|h|}$ and a 1-form $v_i = h_{ij}V^j$. Then, neglecting the finite part of the radiation-reaction force, we obtain the following effective equations of motion of the string:

$$\frac{\delta S_0[x]}{\delta x^\mu} = \frac{\chi}{c}(T^{ij}\nabla_{ij}x_\mu + \nabla_i T^{ij}\partial_j x_\mu)\sqrt{|h|} + \frac{1}{c}F_{\mu\nu}e^i\partial_i x^\nu, \qquad T^{ij} = V^2 h^{ij} - 2V^i V^j, \tag{7}$$

where $\chi = 2\ln(\Lambda/\varepsilon)$ is a dimensionless constant, $F_{\mu\nu}$ is the strength tensor of the external electromagnetic field, and ∇_i is a connection compatible with the metric h_{ij}. The traceless tensor T^{ij} represents the density of the energy–momentum tensor corresponding to Lagrangian (6); i.e.,

$$T^{\text{sing}}_{\mu\nu} = \frac{\chi}{c}\int \delta^4(x - x(\tau,\sigma)) \times T^{ij}\partial_i x_\mu \partial_j x_\nu \sqrt{|h|}\, d\tau d\sigma. \tag{8}$$

The tracelessness of the tensors T^{ij} and $T^{\text{sing}}_{\mu\nu}$ follows from the conformal invariance of Lagrangian (6).

If the free action $S_0[x]$ of the string is reparameterization-invariant, then the equations of motion (7) possess "residual" reparameterization invariance, which implies that the equations are orthogonal to the vector $V^i\partial_i x^\mu$. In addition, we have

$$\nabla_k T^k_i = -2i_V d v_i = -\chi^{-1}\partial_i x^\mu F_{\mu\nu}V^k\partial_k x^\nu. \tag{9}$$

In particular, in the absence of an external field, the last equality and the charge conservation law $\nabla_i V^i = 0$ imply

$$dv = 0, \quad \delta v = 0; \tag{10}$$

i.e., v is a harmonic 1-form. If the closed string has no self-intersections, Eqs. (10) are easily solved. Applying the conformal gauge

$$\dot{x}_\mu x'^\mu = 0, \quad \dot{x}_\mu \dot{x}^\mu = -x'_\mu x'^\mu, \tag{11}$$

where the dot denotes the differentiation with respect to τ and the prime denotes the differentiation with respect to σ, we obtain the following expressions for the general solution to Eqs. (10):

$$\lambda = \lambda_0 + f(\sigma+\tau) + g(\sigma-\tau), \qquad I = I_0 - f(\sigma+\tau) + g(\sigma-\tau). \tag{12}$$

Here, we used more customary notations $\lambda = e^0$ and $I = e^1$; λ_0 and I_0 are arbitrary constants, and f and g are arbitrary 2π-periodic functions. In other words, in the absence of external field, the energy–momentum conservation law implies that the relation between the linear density of charge and the current is given by (12).

Within our approximation, Eqs. (10) represent a mathematical expression for the condition that the string is superconducting: one of these equations states the charge conservation law, and the other states that, in the absence of external fields, the time-variation of the current at a given point of the string is attributed only to the gradient of the linear density of charge. The fulfillment of these equations follows from the requirement that the free action $S_0[x]$ should be reparameterization-invariant. The converse is also true: the superconductivity conditions (10) for an arbitrary configuration of the string imply the reparameterization invariance of its free action.

Using the energy–momentum conservation law (9), we can rewrite the equations of motion of the string in an external field as

$$\frac{\delta S_0[x]}{\delta x^\mu} = \frac{\chi}{c}T^{ij}\nabla_{ij}x_\mu\sqrt{|h|} + \frac{1}{c}\gamma^\rho_\mu F_{\rho\nu}e^i\partial_i x^\nu, \tag{13}$$

where

$$\gamma_{\mu\nu} = \eta_{\mu\nu} - h^{ij}\partial_i x_\mu \partial_j x_\nu.$$

Thus, if the free action of an electrically charged string with a current is reparameterization-invariant, then its effective dynamics in an external electromagnetic field are described by the system of equations (9), (13).

When the contribution of the singular term is sufficiently large, i.e., when the string is sufficiently thin ($\chi \gg 1$) and either the current flowing through it or the linear density of charge are large, one can neglect the left-hand sides of Eqs. (7); in this case, the free effective dynamics of the string are completely determined by the leading contribution of the self-action force of the charged string. We say that such a string is absolutely elastic because its internal structure does not appreciably resist an action.

When the current density increases further, one can also neglect the effect of the external field; then, the effective dynamics of the string are described by the equation

$$T^{ij}\nabla_{ij}x_\mu = 0, \tag{14}$$

provided that v is a harmonic 1-form.

Further, we will solve the system of equations (9), (13) for the model of a ring-shaped absolutely elastic string in an external electromagnetic field and consider the model of an absolutely nonstretchable charged string with a current; for the latter model, we will derive the effective equations of motion and obtain certain particular solutions.

3. A CHARGED RING

As we pointed out in the Introduction, the model of an absolutely elastic charged string describes a high-current beam of charged particles; therefore, it is worthwhile to consider its effective dynamics in an external electromagnetic field. In this section, we consider the effective dynamics of an absolutely elastic charged string in the form of a ring (a circle). Further, we derive equations for external electromagnetic fields that admit stationary states of such a ring. Then, we find solutions to free equations of motion and solve the equations of motion of a uniformly charged ring in an external uniform magnetic field. The last model describes the behavior of a high-current beam of charged particles in a synchrotron.

Consider a gauge that is convenient for further calculations. Introduce coordinates $\{\tau, \sigma\}$ so that the vector density e^i is straightened in these coordinates; i.e., it has the form $e = \{1, 0\}$. Let us show that such coordinates can be introduced without changing the coordinate τ.

Suppose that, in the original coordinates $\{t, l\}$, the vector density e has components $\{\lambda, I\}$; then, in the coordinates $\{\tau, \sigma\}$, we obtain

$$e^0 = \frac{\dot\tau\lambda + \tau' I}{\dot\tau\sigma' - \tau'\dot\sigma}, \quad e^1 = \frac{\dot\sigma\lambda + \sigma' I}{\dot\tau\sigma' - \tau'\dot\sigma}; \tag{15}$$

here, the dots and primes denote the differentiation with respect to t and l, respectively. Setting $e^0 = 1$, $e^1 = 0$, $\dot\tau = 1$, $\tau' = 0$, and $\sigma' \neq 0$, we obtain the following relations for σ:

$$\dot\sigma = -I, \quad \sigma' = \lambda, \tag{16}$$

provided that $\lambda \neq 0$ at this point. Equations (16) are integrable by virtue of the charge conservation law. For example, if $dl = \sqrt{\partial_\sigma \mathbf{x}\partial_\sigma \mathbf{x}}\, d\sigma$ is a length element of the string, then the linear density of charge is represented as

$$|\lambda| = (\partial_\sigma \mathbf{x}\partial_\sigma \mathbf{x})^{-1/2}. \tag{17}$$

In other words, the coordinate σ counts the charge on the string. Next, we will assume that $\lambda \neq 0$ throughout the string.

In addition to the above gauge, we require that

$$x^0(\tau, \sigma) = c\tau \Rightarrow \tau = t. \tag{18}$$

Then, the metric induced on the world sheet of the ring

$$\begin{gathered} x^0 = ct, \\ \mathbf{x}(t, \sigma) = r(t)(\cos\varphi(t, \sigma), \sin\varphi(t, \sigma), 0) \end{gathered} \tag{19}$$

and its inverse are given by

$$h_{ij} = \begin{bmatrix} c^2 - \dot r^2 - r^2\dot\varphi^2 & -r^2\dot\varphi\varphi' \\ -r^2\dot\varphi\varphi' & -r^2\varphi'^2 \end{bmatrix},$$

$$h^{ij} = \begin{bmatrix} \dfrac{1}{c^2 - \dot r^2} & -\dfrac{\dot\varphi}{\varphi'(c^2 - \dot r^2)} \\ -\dfrac{\dot\varphi}{\varphi'(c^2 - \dot r^2)} & -\dfrac{c^2 - \dot r^2 - r^2\dot\varphi^2}{r^2\varphi'^2(c^2 - \dot r^2)} \end{bmatrix}. \tag{20}$$

Hereupon, the prime denotes the differentiation with respect to σ. The determinant of the induced metric is equal to $h = -r^2\varphi'^2(c^2 - \dot r^2)$. The functions $\dot\varphi(t, \sigma)$ and $\varphi'(t, \sigma)$ are smooth and Q-periodic in the variable σ, where $Q = \text{const}$ is the total charge of the ring. The linear density of charge is equal to $\lambda = 1/r\varphi'$; here, we matched the signs of λ and φ'. The fundamental harmonic of $\varphi'(t, \sigma)$ with respect to σ is equal to $2\pi/Q$; in particular, $\varphi' = 2\pi/Q$ if the ring is uniformly charged.

The absolute elasticity of a string implies that the free action $S_0[x]$ of the string is identically zero. In this case, Eqs. (13) are rewritten as

$$T^{ij}\nabla_{ij}x_\mu\sqrt{|h|} = -\gamma^\rho_\mu F_{\rho\nu}e^i\partial_i x^\nu, \tag{21}$$

where the external field is redefined as $F_{\mu\nu} \longrightarrow \chi F_{\mu\nu}$. Throughout this section, the expressions for the electromagnetic fields will contain χ^{-1}. We will also assume that the external field is cylindrically symmetric and that $E_z = H_r = H_\varphi = 0$, where, as usual, the subscripts indicate the projections of a vector onto an appropriate unit vector. Then, Eqs. (9) and (21) are equivalent to the following two equations:

$$\begin{gathered} \ddot\varphi - 2\frac{\dot\varphi'\dot\varphi}{\varphi'} - \varphi''\frac{c^2 - \dot r^2 - r^2\dot\varphi^2}{r^2\varphi'^2} \\ + \frac{\dot r\dot\varphi}{r}\frac{c^2 - \dot r^2 + r\ddot r}{c^2 - \dot r^2} = \frac{\varphi'}{2}(c^2 - \dot r^2)^{1/2}(cE_\varphi - H_z\dot r), \\ \frac{(c^2 - \dot r^2 + r^2\dot\varphi^2)(c^2 - \dot r^2 - r\ddot r)}{r^2\varphi'(c^2 - \dot r^2)^{3/2}} \\ = \frac{r\dot\varphi}{c}(E_\varphi\dot r - cH_z) - \frac{E_r}{c}(c^2 - \dot r^2). \end{gathered} \tag{22}$$

In particular, the first equation implies the equation that defines the variation law for the effective angular momentum of an absolutely elastic charged ring:

$$\frac{d}{dt}\left[\int_0^Q d\sigma\frac{r\dot\varphi}{\sqrt{c^2 - \dot r^2}}\right] = \pi r(cE_\varphi - H_z\dot r). \tag{23}$$

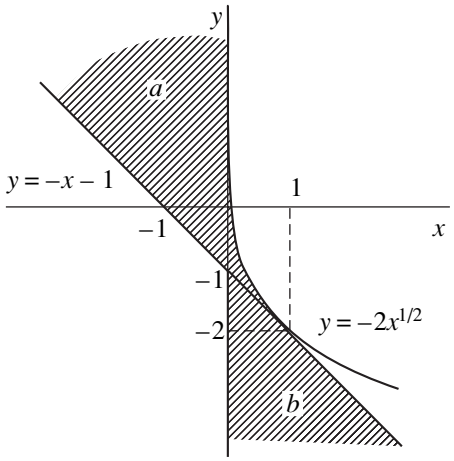

Fig. 1. The region where stationary states of a positively ($\varphi' > 0$) charged ring may exist is crosshatched. $y = r^2\varphi' H_z$, $x = 1 + r^2\varphi' E_r$, and $\dot{\varphi} \geq 0$. The region *a*, which is bounded from below by the straight line $y = -x - 1$, corresponds to the situation when Eq. (26) is taken with sign "+," and the region *b*, which is bounded by the axis *y* and the curve $y = -2x^{1/2}$, corresponds to the branch with sign "–." In the overlap of regions *a* and *b*, there may exist stationary states of the ring that have different angular velocities for the same values of the external field, charge, and ring radius.

Let us consider the stationary states of a charged ring in the external field; i.e., let us set $r(t) = \text{const}$ in Eqs. (22). Then, we obtain

$$\begin{aligned} \ddot{\varphi} - 2\frac{\dot{\varphi}'\dot{\varphi}}{\varphi'} - \varphi''\frac{c^2 - r^2\dot{\varphi}^2}{r^2\varphi'^2} &= \frac{c^2\varphi'}{2}E_\varphi, \\ \frac{c^2 + r^2\dot{\varphi}^2}{r^2 c\varphi'} &= -r\dot{\varphi}H_z - cE_r, \end{aligned} \tag{24}$$

where the external fields are, generally speaking, certain functions of *t*. Formula (23) can be rewritten as

$$\frac{d}{dt}\int_0^Q d\sigma\dot{\varphi} = \pi c^2 E_\varphi. \tag{25}$$

The second equation in (24) implies that

$$\frac{r\dot{\varphi}}{c} = -\frac{r^2\varphi' H_z}{2} \pm \sqrt{\frac{r^4\varphi'^2 H_z^2}{4} - (1 + r^2\varphi' E_r)}. \tag{26}$$

The requirements that the radicand be nonnegative and that the velocity of the string be less than the velocity of light impose constraints on the fields H_z and E_r under which stable states of the ring may exist. These requirements are illustrated graphically in Fig. 1. For example, if there is no electric field and a ring of radius *r* is uniformly charged, then the magnetic field can hold this ring only if

$$|H_z(r)| \geq \frac{|Q|}{\pi r^2}, \tag{27}$$

where *Q* is the total charge of the ring.

Equations (24) are rather complicated in the general case; therefore, we restrict the analysis to a uniformly charged ring ($\varphi'' = 0$) for $E_r = 0$. Then, we have

$$\begin{aligned} \ddot{\varphi} &= \frac{c^2\varphi'}{2}E_\varphi, \quad \varphi' = \frac{2\pi}{Q}, \\ \frac{r\dot{\varphi}}{c} &= -\frac{\pi r^2 H_z}{Q} - \sqrt{\frac{\pi^2 r^4 H_z^2}{Q^2} - 1}, \end{aligned} \tag{28}$$

where we used the fact that the equality $\varphi'' = 0$ implies the equality $\dot{\varphi}' = 0$. The substitution of the expression for $\dot{\varphi}$ into the equation for $\ddot{\varphi}$ in (28) yields equations for the fields that admit such stationary configurations.

For example, in the nonrelativistic limit $r^2\dot{\varphi}^2 \ll c^2$, we obtain the following solution to (28):

$$\begin{aligned} \frac{r\dot{\varphi}(t)}{c} &= -\frac{Q}{2\pi r^2 H_z(t)}, \quad \varphi'(\sigma) = \frac{2\pi}{Q}, \\ 2\pi^2 r^3 c^2 E_\varphi(t) &= Q^2\frac{\dot{H}_z(t)}{H_z^2(t)}, \end{aligned} \tag{29}$$

whereas, in the ultrarelativistic limit $r^2\dot{\varphi}^2 \approx c^2$, Eqs. (28) lead to the equalities

$$\begin{aligned} \frac{r\dot{\varphi}(t)}{c} &= -\frac{Q}{\pi r^2 H_z(t)}, \quad \varphi'(\sigma) = \frac{2\pi}{Q}, \\ \dot{H}_z(t) &= -c\frac{E_\varphi(t)}{r}\left(1 + \frac{\pi r^2}{Q}H_z(t)\right). \end{aligned} \tag{30}$$

Thus, a uniformly charged ring does not change its radius only if the external fields satisfy Eqs. (28) (or Eqs. (29) and (30) in the nonrelativistic and ultrarelativistic cases, respectively), provided, of course, that $E_r = 0$.

Now, we proceed to solving the dynamical equations (22). Consider the case when there are no external fields. The solution of the second equation in (22) yields

$$\begin{aligned} r(t) &= \sqrt{\rho^2 + c^2(t + \tau)^2}, \\ \rho &= r(0)\sqrt{1 - \frac{\dot{r}^2(0)}{c^2}}, \\ \tau &= \frac{r(0)\dot{r}(0)}{c^2}. \end{aligned} \tag{31}$$

Then, the first equation takes the form

$$\ddot{\varphi} - 2\frac{\dot{\varphi}'\dot{\varphi}}{\varphi'} - \frac{\varphi''}{\varphi'^2}\left(\frac{c^2\rho^2}{r^4} - \dot{\varphi}^2\right) + 2\frac{\dot{r}\dot{\varphi}}{r} = 0, \qquad (32)$$

whence

$$\frac{d}{dt}\int_0^Q d\sigma r^2\dot{\varphi} = 0. \qquad (33)$$

If, in addition, we require that $\dot{\varphi}' = 0$, which physically means that all points of the ring rotate with the same velocity, then we obtain the solution

$$r^2(t)\dot{\varphi}(t) = r^2(0)\dot{\varphi}(0), \quad \varphi'(\sigma) = \frac{2\pi}{Q}. \qquad (34)$$

As expected, the equation for φ represents the conservation law of angular momentum.

Solution (31) shows that, after a certain period of time, the ring will expand with a velocity close to the velocity of light; therefore, it is worthwhile to consider the ultrarelativistic limit of Eq. (32); i.e., it is worthwhile to require that $c^2 - \dot{r}^2 - r^2\dot{\varphi}^2 \approx 0$. In this case, we have the following conservation law:

$$\frac{r^2(t)\dot{\varphi}(t,\sigma)}{\varphi'^2(t,\sigma)} = \frac{r^2(0)\dot{\varphi}(0,\sigma)}{\varphi'^2(0,\sigma)}. \qquad (35)$$

This equation can be solved by the method of characteristics (see, for example, [16]). In the particular case when

$$\frac{\dot{\varphi}(0,\sigma)}{\varphi'^2(0,\sigma)} = \lambda^2(0,\sigma)r^2(0)\dot{\varphi}(0,\sigma) = b = \text{const},$$

i.e., when the linear density of the effective angular momentum is the same at all points of the string, we obtain

$$\dot{\varphi}(t,\sigma) = \frac{br^2(0)}{r^2(t)}\varphi'^2(0,\sigma_0(t,\sigma)),$$

$$\sigma = \sigma_0(t,\sigma) - \frac{2br^2(0)}{\rho c} \qquad (36)$$

$$\times\left(\arctan\frac{c(t+\tau)}{\rho} - \arctan\frac{c\tau}{\rho}\right)\varphi'(0,\sigma_0(t,\sigma)).$$

The last equation should be considered an equation for $\sigma_0(t,\sigma)$ for a certain prescribed Q-periodic function $\varphi'(0,\sigma_0)$ whose fundamental harmonic is equal to $2\pi/Q$.

To conclude this section, consider the effective dynamics of a charged ring all of whose points move with the same angular velocity ($\dot{\varphi}' = 0$) in a uniform magnetic field $H_z = \text{const}$. In this case, from (22) and (23) we obtain the system of equations

$$\frac{(c^2 - \dot{r}^2 + r^2\dot{\varphi}^2)(c^2 - \dot{r}^2 - r\ddot{r})}{r^2\varphi'(c^2 - \dot{r}^2)^{3/2}} = -r\dot{\varphi}H_z,$$
$$\frac{r\dot{\varphi}}{\sqrt{c^2 - \dot{r}^2}} + \frac{\pi H_z}{2Q}r^2 = M, \qquad (37)$$

where M is a certain constant defined by the initial data. The first equation implies, in particular, that $\varphi' = 2\pi/Q$. Substituting the expression for $r\dot{\varphi}$ from the second equation into the first, we obtain an equation for the function $r(t)$ alone, which has the form

$$\sqrt{c^2 - v^2} = \frac{k}{r}\left[1 + \left(M - \frac{\pi H_z}{2Q}r^2\right)^2\right], \qquad (38)$$

where $v(r(t)) \equiv \dot{r}(t)$ and k is an integration constant. Let us express the equations of motion in dimensionless coordinates. Introduce $r_0 = |\pi H_z/2Q|^{-1/2}$ and redefine r and t as $r \longrightarrow r_0 r$ and $t \longrightarrow tr_0/c$. For example, the velocity in these coordinates is measured in units of the velocity of light. Then, the equations of motion of a charged ring are expressed as

$$\frac{v^2}{2} + \frac{1}{2}\left[k^2\frac{(1+L^2)^2}{r^2} - 1\right] = 0,$$
$$\dot{\varphi} = k\frac{L(1+L^2)}{r^2}, \qquad (39)$$

where $L = M - \text{sgn}(H_z/Q)r^2$.

The first equation in (39) resembles the equation of motion of a particle of unit mass with zero total energy in the potential field,

$$U(r) = [k^2r^{-2}(1+L^2)^2 - 1]/2,$$

the only difference between these equations being that the form of $U(r)$ depends on the initial data $r(0)$, $v(0)$, and $\dot{\varphi}(0)$ (see Fig. 2). The potential has a single extremum at the point

$$r_{\text{ext}}^2 = \frac{1}{3}\left(\text{sgn}\left(\frac{H_z}{Q}\right)M + \sqrt{4M^2 + 3}\right) \qquad (40)$$

and indefinitely increases as $r \longrightarrow 0$ and $r \longrightarrow \infty$; therefore, for any initial data, the system will oscillate about the equilibrium point r_{ext}. Note that the minimal value of r_{ext} is equal to $\sqrt{2}/2$, which agrees with the results of the

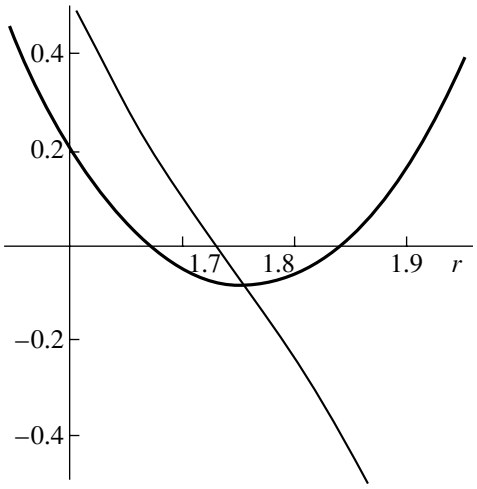

Fig. 2. The functions of the potential $U(r)$ (heavy line) and the angular velocity $r\dot{\varphi}(r)$ (thin line). The diagrams correspond to the initial data $r(0) \approx 1.844$, $v(0) = 0.1$, $M = 3$, and $\mathrm{sgn}(H_z/Q) = 1$. The potential attains its minimum at the point $r_{\mathrm{ext}} \approx 1.755$.

previous analysis of the stationary states (27) of a charged ring.

We can evaluate the ratio of the oscillation frequency ω of the charged ring in the neighborhood of the equilibrium point r_{ext} to the mean angular frequency ω_0 of its rotation. In the harmonic approximation, the first frequency is defined by $U''(r_{\mathrm{ext}})$, and the second, by $|\dot{\varphi}(r_{\mathrm{ext}})|$; hence, we have

$$\frac{\omega}{\omega_0} \approx \frac{1}{r}\frac{(4r^2 - \sqrt{4r^4 - 1})^{1/2}}{(2r^2 - \sqrt{4r^4 - 1})^{3/2}}, \tag{41}$$

where $r = r_{\mathrm{ext}}$. This ratio uniformly increases from 2 to infinity as r_{ext} increases; for large values of r_{ext}, it increases as $8\sqrt{2}r_{\mathrm{ext}}^3$. For example, a threefold increase in the linear dimensions of a synchrotron leads to a ninefold increase in the oscillation frequency for the same mean angular velocities of the high-current beam of particles, other characteristics, such as H_z, Q, and χ, remaining constant.

Thus, an absolutely elastic uniformly charged ring in a uniform magnetic field oscillates about the equilibrium point r_{ext}, $\dot{\varphi}_{\mathrm{ext}}$, according to Eqs. (39), with frequency ω; the ratio of this frequency to the angular frequency of the ring is defined by (41). Therefore, we can expect that, when the energy inflow compensates the energy losses, a high-current beam of charged particles in a synchrotron will also produce radiation at this frequency, in addition to the well-known synchrotron radiation. For example, if one could separate these two types of radiation by certain characteristics and measure the ratio ω/ω_0, then one would determine the equilibrium position r_{ext} in the units of r_0 by formula (41).

Among the disadvantages of the model of an absolutely elastic charged string as applied to the description of a high-current beam of charged particles is the fact that this model does not take into account the radiation reaction due to the synchrotron radiation, which becomes significant at large angular velocities.

4. AN ABSOLUTELY NONSTRETCHABLE STRING

In this section, we consider the dynamics of a thin, absolutely nonstretchable charged string[3] with a current with regard to the first-order correction due to the self-action. We derive equations of motion for such a string, investigate its stationary states in the absence of an external electromagnetic field, and find a class of solutions to the equations of motion for a uncharged string with a current and for a uniformly charged string.

Suppose given a closed string N with coordinates $\{\tau, \sigma\}$, $\sigma \in [0, L)$, that is embedded into the Minkowski space $\mathbb{R}^{3,1}$ by a smooth mapping $x(\tau)$. In relativistic mechanics, the concept of nonstretchability makes sense only in a certain distinguished frame of reference. Let us introduce a 4-vector n^μ, $n^2 = 1$, that characterizes such a frame of reference; then, the action that describes the free dynamics of an absolutely nonstretchable string is given by

$$S_0[x, \varkappa, \kappa^0, \kappa^1] = \int_N d\tau d\sigma(\varkappa(h_{00} + h) + \kappa^0(n^\mu \dot{x}_\mu - c) + \kappa^1 n^\mu x'_\mu), \tag{42}$$

where $\varkappa$, κ^0, and κ^1 are Lagrange multipliers of the constraints that guarantee the nonstretchability of the string. It is obvious that action (42) is not invariant under the change of coordinates $\{\tau, \sigma\}$; hence, there is no constraint (9) on the form of the linear density of charge and current in this case.

The effective equations of motion (7) for a thin absolutely nonstretchable charged string with a current in the distinguished frame of reference $n^\mu = (1, 0, 0, 0)$ are expressed as

$$-\partial_i(L^{ij}\partial_j x_\mu + \kappa^i \delta^0_\mu) = \frac{1}{c}F_{\mu\nu}e^i\partial_i x^\nu,$$

$$L^{ij} \equiv \frac{\chi}{c}\frac{h_{kl}e^k e^l h^{ij} - 2e^i e^j}{\sqrt{|h|}} + 2\varkappa(\delta^i_0\delta^j_0 + hh^{ij}), \tag{43}$$

$$h_{00} = -h, \quad \dot{x}^0 = c, \quad x'^0 = 0.$$

[3] Possibly, a more customary term for the model of an absolutely nonstretchable string is a perfect weightless thread, which comes from mechanics. A detailed description of the theory of an absolutely flexible thread can be found, for example, in [17].

The last three equations in (43) represent the condition of "relativistic nonstretchability." Indeed, the equation $h_{00} = -h$ has the form

$$\mathbf{x}'^2 = 1 - \frac{(\dot{\mathbf{x}} \cdot \mathbf{x}')^2}{c^2 - \dot{\mathbf{x}}^2}, \tag{44}$$

which immediately implies that the coordinate σ is a natural parameter on the string with a correction for relativistic contraction. Since σ ranges from 0 to L, the fulfillment of equality (44) at any point of the string automatically implies that the string of length L is nonstretchable.

Since we consider a closed string, all functions entering Eq. (43) must be periodic in σ. For an open string with free ends, the periodicity condition is replaced by the equality

$$L^{1i}\partial_i x_\mu + \kappa^1 \delta^0_\mu = 0 \tag{45}$$

at the ends of the string.

We will consider the effective dynamics of a closed string in the absence of external electromagnetic fields. The unknown fields κ^i can be obtained from the first two equations in (43) by setting $\mu = 0$. As a result, we are left with four equations for four unknown functions $\mathbf{x}(t, \sigma)$ and $\varkappa(t, \sigma)$:

$$L^{ij}\partial_{ij}\mathbf{x} + \partial_i L^{ij}\partial_j \mathbf{x} = 0, \quad \mathbf{x}'^2 = 1. \tag{46}$$

The consistency condition for this system yields an equation for $\varkappa(t, \sigma)$. The physical meaning of the field $\varkappa(t, \sigma)$ is that it compensates for the forces that stretch (contract) the string.

Let us find the stationary configurations of the string that are consistent with Eqs. (46); i.e., set $\dot{\mathbf{x}} \equiv 0$ in the equations of motion. We can easily show that, in this case, the equations of motion are reduced to the system[4]

$$\begin{aligned} 2\varkappa' + \frac{(\lambda^2)'}{c} + 2\frac{\lambda \dot{I}}{c^3} &= 0, \\ 2\varkappa + \frac{\lambda^2}{c} + \frac{I^2}{c^3} &= 0, \quad \mathbf{x}'^2 = 1. \end{aligned} \tag{47}$$

Taking into account the charge conservation law, $\dot{\lambda} + I' = 0$, we have

$$\varkappa = -\frac{c^2\lambda^2 + I^2}{2c^3}, \quad \dot{\lambda} I + \lambda \dot{I} = 0, \quad \mathbf{x}'^2 = 1. \tag{48}$$

Thus, if the product of the charge density multiplied by the current is independent of time and all points of the string are at rest at the initial moment, then there exists $\varkappa(t, \sigma)$ such that the string retains its initial configuration. The question concerning the stability of such solutions remains open.

Next, consider the nonrelativistic dynamics of a string described by Eqs. (46) in the absence of the current, $I = 0$ (charged dielectric), as well as for $\lambda = 0$ (uncharged conductor). The nonrelativistic limit is understood in the following sense. Let l be a characteristic scale of variation of the field $\mathbf{x}(t, \sigma)$, for example, the length of the string; then, we formally define the order of smallness as follows:

$$\left[\frac{\dot{\mathbf{x}}}{c}\right] \ll 1, \quad [l^k \partial_\sigma^{k+1}\mathbf{x}] = \left[\frac{l\dot{\mathbf{x}}'}{c}\right] = \left[\frac{l\ddot{\mathbf{x}}}{c^2}\right] = 1. \tag{49}$$

In this case, the order of smallness of $\varkappa$ is determined from Eqs. (46).

For a charged dielectric, when $I = 0$ and $\dot{\lambda} = 0$, we obtain the following system of equations from (46) in the nonrelativistic limit:

$$\frac{\lambda^2}{c^2}\ddot{\mathbf{x}} + \Lambda \mathbf{x}'' + \left[\Lambda' + \frac{\ddot{\mathbf{x}} \cdot \mathbf{x}'}{c^2}\Lambda\right]\mathbf{x}' = 0, \quad \mathbf{x}'^2 = 1, \tag{50}$$

where $\Lambda = 2c\varkappa + \lambda^2$. Note that these equations are invariant under the Galilean transformations. Equations (50) can be resolved for the higher derivative only if $\Lambda \neq -\lambda^2$. In this case, we have

$$\frac{\lambda^2}{c^2}\ddot{\mathbf{x}} + \Lambda \mathbf{x}'' + \frac{\lambda^2 \Lambda'}{\lambda^2 + \Lambda}\mathbf{x}' = 0, \quad \mathbf{x}'^2 = 1. \tag{51}$$

The consistency condition for system (51) leads to the following equation for $\Lambda(t, \sigma)$:

$$\left(\frac{\Lambda'}{\lambda^2 + \Lambda}\right)' - \frac{\Lambda \mathbf{x}''^2}{\lambda^2} - \frac{\dot{\mathbf{x}}'^2}{c^2} = 0. \tag{52}$$

Thus, fixing $\Lambda(t, 0)$ and $\Lambda'(t, 0)$, as well as $\mathbf{x}(0, \sigma)$ and $\dot{\mathbf{x}}(0, \sigma)$ subject to the conditions

$$\begin{aligned} \mathbf{x}'(0, \sigma) \cdot \mathbf{x}'(0, \sigma) &= 1, \\ \dot{\mathbf{x}}'(0, \sigma) \cdot \mathbf{x}'(0, \sigma) &= 0, \end{aligned}$$

we can construct a unique solution $\mathbf{x}(t, \sigma)$ that satisfies Eqs. (51). Recall that all the functions must be L-periodic in the variable σ. This condition imposes constraints on the boundary conditions for the function $\Lambda(t, \sigma)$.

One can draw an instantaneous phase portrait for Eq. (52) (see Fig. 3). This portrait shows that, if $\Lambda(t, 0) \neq -\lambda^2$, then $\Lambda(t, \sigma) \neq -\lambda^2$ for all $\sigma \in [0, L)$; i.e., the quantity $1 + \Lambda/\lambda^2$ does not change its sign.

[4] Henceforth, we redefine the Lagrangian multiplier $\varkappa$ as follows: $\varkappa \longrightarrow -\varkappa\chi/c$.

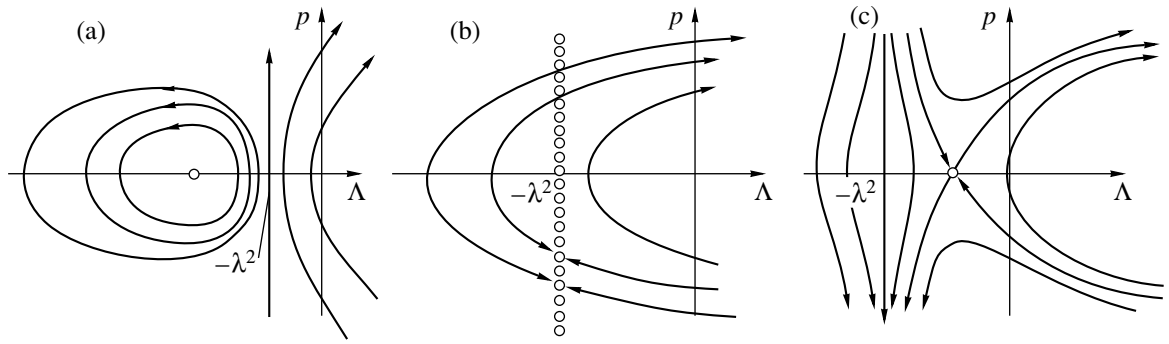

Fig. 3. The instantaneous phase portrait for Eq. (52); here, $p = \Lambda'/(\lambda^2 + \Lambda)$. The circles indicate the singular points of the vector field. (a) $\dot{\mathbf{x}}'^2/c^2 > \mathbf{x}''^2$ (this case can be referred to as the case of small curvature of the string), (b) $\dot{\mathbf{x}}'^2/c^2 = \mathbf{x}''^2$, and (c) $\dot{\mathbf{x}}'^2/c^2 < \mathbf{x}''^2$.

For a uniformly charged string, $\lambda = \text{const}$, Eqs. (51) and (52) are rewritten as

$$\frac{\ddot{\mathbf{x}}}{c^2} - \left(1 - \operatorname{sgn}\left(1 + \frac{\Lambda}{\lambda^2}\right)e^{\Theta}\right)\mathbf{x}'' + \Theta'\mathbf{x}' = 0,$$
$$\Theta'' + \left(1 - \operatorname{sgn}\left(1 + \frac{\Lambda}{\lambda^2}\right)e^{\Theta}\right)\mathbf{x}''^2 - \frac{\dot{\mathbf{x}}'^2}{c^2} = 0, \quad (53)$$

where $\Theta = \ln|1 + \Lambda/\lambda^2|$.

Using similar arguments for an uncharged conductor with current, $\lambda = 0$ and $I' = 0$, we obtain the following equations for $\varkappa \neq 0$ (this condition is an analogue of the inequality $\Lambda \neq -\lambda^2$, which arises in the case of a charged dielectric):

$$\frac{\ddot{\mathbf{x}}}{c^2} + (1 + \operatorname{sgn}(\varkappa)e^{\Theta})\mathbf{x}'' + \Theta'\mathbf{x}' = 0,$$
$$\Theta'' - (1 + \operatorname{sgn}(\varkappa)e^{\Theta})\mathbf{x}''^2 - \frac{\dot{\mathbf{x}}'^2}{c^2} = 0, \quad (54)$$

where $\Theta = \ln|2c^3\varkappa/I^2|$. The instantaneous phase portrait for the second equation in (54) always has the same form (with obvious redefinitions) as the instantaneous phase portrait in the case of a weakly curved charged dielectric (Fig. 3a). Therefore, periodic (in σ) solutions to the second equation in (54) may only exist when $\varkappa(t, 0) < 0$.

Let us find certain particular solutions to the equations obtained. Setting $\Theta'(t, \sigma) = 0$ in Eqs. (53) and (54), we obtain the system

$$\ddot{\mathbf{x}} - u^2(t)\mathbf{x}'' = 0, \quad \dot{\mathbf{x}}'^2 = u^2(t)\mathbf{x}''^2, \quad \mathbf{x}'^2 = 1. \quad (55)$$

In this case, the dynamics of a uniformly charged dielectric and an uncharged conductor are described by the same system of equations. Let us simplify the situation by assuming that $u(t) = \text{const}$. Then, there exists a class of solutions of the form

$$\mathbf{x}(t, \sigma) = \mathbf{V}t + \mathbf{x}_0(\sigma + ut), \quad (56)$$

where $\mathbf{V}$ is a certain constant vector and $\mathbf{x}_0(\sigma)$, $\mathbf{x}_0'^2 = 1$ defines the initial configuration of the string. Substituting the general solution of the wave equation into the remaining two equations, one can show that (56) provides the only possible solutions to Eqs. (55) for $u = \text{const}$. Solutions (56) describe a string that "flows" along itself with velocity u.

5. CONCLUSIONS

We have investigated the effective dynamics of a thin electrically charged string with a current with regard to the leading contribution of the self-action. This approximation has allowed us to describe the effective dynamics of a string in an external electromagnetic field in the form of second-order partial differential equations and obtain their exact solutions for certain simple models of a string in an electromagnetic field of a special form.

We have not analyzed the question concerning the stability of the solutions obtained. This problem may become one of possible directions of further research. Another direction of research may be the study of radiation characteristics of an absolutely elastic charged ring (a high-current beam of charged particles) in an external uniform magnetic field; the existence of such a radiation was discussed at the end of Section 3. Moreover, it would be interesting to find other solutions to the effective equations of motion of a string or to carry out a numerical analysis of these equations.

ACKNOWLEDGMENTS

I am grateful to I.V. Gorbunov and A.A. Sharapov for carefully reading the manuscript and for constructive criticism.

This work was supported by the Russian Foundation for Basic Research (project no. 03-02-17657), by the grant of the President of the Russian Federation for leading scientific schools (project no. NSh-1743.2003.2), by the Federal Agency of Education (project

no. A04-2.9-740), and by the Dynasty Foundation of non-profit programs, supported by the International Center for Fundamental Physics in Moscow.

REFERENCES

1. H. A. Lorentz, *Theory of Electrons* (Teubner, Leipzig, 1909; Gostekhteorizdat, Moscow, 1956).
2. P. A. M. Dirac, Proc. R. Soc. London, Ser. A **167**, 148 (1938).
3. B. S. DeWitt and R. W. Brehme, Ann. Phys. (N.Y.) **9**, 220 (1960).
4. A. O. Barut and N. Unal, Phys. Rev. A **40**, 5404 (1989).
5. P. E. G. Rowe and G. T. Rowe, Phys. Rep. **149**, 287 (1987).
6. B. P. Kosyakov, Teor. Mat. Fiz. **119**, 119 (1999).
7. P. O. Kazinski, S. L. Lyakhovich, and A. A. Sharapov, Phys. Rev. D **66**, 025 017 (2002).
8. P. O. Kazinski and A. A. Sharapov, Class. Quantum Grav. **20**, 2715 (2003).
9. B. Carter, Phys. Lett. B **404**, 246 (1997).
10. A. O. Barut and P. A. Pavšič, Phys. Lett. B **331**, 45 (1994).
11. J. Isberg, U. Lindström, B. Sundborg, and G. Theodoridis, Nucl. Phys. B **411**, 122 (1994).
12. M. B. Hindmarsh and T. W. B. Kibble, Rep. Prog. Phys. **58**, 477 (1995).
13. P. O. Kazinski and A. A. Sharapov, Teor. Mat. Fiz. **143**, 375 (2005).
14. E. Witten, Nucl. Phys. B **249**, 557 (1985).
15. A. Buonanno and T. Damour, Phys. Lett. B **432**, 51 (1998).
16. V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations* (MTsNMO, Moscow, 2002; Springer, Berlin, 1988).
17. D. R. Merkin, *Introduction to Mechanics of Flexible Thread* (Nauka, Moscow, 1980) [in Russian].

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 280–293.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 322–336.

NUCLEI, PARTICLES, FIELDS, GRAVITATION, AND ASTROPHYSICS

Two-Loop Self-Energy Correction in a Strong Coulomb Nuclear Field

V. A. Yerokhin[a,b], P. Indelicato[c], and V. M. Shabaev[a]

[a] *St. Petersburg State University, Ulyanovskaya ul. 1, Petrodvorets, St. Petersburg, 198504 Russia*
[b] *Center for Advanced Studies, St. Petersburg State Polytechnic University, St. Petersburg, 195251 Russia*
[c] *Laboratoire Kastler-Brossel (Unité Mixte de Recherche du CNRS no. 8552), École Normale Supérieure et Université P. et M. Curie, Case 74, 4 pl. Jussieu, F-75252, Paris Cedex 05, France*
e-mail: yerokhin@pcqnt1.phys.spbu.ru

Received March 5, 2005

Abstract—The two-loop self-energy correction to the ground-state energy levels of hydrogen-like ions with nuclear charges $Z \geq 10$ is calculated without the $Z\alpha$ expansion, where α is the fine-structure constant. The data obtained are compared with the results of analytical calculations within the $Z\alpha$ expansion; significant disagreement with the analytical results of order $\alpha^2(Z\alpha)^6$ has been found. Extrapolation is used to obtain the most accurate value for the two-loop self-energy correction for the 1*s* state in hydrogen. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

As the simplest atomic system, a hydrogen-like ion traditionally attracts the rapt attention of both theoreticians and experimenters. In contrast to the hydrogen atom, whose experimental investigation began in the 19th century, heavy hydrogen-like ions have become accessible to experimental study relatively recently. At present, high-accuracy spectroscopic measurements are being made for the heaviest ions up to hydrogen-like uranium. A vivid example is the measurement of the ground-state Lamb shift in the hydrogen-like uranium ion, whose accuracy has increased by a factor of 10 in the past decade. The error of the most accurate experimental result [1] is 13 eV (about 5% of the complete QED contribution); in the immediate future, the accuracy of an experiment is expected to increase to 1 eV. This planned experiment is a stimulus for theoretical studies of QED effects to the second order in fine-structure constant α without the $Z\alpha$ expansion (Z is the nuclear charge). A consistent allowance for these effects is required to theoretically describe heavy hydrogen-like ions with an adequate accuracy.

It can be inferred from the available $Z\alpha$-expansion terms that the two-loop self-energy (SE) correction shown in Fig. 1 makes the largest contribution among the second-order one-electron QED effects. Because of slow convergence of the $Z\alpha$ expansion, the value of this correction based on the available expansion terms and, hence, the complete second-order QED correction are uncertain even for moderately heavy ions. This justifies the necessity of calculating the two-loop SE correction without the $Z\alpha$ expansion.

Calculating this correction to all orders in $Z\alpha$ is also topical for light ions, including hydrogen. Despite considerable efforts directed at calculating the terms of the $Z\alpha$ expansion, the numerically large values of the coefficients found suggest that its convergence is very slow even for $Z = 1$. At present, the theoretical error of the ground-state energy in the hydrogen atom originates mainly from the highest $Z\alpha$-expansion terms for the two-loop SE correction [2]. It should be borne in mind that the difficulty of calculating the expansion coefficients increases rapidly with their order. The recently published results for all contributions of order $\alpha^2(Z\alpha)^6$ [2–4] were a breakthrough in this field. Performing such calculations in the next order will probably be unfeasible in the foreseeable future. Calculating the two-loop SE correction without the $Z\alpha$ expansion is a more realistic alternative to such calculations.

Let us briefly consider the results obtained to date for the $Z\alpha$-expansion coefficients of the two-loop SE correction. The correction to the energy is traditionally expressed in terms of the dimensionless function F defined as follows (in units $\hbar = c = 1$):

$$\Delta E = m\left(\frac{\alpha}{\pi}\right)^2 \frac{(Z\alpha)^4}{n^3} F(Z\alpha), \qquad (1)$$

where n denotes the principal quantum number of the state under consideration. The function F can be repre-

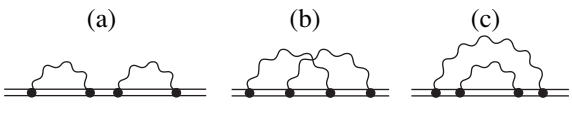

Fig. 1. The two-loop self-energy correction.

1063-7761/05/10102-0280$26.00

sented as a series in $Z\alpha$ and $L = \ln(Z\alpha)^{-2}$:

$$F(Z\alpha) = B_{40} + (Z\alpha)B_{50} + (Z\alpha)^2 \times [L^3 B_{63} + L^2 B_{62} + L B_{61} + G^{\text{h.o.}}(Z\alpha)], \quad (2)$$

where $G^{\text{h.o.}}$ denotes the contribution from all of the highest terms in $Z\alpha$,

$$G^{\text{h.o.}}(Z\alpha) = B_{60} + Z\alpha(\ldots).$$

The $Z\alpha$-expansion terms known to date are given below (see [5] for a review):

$$B_{40}(ns) = 1.409244\ldots, \quad (3)$$

$$B_{50}(ns) = -24.2668(31), \quad (4)$$

$$B_{63}(ns) = -8/27, \quad (5)$$

$$B_{62}(1s) = 16/27 - (16/9)\ln 2, \quad (6)$$

$$B_{61}(1s) = 49.838317, \quad (7)$$

$$B_{60}(1s) = -61.6(9.2). \quad (8)$$

Let us now discuss the calculation of the two-loop SE correction represented by the diagrams in Fig. 1 without using the $Z\alpha$ expansion. It is convenient to divide the contribution of diagram (a) into two parts: reducible and irreducible. The reducible part means the contribution in which the intermediate-state energy in the spectral decomposition of the mean electron propagator is equal to the initial-state energy, while the irreducible part means the residual. It can be shown that the irreducible part remains invariant under covariant gauge transformations. It is commonly analyzed separately and is called the loop-after-loop correction. The main reason for a separate analysis of this contribution is that the calculations for it are much simpler than those for the residual part of the two-loop SE correction, which below we call compact. The loop-after-loop correction was first calculated to all orders in $Z\alpha$ by Mitrushenkov *et al.* [6] for a number of heavy hydrogen-like ions. Subsequently, this correction was also calculated by other teams [7, 8].

The compact part of the two-loop SE correction is represented by the reducible part of the diagram in Fig. 1a as well as by the diagrams in Figs. 1b (overlapping diagram) and 1c (nested diagram). The first attempts to calculate them to all orders in $Z\alpha$ were made relatively recently by Mallampalli and Sapirstein [9] and Goidenko *et al.* [10]. Since the authors of [10] used a noncovariant renormalization procedure in their calculations, the results obtained cannot be considered reliable enough. Mallampalli and Sapirstein used a completely covariant approach, but the actual calculation was performed only for a certain part of the complete contribution. These authors divided the compact part of the two-loop SE correction into three contributions referred to as an *M* term, a *P* term, and an *F* term. (For a detailed discussion of each of these terms, see below.) Practical calculations were performed only for the *M* and *F* terms; the *M* term was calculated only for two ions with $Z = 92$ and 83 (due to the very long computational time). The *P* term was not considered in that paper, since its calculation required developing new computational methods.

The study by Mallampalli and Sapirstein was continued in our paper [11], in which we calculated the *P* term and obtained the complete two-loop SE correction for $Z = 92$ and 83. In our subsequent papers [12, 13], we performed the first full calculation of the two-loop SE correction for the ground states of hydrogen-like ions with $Z \geq 40$. Since the calculation was performed only for relatively high *Z* and since the calculations in these papers were not accurate enough, we did not made a detailed comparison with the analytical calculations within the $Z\alpha$ expansion, and whether the results obtained by different methods agree remained an open question.

The calculations of the two-loop SE correction in the range of low *Z* are complicated greatly by the numerical cancellations, which increase rapidly with decreasing *Z*. In this paper, we present a scheme for calculating this correction that is suitable for calculations in the range of low *Z* and perform an actual calculation for $Z \geq 10$. The data obtained are compared with the results of analytical calculations within the $Z\alpha$ expansion. We conclude that our results agree well with the $Z\alpha$-expansion terms in the orders $\alpha^2(Z\alpha)^4$ and $\alpha^2(Z\alpha)^5$, but disagree with the existing results of order $\alpha^2(Z\alpha)^6$. By extrapolating our data, we obtain the most accurate result for the two-loop SE correction to the ground-state Lamb shift in the hydrogen atom.

2. FORMALISM

Let us now consider in detail the two-loop SE correction. General formulas for the various contributions to this correction can be easily obtained using the method of two-time Green functions [14]. These formulas were first derived in [15].

The expression for the loop-after-loop correction (the irreducible part of the diagram in Fig. 1a) can be written as

$$\Delta E_{LAL} = \sum_n^{\varepsilon_n \neq \varepsilon_a} \frac{\langle a|\gamma^0 \tilde{\Sigma}(\varepsilon_a)|n\rangle\langle n|\gamma^0 \tilde{\Sigma}(\varepsilon_a)|a\rangle}{\varepsilon_a - \varepsilon_n}, \quad (9)$$

where $\tilde{\Sigma}(\varepsilon) = \Sigma(\varepsilon) - \delta m^{(1)}$, $\Sigma(\varepsilon)$ denotes the one-loop SE operator, and $\delta m^{(1)}$ is the one-loop mass renormal-

ization counterterm. The one-loop SE operator is defined as

$$\Sigma(\varepsilon, \mathbf{x}_1, \mathbf{x}_2) = 2i\alpha\gamma^0 \times \int_{-\infty}^{\infty} d\omega D^{\mu\nu}(\omega, \mathbf{x}_{12})\alpha_\mu G(\varepsilon - \omega, \mathbf{x}_1, \mathbf{x}_2)\alpha_\nu, \quad (10)$$

where G is the Green function of the Dirac equation with the Coulomb potential $G(\varepsilon) = [\varepsilon - H(1 - i0)]^{-1}$, H denotes the Dirac Hamiltonian, $D^{\mu\nu}$ is the photon propagator, $\alpha_\mu = (1, \boldsymbol{\alpha})$ is the vector composed of the Dirac matrices, and $\mathbf{x}_{12} = \mathbf{x}_1 - \mathbf{x}_2$. As follows from Eq. (9), the loop-after-loop correction can be calculated by generalizing the numerical scheme developed to calculate the first-order SE correction.

Note that we renormalize the mass by introducing the mass counterterm in the system's Lagrangian (see, e.g., [16]). In this case, each (sub)graph that contributes to the self energy is accompanied by the corresponding contribution from the mass renormalization counterterm.

The contribution from the overlapping (O) diagram (Fig. 1b) is given by the following expression (below, for simplicity, we imply the Feynman gauge of the photon propagator):

$$\Delta E_O = 2i\alpha \int_{-\infty}^{\infty} d\omega_1 \int d\mathbf{x}_1 \ldots d\mathbf{x}_4 D(\omega_1, x_{13}) \times \psi_a^\dagger(\mathbf{x}_1)\alpha_\mu G(\varepsilon_a - \omega_1)\gamma^0 \Lambda^\mu(\varepsilon_a - \omega_1, \varepsilon_a) \times \psi_a(\mathbf{x}_4) - \text{c.t.}, \quad (11)$$

where c.t. denotes the corresponding contribution from the mass renormalization counterterm. The quantity D denotes the scalar part of the photon propagator in the Feynman gauge,

$$D(\omega, x_{12}) = \frac{\exp(i\sqrt{\omega^2 + i0}\,x_{12})}{4\pi x_{12}}, \quad (12)$$

where the choice of the phase of the square root is fixed by the condition $\mathrm{Im}\sqrt{\omega^2 + i0} > 0$ and $x_{12} = |\mathbf{x}_{12}|$. The vertex operator Λ^μ is defined as

$$\Lambda^\mu(\varepsilon_a - \omega_1, \varepsilon_a) = 2i\alpha\gamma^0 \int_{-\infty}^{\infty} d\omega_2 D(\omega_2, x_{24})\alpha_\nu \times G(\varepsilon_a - \omega_1 - \omega_2)\alpha^\mu G(\varepsilon_a - \omega_2)\alpha^\nu. \quad (13)$$

Below, for simplicity, we omit the radial arguments in the functions G, Σ, and Λ in the cases where they can be restored unequivocally. The contribution from the nested (N) diagram (Fig. 1c) can be written as

$$\Delta E_N = 2i\alpha \int_{-\infty}^{\infty} d\omega_1 \int d\mathbf{x}_1 \ldots d\mathbf{x}_4 D(\omega_1, x_{14}) \times \psi_a^\dagger(\mathbf{x}_1)\alpha_\mu G(\varepsilon_a - \omega_1)\gamma^0 \tilde{\Sigma}(\varepsilon_a - \omega_1) \times G(\varepsilon_a - \omega_1)\alpha^\mu(\mathbf{x}_4) - \text{c.t.} \quad (14)$$

Finally, the reducible (red) contribution from the diagram in Fig. 1a is given by the formula

$$\Delta E_{\text{red}} = \Delta E_{SE}\langle a|\gamma^0 \frac{\partial}{\partial\varepsilon}\tilde{\Sigma}(\varepsilon)\Big|_{\varepsilon = \varepsilon_a}|a\rangle, \quad (15)$$

where ΔE_{SE} denotes the first-order SE correction,

$$\Delta E_{SE} = \langle a|\gamma^0\tilde{\Sigma}(\varepsilon_a)|a\rangle.$$

The compact part of the two-loop SE correction can be renormalized using the standard technique of free QED by expanding the bound electron propagators in a power series of the interaction with the nucleus. Analysis of the above expressions (see, e.g., [17]) indicates that each of the contributions (11), (14), and (15) contains ultraviolet (UV) divergences that cancel out in the complete sum. Thus, the sum of all three contributions to the energy shift,

$$\Delta E_{\text{comp}} = \Delta E_N + \Delta E_O + \Delta E_{\text{red}}, \quad (16)$$

is free from the UV divergences. The expressions for the reducible and nested contributions also contain infrared (IR) divergences. These IR divergences (also called reference-state singularities) can emerge in bound-state QED when the intermediate-state energy in the spectral decomposition of the electron propagators is equal to the initial-state energy. A detailed analysis of the emerging IR divergences and a proof of their cancellation can be found in [13].

Let us now consider a general scheme for practical calculations of the corresponding contributions. The main problem is that the expressions containing the bound electron propagators are much more convenient to calculate in coordinate space, while the covariant separation of the UV divergences is usually made in the momentum representation for the free propagators. For the first-order SE correction, it proves to be possible to single out all UV divergences in the form of terms containing no bound electron propagators (the so-called 0 and 1 potential terms; see [18, 19]). These contributions are calculated in the momentum representation, while the finite residual (the many-potential term) is considered in coordinate space. No such separation can be

made in the two-loop case, because there are overlapping UV-diverging blocks in the overlapping and nested diagrams. Nevertheless, the general strategy in this case is still based on the expansion of the bound electron operators in a power series of the interaction with the Coulomb nuclear potential.

At the first step, we choose the subtractions composed of the diagrams that contain the free rather than bound electron propagators in such a way that the point-to-point difference between the corresponding contributions contains no UV divergences. (The point-to-point difference means that these subtractions must be performed at each point, i.e., for the integrand.) Following Mallampalli and Sapirstein [9], we call the derived contribution an *M* term. It is schematically shown in Fig. 2. The *M* term consists of three parts that originate from the nested and overlapping diagrams and the reducible part of the diagram in Fig. 1a. The subtractions were chosen in such a way that each of these three parts contained no UV divergences.

We must now take into account the subtracted diagrams. These diagrams naturally break up into two parts. Some of them contain only the free electron propagators and, hence, can be calculated using the standard technique of free QED. The others contain the bound electron propagators and constitute a bigger problem for calculations. For these diagrams, we introduce additional subtractions that reduce the degree of their divergence. The derived point-to-point difference is called a *P* term [9]; the corresponding UV subtractions are schematically shown in Fig. 3. The *P* term consists of three parts, the first two of which arise from the nested diagram, while the third arises from the overlapping diagram. The latter part must be taken into account twice, which corresponds to two equivalent subtractions in the *M* term. Let us illustrate the achieved reduction in the degree of divergence using the first difference in Fig. 3 as an example. It can be shown that this (point-to-point) UV difference diverges only through the inner SE loop, while the divergence originating from the outer SE loop cancels out.

Finally, let us collect all of the remaining subtractions and call the derived contribution an *F* term (Fig. 4). It consists of the diagrams that contain only the free propagators.

Thus, we divided the compact part of the two-loop SE correction into three contributions. The meaning of this division is that each of these terms is calculated using its own method. The *M* term is calculated in coordinate space using an analytical representation for the bound electron propagators. The *F* term is calculated in momentum space using dimensional regularization and Feynman parametrization. The *P* term is calculated in the mixed coordinate–momentum representation, with the bound electron propagators being calculated using a finite basis set for the Dirac equation and the subsequent numerical Fourier transform.

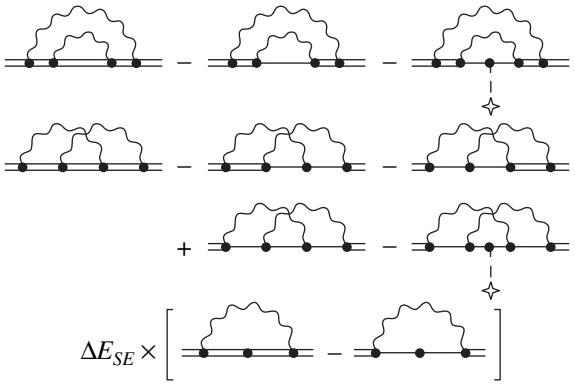

Fig. 2. Schematic view of the UV subtractions for the *M* term. The identity $(\partial/\partial\varepsilon)(\varepsilon - H)^{-1} = -(\varepsilon - H)^{-2}$ is used for the reducible part.

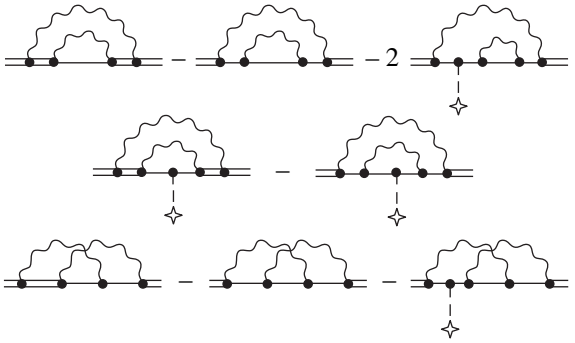

Fig. 3. Schematic view of the UV subtractions for the *P* term.

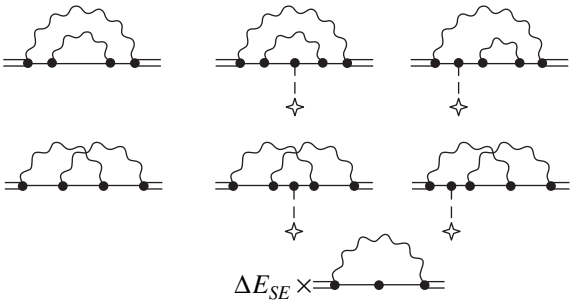

Fig. 4. Graphical representation of the *F* term. The extreme right diagrams in the first two rows must be taken into account twice.

2.1. The M Term: Formulation and Analysis

As follows from Fig. 2, the *M* term consists of three parts that we will call the nested, overlapping, and reducible *M* contributions, respectively:

$$\Delta E_M = \Delta E_{N,M} + \Delta E_{O,M} + \Delta E_{\mathrm{red},M}. \qquad (17)$$

We derive an expression for the nested M contribution from Eq. (14) using additional subtractions,

$$\Delta E_{N,M} = 2i\alpha\int_C d\omega_1\int d\mathbf{x}_1\ldots d\mathbf{x}_4 \times D(\omega_1, x_{14})[\psi_a^\dagger(\mathbf{x}_1)\alpha_\mu G(\varepsilon_a-\omega_1)\gamma^0\Sigma(\varepsilon_a-\omega_1) \times G(\varepsilon_a-\omega_1)\alpha^\mu\psi_a(\mathbf{x}_4) - \text{subtractions}]. \quad (18)$$

The subtractions are defined by the substitution

$$G(\varepsilon_a-\omega_1)\gamma^0\Sigma(\varepsilon_a-\omega_1)G(\varepsilon_a-\omega_1) \longrightarrow G(\varepsilon_a-\omega_1)\gamma^0\Sigma^{(2+)}(\varepsilon_a-\omega_1)G(\varepsilon_a-\omega_1) - G^{(a)}(\varepsilon_a-\omega_1)\gamma^0\Sigma^{(2+)}(\varepsilon_a)G^{(a)}(\varepsilon_a-\omega_1). \quad (19)$$

Here,

$$G^{(a)}(E,\mathbf{x}_1,\mathbf{x}_2) = \sum_{\mu_a}\frac{\psi_a(\mathbf{x}_1)\psi_a^\dagger(\mathbf{x}_2)}{E-\varepsilon_a+i0}, \quad (20)$$

μ_a denotes the projection of the total moment of the states ψ_a in Eq. (20), and $\Sigma^{(2+)}$ is the part of the one-loop SE operator (10) that contains two or more interactions with the Coulomb nuclear potential:

$$\Sigma^{(2+)} = \Sigma - \Sigma^{(0)} - \Sigma^{(1)}, \quad (21)$$

where the superscript denotes the order of the term of the potential expansion. The first part of subtractions (19) cancels the UV divergences, while their second part cancels the IR divergences present in the complete nested contribution (14). A direct analysis indicates that the nested M contribution defined in this way is finite.

Let us now discuss the choice of a contour of integration over ω_1 in Eq. (18) (for simplicity, we assume that a is the ground state). In the initial expression (14), the integration over ω_1 was along the real axis. However, for practical calculations, it is convenient to rotate the contour of integration until it is parallel to the imaginary axis. In our calculations, we used a contour C_{LH} [19, 20] that consists of two parts, low-energy (C_L) and high-energy (C_H). The low-energy part extends from $\varepsilon_0 - i0$ to $-i0$ over the lower edge of the cut of the proton propagator and from $i0$ to $\varepsilon_0 + i0$ over the upper edge of its cut (for details, see [19]). The high-energy part is defined as follows:

$$C_H = [\varepsilon_0 + i0, \varepsilon_0 + i\infty) + (\varepsilon_0 - i\infty, \varepsilon_0 - i0].$$

The parameter ε_0, which divides the contour into two parts, can be chosen arbitrarily within the range $\varepsilon_0 \in (0, m + \varepsilon_a)$. In this paper, we chose this parameter to be $\varepsilon_0 = \varepsilon_a/2$.

To prove the possibility of deforming the contour of integration in Eq. (18), we must show that the SE operator $\Sigma(\varepsilon_a - \omega_1)$ can be analytically continued into the required region of the complex ω_1 plane. Such a proof is presented in [13]. It is also shown in the latter paper that an explicit expression for the SE operator that realizes this analytic continuation can be derived from Eq. (10) by substituting C_{LH} for the contour of integration in ω over $(-\infty, \infty)$. Thus, we conclude that Eq. (18) remains valid if we substitute C_{LH} for the contour of integration over ω_1 and take the analytic continuation of the SE operator given above.

We derive an expression for the overlapping M contribution from Eq. (11) using additional subtractions,

$$\Delta E_{O,M} = 2i\alpha\int_C d\omega_1\int d\mathbf{x}_1\ldots d\mathbf{x}_4 \times D(\omega_1, x_{13})[\psi_a^\dagger(\mathbf{x}_1)\alpha_\mu G(\varepsilon_a-\omega_1)\gamma^0\Lambda^\mu(\varepsilon_a-\omega_1,\varepsilon_a) \times \psi_a(\mathbf{x}_4) - \text{subtractions}], \quad (22)$$

where the vertex operator is defined by Eq. (13) and the subtractions are defined by the substitution (see Fig. 2)

$$G_1G_2G_3 \longrightarrow G_1G_2G_3 - G_1G_2^{(0)}G_3^{(0)} - G_1^{(0)}G_2^{(0)}G_3 + G_1^{(0)}G_2^{(0)}G_3^{(0)} - G_1^{(0)}G_2^{(1)}G_3^{(0)}. \quad (23)$$

Here, $G^{(0)}$ denotes the free Green function of the Dirac equation, and $G^{(1)}$ is the first-order term in the expansion of the Coulomb Green function in a power series of the Coulomb potential V_C:

$$G^{(1)}(E,\mathbf{x}_1,\mathbf{x}_2) = \int d\mathbf{z}G^{(0)}(E,\mathbf{x}_1,\mathbf{z})V_C(z)G^{(0)}(E,\mathbf{z},\mathbf{x}_2). \quad (24)$$

It can be shown that expression (22) is finite.

Just as for the nested M contribution, it is convenient to numerically calculate expression (22) using contour C_{LH} for integration over ω_1. The fact that the vertex operator $\Lambda^\mu(\varepsilon_a - \omega_1, \varepsilon_a)$ admits of the required analytic continuation into the complex ω_1 plane is proven in the same way as for the SE operator. An explicit expression for the analytic continuation of the vertex operator can be conveniently chosen by substituting C_{LH} for the contour of integration in Eq. (13).

Finally, an expression for the reducible M contribution can be derived from Eq. (15) with the correspond-

ing subtractions. Using the explicit expression for the SE operator, we obtain

$$\Delta E_{\mathrm{red},M} = -2i\alpha\Delta E_{SE}\int_C d\omega\int d\mathbf{x}_1 d\mathbf{x}_2 d\mathbf{x}_3 D(\omega, x_{13}) \times \psi_a^\dagger(\mathbf{x}_1)\alpha_\mu[G(\varepsilon_a-\omega)G(\varepsilon_a-\omega) - \text{subtractions}] \times \alpha^\mu\psi_a(\mathbf{x}_3), \quad (25)$$

where the subtractions are defined by the substitution

$$G_1G_2 \longrightarrow G_1G_2 - G_1^{(0)}G_2^{(0)} - G_1^{(a)}G_2^{(a)}. \quad (26)$$

Here, the first and second parts of the subtractions cancel the UV and IR divergences, respectively. The choice of a contour of integration in Eq. (25), along with the numerical calculation of this contribution, is made in the same way as for the first-order SE correction [19].

2.2. The P Term: Formulation and Analysis

A general scheme of the UV subtractions in the *P* term is presented in Fig. 3. The *P* term consists of three parts,

$$\Delta E_P = \Delta E_{N1,P} + \Delta E_{N2,P} + 2\Delta E_{O,P}, \quad (27)$$

which correspond to the three rows in Fig. 3. We will refer to the individual contributions as the first nested *P* term, the second nested *P* term, and the overlapping *P* term, respectively. The latter contribution must be taken into account twice, which corresponds to two equivalent subtractions in the *M* term.

The main problem in calculating the *P* term is that the subtractions shown in Fig. 3 do not cancel all UV divergences. For this reason, we will write the general expressions for the *P* contributions in the mixed coordinate–momentum representation. In this case, we will calculate the UV-diverging part in the momentum representation (which facilitates the covariant separation of the UV divergences) and the remaining part of the expression in the coordinate representation. Using the coordinate representation significantly facilitates the analytical calculation of the integrals over angular variables and the subsequent numerical calculation of the derived expressions.

The general expression for the first nested *P* term is

$$\Delta E_{N1,P} = 2i\alpha\int_C d\omega\int\frac{d\mathbf{p}}{(2\pi)^3}\int d\mathbf{x}_1 d\mathbf{x}_2 \times D(\omega, x_{12})\psi_a^\dagger(\mathbf{x}_1)\alpha_\mu[G(E,\mathbf{x}_1,\mathbf{p}) \times \gamma^0\tilde{\Sigma}^{(0)}(E,\mathbf{p})G(E,\mathbf{p},\mathbf{x}_2) - \text{subtractions}]\alpha^\mu\psi_a(\mathbf{x}_2), \quad (28)$$

where $E \equiv \varepsilon_a - \omega$, $\Sigma^{(0)}$ is the free SE operator, $\tilde{\Sigma}^{(0)} = \Sigma^{(0)} - \delta m^{(1)}$, the subtractions are defined by the substitution

$$G\tilde{\Sigma}^{(0)}(E)G \longrightarrow G\tilde{\Sigma}^{(0)}(E)G - G^{(0)}\tilde{\Sigma}^{(0)}(E)G^{(0)} - G^{(1)}\tilde{\Sigma}^{(0)}(E)G^{(0)} - G^{(0)}\tilde{\Sigma}^{(0)}(E)G^{(1)} - G^{(a)}\tilde{\Sigma}^{(0)}(\varepsilon_a)G^{(a)}, \quad (29)$$

and the Green function in the mixed representation is defined as

$$G(\varepsilon,\mathbf{x}_1,\mathbf{p}) = \int d\mathbf{x}_2 \exp(i\mathbf{p}\cdot\mathbf{x}_2)G(\varepsilon,\mathbf{x}_1,\mathbf{x}_2), \quad (30)$$

$$G(\varepsilon,\mathbf{p},\mathbf{x}_2) = \int d\mathbf{x}_1 \exp(-i\mathbf{p}\cdot\mathbf{x}_1)G(\varepsilon,\mathbf{x}_1,\mathbf{x}_2). \quad (31)$$

Note that the last part of subtractions (29) cancels the IR divergences.

The second nested *P* contribution is given by the expression

$$\Delta E_{N2,P} = 2i\alpha\int_C d\omega\int\frac{d\mathbf{p}_1}{(2\pi)^3}\frac{d\mathbf{p}_2}{(2\pi)^3} \times \int d\mathbf{x}_1 d\mathbf{x}_2 D(\omega, x_{12})V_C(\mathbf{q})\psi_a^\dagger(\mathbf{x}_1) \times \alpha_\mu[G(E,\mathbf{x}_1,\mathbf{p}_1)\gamma^0\Gamma^0(E,\mathbf{p}_1;E,\mathbf{p}_2) \times G(E,\mathbf{p}_2,\mathbf{x}_2) - \text{subtractions}]\alpha^\mu\psi_a(\mathbf{x}_2), \quad (32)$$

where $\mathbf{q} = \mathbf{p}_1 - \mathbf{p}_2$, Γ^μ denotes the free vertex operator (given by (13) in the case of zero external field), and the subtractions are defined by the substitution

$$G\Gamma^0(E,E)G \longrightarrow G\Gamma^0(E,E)G - G^{(0)}\Gamma^0(E,E)G^{(0)} - G^{(a)}\Gamma^0(\varepsilon_a,\varepsilon_a)G^{(a)}. \quad (33)$$

Again, the last part of subtractions (33) cancels the IR divergences.

The expression for the overlapping *P* contribution is

$$\Delta E_{O,P} = -2i\alpha\int_C d\omega\int\frac{d\mathbf{p}_1}{(2\pi)^3}\frac{d\mathbf{p}_2}{(2\pi)^3} \times \int d\mathbf{x}_1\frac{\exp(-i\mathbf{q}\cdot\mathbf{x}_1)}{\omega^2-\mathbf{q}^2+i0}\psi_a^\dagger(\mathbf{x}_1)\alpha_\mu G^{(2+)}(E,\mathbf{x}_1,\mathbf{p}_1) \times \gamma^0\Gamma^\mu(E,\mathbf{p}_1;\varepsilon_a,\mathbf{p}_2)\psi_a(\mathbf{p}_2), \quad (34)$$

where $G^{(2+)}$ denotes the part of the Coulomb Green function that contains two or more interactions with the Coulomb field: $G^{(2+)} = G - G^{(0)} - G^{(1)}$.

Let us now discuss the divergences in the expressions for the P contributions. Using dimensional regularization, we first separate out the UV-diverging parts of the free one-loop operators:

$$\tilde{\Sigma}^{(0)}(p) = B^{(1)}(\not{p} - m) + \Sigma_R^{(0)}, \tag{35}$$

$$\Gamma^{\mu}(p_1, p_2) = L^{(1)}\gamma^{\mu} + \Gamma_R^{\mu}(p_1, p_2), \tag{36}$$

where the subscript R denotes the UV-finite part of the corresponding contribution. Using these formulas, the Ward identity ($L^{(1)} = -B^{(1)}$), and the Dirac equation, we obtain expressions for the UV-diverging parts of the P contributions:

$$\begin{aligned}&\Delta E_{N1,P} + \Delta E_{N2,P}\\ &= B^{(1)}\Delta E_{\text{many},D} + \Delta E^R_{N1,P} + \Delta E^R_{N2,P},\end{aligned} \tag{37}$$

$$2\Delta E_{O,P} = 2L^{(1)}\Delta E_{\text{many},D} + 2\Delta E^R_{O,P}. \tag{38}$$

Here, $\Delta E_{\text{many},D} = \langle a|\gamma^0\Sigma^{(2+)}(\varepsilon_a)|a\rangle$ is the many-potential part of the first-order SE correction, and the subscript D indicates that this contribution must be calculated in space of extended dimension D, since it is multiplied by a diverging renormalization constant. (Fortunately, there is no need for its practical calculation, because it cancels out in the sum with the F term.) Expressions for $\Delta E^R_{N1,P}$, $\Delta E^R_{N2,P}$, and $\Delta E^R_{O,P}$ can be derived from Eqs. (28), (32), and (34) using the substitutions $\tilde{\Sigma} \longrightarrow \Sigma_R$ and $\Gamma^{\mu} \longrightarrow \Gamma^{\mu}_R$.

It should be noted that our choice of renormalization constants $B^{(1)}$ and $L^{(1)}$ differs from the standard choice (see, e.g., [16]). In Eqs. (35) and (36), we single out only the UV-diverging part of the one-loop operators. In this case, the constants $B^{(1)}$ and $L^{(1)}$ (in space of dimension $D = 4 - 2\epsilon$) are given by the following explicit expression:

$$B^{(1)} = -L^{(1)} = \frac{\alpha C_\epsilon}{4\pi\epsilon}, \tag{39}$$

where the constant C_ϵ is

$$C_\epsilon = \Gamma(1+\epsilon)(4\pi)^\epsilon\left(\frac{\mu^2}{m^2}\right)^\epsilon \tag{40}$$

and μ is a unit parameter that has the dimensions of mass and that was introduced so that the term with the interaction in the system's Lagrangian has the correct dimensions at $D \neq 4$. This choice of renormalization constants allows the emergence of IR divergences in the expressions for $\Delta E^R_{N1,P}$, $\Delta E^R_{N2,P}$, and $\Delta E^R_{O,P}$ to be avoided.

Taking into account the Ward identity, we find that the P term can be represented by the sum

$$\begin{aligned}\Delta E_P = {} & L^{(1)}\Delta E_{\text{many},D} + \Delta E^R_{N1,P}\\ & + \Delta E^R_{N2,P} + 2\Delta E^R_{O,P},\end{aligned} \tag{41}$$

where only the first term diverges.

It can be shown that the IR part of subtractions (29) and (33) in the P term cancels out completely with the corresponding part of subtractions (19) and (26) in the M term. Therefore, no need arises to consider the subtracted IR-diverging terms separately.

For the convenience of numerically calculating the renormalized expressions (28), (32), and (34), the contour of integration over ω should be rotated until it is parallel to the imaginary axis. In contrast to contour C_{LH}, which was used to calculate the M term, we used the standard Wick rotation of the contour $\omega \longrightarrow i\omega$ to calculate the P contributions. This deformation of the contour gives rise to the so-called pole contributions that originate from the intermediate states in the electron propagators with the energy $\varepsilon_n = \varepsilon_a$. A detail discussion of this method of contour rotation and the expressions emerging in this case can be found in [11].

2.3. The F Term: Formulation and Analysis

The set of diagrams that contribute to the F term is shown in Fig. 4. Recall that each (sub)graph contributing to the self-energy is accompanied by the corresponding mass renormalization counterterm. The F term is peculiar in that the corresponding diagrams contain only the free electron propagators. Therefore, it was calculated in momentum space using the standard technique of free QED. The UV divergences were regularized by passing to space of extended dimension $D = 4 - 2\epsilon$.

The diagrams presented in Fig. 4 can be divided into three parts. The first part consists of the nested and overlapping diagrams containing no interaction with the Coulomb field in the electron propagators. We will call it the 0-potential F term; the corresponding contributions will be marked by the subscript zero. The second part consists of the diagrams containing one Coulomb interaction; it is called the 1-potential F term (one). Finally, the remaining part is called the reducible F contribution (red). Thus,

$$\Delta E_F = \Delta E^{(2)}_{\text{zero}} + \Delta E^{(2)}_{\text{one}} + \Delta E^{\text{zero}}_{\text{red}}. \tag{42}$$

The expression for the 0-potential F contribution

can be written as

$$\Delta E^{(2)}_{\text{zero}} = \int \frac{d\mathbf{p}}{(2\pi)^3} \bar{\psi}_a(\mathbf{p}) \tilde{\Sigma}^{(2)}_{\text{zero}}(p) \psi_a(\mathbf{p}), \tag{43}$$

where $p = (\varepsilon_a, \mathbf{p})$ is the 4-dimensional momentum, $\tilde{\Sigma}^{(2)}_{\text{zero}} = \Sigma^{(2)}_{\text{zero}} - \delta m^{(2)}$, $\Sigma^{(2)}_{\text{zero}}$ denotes the free two-loop SE operator, and $\delta m^{(2)}$ is the two-loop mass renormalization counterterm. The free two-loop SE operator consists of two parts that arise from the nested (N) and overlapping (O) diagrams. The corresponding contributions are given by the expressions

$$\Sigma^{(2N)}_{\text{zero}}(p) = -4\pi i \alpha \mu^{2\epsilon} \int \frac{d^D k}{(2\pi)^D} \frac{1}{k^2} \gamma_\sigma \times \frac{1}{\not p - \not k - m} \tilde{\Sigma}^{(0)}_D(p-k) \frac{1}{\not p - \not k - m} \gamma^\sigma, \tag{44}$$

$$\Sigma^{(2O)}_{\text{zero}}(p) = -4\pi i \alpha \mu^{2\epsilon} \int \frac{d^D k}{(2\pi)^D} \frac{1}{k^2} \gamma_\sigma \times \frac{1}{\not p - \not k - m} \Gamma^\sigma_D(p-k, p), \tag{45}$$

where $\not p = p_\mu \gamma^\mu$, $\tilde{\Sigma}^{(0)}_D = \Sigma^{(0)}_D - \delta m^{(1)}$, $\Sigma^{(0)}_D$ and $\Gamma^\sigma_D(p_1, p_2)$ denote the free one-loop SE and vertex operators, and the subscript D emphasizes that they must be calculated in space of extended dimension. The general structure of the UV divergences in the free two-loop SE operator is given by the expression

$$\tilde{\Sigma}^{(2)}_{\text{zero}}(p) = (\not p - m) B^{(2)} + \frac{\alpha C^2_\epsilon}{4\pi\epsilon} \Sigma^{(0)}_{R,4}(p) + \Sigma^{(2)}_{\text{zero},R}(p), \tag{46}$$

where $\Sigma^{(0)}_{R,4}$ denotes the renormalized free one-loop SE operator calculated in space of physical dimension, and the two-loop renormalization constant is

$$B^{(2)} = \frac{\alpha^2 C^2_\epsilon}{16\pi^2} \left(-\frac{1}{2\epsilon^2} + \frac{3}{4\epsilon} \right). \tag{47}$$

The one-potential F term can be expressed in terms of the corresponding contribution in the two-loop SE operator as follows:

$$\Delta E^{(2)}_{\text{one}} = \int \frac{d\mathbf{p}_1 d\mathbf{p}_2}{(2\pi)^6} \bar{\psi}_a(\mathbf{p}_1) V_C(\mathbf{q}) \times \Sigma^{(2)}_{\text{one}}(p_1, p_2) \psi_a(\mathbf{p}_2), \tag{48}$$

where $p_1 = (\varepsilon_a, \mathbf{p}_1)$, $p_2 = (\varepsilon_a, \mathbf{p}_2)$, and $\mathbf{q} = \mathbf{p}_1 - \mathbf{p}_2$. The one-potential two-loop SE operator $\Sigma^{(2)}_{\text{one}}$ is represented by four nonequivalent diagrams in Fig. 4, which can be derived from two free diagrams by adding the interaction with the Coulomb field at all the possible positions. When the interaction is added to the central propagator, we will call this contribution a ladder one; otherwise, it will be called a side one. The side contributions must be taken into account twice to allow for the equivalent diagrams. Let us now give the general expressions for the one-potential contributions:

$$\Sigma^{(2N)}_{\text{lad}}(p_1, p_2) = -4\pi i \alpha \mu^{2\epsilon} \int \frac{d^D k}{(2\pi)^D} \frac{1}{k^2} \gamma_\sigma \times \frac{1}{\not p_1 - \not k - m} \Gamma^0_D(p_1 - k, p_2 - k) \frac{1}{\not p_2 - \not k - m} \gamma^\sigma, \tag{49}$$

$$\Sigma^{(2N)}_{\text{side}}(p_1, p_2) = -8\pi i \alpha \mu^{2\epsilon} \int \frac{d^D k}{(2\pi)^D} \frac{1}{k^2} \times \gamma_\sigma \frac{1}{\not p_1 - \not k - m} \tilde{\Sigma}^{(0)}_D(p_1 - k) \frac{1}{\not p_1 - \not k - m} \times \gamma^0 \frac{1}{\not p_2 - \not k - m} \gamma^\sigma, \tag{50}$$

$$\Sigma^{(2O)}_{\text{lad}}(p_1, p_2) = (4\pi i \alpha \mu^{2\epsilon})^2 \int \frac{d^D k}{(2\pi)^D} \frac{d^D l}{(2\pi)^D} \frac{1}{k^2} \frac{1}{l^2} \times \gamma_\sigma \frac{1}{\not p_1 - \not k - m} \gamma_\alpha \frac{1}{\not p_1 - \not k - \not l - m} \times \gamma^0 \frac{1}{\not p_2 - \not k - \not l - m} \gamma^\sigma \frac{1}{\not p_2 - \not l - m} \gamma^\alpha, \tag{51}$$

$$\Sigma^{(2O)}_{\text{side}}(p_1, p_2) = -8\pi i \alpha \mu^{2\epsilon} \int \frac{d^D k}{(2\pi)^D} \frac{1}{k^2} \gamma_\sigma \times \frac{1}{\not p_1 - \not k - m} \gamma^0 \frac{1}{\not p_2 - \not k - m} \Gamma^\sigma_D(p_2 - k, p_2). \tag{52}$$

We also give the general structure of the UV divergences in the one-potential term:

$$\Sigma^{(2)}_{\text{one}}(p_1, p_2) = \gamma^0 L^{(2)} + \frac{\alpha C^2_\epsilon}{4\pi\epsilon} \Gamma^0_{R,4}(p_1, p_2) + \Sigma^{(2)}_{\text{one},R}(p_1, p_2), \tag{53}$$

where $\Gamma^\mu_{R,4}(p_1, p_2)$ is the free vertex operator calculated in space of dimension 4, and the renormalization constant $L^{(2)}$ is related to the constant $B^{(2)}$ by the Ward identity $L^{(2)} = -B^{(2)}$.

Finally, let us consider the last part of the F term. The reducible F contribution can be written as

$$\Delta E^{\rm zero}_{\rm red} = \Delta E_{SE,D}\langle a|\gamma^0\frac{\partial}{\partial p^0}\Sigma^{(0)}_D(p)\Big|_{p^0=\varepsilon_a}|a\rangle, \qquad (54)$$

where $\Delta E_{SE,D}$ is the first-order SE correction calculated in space of extended dimension. By separating out explicitly the contributions in the limit $\epsilon \longrightarrow 0$, we obtain

$$\Delta E^{\rm zero}_{\rm red} = -\frac{\alpha C_\epsilon^2}{4\pi\epsilon}(\Delta E_{\rm zero}+\Delta E_{\rm one}) + B^{(1)}\Delta E_{{\rm many},D}+\Delta E^{\rm zero}_{{\rm red},R}, \qquad (55)$$

where

$$\Delta E_{\rm zero} = \langle a|\gamma^0\Sigma^{(0)}_{R,4}|a\rangle$$

and

$$\Delta E_{\rm one} = \langle a|V_C(q)\gamma^0\Gamma^0_{R,4}|a\rangle$$

are the 0- and 1-potential parts of the one-loop SE correction.

Finally, let us gather together the 0-potential, 1-potential, and reducible F contributions and write the final result as

$$\Delta E_F = B^{(1)}\Delta E_{{\rm many},D}+\Delta E^{(2)}_{{\rm zero},R} + \Delta E^{(2)}_{{\rm one},R}+\Delta E^{\rm zero}_{{\rm red},R}. \qquad (56)$$

We used the Ward identity and the Dirac equation to derive this expression. The diverging part of Eq. (56) cancels out in the sum with the P term (see (41)).

3. THE SCHEME OF CALCULATIONS

The general scheme for calculating the individual contributions to the two-loop SE correction is described in [13]. The goal of this paper is to calculate this correction in the range of low nuclear charges Z, where the calculations are complicated significantly by the emerging numerical cancellations. In this section, we describe the main points that allowed us to achieve a relatively high accuracy of calculations in the range of low Z.

The calculation of the M term is most exacting from the standpoint of computational resources. This is because after the integrations over angular variables in Eqs. (18) and (22) and allowance for the selection rules, the corresponding expressions are represented as a double infinite sum over partial waves. To achieve the required accuracy of the result and to keep the total time of the calculation within reasonable limits, we must separate the parts that contain the dominant contribution in the limit of low Z and that can be calculated more easily than the complete expressions for $\Delta E_{N,M}$ and $\Delta E_{O,M}$.

For the nested M contribution, this part can be derived from Eq. (18) by the substitution

$$G(\varepsilon_a-\omega_1)\gamma^0\Sigma(\varepsilon_a-\omega_1)G(\varepsilon_a-\omega_1)\longrightarrow G^{(a)}(\varepsilon_a-\omega_1) \times\gamma^0[\Sigma^{(2+)}(\varepsilon_a-\omega_1)-\Sigma^{(2+)}(\varepsilon_a)]G^{(a)}(\varepsilon_a-\omega_1). \qquad (57)$$

After this substitution, the double summation over partial waves is reduced to a single summation, and the general structure of the expression is simplified; as a result, the corresponding numerical calculation can be performed very efficiently. For the overlapping M term, we chose the dominant contribution as follows. Denoting the relativistic angular parameters of the first, second, and third electron propagators in Eq. (22) by κ_1, κ_2, and κ_3, respectively, and the angular momenta of the first and second virtual photons by l_1 and l_2, we separated out the contribution with $\kappa_1=\kappa_a$, $\kappa_2=\kappa_3$, and $l_1=0$ and the (identical) contribution with $\kappa_3=\kappa_a$, $\kappa_1=\kappa_2$, and $l_2=0$. Clearly, these contributions contain only the single summation over partial waves and, hence, can be calculated with a better accuracy than the complete expression.

Let us now consider the general scheme for calculating the P term. To reduce the expression for the first nested P contribution to a form convenient for a numerical calculation, we rewrite it as

$$\Delta E^R_{N1,P} = 2i\alpha\int_C d\omega\int\frac{d\mathbf{p}}{(2\pi)^3} \times\Bigg[\sum_{n_1,n_2}\frac{\langle an_2|\alpha_\mu\alpha^\mu D(\omega)|n_1a\rangle}{(\varepsilon_a-\omega-\varepsilon_{n_1})(\varepsilon_a-\omega-\varepsilon_{n_2})}\psi^\dagger_{n_1}(\mathbf{p}) \times\gamma^0\Sigma^{(0)}_R(\varepsilon_a-\omega,\mathbf{p})\psi_{n_2}(\mathbf{p})-{\rm subtractions}\Bigg]. \qquad (58)$$

Here, we use the spectral representation for the electron propagators, and the subtractions are defined by Eq. (29). Let us now introduce the matrix S:

$$S(\omega,\mathbf{x}_1,\mathbf{x}_2) = \sum_{n_1n_2}\frac{\langle an_2|\alpha_\mu\alpha^\mu D(\omega)|n_1a\rangle}{(\varepsilon_a-\omega-\varepsilon_{n_1})(\varepsilon_a-\omega-\varepsilon_{n_2})} \times\psi_{n_2}(\mathbf{x}_1)\psi^\dagger_{n_1}(\mathbf{x}_2). \qquad (59)$$

Its Fourier transform is defined as

$$S(\omega,\mathbf{p}_1,\mathbf{p}_2) = \int d\mathbf{x}_1d\mathbf{x}_2\exp(i\mathbf{p}_2\cdot\mathbf{x}_2-i\mathbf{p}_1\cdot\mathbf{x}_1) \times S(\omega,\mathbf{x}_1,\mathbf{x}_2). \qquad (60)$$

We can now write Eq. (58) as

$$\Delta E^R_{N1,P} = 2i\alpha\int_C d\omega\int\frac{d\mathbf{p}}{(2\pi)^3}\{\mathrm{Tr}[S(\omega,\mathbf{p},\mathbf{p}) \times\gamma^0\Sigma^{(0)}_R(\varepsilon_a-\omega,\mathbf{p})] - \text{subtractions}\}. \quad (61)$$

After the analytical integrations over angular variables, we used this expression for our numerical calculation. The matrix S was constructed initially in the coordinate representation and subsequently in the momentum representation using a finite basis set for the Dirac equation constructed from B-splines [21]. The use of a basis set satisfying the dual kinetic balance relations [22] was an important point for achieving the required accuracy of our calculations in the range of low Z.

To facilitate the numerical calculation, the expression for the second nested P contribution should be written in a form similar to (61):

$$\Delta E^R_{N2,P} = 2i\alpha\int_C d\omega\int\frac{d\mathbf{p}_1}{(2\pi)^3}\frac{d\mathbf{p}_2}{(2\pi)^3}V_C(q) \times\{\mathrm{Tr}[S(\omega,\mathbf{p}_2,\mathbf{p}_1) \times\gamma^0\Gamma^0_R(\varepsilon_a-\omega,\mathbf{p}_1;\varepsilon_a-\omega,\mathbf{p}_2)] - \text{subtractions}\}, \quad (62)$$

where the subtractions are defined by Eq. (33). After the analytical integrations over angular variables, we used this expression for our numerical calculation. In contrast to the calculations for $\Delta E^R_{N1,P}$, this calculation contains two additional numerical integrations (a triple integral over p_1, p_2, and $\xi = \hat{\mathbf{p}}_1\cdot\hat{\mathbf{p}}_2$ instead of a single integral over p). The range of large momenta p_1 and p_2 presents a considerable difficulty in performing the numerical integrations. The reason is that the Green function constructed using a finite basis set in the coordinate representation is not smooth enough (compared to the exact Green function). After applying the numerical Fourier transform, this nonsmoothness manifests itself in the form of oscillations at large momenta, which, in turn, leads to an insufficient stability of the numerical integrations in this range. This problem was solved by choosing the subtractions whose behavior at large momenta approached the behavior of the integrand in Eq. (62), but whose structure was simpler and admitted of a more accurate numerical calculation. Our chosen set of subtractions is given by the following substitution in Eq. (62):

$$\Gamma^0_R(\varepsilon_a-\omega,\mathbf{p}_1;\varepsilon_a-\omega,\mathbf{p}_2) \longrightarrow \Gamma^0_R(\varepsilon_a-\omega,\mathbf{p}_1;\varepsilon_a-\omega,\mathbf{p}_2) - \Gamma^0_R(\varepsilon_a,\mathbf{p}_1;\varepsilon_a,\mathbf{p}_1) - \frac{1}{2}[\Gamma^0_R(\varepsilon_a-\omega,\mathbf{p}_1;\varepsilon_a-\omega,\mathbf{p}_1) - \Gamma^0_R(\varepsilon_a,\mathbf{p}_1;\varepsilon_a,\mathbf{p}_1)] - \frac{1}{2}[\Gamma^0_R(\varepsilon_a-\omega,\mathbf{p}_2;\varepsilon_a-\omega,\mathbf{p}_2) - \Gamma^0_R(\varepsilon_a,\mathbf{p}_2;\varepsilon_a,\mathbf{p}_2)]. \quad (63)$$

Clearly, this difference (as a function of p_1 and p_2) decreases well at $p_1\approx p_2$ and $p_{1,2}\gg\omega$. At the same time, each of the subtracted terms is easier to calculate numerically than the initial expression (62).

We numerically calculated the overlapping P term using Eq. (34) after the analytical integrations over angular variables. To improve the stability of the numerical integrations in the range of large momenta p and q, we used the subtraction

$$\Gamma^0_R(\varepsilon_a-\omega,\mathbf{p}_1;\varepsilon_a,\mathbf{p}_2) \longrightarrow \Gamma^0_R(\varepsilon_a-\omega,\mathbf{p}_1;\varepsilon_a,\mathbf{p}_2) - \Gamma^0_R(\varepsilon_a,\mathbf{p}_1;\varepsilon_a,\mathbf{p}_2). \quad (64)$$

The numerical calculation of the subtracted contribution is facilitated greatly by the fact that the integration for it over ω in Eq. (34) is performed analytically.

4. RESULTS AND DISCUSSION

Table 1 presents our numerical results for the individual contributions to the two-loop SE correction to the ground-state energy of hydrogen-like ions with nuclear charges $Z \geq 10$. The calculations were performed in the Feynman gauge for the point nuclear charge model. In this paper, we calculated the correction for $Z = 10$–30 and significantly increased the accuracy for $Z = 40$, 50, and 60 compared to our previous calculations of this correction [12, 13]. The numerical data for $Z \geq 70$ in Table 1 were taken from [13]. The results of our calculations to all orders in parameter $Z\alpha$ can be compared with the data obtained within the $Z\alpha$ expansion. This comparison is shown in Fig. 5, where the dashed line indicates the contribution from all of the known expansion terms (Eqs. (3)–(8)), while the solid line and circles represent the results of our calculations (in units of the dimensionless function $F(Z\alpha)$ defined by Eq. (1)). We see that the numerical data show a smooth Z dependence and visually tend to the well-known result as $Z \longrightarrow 0$. Note that the physical $(\sim(Z\alpha)^4)$ dependence of the complete correction arises in our calculations from a delicate cancellation between the individual contributions (see Table 1). The observed agreement between the numerical values for the complete correction with the first $Z\alpha$-expansion terms is an argument for the reliability of our results.

To make a more detailed comparison of the numerical data with the first $Z\alpha$-expansion terms, let us sepa-

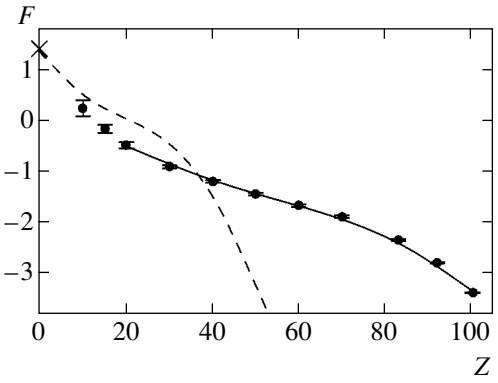

Fig. 5. The two-loop SE correction in all orders in $Z\alpha$ (circles and solid line) and the results obtained within the $Z\alpha$ expansion (dashed line).

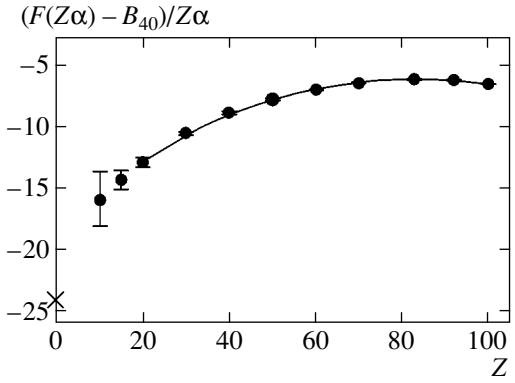

Fig. 6. Numerical results for the function $\tilde{F}(Z\alpha)$ defined by Eq. (65). The overlapping on the vertical axis denotes the analytical value of this function at $Z = 0$.

rate the contribution of all orders starting from $(Z\alpha)^5$ from our results:

$$\tilde{F}(Z\alpha) = \frac{F(Z\alpha) - B_{40}}{Z\alpha} = B_{50} + Z\alpha(\ldots). \qquad (65)$$

Our numerical results for the function $\tilde{F}(Z\alpha)$ are presented in Fig. 6. The overlapping on the vertical axis denotes the analytical value of $\tilde{F}(0)$. A quantitative comparison with the analytical results was made by extrapolating the data to the point with $Z = 0$. For the extrapolation, we chose a procedure that was first used by Mohr [23] and has recently been described in detail in [24]. This extrapolation applied to our data (the points with $Z = 10$ and $Z = 15$ were not used because of the large numerical error) reproduces the analytical value for the coefficient B_{50} with an accuracy of 15%. This relatively large extrapolation error results from the presence of powers of $\ln(Z\alpha)$ in the order $(Z\alpha)^6$ and the absence of numerical results in the range of low Z. Based on the above analysis, we may conclude that our data confirm the analytical results for the expansion coefficients of order $(Z\alpha)^4$ and $(Z\alpha)^5$.

To make a comparison with the analytical calculations in the next order in $Z\alpha$, let us separate the residual $G^{\text{h.o.}}(Z\alpha)$ defined by Eq. (2) from our results. The

Table 1. The individual contributions to the two-loop SE correction

Z	ΔE_{LAL}	ΔE_F^R	ΔE_P^R	ΔE_M	Sum
10	−0.3577	822.14(2)	−721.34(12)	−100.19(10)	0.25(16)
15	−0.4951	292.902(13)	−235.205(70)	−57.366(48)	−0.164(85)
20	−0.6015	136.911(7)	−102.026(55)	−34.764(16)	−0.481(58)
30	−0.7565	44.729(3)	−29.410(25)	−15.465(5)	−0.903(26)
40	−0.8711	19.505(3)	−11.575(30)	−8.253(5)	−1.194(31)
			−11.41(15)[a]	−8.27(18)[a]	−1.05(23)[a]
50	−0.9734	10.025(2)	−5.488(26)	−5.001(3)	−1.437(26)
			−5.41(8)[a]	−4.99(6)[a]	−1.34(10)[a]
60	−1.082	5.723(1)	−2.970(18)	−3.341(2)	−1.670(18)
			−2.93(4)[a]	−3.342(21)[a]	−1.63(4)[a]
70	−1.216	3.497(1)	−1.757(25)	−2.412(11)	−1.888(27)
83	−1.466	1.938	−1.057(13)	−1.764(4)	−2.349(14)
92	−1.734	1.276	−0.812(10)	−1.513(3)	−2.783(10)
100	−2.099	0.825	−0.723(7)	−1.384(3)	−3.381(8)

Note: All of the data are given in units of $F(Z\alpha)$. [a] The data from [13].

numerical data for this function are shown in Fig. 7. We see that the visual limit of $G^{\text{h.o.}}(Z\alpha)$ at low Z is approximately twice that obtained by Pachucki and Jentschura [2] for the coefficient $B_{60} = -62(9)$. Thus, we may conclude that our data disagree with the available analytical results in the order $(Z\alpha)^6$. At present, we cannot unambiguously determine which of the coefficients (B_{60}, B_{61}, B_{62}, or B_{63}) is responsible for the disagreement. Already after the preliminary results of our calculations had been obtained, the result from [3] (Eq. (7)) for the coefficient B_{61} was found by Pachucki [25] to be incomplete. The missing contribution to B_{61} is currently being calculated. It may well be that this extra term will restore the agreement between the results of different methods for calculating the two-loop SE correction.

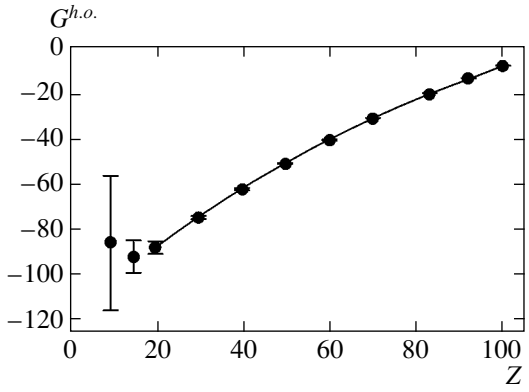

Fig. 7. Numerical results for the function $G^{\text{h.o.}}(Z\alpha)$ defined by Eq. (2).

Our results for the higher order contribution $G^{\text{h.o.}}(Z\alpha)$ can be used to improve the existing results for the two-loop SE correction in ions with $Z < 10$. The hydrogen atom is particularly important from a practical point of view. Applying the extrapolation procedure described above to the data with $Z \geq 20$ in Fig. 7 yields the following result for $Z = 1$:

$$G^{\text{h.o.}}(1\alpha) = -127 \pm 30\%. \qquad (66)$$

This error was obtained by assuming that the possible additional contribution to the coefficient B_{61} does not exceed 10. Result (66) differs significantly from the previous estimate of the higher order contribution obtained in [2] within the $Z\alpha$ expansion:

$$G^{\text{h.o.}}(1\alpha; \text{old}) = -61.6 \pm 15\%. \qquad (67)$$

This difference changes the theoretical ground-state Lamb shift in the hydrogen atom by 7 kHz.

The contributions from the various corrections to the ground-state Lamb shift in several hydrogen-like ions are given in Table 2. By tradition, the term "Lamb shift" is defined somewhat differently for light and heavy ions. This is because for light ions it is natural to consider the shift relative to the Dirac energy level with the correct dependence on the reduced mass (thereby

Table 2. The individual contributions to the ground-state Lamb shift in hydrogen-like ions

	$Z = 1$	$Z = 54$	$Z = 79$	$Z = 92$
QED effects, $\sim\alpha$	8168514	43.841	156.48	266.42
Two-loop SE correction, $\sim\alpha^2$	2327(4)	−0.097(7)	−0.67(1)	−1.56(1)
Other QED effects, $\sim\alpha^2$	−1603	−0.005(9)	0.08(11)	0.30(33)
QED effects, $\sim\alpha^3$	2			
Finite nuclear size	1163(32)[a]	3.181(7)	49.11(11)	198.33(52)
	1253(50)[b]			
Recoil	2389	0.187	0.33	0.46
Nuclear polarization	5			−0.19(9)
Complete theoretical value	8172797(32)(4)[a]	47.107(13)	205.33(16)	463.76(62)
	8172888(50)(4)[b]			
Experimental value	8172837(22)[c]	54(10)[d]	202(8)[e]	468(13)[f]
				460.2(4.6)[g]

Note: The data are given in kHz in the second column and in eV in the third, fourth, and fifth columns. [a] $r_p = 0.862(12)$ Fm [26]; [b] $r_p = 0.895(18)$ Fm [27]; [c] a combination of results from [28, 29]; [d] data from [31]; [e] data from [32]; [f] data from [1]; [g] data from [33].

taking into account much of the recoil effect). For heavy ions, it makes no sense to separate out the nonrelativistic part of the recoil effect. Thus, to obtain the ground-state energy of the hydrogen atom, we must add the following to the Lamb shifts listed in Table 2:

$$E_M = M + [f(n, j) - 1]m_r - [f(n, j) - 1]^2 \frac{m_r^2}{2M}, \quad (68)$$

where $M = m + m_N$ is the atomic mass, m_N is the nuclear mass, and $m_r = mm_N/(m + m_N)$ is the reduced mass. The function $f(n, j)$ is given by the expression

$$f(n, j) = \left[1 + \frac{(Z\alpha)^2}{(n - \delta)^2}\right]^{-1/2}, \quad (69)$$

where n is the principal quantum number,

$$\delta = |\kappa| - [\kappa^2 - (Z\alpha)^2]^{1/2}, \quad (70)$$

and κ is the relativistic angular parameter of the Dirac equation. To obtain the value for the ground-state energy level of the electron in heavy hydrogen-like ions, the following must be added to the Lamb shifts listed in Table 2:

$$E_D = mf(n, j). \quad (71)$$

An overview of the current status of the theory for energy levels of the hydrogen atom can be found in [5]. The two-loop SE correction for $Z = 1$ given in Table 2 was calculated by taking into account the result obtained for $G^{\text{h.o.}}(1\alpha)$ in this paper. The error of the correction corresponds to that in Eq. (66). The nuclear size effect is given for two proton charge radii: r_p = 0.862(12) Fm, which was obtained in the original paper [26], and r_p = 0.895(18) Fm, which was obtained from a recent analysis of the available experimental data on electron–proton collisions [27]. Note that we do not use the more accurate value for the proton radius recommended in [5], since it is based largely on the comparison of the theoretical and experimental Lamb shifts in hydrogen. Note also that the QED corrections given in Table 2 for hydrogen include a part of the recoil effect that is taken into account by introducing the reduced mass in the corresponding formulas (for details, see [5]). The complete theoretical Lamb shifts for hydrogen contain two errors: the first corresponds to the uncertainty in the experimental proton charge radius, while the second is a purely theoretical error and originates from the two-loop SE correction. Since the first error is much larger than the second error, we can extract the proton charge radius from the comparison of the theoretical and experimental Lamb shifts. This method of determination yields

$$r_p = 0.877(8)\ \text{Fm}, \quad (72)$$

which is very close to the value of r_p = 0.875(7) recommended in [5]. The proton radius can be determined more accurately from a muonic hydrogen experiment [30].

Table 2 also gives the contributions from the individual corrections to the ground-state Lamb shift for several heavy hydrogen-like ions. A detailed analysis of each of these contributions can be found in our previous paper [13]. In this paper, we present updated results for the nuclear size effect obtained from the most recent data for the charge radii of nuclei and their errors [34].

ACKNOWLEDGMENTS

We thank K. Pachucki and U. Jentschura for helpful discussions. This work was supported by the INTAS (grant no. 03-55-1442), the Russian Foundation for Basic Research (project no. 04-02-17574), and the Dynasty Foundation. The calculations were performed in part at the French CINES and IDRIS Computational Centers.

REFERENCES

1. T. Stöhlker, P. H. Mokler, F. Bosch, *et al.*, Phys. Rev. Lett. **85**, 3109 (2000).
2. K. Pachucki and U. D. Jentschura, Phys. Rev. Lett. **91**, 113005 (2003).
3. K. Pachucki, Phys. Rev. A **63**, 042503 (2001).
4. U. D. Jentschura, Phys. Rev. A **70**, 052108 (2004).
5. P. J. Mohr and B. N. Taylor, Rev. Mod. Phys. **77**, 1 (2005).
6. A. Mitrushenkov, L. Labzowsky, I. Lindgren, *et al.*, Phys. Lett. A **200**, 51 (1995).
7. S. Mallampalli and J. Sapirstein, Phys. Rev. Lett. **80**, 5297 (1998).
8. V. A. Yerokhin, Phys. Rev. A **62**, 012508 (2000).
9. S. Mallampalli and J. Sapirstein, Phys. Rev. A **57**, 1548 (1998).
10. I. Goidenko, L. Labzowsky, A. Nefiodov, *et al.*, in *The Hydrogen Atom. Precision Physics of Simple Atomic Systems*, Ed. by S. G. Karshenboim *et al.* (Springer, Berlin, 2001), p. 619.
11. V. A. Yerokhin and V. M. Shabaev, Phys. Rev. A **64**, 062507 (2001).
12. V. A. Yerokhin, P. Indelicato, and V. M. Shabaev, Phys. Rev. Lett. **91**, 073001 (2003).
13. V. A. Yerokhin, P. Indelicato, and V. M. Shabaev, Eur. Phys. J. D **25**, 203 (2003).
14. V. M. Shabaev, Phys. Rep. **356**, 119 (2002).
15. R. Mills and N. Kroll, Phys. Rev. **98**, 1489 (1955).
16. C. Itzykson and J. B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980; Mir, Moscow, 1984).
17. A. Fox and D. R. Yennie, Ann. Phys. (N.Y.) **81**, 438 (1973).
18. N. J. Snyderman, Ann. Phys. (N.Y.) **211**, 43 (1991).

19. V. A. Yerokhin and V. M. Shabaev, Phys. Rev. A **60**, 800 (1999).
20. P. J. Mohr, Ann. Phys. (N.Y.) **88**, 26 (1974).
21. C. de Boor, *A Practical Guide to Splines* (Springer, New York, 1978).
22. V. M. Shabaev, I. I. Tupitsyn, V. A. Yerokhin, *et al.*, Phys. Rev. Lett. **93**, 130405 (2004).
23. P. J. Mohr, Phys. Rev. Lett. **34**, 1050 (1975).
24. E.-O. Le Bigot, U. D. Jentschura, P. J. Mohr, *et al.*, Phys. Rev. A **68**, 042101 (2003).
25. K. Pachucki, private communication (2004).
26. G. G. Simon, C. Schmidt, F. Borkowski, and V. H. Walther, Nucl. Phys. A **333**, 381 (1980).
27. I. Sick, Phys. Lett. B **576**, 62 (2003).
28. A. Huber, T. Udem, B. Gross, *et al.*, Phys. Rev. Lett. **80**, 468 (1998).
29. C. Schwob, L. Jozefowski, B. de Beauvoir, *et al.*, Phys. Rev. Lett. **82**, 4960 (1999).
30. J. P. Briand, P. Indelicato, A. Simionovici, *et al.*, Europhys. Lett. **9**, 225 (1989).
31. H. F. Beyer, D. Liesen, F. Bosch, *et al.*, Phys. Lett. A **184**, 435 (1994).
32. A. Gumberidze, T. Stöhlker, D. Banaś, *et al.*, Gesellschaft für Schwerionenforschung Scientific Report 2003, Ed. by U. Grundinger (Darmstadt, 2004), p. 99.
33. D. Taqqu, F. Biraben, C. A. N. Conde, *et al.*, Hyperfine Interact. **119**, 311 (1999).
34. I. Angeli, At. Data Nucl. Data Tables **87**, 185 (2004).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 294–298.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 337–343.

ORDER, DISORDER, AND PHASE TRANSITIONS IN CONDENSED SYSTEMS

Monte Carlo Simulation of a Random-Field Ising Antiferromagnet

V. V. Prudnikov and V. N. Borodikhin

Omsk State University, pr. Mira 55, Omsk, 644077 Russia

e-mail: prudnikv@univer.omsk.su

Received January 11, 2005

Abstract—Phase transitions in the three-dimensional diluted Ising antiferromagnet in an applied magnetic field are analyzed numerically. It is found that random magnetic field in a system with spin concentration below a certain threshold induces a crossover from second-order phase transition to first-order transition to a new phase characterized by a spin-glass ground state and metastable energy states at finite temperatures. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

Critical behavior of disordered systems with quenched disorder has been the subject of much theoretical and experimental interest, because the presence of quenched defects in most real solids modifies their thermodynamic characteristics, including critical behavior. It is well known that quenched disorder manifests itself by temperature fluctuations in ferro- and antiferromagnetic systems in the absence of external magnetic field or by magnetic field fluctuations in antiferromagnets in uniform magnetic field.

In the former case, quenched disorder affects the properties of only those homogeneous magnetic materials whose specific heat is divergent at the critical point [1]. Otherwise, the presence of defects does not change the critical behavior of magnets. This criterion applies only when the effective Hamiltonian near the critical point is isomorphic to the Ising model Hamiltonian. Disorder-induced critical behavior of the Ising model was analyzed in numerous recent studies [2]. For dilute Ising-like systems, it was found that theoretical calculations are in good agreement with experimental results and Monte Carlo simulations.

Despite extensive theoretical and experimental studies of random-field magnets conducted over the past twenty years [3], very few facts concerning their behavior have been established. In particular, the nature of phase transition in the random-field Ising model remains unclear, and the currently available theoretical results in this area disagree with experiment. The only theoretically proved fact is that the upper critical dimension for this phase transition is six (i.e., critical phenomena in systems of higher dimension are described by mean field theory) [3], whereas the critical dimension is four for homogeneous systems. While it had been argued that the lower critical dimension d_l can be both $d_l = 2$ [4] and $d_l = 3$ [5] (i.e., there is long-range order at finite temperatures if the system's dimension is higher), specialists came to the conclusion that $d_l = 2$ after the publication of [6, 7]. However, the nature of phase transition in the three-dimensional random-field Ising model remains unclear. According to [8, 9], it is a first-order phase transition even at very low random-field strengths; according to [10, 11], it is a second-order transition.

The effect of random fields on the behavior of magnetic systems is described by using two qualitatively equivalent models: the ferromagnetic random-field Ising model (RFIM) [12, 13] and the Ising diluted antiferromagnets in a field (DAFF) [14]. Real random-field magnets are antiferromagnets with quenched nonmagnetic impurities. Their behavior includes manifestations of both antiferromagnetic interaction between nearest neighbor atoms and ferromagnetic interaction between next-nearest neighbor atoms. The structure of an antiferromagnet can be represented as several interpenetrating ferromagnetic sublattices such that the total magnetization of the antiferromagnet is zero even though each ferromagnetic sublattice is magnetically ordered at a temperature below the Neél temperature. Examples of two-sublattice antiferromagnets are the following materials: NiO, MnO, Fe_2O_3, and MnF_2. Examples of random-field magnets include the uniaxial Ising-like antiferromagnets MnF_2 and FeF_2 diluted with zinc atoms in an external magnetic field [15].

2. MODEL

In this study, a Monte Carlo method is used to simulate the thermodynamic behavior of a diluted antiferromagnetic Ising model in an applied magnetic field on the simple cubic lattice by taking into account next-

nearest-neighbor interaction. The Hamiltonian of the model has the form

$$H = J_1 \sum_{ij} p_i p_j \sigma_i \sigma_j + J_2 \sum_{ik} p_i p_k \sigma_i \sigma_k + \mu h \sum_i p_i \sigma_i, \quad (1)$$

where $\sigma_i = \pm 1$ is the spin located at site i; μ is the Bohr magneton; $J_1 = 1$ and $J_2 = -1/2$ characterize antiferromagnetic nearest-neighbor and ferromagnetic next-nearest-neighbor exchange couplings, respectively; h is the strength of the uniform magnetic field; and p_i and p_j are random variables characterized by the distribution function

$$P(p_i) = p\delta(p_i - 1) + (1 - p)\delta(p_i), \quad (2)$$

which are introduced to describe quenched nonmagnetic impurity atoms vacancies distributed over the lattice and characterized by the concentration $c_{\text{imp}} = 1 - p$, where p is spin concentration. For $p = 1.0$, the model with competing interactions has been studied by Monte Carlo methods for over twenty years [16, 17]. The first study of effects of disorder on critical behavior based on this model was presented in [18]. For the DAFF mentioned above [13], competition between ferromagnetic order parameters was not taken into account. This model provides the most realistic physical representation. Since the strength of random-field effects is determined by impurity concentration and external field strength both in the model and in real magnets, the parameters of the model can be compared to those of real physical experiments on Ising diluted antiferromagnets. However, an analogous comparison of the random field with the impurity concentration in a sample and the applied field strength is difficult to perform for the ferromagnetic random-field Ising model (RFIM), which is most widely used in numerical simulations. Therefore, random field variation in RFIM cannot be quantitatively compared with structural disorder in real systems, which is shown here to be the key factor that controls phase transitions.

An antiferromagnet is characterized by the staggered magnetization M_{stg} defined as the difference of the magnetizations of the two sublattices, which plays the role of an order parameter. To determine the type of phase transition, we calculate the Binder cumulant [19]

$$U = \frac{1}{2}\left(3 - \frac{[\langle M_{\text{stg}}^4 \rangle]}{[\langle M_{\text{stg}}^2 \rangle]^2}\right), \quad (3)$$

where angle and square brackets denote statistical averaging and averaging over disorder realizations. The calculation of the cumulant is a good test for the order of phase transition: the cumulants plotted versus temperature have a distinct point of intersection in the case of second-order transition, whereas those corresponding to first-order phase transition have a characteristic shape and do not intersect.

We also examine spin-glass states. It is well known that spin glasses are characterized by transition to a phase with an infinite number of metastable states separated by potential barriers in the thermodynamic limit [20]. The complex magnetic ordering in such systems can be described in terms of the spin-glass order parameter

$$q_s = \frac{1}{pL^3}[\langle \sigma_i^\alpha \sigma_i^\beta \rangle], \quad (4)$$

where α and β refer to the spin configurations corresponding to replicas of the simulated disordered system characterized by equal temperatures, but different initial disorder realizations.

To obtain correct values of thermodynamic characteristics of critical behavior, both statistical averaging and averaging over disorder realizations must be performed only after the system has thermalized. Critical behavior of disordered systems is characterized by anomalously long relaxation times, which rapidly increase with the size of the simulated system. To reach equilibrium at near-critical temperatures and determine the corresponding thermodynamic characteristics, the system was quenched with a temperature step of $\Delta T = 0.1$ starting from a temperature at which no metastable states had been obtained in any sweep. At each temperature step, a relaxation regime was computed in 5000 steps and averaging was performed in 10000 steps by using the spin configuration obtained at the preceding step as an initial condition. This procedure was executed to obtain a stable equilibrium at each temperature and avoid metastable states [18].

For each lattice size L, thermodynamic characteristics were computed for constant h and p by ensemble-averaging the results of five sweeps executed for different initial spin configurations corresponding to a particular disorder realization and then averaging over 10 to 20 different disorder realizations.

3. RESULTS

We examined the temperature dependence of several thermodynamic characteristics of three-dimensional Ising antiferromagnets in a wide range of impurity concentrations for systems having a size varying from $L = 8$ to $L = 64$ in applied magnetic fields of a strength between $h = 1$ and $h = 4$.

Our analysis revealed several intervals of p corresponding to different behavior for each value of h. Second-order transition between paramagnetic and ferromagnetic phases is observed at $T_c(h, p)$ when $p_u < p < 1$ [18], where p_u is the vacancy percolation

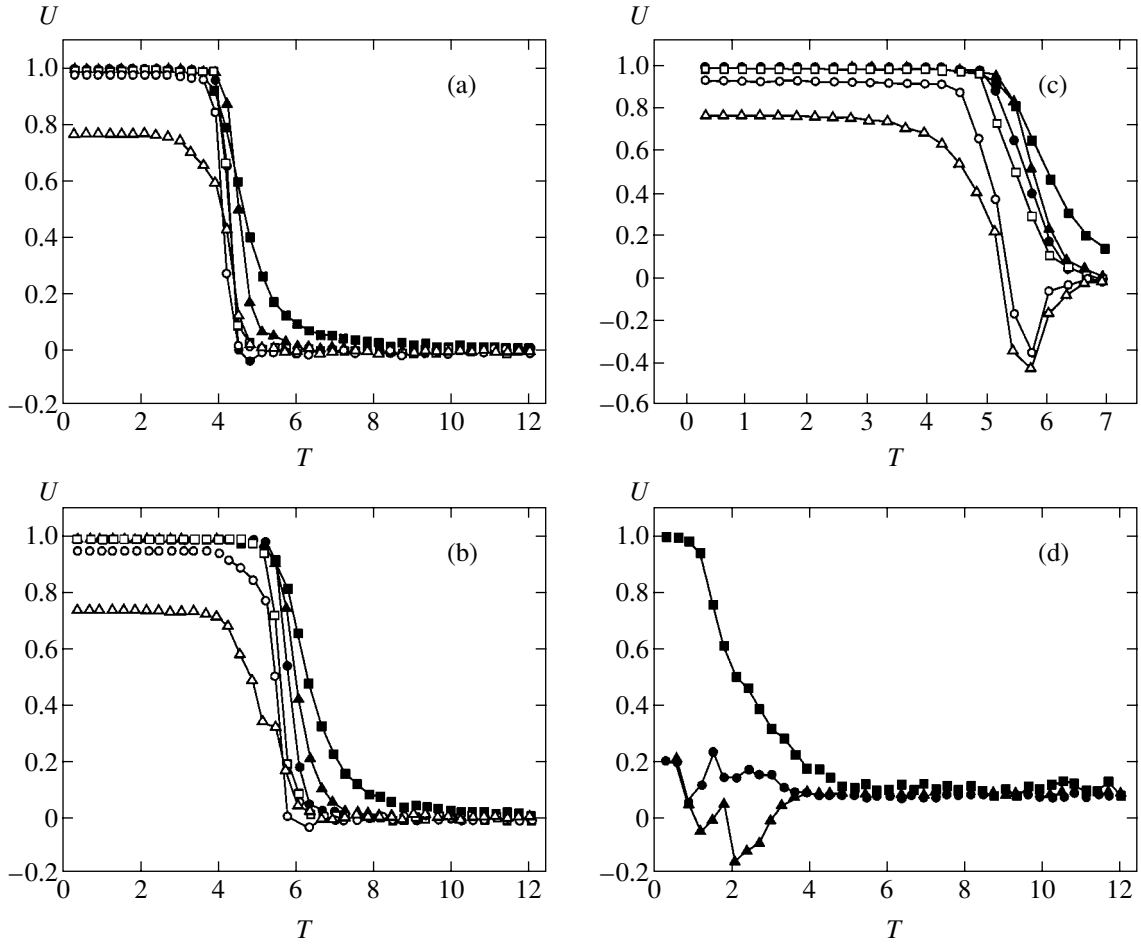

Fig. 1. Temperature dependence of the Binder cumulant on lattices with $L = 8$ (■), 16 (▲), 24 (●), 32 (□), 48 (○), and 64 (△): (a) $p = 0.5$, $h = 1$; (b) $p = 0.725$, $h = 3$; (c) $p = 0.8$, $h = 4$; (d) $p = 0.5$, $h = 3$.

threshold ($p_u = 0.83$ for the present model). When $p_c < p < p_u$, where p_c is the magnetic percolation threshold ($p_c = 0.17$ for the present model), there exist such $p(L', h)$ that the computed quantities exhibit behavior characteristic of second- and first-order phase transition if $p > p(L', h)$ and $p < p(L', h)$, respectively, on lattices with $L < L'$. The value of $p(L', h)$ increases with h and L', approaching the threshold $p_u = 0.83$.

This size-dependent behavior is explained by the existence of interpenetrating spin and vacancy clusters whose fractal dimensions vary between 0 and 3, depending on spin concentration. Therefore, the size-dependent parameterization of transition from long-range order to domain structure with characteristic size L_c by

$$\frac{h_r}{J(L)} = \frac{h_r}{JL^{(2-d_f)/2}}, \quad L_c \approx \left(\frac{J}{h_r}\right)^{2/(2-d_f)} \qquad (5)$$

proposed for Ising-like systems in [21], where h_r is the random-field amplitude, J is the exchange coupling, and d_f is interpreted as the fractal dimension of the spin cluster, can be used to predict that antiferromagnetic long-range order breaks down at $d_f < 2$.

Figures 1–4 illustrate the existence of boundaries separating spin-concentration intervals characterized by different strength of random-field effects for systems with $L \le 64$ in applied magnetic fields of strength between $h = 1$ and $h = 4$.

Figure 1 shows the temperature-dependent Binder cumulants calculated for several lattices with $p = 0.5$ for $h = 1$, with $p = 0.5$ and 0.725 for $h = 3$, and with $p = 0.8$ for $h = 4$. For spin concentrations close to p_u, the Binder cumulants do not intersect only if $L \ge 64$. When $p = 0.5$ and $h = 3$, no intersection of Binder cumulants is observed for lattices of all sizes used in the computations. Comparing Figs. 1a–1c, we see that the size-dependent change in the behavior of Binder cumulants due to the increase in field strength from $h = 1$ to $h = 4$ (increasing random-field effects) corresponds to the spin concentration increasing from $p = 0.5$ to $p = 0.8$.

For systems with $p < p(L', h)$, the behavior of $M_{stg}(T)$ (Fig. 2) strongly depends on the lattice size for all val-

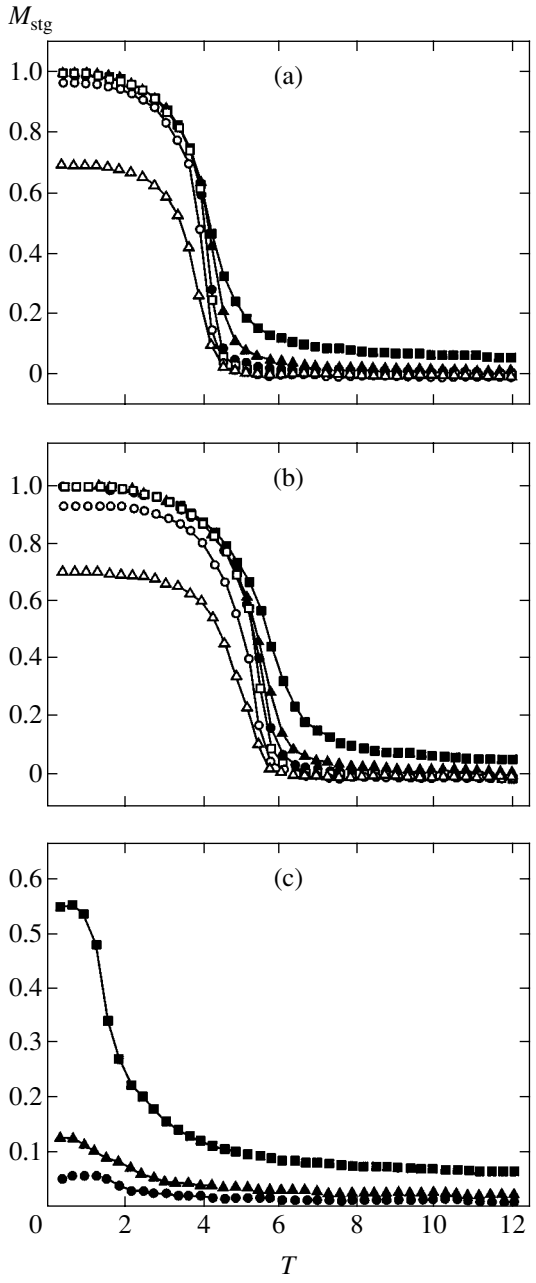

Fig. 2. Temperature dependence of staggered magnetization on lattices with $L = 8$ (■), 16 (▲), 24 (●), 32 (□), 48 (○), and 64 (△): (a) $p = 0.5$, $h = 1$; (b) $p = 0.725$, $h = 3$; (c) $p = 0.5$, $h = 3$.

ues of h used in the computations. The decrease in staggered magnetization with increasing L points to the absence of an antiferromagnetic ground state. Furthermore, the insignificant increase in total magnetization M with increasing L (Fig. 3a) indicates that the system breaks up into antiferromagnetic domains of size $L < L'$ with nearly compensated magnetizations. As the random-field effects increase with impurity concentration and applied magnetic field, both number and size of antiferromagnetic domains increases (Fig. 2c) and both

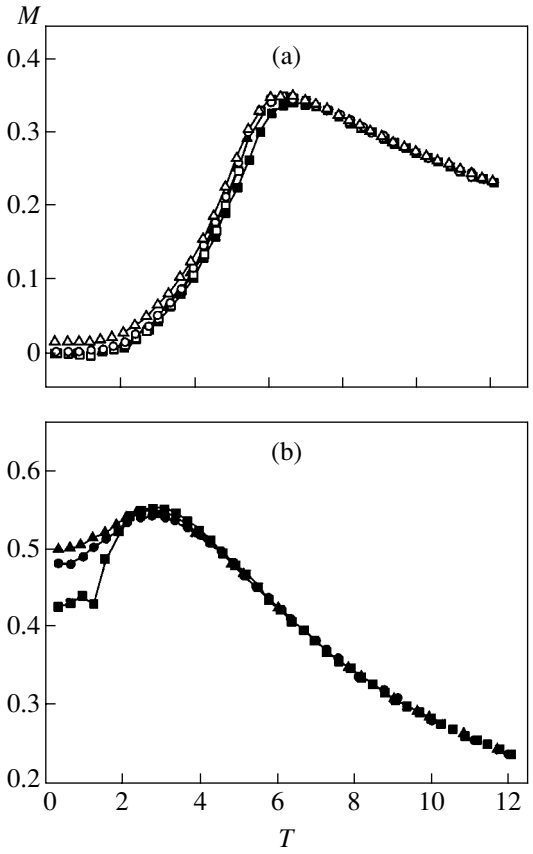

Fig. 3. Temperature dependence of total magnetization on lattices with $L = 8$ (■), 16 (▲]), 24 (●), 32 (□), 48 (○), and 64 (△): (a) $p = 0.725$, $h = 3$; (b) $p = 0.5$, $h = 3$.

number and size of ferromagnetic ones increases (Fig. 3b), while it holds that $M_{stg} + M < 1$.

To further elucidate the properties of systems with $p_c < p < p_u$, we examined the temperature dependence of the spin-glass order parameter. The results obtained for several disorder realizations are shown in Fig. 4. The graphs demonstrate that a spin-glass phase with "frozen" configuration of magnetic moments is obtained as temperature approaches zero. Thus, a random magnetic field induces transition from antiferromagnetic to spin-glass ground state in the Ising model with competing interactions when $p < p_u$. At finite temperatures, the corresponding change in the state of a disordered system is a first-order transition from a paramagnetic to a mixed phase. When the spin concentration is high, the latter consists of antiferromagnetic domains separated by spin-glass regions. With decreasing spin concentration, the number and size of antiferromagnetic domains decrease and the number and size of ferromagnetic domains increases, while the volume fraction occupied by the spin-glass phase decreases.

We used the temperature and field dependence of magnetization, internal energy, and specific heat to calculate the first-order phase transition lines. The T–p

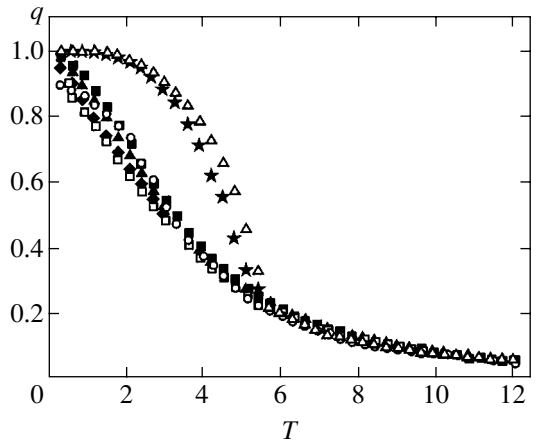

Fig. 4. Temperature dependence of spin-glass order parameter: $h = 3$; $L = 24$; $p = 0.2$ (■), 0.3 (▲), 0.4 (◆), 0.5 (□), 0.6 (○), 0.7 (★), and 0.725 (△).

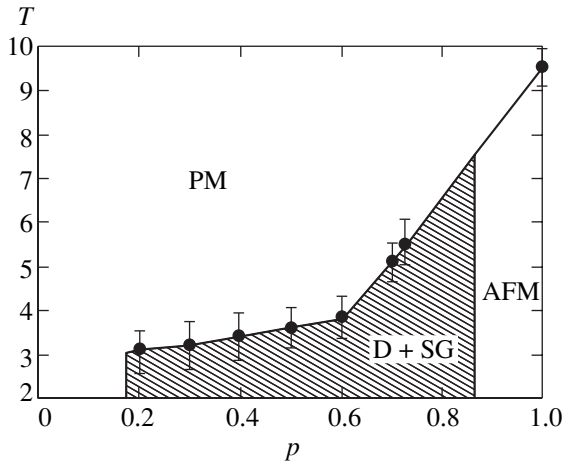

Fig. 5. Phase diagram for random-field Ising antiferromagnet at $h = 3$: PM = paramagnet; AFM = antiferromagnet; D+SG = domain structure and spin glass.

phase diagram shown in Fig. 5 summarizes the results obtained for $h = 3$.

4. CONCLUSIONS

The Monte Carlo simulations of thermodynamics of the three-dimensional random-field Ising model performed in this study demonstrate second-order phase transition from paramagnetic to antiferromagnetic state when the spin concentration is higher than p_u and first-order phase transition from paramagnetic to mixed phase consisting of antiferromagnetic and ferromagnetic domains separated by spin-glass domains when $p_c < p < p_u$, where p_u and p_c are vacancy and magnetic percolation thresholds, respectively. When the spin concentration is high, the system consists of antiferromagnetic domains separated by spin-glass regions. With decreasing spin concentration or increasing applied magnetic field strength, both the number and size of antiferromagnetic domains decrease, both the number and size of ferromagnetic domains increase, and the volume fraction of the spin-glass phase decreases. It is shown that random magnetic field induces a transition from antiferromagnetic to spin-glass ground state when $p_c < p < p_u$ in the three-dimensional random-field Ising model with competing interactions analyzed in this study.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 04-02-17524 and 04-02-39000) and by the Ministry of Education of the Russian Federation (grant no. UR 01.01.230).

REFERENCES

1. A. B. Harris, J. Phys. C **7**, 1671 (1993).
2. R. Folk, Y. Holovatch, and T. Yavorskiĭ, Usp. Fiz. Nauk **173**, 175 (2003) [Phys. Usp. **46**, 169 (2003)].
3. V. S. Dotsenko, Usp. Fiz. Nauk **165**, 481 (1995) [Phys. Usp. **38**, 457 (1995)].
4. Y. Imry and S.-K. Ma, Phys. Rev. Lett. **35**, 1399 (1975).
5. G. Parisi and N. Sourlas, Phys. Rev. Lett. **43**, 744 (1979).
6. J. Z. Imbrie, Phys. Rev. Lett. **53**, 1747 (1984).
7. J. Bricmont and A. Kupiainen, Phys. Rev. Lett. **59**, 1829 (1987).
8. A. P. Young and M. Nauenberg, Phys. Rev. Lett. **54**, 2429 (1985).
9. H. Rieger and A. P. Young, J. Phys. A **26**, 5279 (1993).
10. A. T. Ogielski and D. A. Huse, Phys. Rev. Lett. **56**, 1298 (1986).
11. A. T. Ogielski, Phys. Rev. Lett. **57**, 1251 (1986).
12. J. Cardy, Phys. Rev. B **29**, 505 (1984).
13. D. P. Belanger and A. P. Young, J. Magn. Magn. Mater. **100**, 272 (1991).
14. G. S. Grest, C. M. Soukoulis, and K. Levin, Phys. Rev. B **33**, 7659 (1986).
15. F. Ye, L. Zhou, S. Larochelle, *et al.*, Phys. Rev. Lett. **89**, 157 202 (2002).
16. D. P. Landau, Phys. Rev. Lett. **28**, 449 (1972).
17. H. Muller-Krumbhaar and D. P. Landau, Phys. Rev. B **14**, 2014 (1976).
18. V. V. Prudnikov, O. N. Markov, and E. V. Osintsev, Zh. Éksp. Teor. Fiz. **116**, 953 (1999) [JETP **89**, 508 (1999)].
19. K. Binder, Z. Phys. B **43**, 119 (1981).
20. V. S. Dotsenko, Usp. Fiz. Nauk **163** (6), 1 (1993) [Phys. Usp. **36**, 455 (1993)].
21. E. Pytte, Y. Imry, and D. Mukamel, Phys. Rev. Lett. **46**, 1173 (1981).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 299–304.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 344–350.

ORDER, DISORDER, AND PHASE TRANSITIONS IN CONDENSED SYSTEMS

Dynamic Critical Behavior in Models of Ferromagnetic Gadolinium

A. K. Murtazaev* and V. A. Mutailamov

Institute of Physics, Dagestan Scientific Center, Russian Academy of Sciences, Makhachkala, 367003 Russia

**e-mail: m_akai@iwt.ru*

Received February 4, 2005

Abstract—A numerical technique combining Monte Carlo and molecular dynamics simulations is used for the first time to examine the complex critical dynamics of models of ferromagnetic gadolinium in which both strong exchange interactions and relativistic effects of several different types are taken into account. A finite-size scaling technique is used to calculate the corresponding dynamic critical exponents. The role played by isotropic dipole–dipole interaction in the critical behavior of gadolinium is evaluated. The results obtained provide an explanation for the anomalous dynamic critical behavior of gadolinium. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

Analysis of dynamic critical properties is a major problem in statistical physics and theory of phase transitions [1–3]. The substantial progress achieved to date in this field became possible mainly through theoretical and experimental studies. Nevertheless, the development of a rigorous and consistent theory of dynamic critical phenomena based on microscopic Hamiltonians remains a challenging problem in the modern theory of phase transitions and critical phenomena [1, 4].

Current knowledge of critical dynamics has mostly been gained in mode-coupling and dynamic scaling theories [1–3], which were developed independently on the basis of essentially different ideas. However, many predictions of these theories are in good agreement. In [3], universality classes of dynamic critical behavior were proposed in the framework of the dynamic scaling theory. It was shown that the dynamic universality classes depend not only on the space dimension d, the number n of order-parameter components, the interaction range, and the symmetry of the Hamiltonian, but also on the energy and order-parameter conservation laws [3]. Classification has played an important role in understanding critical dynamics and is being used to this day. The underlying dynamic scaling hypothesis provides an adequate characterization of dynamic critical behavior. However, the classification is not flawless: its basic ideas are not supported by solid evidence, even though the ensuing predictions are consistent with many, but not all, experimental findings. Moreover, the classification is not complete.

The critical dynamics of magnetically ordered crystals, especially ferromagnets, is characterized by a great diversity and complexity due to the importance of both exchange interaction and relatively weak relativistic effects (such as anisotropy and dipole–dipole interaction). The most essential factor of the latter kind is dipole–dipole interaction, which plays an increasingly important role as the critical point is approached. Note that the aforementioned classification of universal dynamic critical behavior does not allow for any effect due to dipole–dipole interaction. Further analyses (see [5–7]) have shown that theories taking into account dipole–dipole interaction predict dynamics of two types, normal and stiff, each characterized by a specific set of critical exponents.

There is no conclusive experimental evidence, because the available data are mutually inconsistent [4]. Experimental validation is further complicated by the fact that exchange, anisotropy, and dipole–dipole interaction can contribute to the behavior of a real material simultaneously, in which case critical dynamics obviously depend on their relative strengths. Moreover, since there exist crossover regions near the critical point, the actual critical behavior may change as the critical point is approached. The real behavior due to relativistic effects is even more diverse: in particular, anisotropy can be uniaxial, cubic, etc., while dipole–dipole interaction can be either isotropic or anisotropic. Simultaneous influence of these factors may partially explain the inconsistency of experimental data on the dynamic critical properties of magnetically ordered materials.

It is clear that experimental studies can hardly elucidate the current discrepancy between observations and predictions, because high-precision measurement of critical parameters is a very difficult task. Moreover, since almost any experimental result is due to the combined effects of all factors, their individual strengths and contributions are practically impossible to single out. The problem is also unlikely to be amenable to rig-

orous theoretical analysis in view of enormous mathematical difficulties.

In recent studies, methods of computational physics have played an increasingly important role in dealing with complex issues of this kind. Highly accurate and reliable calculations of critical parameters can be performed by applying these methods, at least, in studies of static critical behavior [8]. In particular, the advantages of Monte Carlo and molecular dynamics simulations include not only rigorous mathematical foundations and error control within their respective frameworks, but also the possibility of evaluating the relative importance of individual parameters.

The key characteristics of critical dynamics are the critical exponent w for the relaxation time τ and the dynamic critical exponent z:

$$\tau \propto |\varepsilon|^{-w}, \quad \tau \propto \xi^{z},$$

where

$$\varepsilon = |T - T_c|/T_c, \quad \xi = (T/T_c - 1)^{-\nu}.$$

In the mid-1990s, a new technique was proposed for calculating the dynamic critical exponent z, based on dynamic finite-size scaling theory [9] and a special procedure for determining the characteristic frequency ω_c [10–12].

In this paper, methods of computational physics are used to analyze critical dynamics in models of ferromagnetic gadolinium. The key questions to be answered here can be formulated as follows.

I. How does isotropic dipole–dipole interaction affect dynamic critical behavior?

II. Is there any difference in critical dynamics between different directions in noncubic crystals?

III. Can the computational technique employed in this study be used to expose the dependence of critical dynamics on weak factors, such as dipole–dipole interaction?

The present analysis of models of gadolinium is motivated by the following observations.

1. The static critical behavior of gadolinium is strongly affected by isotropic dipole–dipole interaction [13, 14].

2. The controversial results obtained in extensive experimental studies of the critical dynamics of ferromagnetic gadolinium [4, 15–18] do not provide sufficient basis for any final conclusion.

3. The complex dynamic critical behavior of gadolinium has been analyzed in a number of theoretical studies [19, 20].

4. The dynamic critical behavior of gadolinium is of interest per se, because it is determined by the combined effects of three factors: exchange interaction, magnetocrystalline anisotropy, and isotropic dipole–dipole interaction.

5. The static critical behavior of gadolinium has been thoroughly examined in both experimental [13] and numerical [14] studies, and their results provide a good basis for an analysis of critical dynamics.

Note that the technique employed in the present analysis of critical dynamics was previously used in our studies of the dynamic critical behavior of the classical Heisenberg model [21] and models of the multisublattice antiferromagnet Cr_2O_3 [21, 22].

2. MODEL

Gadolinium is a rare-earth metal with a hexagonal close-packed lattice. The Curie temperature of the order–disorder second-order phase transition between ferromagnetic and paramagnetic states in gadolinium is $T_C = 293$ K. In a previous study, we proposed models of ferromagnetic gadolinium that take into account all basic properties of these materials [14] and examined their static behavior. A model Hamiltonian for gadolinium can be represented as follows [14]:

$$H = -\frac{1}{2}\sum_{i,j} J(\mathbf{S}_i \cdot \mathbf{S}_j) - D\sum_i (\mathbf{S}_i^z)^2 - D_{\text{dip}}\sum_i \langle \mathbf{M} \rangle \cdot \mathbf{S}_i, \quad |\mathbf{S}_i| = 1. \tag{1}$$

Here, the first term represents exchange interaction between a gadolinium ion and its 12 nearest neighbors, with $J > 0$; the second term, easy-axis anisotropy along the hexagonal axis, characterized by an anisotropy constant D; the third one, isotropic dipole–dipole interaction with coupling constant D_{dip}; $\mathbf{M}$ is magnetization. We considered two models: in model I, only exchange interaction and anisotropy were taken into account; in model II, the dipole–dipole interaction was also included.

According to laboratory experiments [13], $D/J = 1.41 \times 10^{-4}$ and $D_{\text{dip}}/J = 1.35 \times 10^{-2}$. The coordinate system was defined so that the z axis was aligned with the anisotropy axis, i.e., with the hexagonal axis of the crystal. The values of critical temperature were obtained in Monte Carlo simulations of the static critical behavior of these models [14].

3. TECHNIQUE

Our approach relies on dynamic scaling theory [1] and calculation of space- and time-displaced spin–spin

correlation functions

$$C^k(\mathbf{r}_{12}, t) = \langle S^k_{\mathbf{r}_1}(t) S^k_{\mathbf{r}_2}(0)\rangle - \langle S^k_{\mathbf{r}_1}(t)\rangle \langle S^k_{\mathbf{r}_2}(0)\rangle, \quad (2)$$

where $\mathbf{r}_{1,2} = \mathbf{r}_1 - \mathbf{r}_2$, $S^k_{\mathbf{r}_1}(t)$ ($k = x, y, z$) is the spin at the site $\mathbf{r}_1$ at an instant t, $S^k_{\mathbf{r}_2}(0)$ is the spin at the site $\mathbf{r}_2$ at the initial moment ($t = 0$), and angle brackets denote ensemble averaging. At $T \approx T_C$ and sufficiently long times, the second term on the right-hand side of (2) can be neglected [11].

The spacetime Fourier transform of (2) defines the dynamic structure factor

$$S^k(\mathbf{q}, \omega) = \int d\mathbf{r} \int_{-\infty}^{\infty} C^k(\mathbf{r}, t) \exp[-i(\mathbf{q}\cdot\mathbf{r} - \omega t)] dt, \quad (3)$$

where $\mathbf{q}$ is the wavevector and ω is frequency. In the general case, the dynamic structure factor is proportional to the directly measurable neutron scattering function, being shifted by a constant frequency [1].

In accordance with the dynamic scaling hypothesis, the characteristic frequency is the median frequency defined by the relation

$$\int_{-\omega_c(\mathbf{q},\xi)}^{\omega_c(\mathbf{q},\xi)} S^k(\mathbf{q}, \omega) d\omega = \frac{1}{2}\int_{-\infty}^{\infty} S^k(\mathbf{q}, \omega) d\omega. \quad (4)$$

In the general case, the characteristic frequency depends on wavevector $\mathbf{q}$ and correlation length ξ. In dynamic scaling theory [1], it is postulated that

$$\omega_c(\mathbf{q}, \xi) = q^z \Omega(q\xi), \quad (5)$$

where Ω is an unknown homogeneous function of $q\xi$.

In the models considered here, expression (3) for a system of size L simulated at the critical point over a finite time interval t_{cutoff} is represented as [10]

$$S^k(\mathbf{q}, \omega) = \frac{1}{\sqrt{2\pi}} \sum_{\mathbf{r}_1, \mathbf{r}_2} \exp[i\mathbf{q}\cdot(\mathbf{r}_1 - \mathbf{r}_2)] \times \int_{-t_{\text{cutoff}}}^{t_{\text{cutoff}}} \exp(i\omega t) C^k(\mathbf{r}_1 - \mathbf{r}_2, t) dt, \quad (6)$$

and Eq. (5) becomes

$$\omega_c \propto L^{-z} \Omega'(qL), \quad (7)$$

where $q = 2\pi m/L$ ($m = \pm 1, \pm 2, \ldots, \pm L$). The sum in (6) corresponds to an integral over the space discretized into a lattice. Relation (7) is used to evaluate z in practical calculations. The function Ω' is not known, but it is known to depend only on the product of q and L. Therefore, the dynamic critical exponent can be evaluated by keeping it constant while varying the lattice size.

In each case considered in this study, the wavevector was aligned with a certain crystallographic axis, averaging over spins lying in alternating planes perpendicular to this axis was performed, and the averaged spins were used to calculate spin–spin correlation functions. In the case of a simple cubic lattice, these are crystallographic planes. The geometry corresponding to the hexagonal lattice of gadolinium is more complicated.

The correlation functions were computed by using the system of differential equations of spin dynamics:

$$\frac{\partial \mathbf{S}_i}{\partial t} = \mathbf{S}_i \times \mathbf{h}_{\text{loc}}, \quad t = t' \frac{g\mu_B}{J\gamma}, \quad |\mathbf{S}_i| = 1, \quad (8)$$

where g is the Landé factor, γ is the gyromagnetic ratio, and $\mathbf{h}_{\text{loc}}$ is the local field defined by (1). Before solving system (8), the standard Metropolis algorithm is executed to thermalize the system at the critical temperature [23].

We simulated systems with L and number N of magnetic sites varied from 4 to 18 and from 128 to 11664, respectively, using periodic boundary conditions and calculating the three spin components for $\mathbf{q}$ aligned with three crystallographic axis. Note that the distances between the atomic planes are not equal when the wavevector is aligned with the *a* or *b* axis, whereas they are equal to $c/2$ in the case of alignment with the *c* axis. The longest simulation time was $t_{\text{cutoff}} = 130$ (in arbitrary units), and the time step in computing the differential equations was $\Delta t = 0.01$. System (8) was solved by the fourth-order accurate Runge–Kutta method. Correlation functions (2) were computed by performing up to $n = 350$ averaging operations. It was found that the use of $\Delta t = 0.005$ and $n = 700$ did not lead to any significant change in numerical results.

The spin dynamics equations were solved by using several (up to ten) initial equilibrium configurations for each lattice size L, and the resulting characteristic frequencies ω_c were averaged.

Correlation functions (2) were computed without using any normalization condition.

4. RESULTS

Figure 1 shows the dynamic structure factor $S^k(\mathbf{q}, \omega)$ calculated as a function of frequency in model I for systems with different N by using the x components of spins in the case when the wavevector is aligned with the *a* axis. The figure demonstrates that the maximum of the dynamic structure factor shifts toward smaller values of ω with increasing number of magnetic sites,

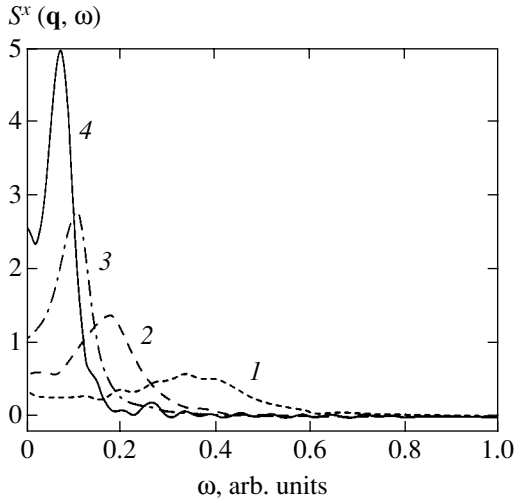

Fig. 1. Dynamic structure factor versus frequency in model I with $qL = 2\pi$ and $N = 432$ (*1*), 1024 (*2*), 2000 (*3*), and 3456 (*4*).

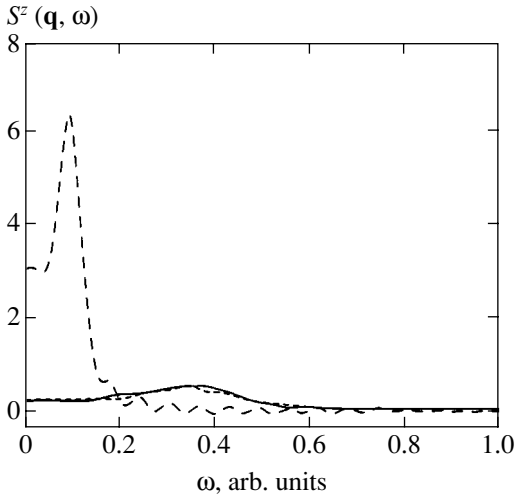

Fig. 2. Dynamic structure factor versus frequency in model I with $N = 432$ and $qL = 2\pi$ for wavevector aligned with the crystallographic axes *a* (short-dash curve), *b* (solid curve), and *c* (long-dash curve).

while the absolute value of the maximum increases. The latter effect should be explained by the fact that correlation functions (2) were not normalized. However, the characteristic frequency is determined by the shape of the dynamic structure factor rather than its absolute value.

Figure 2 shows the dynamic structure factors calculated for different wavevector directions in model I by using the *z* components of spins. Expectably, the shapes of the structure factors calculated for wavevectors aligned with the *a* and *b* axes are nearly identical and different from those calculated in the case when the wavevector **q** is aligned with the hexagonal *c* axis.

Figure 3 illustrates the dependence of ω_c on L in model I computed for different wavevector alignments by using the *x* components of spins and averaging the characteristic frequency over several initial configurations. These results demonstrate similarity between the structure factors corresponding to characteristic frequencies ω_c for **q** aligned with the *a* and *b* axes and their substantial difference from those for **q** aligned with the *c* axis.

The results obtained for model II are qualitatively similar to those presented in Figs. 1–3, but exhibit some quantitative differences.

The table summarizes the dynamic critical exponents *z* evaluated by using the three spin components in both models and substituting the results illustrated by Fig. 3 into (7) for **q** aligned with three crystallographic axes.

In model I, the values of *z* calculated by using the *x* and *y* spin components for $\mathbf{q} \parallel a$ and $\mathbf{q} \parallel b$ are in good agreement with theoretical predictions for anisotropic magnetic systems ($z = 2$, model *A* in [3]). However, the values of the dynamic critical exponent determined by using the *z* components for $\mathbf{q} \parallel a$ and for $\mathbf{q} \parallel b$ lie between those predicted theoretically for anisotropic magnetic systems ($z = 2$, model *A*) and isotropic ferromagnets ($z = 2.5$, model *J* in [3]). Note that the values of *z* calculated by using all three spin components lie between $z = 2$ (model *A*) and $z = 2.5$ (model *J*).

The values of *z* calculated by using the *x* and *y* spin components in model II are higher by approximately 0.2, and those calculated by using the *z* spin components are higher by amounts varying from 0.05 to 0.2.

It is obvious that the difference between the values of *z* predicted in models I and II is due solely to the contribution of dipole–dipole interaction in model II, because all calculations were otherwise similar. Note

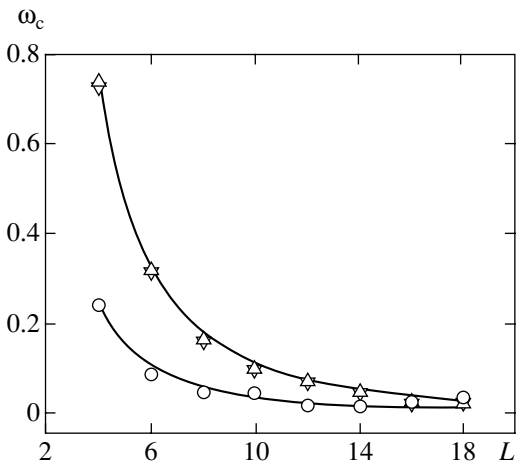

Fig. 3. Characteristic frequency versus lattice size in model I with $qL = 2\pi$ for wavevector aligned with the crystallographic axes *a* (△), *b* (▽), and *c* (○).

Dynamic critical exponent z for three different wavevector directions and three spin projections in models of ferromagnetic gadolinium

	$\mathbf{q} \parallel a$	$\mathbf{q} \parallel b$	$\mathbf{q} \parallel c$
Model I			
$k = x$	2.07 ± 0.06	2.11 ± 0.06	2.28 ± 0.06
$k = y$	2.06 ± 0.06	2.08 ± 0.06	2.36 ± 0.06
$k = z$	2.30 ± 0.06	2.24 ± 0.06	2.37 ± 0.06
Model II			
$k = x$	2.29 ± 0.06	2.25 ± 0.06	2.47 ± 0.06
$k = y$	2.26 ± 0.06	2.27 ± 0.06	2.49 ± 0.06
$k = z$	2.35 ± 0.06	2.35 ± 0.06	2.54 ± 0.06

also that all calculated critical exponents z have reasonable values.

It is obvious that some of our results are impossible to explain within the framework of the scheme proposed in [3], and so are some results obtained in experimental studies of critical dynamics of gadolinium [4, 15–20]. As noted in the Introduction, the classification of universal critical behavior proposed in [3] ignores certain additional factors, such as dipole–dipole interaction. Therefore, some new universality classes may have to be introduced. Even more complex behavior must be observed when several interactions of different types contribute simultaneously to critical dynamics, in addition to strong exchange interactions. This behavior is actually observed both in real materials and in our model II.

Furthermore, the dynamic critical behavior of a noncubic crystal must depend on direction, as suggested by the data listed in the table. This conjecture is corroborated by the results presented in [20], where it was demonstrated that longitudinal and transverse scaling functions analogous to Ω in (5) exhibit totally different behavior depending on temperature (or crossover) region. Several crossover scenarios depending on temperature regime, anisotropy, and dipole–dipole interaction were described in [20] for Heisenberg models. In that study, the effects due to anisotropy and dipole–dipole interaction were treated separately because of enormous theoretical difficulties. In our model II, these effects are taken into account simultaneously.

We believe that many special characteristics exposed here, as well as the anomalous dynamic critical behavior of real gadolinium, can be explained by the combined effects of all factors. Note that a new dynamic universality class, model J^*, has been proposed for gadolinium in [19].

5. CONCLUSIONS

A numerical approach combining Monte Carlo and molecular dynamics simulations is used to examine the critical dynamics of real ferromagnetic gadolinium.

1. Dynamic critical behavior of noncubic crystals depends on the direction of wavevector relative to crystallographic axes.

2. Isotropic dipole–dipole interaction strongly contributes to the critical dynamics of magnetic systems.

3. The critical behavior of gadolinium that is difficult to explain theoretically should be attributed to the effects of anisotropy and dipole–dipole interaction combined with strong exchange interaction.

4. It is obvious that the developing application of methods of computational physics to dynamic critical phenomena not only leads to valuable results, but also reveals subtle effects, such as those due to dipole–dipole interaction.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 04-02-16487; under the State Program for Support of Leading Science Schools, grant 2253.2003.2; under the Federal Program “Integratsiya,” grant I0228; and by the Foundation for Support of Domestic Science (A.K. Murtazaev).

REFERENCES

1. H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Clarendon, Oxford, 1971; Mir, Moscow, 1973).
2. S. Ma, *Modern Theory of Critical Phenomena* (Benjamin, Reading, Mass., 1976; Mir, Moscow, 1980).
3. P. C. Hohenberg and B. I. Halperin, Rev. Mod. Phys. **49**, 435 (1977).
4. I. K. Kamilov and Kh. K. Aliev, Usp. Fiz. Nauk **168**, 953 (1998) [Phys. Usp. **41**, 865 (1998)].
5. S. V. Maleev, Preprint No. 1038, LIYaF AN SSSR (Inst. of Nuclear Physics, USSR Academy of Sciences, Leningrad, 1985).
6. S. V. Maleev, Preprint No. 1039, LIYaF AN SSSR (Inst. of Nuclear Physics, USSR Academy of Sciences, Leningrad, 1985).
7. S. V. Maleev, Zh. Éksp. Teor. Fiz. **66**, 1809 (1974) [Sov. Phys. JETP **39**, 889 (1974)].
8. I. K. Kamilov, A. K. Murtazaev, and Kh. K. Aliev, Usp. Fiz. Nauk **169**, 773 (1999) [Phys. Usp. **42**, 689 (1999)].
9. M. Suzuki, Prog. Theor. Phys. **58**, 1142 (1977).
10. K. Chen and D. P. Landau, Phys. Rev. B **49**, 3266 (1994).
11. D. P. Landau, Physica A (Amsterdam) **205**, 41 (1994).
12. D. P. Landau and M. Krech, J. Phys.: Condens. Matter **11**, R179 (1999).

13. Kh. K. Aliev, I. K. Kamilov, and O. M. Omarov, Zh. Éksp. Teor. Fiz. **94** (11), 153 (1988) [Sov. Phys. JETP **67**, 2262 (1988)].

14. A. K. Murtazaev, I. K. Kamilov, and M. A. Magomedov, Zh. Éksp. Teor. Fiz. **120**, 1535 (2001) [JETP **93**, 1330 (2001)].

15. Kh. K. Aliev, I. K. Kamilov, Kh. I. Magomedgadzhiev, and M.-G. K. Omarov, Zh. Éksp. Teor. Fiz. **95**, 1896 (1989) [Sov. Phys. JETP **68**, 1096 (1989)].

16. A. R. Chowdhury, G. S. Collins, and C. Hohenemser, Phys. Rev. B **30**, 6277 (1984).

17. G. S. Collins, A. R. Chowdhury, and C. Hohenemser, Phys. Rev. B **33**, 4747 (1986).

18. A. R. Chowdhury, G. S. Collins, and C. Hohenemser, Phys. Rev. B **33**, 5070 (1986).

19. E. Frey, F. Schwabl, S. Hennenberg, *et al.*, Phys. Rev. Lett. **79**, 5142 (1997).

20. S. Hennenberg, E. Frey, P. G. Maier, *et al.*, Phys. Rev. B **60**, 9630 (1999).

21. A. K. Murtazaev, V. A. Mutailamov, I. K. Kamilov, *et al.*, J. Magn. Magn. Mater. **258**–**259**, 48 (2003).

22. A. K. Murtazaev, V. A. Mutaĭlamov, and K. Sh. Khizriev, Izv. Ross. Akad. Nauk, Ser. Fiz. **68**, 734 (2004).

23. *Monte Carlo Methods in Statistical Physics*, Ed. by K. Binder (Springer, Berlin, 1979; Mir, Moscow, 1982).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 305–316.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 351–363.

ORDER, DISORDER, AND PHASE TRANSITIONS IN CONDENSED SYSTEMS

Short-Range Ferromagnetism and Transport Properties of Amorphous $(Gd, Y)_xSi_{1-x}$ Alloys¶

S. Caprara[a], V. V. Tugushev[b], and N. K. Chumakov[b]

[a]*Dipartimento di Fisica, Università di Roma "La Sapienza," Istituto Nazionale per la Fisica della Materia, SMC and UdR di Roma 1, 00185 Rome, Italy*

[b]*Russian Research Centre Kurchatov Institute, Moscow, 123182 Russia*

e-mail: vvtugushev@mail.ru; chumakov@imp.kiae.ru

Received January 31, 2005

Abstract—We present a theoretical description and electrical conductivity measurements for amorphous $(Gd, Y)_xSi_{1-x}$ alloys with $0.1 < x < 0.2$. In our model, we take into account the strong topological disorder in the system, causing the appearance of regions with higher electron density (electron "drops") around nanoscale structural defects enriched with rare-earth ions ("clusters"). We calculate the local density of electron states in the drops and in the matrix and establish the criterion for local instability to ferromagnetism. In the framework of the "local phase transition" approach, we find that short-range ferromagnetic order is more favorable inside the drops than in the matrix and exists in a wide temperature range. We analyze recent measurements of the temperature and magnetic-field dependence of the electrical conductivity in these systems and show that the spin polarization of the electron states in the drops enhances the tendency towards the metal–insulator transition. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

The anomalous transport and magnetic properties of amorphous (a-) RE_xSi_{1-x} alloys (with RE = Gd, Tb, Y, and $0.1 < x < 0.2$) have been the object of controversial debate in recent years. The standard approach to these systems, described as disordered magnetic semiconductors, is unable to account for various peculiarities in a wide range of temperatures and magnetic fields and, in particular, for transformations in the electronic and magnetic structure. Various experiments [1–5] reveal that the presence of doped magnetic moments in a strongly disordered semiconductor can combine features of the usual doping-driven metal–insulator transition in amorphous systems with the physics of the temperature- and field-driven magnetic (spin glass) transition. The competition between structural and magnetic disorder, which is responsible for the features observed at low temperatures in the magnetic and transport properties of a-RE_xSi_{1-x}, has been analyzed in [6, 7] within the framework of the Anderson–Mott transition driven by spin disorder. ESR and dc-magnetization results show (see [8]) that RE is incorporated as a trivalent ion (RE^{3+}) in the a-Si matrix. Two (*s*–*d*) electrons of RE form saturated bonds with (*s*–*p*) electrons of neighbored Si, while the third (*s*–*d*) electron remains itinerant and participates in the conductivity; below, we consider RE as the one-electron donor in the amorphous silicon host.

So far, the a-RE_xSi_{1-x} alloy was considered as a completely disordered, heavily doped, magnetic semiconductor, and the role of the short-range structural and magnetic order was not discussed. However, as a rule, different kinds of disorder exist in such amorphous alloys [9]. The compositional disorder at the atomic scale distances, associated with dangling bonds, vacancies, and substitutional and interstitial centers, can be qualitatively described within a model of point defects in a regular crystal lattice. The structural (topological) disorder at nanoscale distances, which is originated by dislocations or inclusions, has to be described in a different way, within a model of continuous defects with a short-range order, embedded into a completely disordered effective medium (matrix).

As we argue below, sharp spatial fluctuations of the RE concentration play an important role in a-RE_xSi_{1-x} in a wide temperature range, far above both the paramagnet–spin glass and the metal–insulator phase transitions (see, e.g., [9] for a discussion on the role of the so-called compositional disorder in amorphous semiconducting alloys). Experiments [1–5] revealed five distinct temperature regimes, characterized by different magnetic and transport properties. For instance, at $T > 70$ K, the temperature dependence of the electrical conductivity σ in a-Gd_xSi_{1-x} is similar to that of its nonmagnetic structural analog a-Y_xSi_{1-x}. At $T < 50$–70 K, a significant difference in their behavior has been observed instead, and the conductivity diminishes with decreasing temperature more rapidly in a-Gd_xSi_{1-x} than in a-Y_xSi_{1-x}. This fact points to the magnetic

¶ The text was submitted by the authors in English.

nature of the phenomenon. In the temperature range $5 < T < 50$–70 K, the low-field magnetization qualitatively obeys the Curie–Weiss law, although with a small Curie constant and the effective temperature Θ; large negative magnetoresistance is found at $T < 50$ K. At $T < 5$ K, the material shows a spin glass freezing. Samples that are metallic at high temperature show a tendency towards the metal–insulator transition at low temperature.

To explain these properties, the authors of [10] proposed that the strong structural disorder of the system favors the formation of clusters, i.e., nanoscale structural defects with an enhanced concentration of RE ions, leading to a redistribution of the electron density, such that regions with the higher electron density (electron "drops") appear within the a-RE_xSi_{1-x} matrix. Magnetic ordering inside the drops, more favorable than in the matrix, was predicted. To verify the magnetic state of the drops, a "local" experimental method, electron spin resonance (ESR), was proposed, together with conductivity and Hall-effect measurements. Preliminary ESR results were reported in [10] and allowed a rough estimate of some parameters of the drops. As is shown below, the typical radius of a drop for a-Gd_xSi_{1-x} is $r_D \approx 4.5$–6 Å, corresponding to a volume of $v_D = 4\pi r_D^3/3 \approx 400$–$800$ Å^3; the number of RE ions inside a cluster is $\kappa_D \approx 10$–13; the volume fraction occupied by the drops is $f \approx 0.05$–0.1. The short-range ferromagnetic order inside the drops develops at a temperature $T \approx 100$ K and becomes saturated in the temperature range $50 < T < 100$ K; for 2–$5 < T < 50$ K, the magnetic moments of different drops are uncorrelated, but at $T < 2$–5 K, they are frozen by a spin glass transition in the matrix.

In this paper, we study the a-RE_xSi_{1-x} system in the temperature regime far above the metal–insulator and paramagnet–spin glass transitions, i.e., at $T \gg 2$–5 K for the a-Gd_xSi_{1-x} alloy. We describe the disordered amorphous magnetic semiconductor within a model similar to the one adopted in [6, 7], taking the short-range structural, electronic, and magnetic correlations into account in a semiphenomenological way, within the so-called "local phase transition" approach [11]. To describe the effective electron potential and the charge and spin density distributions of electrons in the drops embedded into the a-RE_xSi_{1-x} matrix, we define the corresponding "order parameters." To obtain the ground-state electron density, we introduce a self-consistency equation in the form of a local electrical neutrality condition for an isolated drop. We also derive the criterion for a ferromagnetic instability and calculate the temperature of the local ferromagnetic transition inside a drop. Finally, we discuss some experimental findings on the behavior of the electrical conductivity as a function of temperature and magnetic field and their correspondence to the predictions of our theory.

2. THE MODEL

a-RE_xSi_{1-x} alloys are systems with a rather complicated topological and compositional disorder. Itinerant electrons move in the crystal potential consisting of a periodic and a disordered part, the latter having components with very different characteristic length scales. Together with the local part, conventional for all amorphous alloys, provided by the potentials of Si dangling bonds and isolated RE ions included in the a-Si network, there is also a continuous part of the disordered potential. We suppose that it is formed in the vicinity of the RE clusters by the Coulomb "tails" of the potential of charged RE ions. Obviously, to make an analytic treatment possible, we need to simplify the real distribution of the crystal potential within some reasonable modeling, which we discuss in what follows.

We consider a set of structurally isolated clusters embedded into a weakly disordered matrix. The matrix is assumed quasihomogeneous on length scales exceeding the interatomic distances but small compared with the characteristic cluster size and the intercluster distance. We assume that the electron structure of the matrix averaged over the realizations of the local random potential is qualitatively described in terms of quasiperiodic electron states with a finite lifetime. Following [9], we write the electron Hamiltonian of our system as a one-band model in the **k**–**r** representation,

$$\mathcal{H} = \sum_{\mathbf{k},\alpha} E(\mathbf{k}) c^\dagger_{\mathbf{k},\alpha} c_{\mathbf{k},\alpha} + \int d\mathbf{r} \sum_{\alpha,\beta} [U_{\alpha\beta}(\mathbf{r}) + \Phi(\mathbf{r})\delta_{\alpha\beta}] \Psi^\dagger_\alpha(\mathbf{r}) \Psi_\beta(\mathbf{r}), \quad (1)$$

where $\mathbf{k} = -i\partial/\partial\mathbf{r}$ is the quasimomentum and $E(\mathbf{k})$ is the Bloch band dispersion of an ideal periodic lattice. The operator $c^{(\dagger)}_{\mathbf{k},\alpha}$ annihilates (creates) an electron in the Bloch state labeled by $\mathbf{k}$ with spin projection α, and the operator $\Psi^{(\dagger)}_\alpha(\mathbf{r})$ annihilates (creates) an electron at the point $\mathbf{r}$ with spin projection α.

The local part of the disordered potential has the form

$$U_{\alpha\beta}(\mathbf{r}) = \sum_i (\mathcal{V}\delta_{\alpha\beta} + \mathcal{J}\mathbf{S}_i \cdot \boldsymbol{\sigma}_{\alpha\beta}) \delta(\mathbf{r} - \mathbf{r}_i), \quad (2)$$

where $\mathcal{V}$ and $\mathcal{J}$ are the Coulomb and exchange couplings of the electrons with the impurities, respectively, $\mathbf{S}_i$ is the local spin vector, and $\boldsymbol{\sigma}$ is the vector of Pauli matrices. The sum in Eq. (2) ranges over the positions of the impurities located at the lattice sites $\mathbf{r}_i$, which are randomly distributed.

The continuous part of the potential is nonzero only inside the clusters and can be written as

$$\Psi(\mathbf{r}) = \sum_j \Phi_j(\mathbf{r}),$$

where $\Phi_j(\mathbf{r})$ is an effective "envelope" Coulomb potential of the jth cluster. In principle, the equation for $\Phi(\mathbf{r})$ has to be derived and solved self-consistently with the charge redistribution in the system. However, in our model, for simplicity, we take $\Phi_j(\mathbf{r}) = \varphi$, independent of $\mathbf{r}$, inside the jth cluster, and $\Phi_j(\mathbf{r}) = 0$ elsewhere. Within this simple approximation, φ occurs as a local shift of the bulk chemical potential μ inside a drop, $\mu_D = \mu + \varphi$.

To characterize the drops, we have to specify their properties. We let N_D and N_M denote the total number of RE ions in the clusters and in the matrix, respectively, with the total number of RE ions $N = N_D + N_M$ being fixed. The total volume occupied by the drops is V_D and the volume of the matrix is V_M, the total volume of the system $V = V_D + V_M$ being fixed. The volume fraction occupied by the drops is denoted by $f \equiv V_D/V < 1$. The RE ion density in the clusters is $n_D \equiv N_D/V_D = \gamma n$, where $\gamma > 1$ is the enhancement factor and $n \equiv N/V$ is the nominal concentration of RE ions. Because $V_D = fV$, we have $N_D = f\gamma N$, $N_M = (1 - f\gamma)N$, and using $V_M = (1 - f)V$, we can calculate the RE ion density in the matrix as

$$n_M \equiv \frac{N_M}{V_M} = \frac{1 - f\gamma}{1 - f} n.$$

To proceed further, we have to make some assumptions about the drops. For simplicity, we assume the drops to be equal and spherical, with a radius r_D and volume $v_D = 4\pi r_D^3/3$. The number of RE ions in a single cluster is then $\kappa_D = n_D v_D$, and the excess of RE ions with respect to the matrix is

$$\Delta\kappa \equiv (n_D - n_M) v_D = \frac{\gamma - 1}{1 - f}\kappa,$$

where $\kappa \equiv n v_D$ is the nominal number of RE ions in a single cluster. Thus, we have

$$\gamma = 1 + (1 - f)\frac{\Delta\kappa}{\kappa}, \tag{3}$$

and in what follows, we assume that $\Delta\kappa$ is a parameter of our model, which is possibly determined by the alloy growing conditions. It is related to the number of RE ions in a cluster by $\kappa_D = \gamma\kappa = \kappa + (1 - f)\Delta\kappa$.

We still have to find a connection between f and v_D. Let $\mathcal{N}_D$ be the total number of drops. Then, $fV = V_D = \mathcal{N}_D v_D$, i.e., $f = \mathcal{N}_D v_D/V$. Assuming that the number of drops per unit volume $\mathcal{N}_D/V$ is technologically fixed, we have f proportional to the volume of a drop v_D, i.e., to the nominal number of RE ions in a cluster, $\kappa = n v_D$. We write $f = \lambda\kappa$, with

$$\lambda \equiv \frac{\mathcal{N}_D}{N} = \frac{\mathcal{N}_D}{Vn}$$

viewed as a parameter. We must find a physical condition to determine κ and, hence, all the drop parameters. As we show in Section 3.1, this is the electrical neutrality condition for an isolated drop.

The potential φ, which determines the position of the local chemical potential in the drops $\mu_D = \mu + \varphi$, can be qualitatively estimated as an average electrostatic potential inside a drop of radius r_D,

$$\varphi = \frac{Ze^2}{\epsilon r_D}\Delta\kappa. \tag{4}$$

Here, e is the electron charge, we take the RE ion as a donor with the effective uncompensated positive charge $Z|e|$, and ϵ is the static dielectric constant of the system. For a-Gd_xSi_{1-x}, we have $Z = 1$ (see Section 1), $\epsilon \approx$ 12–15, the bandwidth of the itinerant electron band is $W \approx$ 6–8 eV, and the average volume of the elementary cell is $a^3 \approx 20$ Å^3; i.e., the average lattice spacing is $a \approx$ 2.7 Å. Thus, for a nominal chemical composition $x =$ 0.14, the average concentration of RE ions is $n \approx 7 \times 10^{21}$ cm^{-3} [4, 5]. From the experimental results [10], we can estimate $\Delta\kappa \approx$ 7–9, $\kappa \approx$ 3–4, $\lambda \approx$ 0.01–0.03. This gives the estimate $\mathcal{N}_D/V = f/v_D \approx (6\text{–}25) \times 10^{19}$ cm^{-3} for the number of drops per unit volume.

3. LOCAL DENSITY OF ELECTRON STATES AND BASIC EQUATIONS

In this section, we derive the equations that fix all the parameters of our model. We start by calculating the local density of states (DOS) in the matrix and in the drops, in the paramagnetic phase, through the usual expression [9] $\rho(\varepsilon) = \pi^{-1}\mathrm{Im}\langle G_A(\mathbf{r}, \mathbf{r}; \varepsilon)\rangle$, where ε is the electron energy and $G_A(\mathbf{r}, \mathbf{r}; \varepsilon)$ is the advanced one-particle Green function associated with Hamiltonian (1), averaged over the realizations of disordered potential (2). Assuming that magnetic order is absent, i.e., $\langle \mathbf{S}_i \rangle = 0$ everywhere in the system, we obtain the expression

$$\rho(\varepsilon) = \mathrm{Im}\int_{-\infty}^{+\infty} \frac{\rho_0(z)}{\varepsilon - z - \Sigma_A(\varepsilon)}\frac{dz}{\pi}, \tag{5}$$

where the energy ε is measured from the center of the band of the ideal lattice. The function $\rho_0(z)$ in Eq. (5) is the DOS corresponding to the electron spectrum $E(\mathbf{k})$

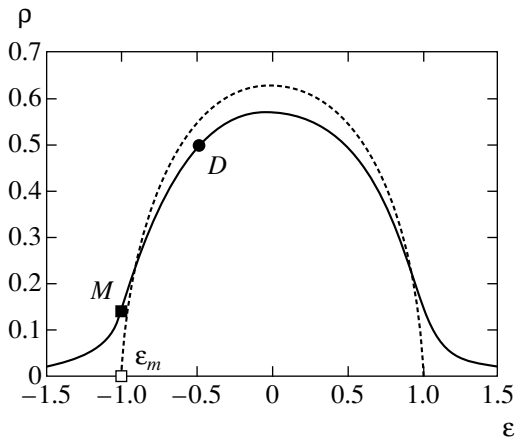

Fig. 1. Dashed line: the semielliptic DOS (Eq. (6)) of the ideal lattice. Solid line: the DOS resulting from the inclusion of "local" impurity effects, resulting in a broadening of the semielliptic DOS (Eq. (7)) with $\Gamma = 0.1$. The empty square marks the position of the mobility edge assumed at the bottom of the ideal band, $\varepsilon_m = -1$. The black square, labeled by *M*, and the black circle, labeled by *D*, mark the value of the DOS in the matrix and in the drops, respectively, for the set of parameters adopted in the text.

of the ideal lattice, and for definiteness, we adopt the semielliptic form

$$\rho_0(z) = \begin{cases} \frac{2}{\pi}\sqrt{1-z^2}, & |z| \le 1, \\ 0, & |z| > 1. \end{cases} \tag{6}$$

We take energy and length units such that half the bandwidth $W/2$ and the size of the elementary cell a are equal to one. The advanced self-energy $\Sigma_A(\varepsilon)$ is obtained by averaging over the realizations of the disordered potential $U(\mathbf{r})$ in Eq. (2) within some approximation scheme. As is customary, we include the average $\langle U(\mathbf{r})\rangle$ into the chemical potential μ; if we then assume that the impurities giving rise to the random potential in Eq. (2) are uncorrelated over different impurity sites $\mathbf{r}_i$, we find $\langle [U(\mathbf{r})]^2\rangle \approx [\mathcal{V}^2 + S(S+1)\mathcal{J}^2]n_{\text{imp}}$ in the noncrossing Born approximation, where n_{imp} is the concentration of impurities. Explicit results for $\Sigma(\varepsilon)$ were obtained in [6, 7] by means of numerical calculations for different values of the scattering parameter $\langle [2U/W]^2\rangle$. For our qualitative purposes, it is sufficient to assume that the fluctuations of potential (2) lead to a simple homogeneous broadening of the ideal semielliptic DOS (6) with a finite inverse lifetime 2Γ proportional to the scattering parameter. Thus, the resulting local DOS in our model is characterized by a tail of localized states (see Fig. 1). Using Eq. (5) with $\Sigma_A(\varepsilon) = i\Gamma$ and with $\rho_0(z)$ given by Eq. (6), we find

$$\rho(\varepsilon) = \frac{2}{\pi}\left(\sqrt{\sqrt{\mathcal{R}^2 + \varepsilon^2\Gamma^2} - \mathcal{R}} - \Gamma\right), \tag{7}$$

where $\mathcal{R} \equiv (\varepsilon^2 - \Gamma^2 - 1)/2$.

3.1. The Paramagnetic Phase

We start our analysis by discussing the properties of the paramagnetic phase. Hereafter, the subscript p indicates that the corresponding quantity is evaluated in the paramagnetic phase whenever this is expected to have a different value in the phase with a short-range ferromagnetic order, which is discussed in Section 3.2.

The chemical potential of the system in the paramagnetic phase, μ_p, is fixed by the condition of conservation of the average number of electrons,

$$\begin{aligned} &2(1-f_p)\int_{-\infty}^{+\infty}\rho(\varepsilon)f(\varepsilon-\mu_p)d\varepsilon \\ &+2f_p\int_{-\infty}^{+\infty}\rho(\varepsilon)f(\varepsilon-\mu_p-\varphi_p)d\varepsilon = x, \end{aligned} \tag{8}$$

where the factors 2 account for the spin degeneracy, $f_p = \lambda\kappa_p$ is the volume fraction occupied by the drops in the paramagnetic phase, $f(z) = [\exp(z/T) + 1]^{-1}$ is the Fermi–Dirac distribution function at the temperature T (in energy units), and $x = na^3$ is the nominal RE content of the alloy (here, RE ions are taken as donors with $Z = 1$, such that the number of doped electrons in the conduction band equals the number of RE ions). For simplicity, we assume that the inverse lifetime Γ is the same in the matrix and in the drops, although this assumption plays no role in the following derivation, and we could even adopt a different DOS $\rho_{M,D}(\varepsilon)$ in the matrix and in the drops.

As discussed in Section 2, the excess of RE ions inside a cluster, $\Delta\kappa$, causes an increase in the electron density with respect to the matrix, which is controlled by the average potential φ_p in Eq. (4) with $r_D \longrightarrow r_{D,p}$ in the paramagnetic phase. We assume that the electrical neutrality condition is satisfied for an isolated drop, ensuring that the excess of the RE ion density is screened by the corresponding excess of the electron density,

$$2\int_{-\infty}^{+\infty}\rho(\varepsilon)f(\varepsilon-\mu_p-\varphi_p)d\varepsilon = \gamma_p x, \tag{9}$$

where γ_p is the density enhancement factor (Eq. (3)) calculated in the paramagnetic phase (i.e., with $f \longrightarrow f_p$, $\kappa \longrightarrow \kappa_p$). With Eq. (9), we can rewrite Eq. (8) in the simpler form

$$2\int_{-\infty}^{+\infty}\rho(\varepsilon)f(\varepsilon-\mu_p)d\varepsilon = (1-\lambda\Delta\kappa)x, \tag{10}$$

whence it is evident that the chemical potential μ_p is uniquely determined in terms of the parameters of our

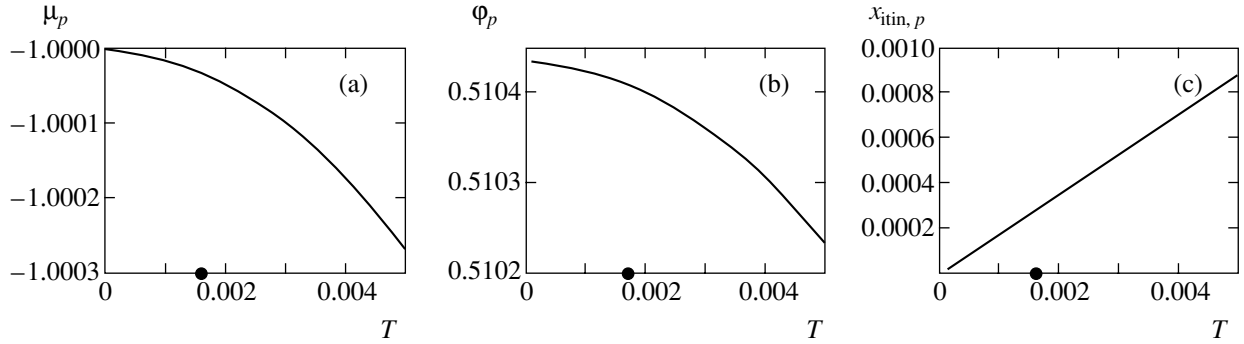

Fig. 2. (a) Temperature dependence of the chemical potential μ_p in the paramagnetic phase as a result of the numerical solution of Eq. (10) at finite temperature. (b) Temperature dependence of the Coulomb shift φ_p in the paramagnetic phase as resulting from the numerical solution of Eq. (9) at finite temperature and with $\mu_p(T)$ previously determined. (c) Temperature dependence of the fraction of itinerant electrons in the paramagnetic phase, $x_{\mathrm{itin},\,p}$, calculated according to Eq. (11). In all three panels, the black circle on the temperature axis marks the transition point to the phase with a short-range ferromagnetic order in the drops.

model. Once μ_p is obtained from Eq. (10), Eq. (9) contains κ_p (or, equivalently, $r_{D,\,p}$) as the only variable and can be easily solved by means of standard numerical methods.

The simultaneous numerical solution of Eqs. (9) and (10) at $T = 0$, e.g., for $x = 0.14$, $W = 8$ eV, $\epsilon = 12$, $\Delta\kappa = 9$, and $\lambda \approx 0.029$, with $\Gamma \approx 0.1$ (chosen as in [6, 7] to fix the chemical potential at the mobility edge, $\mu_p = -1$; see below) yields $\kappa_p \approx 3.1$, which gives $f_p \approx 0.091$ and $\gamma_p \approx 3.34$. Hence, the number of RE ions in a cluster is $r_{D,\,p} \approx 10.4$, the volume of a drop is $v_{D,\,p} \approx 440$ Å^3, and the radius of a drop is $r_{D,\,p} \approx 4.7$ Å. The value of the Coulomb shift of the chemical potential in the drops is $\varphi_p \approx 0.51$ (which corresponds to an energy of approximately 2 eV). The local DOS at the Fermi level is $\rho(\mu_p) \approx 0.14$ and $\rho(\mu_p + \varphi_p) \approx 0.50$ in the matrix and in the drops, respectively (the maximum value for the DOS is $\rho_{\max} \approx 0.58$ for the chosen set of parameters; see Fig. 1).

So far, we have discussed only the paramagnetic phase of the system, and, hence, our results are valid for both the magnetic alloy a-Gd_xSi_{1-x} (at $T > T_D$; see Section 3.2) and the nonmagnetic alloy a-Y_xSi_{1-x}.

For small f_p well below the percolation limit of the drops, the electron states within the drops are localized in the volume v_D and are separated from the matrix by a surface energy barrier, which determines the excitation energy of a drop, $\mathscr{E}_D$. Also, the electron states within the tail of the DOS of the matrix are localized at the scale of interatomic distances. Therefore, at $T \ll \mathscr{E}_D$, the fraction of itinerant electrons within the elementary cell in our system can be estimated as

$$x_{\mathrm{itin},\,p} = 2(1 - f_p)\int_{\varepsilon_m}^{+\infty} \rho(\varepsilon) f(\varepsilon - \mu_p) d\varepsilon, \qquad (11)$$

where ε_m is the mobility edge, which depends on the scattering potential. We assume, for simplicity, that it is located at the bottom of the ideal lattice band, $\varepsilon_m = -1$. Although, strictly speaking, one should define the position of the mobility edge self-consistently, from the calculation of the two-particle Green function of the system, our simplifying assumption does not play a relevant role.

The variation of $x_{\mathrm{itin},\,p}$ with increasing temperature in the paramagnetic phase, together with a tiny variation of the chemical potential μ_p and of the Coulomb shift φ_p, is reported in Fig. 2 for the chosen set of parameters. As can be seen, because the Fermi level was fixed at the mobility edge at $T = 0$, the itinerant electrons are thermally excited from the localized states in the tails of the DOS at finite temperature, and their density increases almost linearly with T. We note that $x_{\mathrm{itin},\,p}$ is at most about $0.006x$ at the highest temperature reported in Fig. 2 (which corresponds to $T \approx 200$ K), and, therefore, itinerant electrons are a tiny fraction of all the electrons in the system, the majority being localized into the DOS tails. Therefore, whereas the nominal density of doped electrons is $n \approx 7 \times 10^{21}$ cm^{-3}, the density of thermally excited itinerant electrons is, e.g., $n_{\mathrm{itin},\,p} \approx 4 \times 10^{19}$ cm^{-3} at $T \approx 200$ K.

3.2. The Phase with a Short-Range Ferromagnetic Order

In this section, we treat the exchange part of Hamiltonian (1) in the mean-field approximation, supposing $\langle \mathbf{S}_i \rangle = 0$ in the matrix and $\langle \mathbf{S}_i \rangle \neq 0$ inside the drops. This assumption is quite reasonable in a wide temperature range, because the local DOS at the Fermi level is larger in the drops than in the matrix (see Fig. 1), and, therefore, the condition for ferromagnetic ordering in the presence of an exchange coupling between magnetic RE ions and electrons is more easily realized in the

drops. The magnetic RE ions inside the clusters experience the effective magnetic field

$$H_{\text{eff}} = \mathcal{J}\int_{-\infty}^{+\infty}[\rho(\varepsilon+m)-\rho(\varepsilon-m)] \times f(\varepsilon-\mu-\varphi)d\varepsilon, \quad (12)$$

where $m \equiv \mathcal{J}x_D\langle S_z\rangle$, $x_D = \gamma x$ is the concentration of magnetic (e.g., Gd) RE ions per unit cell in a cluster, the index z defines the direction of the local quantization axis, and the average value of the spin at the RE site is defined self-consistently as

$$\langle S_z\rangle = S\mathcal{B}_S\left(\frac{SH_{\text{eff}}}{T}\right),$$

where

$$\mathcal{B}_S(y) = \frac{2S+1}{2S}\cosh\left(\frac{2S+1}{2S}y\right) - \frac{1}{2S}\cosh\left(\frac{1}{2S}y\right)$$

is the Brillouin function for spin S. For Gd ions, $S = 7/2$.

The above equations should be solved simultaneously, together with the equation for the chemical potential μ,

$$2(1-f)\int_{-\infty}^{+\infty}\rho(\varepsilon)f(\varepsilon-\mu)d\varepsilon + f\int_{-\infty}^{+\infty}[\rho(\varepsilon+m)+\rho(\varepsilon-m)]f(\varepsilon-\mu-\varphi)d\varepsilon = x, \quad (13)$$

which corresponds to the conservation of the average number of electrons in the phase with a short-range ferromagnetic order within the drops, and the equation for κ, which enforces charge neutrality for an isolated drop,

$$\int_{-\infty}^{+\infty}[\rho(\varepsilon+m)+\rho(\varepsilon-m)]f(\varepsilon-\mu-\varphi)d\varepsilon = \gamma x, \quad (14)$$

where γ is the density enhancement factor defined in Eq. (3). As discussed above, we assume that the number of excess RE ions $\Delta\kappa$ does not change in passing from the paramagnetic phase to the phase with a short-range ferromagnetic order, whereas the radius of the drops changes from $r_{D,p}$ to r_D (i.e., the nominal number of RE ions within the clusters changes from κ_p to κ). The volume fraction occupied by the drops is $f = \lambda\kappa$, where λ is the same as in the paramagnetic phase, assuming that the number of drops per unit volume does not change across the local ferromagnetic transition.

Using Eq. (14), we can rewrite Eq. (13) as

$$2\int_{-\infty}^{+\infty}\rho(\varepsilon)f(\varepsilon-\mu)d\varepsilon = (1-\lambda\Delta\kappa)x, \quad (15)$$

which coincides with Eq. (10). Therefore, it is evident that, for a given set of parameters, the chemical potential has the same value as in the paramagnetic phase at the same temperature, $\mu(T) = \mu_p(T)$.

The typical value of the exchange potential in the units of $W/2$ is $\mathcal{J}S \approx 0.1$–$0.2 \ll \varphi$. In what follows, we take m as a small expansion parameter and seek solutions of the above self-consistency equations that are close to the solutions in the paramagnetic phase, $r_D = r_{D,p} + \eta$, with $\eta \ll r_{D,p}$. We observe that $\varphi \approx \varphi_p - e^2\Delta\kappa\eta/\epsilon r_{D,p}^2 \equiv \varphi_p + \zeta$, with $\zeta \equiv -\eta\varphi_p/r_{D,p} \ll \varphi_p$. The volume of a drop changes as $v_D \approx v_{D,p}(1 + 3\eta/r_{D,p})$, and, hence, $\kappa = nv_D \approx \kappa_p(1 + 3\eta/r_{D,p})$, $f = \lambda\kappa \approx f_p(1 + 3\eta/r_{D,p})$, and $\gamma \approx \gamma_p(1 + 3\Delta\kappa\eta/\gamma_p\kappa_p r_{D,p})$.

Now, we expand the DOS and the Fermi–Dirac distribution function as

$$\rho(\varepsilon+\sigma m) \approx \rho(\varepsilon) + \rho'(\varepsilon)\sigma m + \frac{1}{2}\rho''(\varepsilon)m^2 + \frac{1}{6}\rho'''(\varepsilon)\sigma m^3 \quad (\sigma = \pm 1),$$

$$f(\varepsilon-\mu_p-\varphi) \approx f(\varepsilon-\mu_p-\varphi_p) - f'(\varepsilon-\mu_p-\varphi_p)\zeta$$

(here and in what follows, the prime is a short notation for the derivative with respect to ε). Then, Eq. (14) for charge neutrality, at this order of approximation, gives

$$-2\left[\int_{-\infty}^{+\infty}\rho(\varepsilon)f'(\varepsilon-\mu_p-\varphi_p)d\varepsilon\right]\zeta + \left[\int_{-\infty}^{+\infty}\rho''(\varepsilon)f(\varepsilon-\mu_p-\varphi_p)d\varepsilon\right]m^2 = \frac{3x\Delta\kappa}{\kappa_p\varphi_p}\zeta, \quad (16)$$

i.e., $L\zeta + Mm^2 = 0$, where the coefficients L and M are calculated in the paramagnetic phase. The coefficient M is reexpressed in a more suitable form via integration by parts that transfers the derivative with respect to ε from ρ to the Fermi–Dirac distribution function f. At low temperature $T \ll \varphi_p$, we find $L = 2\rho(\mu_p + \varphi_p) - (3x\Delta\kappa/\kappa_p\varphi_p)$ and $M = \rho'(\mu_p + \varphi_p)$.

It is evident that $\zeta \sim m^2$, as expected, because the corrections to the Coulomb shift cannot depend on the

sign of the magnetization. Finally, expression (12) for the effective field up to $O(m^3)$ becomes

$$H_{\text{eff}} \approx 2\mathcal{J}\left[\int_{-\infty}^{+\infty}\rho'(\varepsilon)f(\varepsilon-\mu_p-\varphi_p)d\varepsilon\right]m + \frac{\mathcal{J}}{3}\left[\int_{-\infty}^{+\infty}\rho'''(\varepsilon)f(\varepsilon-\mu_p-\varphi_p)d\varepsilon\right]m^3 - 2\mathcal{J}\left[\int_{-\infty}^{+\infty}\rho'(\varepsilon)f'(\varepsilon-\mu_p-\varphi_p)d\varepsilon\right]m\zeta,$$

i.e., $H_{\text{eff}} \approx Am + Bm^3 + Cm\zeta$, where the coefficients A, B, and C are calculated in the paramagnetic phase. The coefficients A and B are reexpressed in a more appropriate form via integration by parts that transfers the derivative with respect to ε from ρ to f. At low temperature, $A = 2\mathcal{J}\rho(\mu_p + \varphi_p)$, $B = \mathcal{J}\rho''(\mu_p + \varphi_p)/3$, and $C = 2\mathcal{J}\rho'(\mu_p + \varphi_p) = 2\mathcal{J}M$.

We find the solution of Eqs. (12)–(15) near the local phase transition, i.e., at temperatures close to the local Curie point of the drops T_D (which is defined below), where our expansion in powers of m and ζ is valid. We must expand the Brillouin function, observing that $\cosh(y) \approx 1/y + y/3 - y^3/45$, i.e.,

$$\mathcal{B}_S(y) \approx \frac{S+1}{3S}y - \frac{2S^3 + (2S+1)^2}{90S^3}y^3.$$

Then, the self-consistency equation for m in the phase with a short-range ferromagnetic order for $T \lesssim T_D$ (T_D is defined below), at the same order of approximation, becomes

$$1 = \mathcal{J}x\gamma_p S\left\{\frac{S+1}{3T}[A + Bm^2 + (C+D)\zeta] - \frac{2S^3 + (2S+1)^2}{90T^3}A^3m^2\right\}, \quad (17)$$

where $D = 3A\Delta\kappa/\gamma_p\kappa_p\varphi_p$ accounts for the variation of γ (i.e., of $x_D = x\gamma$) in entering the phase with the short-range ferromagnetic order. Equation (17) must be solved together with Eq. (16) to yield ζ and m^2.

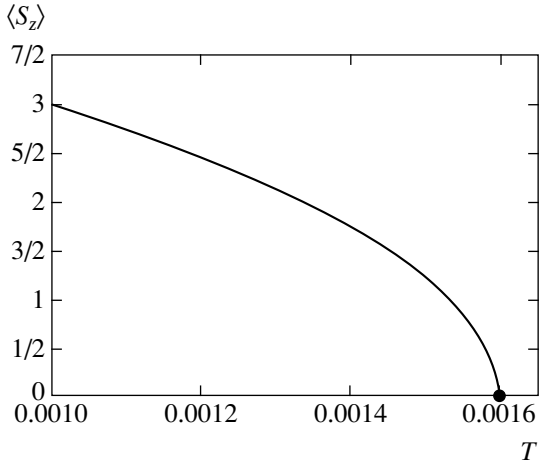

Fig. 3. Temperature dependence of the magnetization of a Gd ion in a cluster, $\langle S_z\rangle \equiv m/\mathcal{J}x_D$, obtained by solving Eq. (18) for T close to T_D. The values of the parameters are given in the text. The local transition point at $T = T_D = 0.0016$ is marked by a black circle.

It is evident that the local transition temperature for the drops is $T_D = \mathcal{J}A^*S(S+1)/3$, where $A^* \equiv Ax\gamma_p = Aa^3\kappa_D/v_D$. We note that, in our simple mean-field approach and within the approximation of isolated drops, the full dependence of the transition temperature T_D on the mean RE concentration x might not be correctly described, because κ_D and v_D are the local parameters of a drop, which, within our model, are self-consistently determined by x and by the cluster parameters $\Delta\kappa$ and λ, assumed fixed. A more developed theory has to account for both charge and spin correlations in the system (which may introduce a dependence of $\Delta\kappa$ and λ on x), as well as an exchange between moments of different drops, to describe the correct dependence of T_D on x. However, this quantitative description is beyond the scope of our paper.

For $T > T_D$, Eq. (17) has no real solutions and $m = 0$. For $T < T_D$, the ferromagnetic solution within the drops becomes stable. From Eq. (16), we find $\zeta = -Mm^2/L$, and substituting this in Eq. (17), we obtain the equation for m^2 for $T \lesssim T_D$,

$$m^2 = \frac{\frac{1}{3}\mathcal{J}Ax\gamma_p S(S+1) - T}{\mathcal{J}x\gamma_p S\left[(S+1)\frac{M(C+D)-BL}{3L} + \frac{2S^3+(2S+1)^2}{90T^2}A^3\right]} \equiv \frac{T_D - T}{P + \frac{Q}{T^2}} \approx \frac{T_D - T}{P + \frac{Q}{T_D^2}}, \quad (18)$$

where P and Q are constants which depend on the parameters calculated in the paramagnetic phase and whose expression can be easily deduced from Eq. (18). For a-$Gd_{0.14}Si_{0.86}$, we have $\rho'(\mu_p + \varphi_p) \approx 0.35$, $\rho''(\mu_p + \varphi_p) \approx -0.91$, and taking $\mathcal{J} \approx 0.026$ (which corresponds to an energy of 0.1 eV, a typical exchange energy in

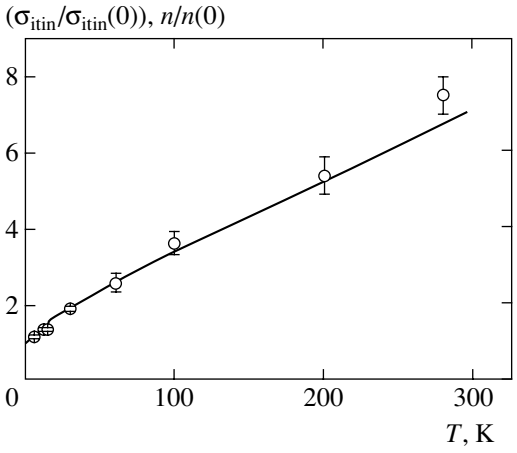

Fig. 4. Temperature dependence of the itinerant electron conductivity (solid line) and concentration (open circles) for the nonmagnetic $Y_{0.17}Si_{0.83}$ sample.

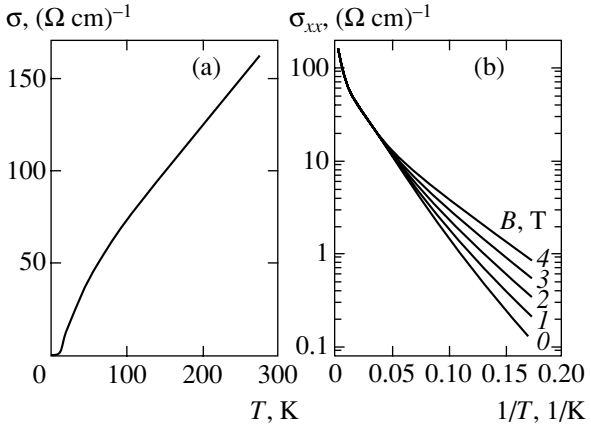

Fig. 5. (a) Conductivity vs. temperature dependence for the $Gd_{0.135}Si_{0.865}$ sample no. 5 at zero magnetic field. (b) Conductivity vs. $1/T$ at various magnetic fields for the same sample.

wideband magnetic semiconductors), we find $T_D = 0.0016$ (which corresponds to a temperature of 70 K) and $P + Q/T_D^2 \approx 0.45$. The magnetization of the magnetic RE ions within the clusters, $\langle S_z \rangle \approx m/\mathcal{J}x_D \approx m/\mathcal{J}x\gamma_p$, near the local transition point T_D is reported in Fig. 3.

We point out that our results are correct only within the mean-field approximation for the RE spin density. The finite volume of the drops causes a "tail" of fluctuations of the magnetization to occur at $T > T_D$, in the temperature range $(T - T_D)/T_D \sim (r_D/a)^{-2} \approx 0.1$–$0.2$.

4. EXPERIMENTS AND DISCUSSION

The X-ray study of the local structure of a-Gd_xSi_{1-x} revealed a strong local distortion of the matrix around Gd ions, as well as the absence of fluctuations on macroscopic scales in the system [12]. Detailed conductivity and tunneling measurements revealed the coexistence of metallic and semiconducting domains (micro- or mesoscopic), identifying the percolation nature of electron transport at low temperatures, near the metal–insulator transition [13]. Here, we consider some interesting experimental results obtained at temperatures far above the metal–insulator transition and discuss their correspondence to the predictions of our theory. The results are obtained using the samples prepared in Prof. F. Hellman's laboratory by a technique described previously [2]. Amorphous films of a-$(Gd, Y)_xSi_{1-x}$, 100–500 nm thick, were grown by e-beam coevaporation on Si/SiN substrates. Magnetotransport measurements were carried out in the temperature range 5–300 K in magnetic fields up to 4 T using the van der Pauw and standard Hall bar technique.

Experiments [1–5] have clearly shown that the electrical conductivity σ increases almost linearly with the temperature T in a-Y_xSi_{1-x} at $T > 2$–5 K and in a-Gd_xSi_{1-x} at $T > 50$–70 K. This dependence is well described by the expression

$$\sigma_p(T) \approx \sigma_0 + \sigma_{itin,\,p}(T), \tag{19}$$

where σ_0 is constant and $\sigma_{itin,\,p}(T)$ depends on the temperature. To explain these results, we propose that, in a-RE_xSi_{1-x} alloys, the electrically neutral drops play a significant role in the itinerant electron transport. The constant part σ_0 is associated with the tunneling between the drops and the matrix through the surface barrier, while $\sigma_{itin,\,p}(T) = x_{itin,\,p}(T)\nu_p(T)$, where $\nu_p(T)$ is the itinerant electron mobility in the paramagnetic state. To clarify the role of the itinerant electron concentration in the temperature dependence of the electrical conductivity, Hall effect measurements were carried out.

In Fig. 4, we present the experimentally determined itinerant electron concentration $n(T)$ for nonmagnetic a-Y_xSi_{1-x}, which linearly increases with temperature in the whole temperature range. According to our model, this implies that the dependence of $\nu_p(T)$ on T is weak enough, as a result of the scattering of itinerant electrons on electrically neutral centers. The temperature dependence of the itinerant electron electrical conductivity presented in Fig. 4 is built by taking into account some shunting tunneling conductivity σ_0, which does not contribute to the Hall effect. The data show that $\sigma_{itin,\,p}(T)/\sigma_{itin,\,p}(0) \approx n(T)/n(0)$ and confirm that the variation of $\sigma_{itin,\,p}(T)$ is produced by $n(T)$, which increases linearly with T.

For the a-Gd_xSi_{1-x} magnetic alloys, the increase of $\sigma_p(T)$ with T is strongly nonlinear at T below 50–70 K and becomes quasilinear only above this temperature (see Fig. 5). We attribute this nonlinearity to complex magnetic transformations in the system. Consistently with the theory developed in Section 2, for $T < T_D \approx$

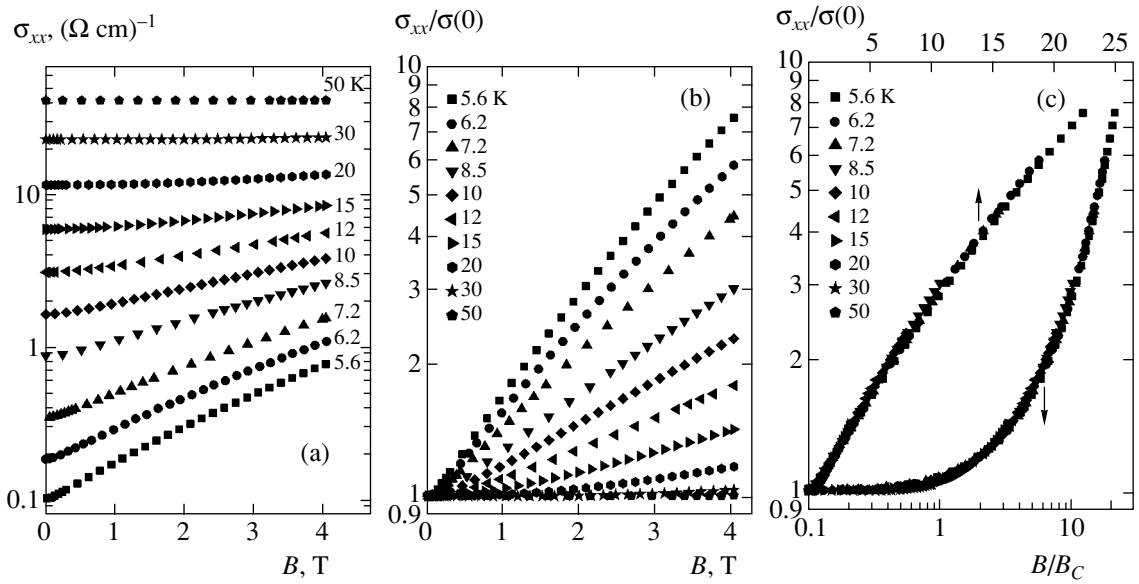

Fig. 6. (a) Conductivity vs. magnetic field dependence for the $Gd_{0.135}Si_{0.865}$ sample no. 5 at different temperatures. (b) The same dependence when the conductivity is rescaled by the zero-field value $\sigma(0)$. (c) The same dependence when the magnetic field is rescaled by the critical value B_C, which evidences the data collapse.

50–70 K, when the short-range ferromagnetic order is formed in the alloys containing magnetic RE atoms, two factors modify the temperature dependence of the conductivity $\sigma(T)$.

First, the exchange scattering of electrons occurs on the noncorrelated magnetic moments of the drops, which shifts the mobility edge ε_m in Eq. (11) upwards,

$$\frac{\Delta\varepsilon_m}{\Gamma} \sim \beta(T) \equiv \frac{\lambda(\mathcal{J}\kappa_D\langle S_z\rangle)^2}{\mathcal{V}^2},$$

thus decreasing the concentration of itinerant electrons in the phase with a short-range ferromagnetic order, $x_{\text{itin}}(T)$.

Second, a dependence of the mobility $\nu(T)$ of the itinerant electrons on T appears, which is qualitatively described as a superposition of the potential and exchange mechanisms for electron scattering on the disordered magnetic drops with a characteristic size that is small compared to the electron mean free path, $\nu(T)/\nu_p(T) \sim 1 - \beta(T)$. Thus, the reduction of $\sigma(T)$ with decreasing temperature is driven, in principle, by both mechanisms.

The parameter $\beta(T)$ is zero at $T > T_D$ and may be on the order of 0.01–0.1 at $T \ll T_D$ if $\mathcal{J}S/\mathcal{V} \approx 0.1$–$0.3$, $\kappa_D \approx$ 10–12, and $\lambda \approx 0.01$–0.03. From the data in Fig. 5, we conclude that the variation of the itinerant electron concentration plays the major role in our system, and the variation of the itinerant electron mobility can be neglected in the following discussion. In any case, the appearance of a short-range ferromagnetic order obviously enhances the tendency towards the metal–insulator transition.

We briefly discuss the variation of the conductivity σ of our system as a function of temperature and of external magnetic field **B**. At $T < 50$–70 K, a strong exponential dependence of the conductivity on the temperature (see Fig. 5b) and on the magnetic field (see Fig. 6) is observed. There are two regimes of magnetic fields characterized by a different behavior of $\sigma(B)$. In a magnetic field less than some critical value B_C, the conductivity slightly depends on the magnetic field. At $B > B_C$, the aforementioned exponential dependence of σ on B is observed (see Fig. 6b); the critical value B_C increases with increasing temperature (see Fig. 7). B_C is determined by the intersection of the local fit of the

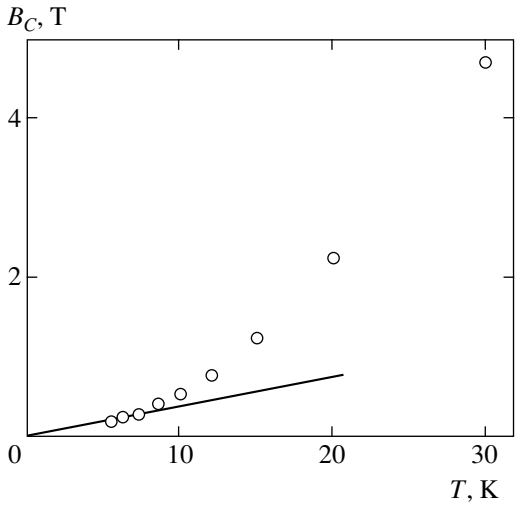

Fig. 7. Temperature dependence of the critical magnetic field B_C for the $Gd_{0.135}Si_{0.865}$ sample no. 5.

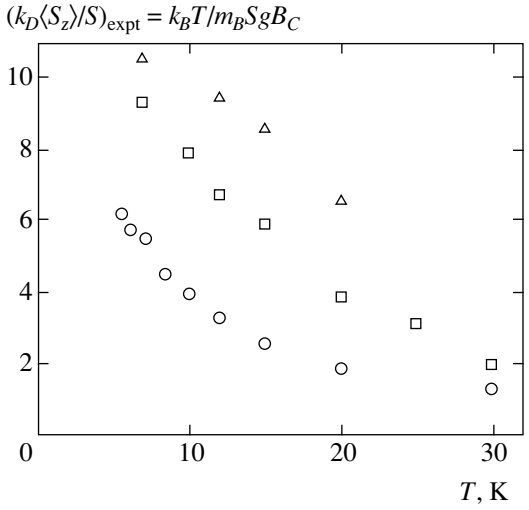

Fig. 8. Temperature dependence of the mean magnetic moment of the clusters for the $Gd_{0.135}Si_{0.865}$ sample no. 5 (open circles), for the $Gd_{0.14}Si_{0.86}$ sample no. 3 (open squares), and for the $Gd_{0.145}Si_{0.855}$ sample no. 7 (open triangles).

experimental $\log[\sigma/\sigma(0)]$ as a function of B with the line $\sigma/\sigma(0) = 1$ at low temperatures. For high temperatures, B_C is the scaling parameter from Fig. 6c. We suppose that B_C corresponds to the critical magnetic field that aligns the magnetic moments of the drops. At low magnetic field, the magnetic energy of the drop $\kappa_D\langle S_z\rangle m_B gB$ (here, m_B is the Bohr magneton and g is the gyromagnetic ratio) is smaller than the thermal energy k_BT, and the magnetic moments of different drops are disordered. This leads to an additional fluctuation potential in the system and raises the mobility edge. Magnetic fields larger than B_C align the magnetic moments of different drops and eliminate this scattering channel. This reduces the mobility edge and increases the itinerant electron concentration, leading to an increase in conductivity.

It is important that the value of the magnetic moment of the drop obtained from the low-temperature part of the curve $B_C(T)$ allows us to estimate the average number of Gd atoms in the cluster, κ_D. The experimentally determined value of $(\kappa_D\langle S_z\rangle/S)_{\text{expt}} = k_BT/m_BSgB_C$ is shown in Fig. 8. At low T, when $\langle S_z\rangle \approx S$, we obtain $\kappa_{D,\text{ expt}} \approx 10$, which is consistent with the prediction of our theory. The experimentally determined κ_D values are close for different samples. We suppose that, if the clusters arise during the sample growth, their size may depend on the synthesis conditions.

We note that the magnetic-field dependence of the conductivity has a universal form for different temperatures. The experimental dependences of $\sigma/\sigma(0)$ vs. (B/B_C) (here, $\sigma(0)$ is the zero-field conductivity) for different temperatures are presented in Fig. 6c. The experimentally observed behavior of $\sigma/\sigma(0)(B/B_C)$ obeys the law

$$\frac{\sigma}{\sigma(0)} \approx \exp\left(\frac{B}{B_C} - 1\right) \tag{20}$$

at $B > B_C$, where $B_C = k_BT/M$ and $M = m_B\langle S_z\rangle g\kappa_{D,\text{ expt}}$. At $B \gg B_C$, this gives

$$\frac{n}{n(0)} \approx \exp\left(\frac{B}{B_C}\right), \tag{21}$$

where $n(n(0))$ is the density of itinerant electrons (at zero field). Our explanation of this result is as follows. We suppose that the Zeeman splitting in the matrix leads to a downward (upward) shift of the bottom of the itinerant electron spin-up (spin-down) subband with respect to the Fermi level. At high B, the full splitting regime sets in when the spin-down subband remains empty, the local Fermi energy measured from the bottom of the spin-up subband rises linearly with B; and the itinerant electron concentration increases according to Eq. (11).

Our experiments, as well as previous data [1–5], have shown that the application of a strong magnetic field B suppresses the tendency towards the metal–insulator transition and even induces an insulator-to-metal transformation in some a-RE_xSi_{1-x} alloys with low RE concentration. This fact is naturally explained within our model if we take into account either the increase of the itinerant electron concentration or the suppression of the electron exchange scattering on the magnetic drops provided by their coherent orientation in the magnetic field.

What external influence, besides the magnetic field, may increase the itinerant electron concentration in the studied system? A way to vary the electron concentration is to increase the current I through the sample. To provide a more uniform current density distribution over the sample, we used the standard Hall bar geometry of measurements. The conductance G and relative conductivity $\sigma/\sigma(0)$ dependences on the current at different temperatures are presented in Fig. 9, where $\sigma(0)$ is the conductivity at zero current limit. (G is used because of the small sample size and not well-defined geometrical factor for σ calculation.) We note that these dependences are analogous to such dependences vs. magnetic field shown in Fig. 10. These figures clearly demonstrate that the current effect on the system is analogous to the influence of the external magnetic field. We suppose that a current flow I through the sample enhances the effective exchange between the magnetic moments of disordered drops, because it increases the itinerant electron concentration. If I exceeds some critical value I_C, determined by the same procedure as B_C, all the drops on the percolation path become magnetically ordered, which leads to a suppression of

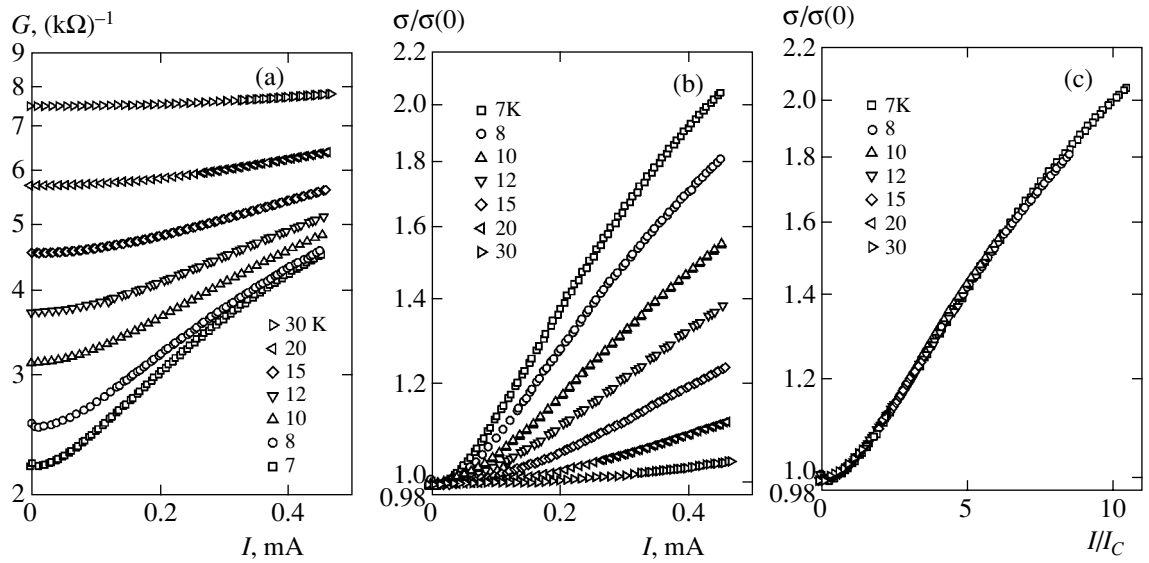

Fig. 9. (a) Conductance as a function of the current through the $Gd_{0.14}Si_{0.86}$ sample no. 3 at different temperatures. (b) The same dependence when the conductivity is rescaled by the zero-current value $\sigma(0)$. (c) The same dependence when the current is rescaled by the critical value I_C to put in evidence the data collapse.

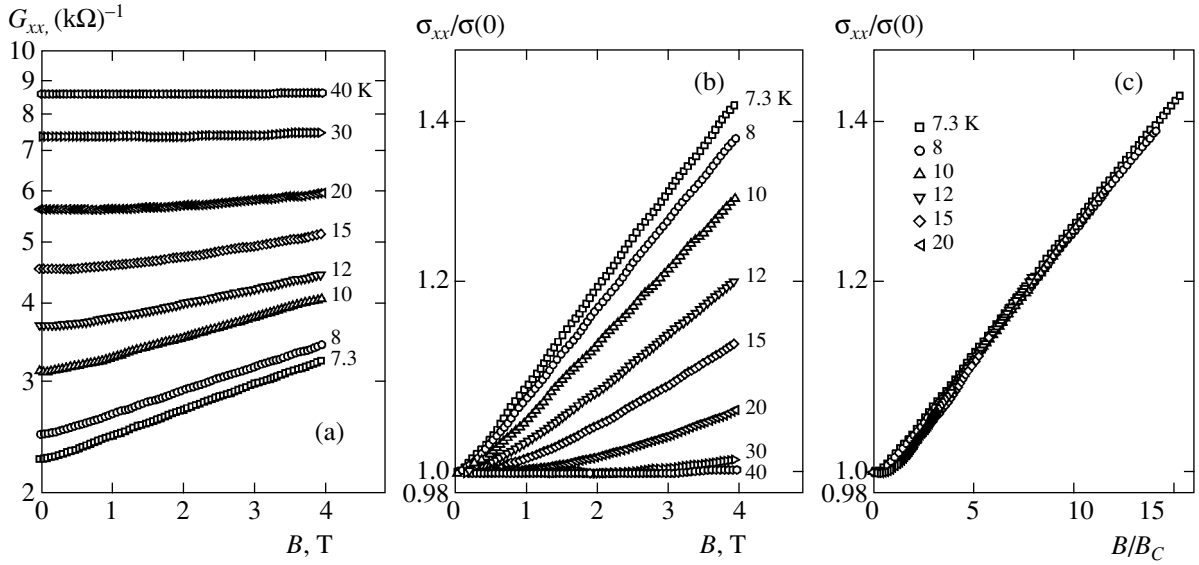

Fig. 10. (a) Conductance vs. magnetic field dependence for the $Gd_{0.14}Si_{0.86}$ sample no. 3 at different temperatures. (b) The same dependence when the conductivity is rescaled by the zero-field value $\sigma(0)$. (c) The same dependence when the field is rescaled by the critical value B_C to put in evidence the data collapse.

the fluctuation potential of the magnetic disorder. Increasing the current also leads to the rise of the itinerant electron concentration and to the reduction of the activation energy between the Fermi level and the mobility edge, which is consistent with Eq. (11).

5. CONCLUSIONS

We presented the theoretical description and electrical conductivity measurements for amorphous a-$(Gd, Y)_xSi_{1-x}$ alloys with $0.1 < x < 0.2$. We took the strong topological disorder in the system into account: in our approach, the nanoscale structural defects enriched with rare-earth ions ("clusters") cause the appearance of regions with higher electron density (electron "drops"). The value of the local DOS at the Fermi level in the drops significantly exceeds the value of the DOS at the Fermi level in the matrix, and, therefore, a short-range ferromagnetic order appears in the drops below some characteristic temperature T_D. We estimated T_D in the "local phase transition" approach and analyzed measurements of the temperature and magnetic-field dependence of the electrical conductivity in amorphous $(Gd, Y)_xSi_{1-x}$ alloys, in the frame-

work of the drop description. We obtained a qualitative agreement between the experimental results and the theoretical predictions. Further ESR measurements, scanning electron microscopy with polarization analysis (SEMPA), and neutron diffraction (ND) experiments are necessary to reveal details of the electron and magnetic structure of the drops.

In our theoretical model, we have neglected the low-temperature effects leading to the metal–insulator transition and associated with the Mott–Hubbard [6, 7] or percolation [13] mechanisms. Nevertheless, our experimental results are in accordance with the conclusions in [13] about the percolation character of the electron transport and the metal–insulator transition in the studied system.

ACKNOWLEDGMENTS

We are grateful to Prof. F. Hellman and Dr. E. Helgren for sample preparation and critical remarks, and to Profs. C. Di Castro, C. Castellani, and M. Grilli for valuable discussions. N.C. and V.T. were supported in part by the CRDF (grant no. RP2-2402-MO-02). V.T. was supported in part by the invited-professor Program of the Center for Statistical Mechanics and Complexity (SMC) of the Istituto Nazionale per la Fisica della Materia (INFM), Unità di Ricerca (UdR) di Roma 1. S.C. was supported by the Italian MIUR, Cofin 2003, prot. 2003020239_006.

REFERENCES

1. F. Hellman, M. Q. Tran, A. E. Gebala, *et al.*, Phys. Rev. Lett. **77**, 4652 (1996).
2. W. Teizer, F. Hellman, and R. C. Dynes, Phys. Rev. Lett. **85**, 848 (2000).
3. F. Hellman, D. R. Queen, R. M. Potok, and B. L. Zink, Phys. Rev. Lett. **84**, 5411 (2000).
4. B. L. Zink, E. Janod, K. Allen, and F. Hellman, Phys. Rev. Lett. **83**, 2266 (1999).
5. B. L. Zink, V. Preisler, D. R. Queen, and F. Hellman, Phys. Rev. B **66**, 195 208 (2002).
6. S. Kumar and P. Majumdar, Int. J. Mod. Phys. B **15**, 2683 (2001).
7. P. Majumdar and S. Kumar, Phys. Rev. Lett. **90**, 237 202 (2003).
8. M. Sercheli, C. Rettori, and A. Zanatta, Braz. J. Phys. **32**, 409 (2002).
9. I. M. Lifshitz, S. A. Gradescul, and L. A. Pastur, *Introduction to the Theory of Disordered Systems* (Nauka, Moscow, 1982; Wiley, New York, 1988).
10. N. K. Chumakov, S. V. Gudenko, V. V. Tugushev, *et al.*, J. Magn. Magn. Mater. **272**–**276**, 1351 (2004).
11. A. P. Levanyuk, V. V. Osipov, A. S. Sigov, and A. A. Sobyanin, Sov. Phys. JETP **49**, 176 (1979).
12. D. Haskel, J. W. Freeland, J. Cross, *et al.*, Phys. Rev. B **67**, 115207 (2003).
13. L. Bokacheva, W. Teizer, F. Hellman, and R. C. Dynes, Phys. Rev. B **69**, 235 111 (2004).

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 317–321.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 364–369.

ORDER, DISORDER, AND PHASE TRANSITIONS IN CONDENSED SYSTEMS

Proton Ordering in Ice at an Ice–Metal Interface

I. A. Ryzhkin[a] and V. F. Petrenko[b]

[a]*Institute of Solid-State Physics, Russian Academy of Sciences, Chernogolovka, Moscow oblast, 142432 Russia*
[b]*Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire 03755 USA*
e-mail: ryzhkin@issp.ac.ru

Received October 25, 2004

Abstract—The structure of the proton sublattice of ice at an ice–metal interface is analyzed by solving the Ginzburg–Landau equation for an order parameter describing the proton ordering under an appropriate boundary condition [1, 2]. When the interaction between protons and the substrate is weak, the ice rules that govern proton order are weaker at the interface as compared to bulk ice, but to a lesser extent than at the free ice surface. In the case of strong proton–substrate interaction (clean interface and/or high conductivity of the substrate), the ice rules are stronger at the interface as compared to bulk ice, which corresponds to a more ordered proton sublattice. The latter case corresponds to a lower concentration of defects in the proton sublattice, which determine important properties of ice, such as adhesion, electrical conductivity, plasticity, and electric field distribution near the interface. A qualitative correlation is described between electrical properties of the substrate and mechanical properties of the interface, including adhesion and friction. © *2005 Pleiades Publishing, Inc.*

Ice is one of the most widespread natural materials. Its unusual mechanical, electrical, thermal, and especially surface properties determine various aspects of our everyday life. In particular, we can hardly overestimate correct understanding and controllability of ice properties, such as adhesion, friction, charge separation, sintering, and fluidity. A detailed survey of recent results in the physics of ice can be found in [3].

Most of these properties are related to processes taking place at the free ice surface or ice–substrate interface. The studies of these processes were pioneered by Faraday in the 19th century [4]. In one of his experiments, he brought into contact two ice balls suspended on strings and observed the fast growth of a neck between the balls. The corresponding self-diffusion coefficient of water molecules was so high that Faraday hypothesized that the ice surface is covered by a quasi-liquid layer.

Shortly afterwards, Thomson put forward a new hypothesis about the nature of the quasi-liquid layer. He speculated that the quasi-liquid layer is just a film of ordinary water [5]. Its presence was explained by the special behavior of the phase diagram of water, namely, by the lowering of the melting point with increasing pressure. According to Thomson, the mechanical pressure between the balls induces local melting of ice and explains its anomalously high self-diffusion coefficient. Thomson's theory became widely known because it explained the sliding of skates at moderately low temperatures. In fact, his theory fails to provide quantitative explanation for the sintering of ice balls or sliding motion. The former conclusion was inferred by Faraday from his careful experiments on extremely weak mechanical stress between the balls. The latter is supported by available characteristics of the phase diagram of water. Actually, the pressure exerted by an adult person on ice cannot lower the melting temperature by more than 0.1°C, i.e., cannot explain sliding at –1°C or lower temperatures.

Recent studies have provided substantial evidence in support of Faraday's theory. In particular, the results of proton channeling experiments [6], X-ray diffraction [7] and NMR studies [8, 9], and measurements of low-frequency electrical conductivity [10–12] can be explained only by the existence of a surface layer whose properties substantially differ from those of ordinary water. For example, knowing the surface conductance and layer thickness, we find that the conductivity of ice is $10^{-1}\ \Omega^{-1}\ m^{-1}$, which is higher than the conductivity of ordinary water by six orders of magnitude. The theoretical model proposed in [2] explains the existence of a quasi-liquid layer near the free ice surface. It is important that the model is based on the same Hamiltonian of the proton sublattice that has been used to predict the bulk order governed by the ice rules [1]. The existence of free surface is taken into account by introducing a boundary condition for order parameter that reflects the fact that surface protons have fewer neighbors as compared to bulk protons. The solution to an appropriate Ginzburg–Landau equation subject to this boundary condition predicts a decrease in order parameter near the surface. This corresponds to an increase in both concentration of defects in the proton sublattice (local violation of the ice rules) and surface conductivity due to surface diffusivity. The increase in defect concentration to a certain critical level, in turn, can lead to instability of the oxygen sublattice.

In this paper, we apply a similar approach to analyzing proton ordering at an ice–metal interface. Our study provides qualitative explanations of surface conductivity, adhesion force, and friction coefficient of the interface. For consistency of presentation, we briefly describe the model and present the basic equations. Next, we derive boundary conditions at an ice–metal interface and solve the Ginzburg–Landau equation. Finally, we discuss the results obtained, feasible experiments, and practical applications.

In ordinary ice, protons can be located on lines connecting oxygen atoms at a distance of one angstrom from them. Since the hydrogen bond length is 2.76 Å, there are two proton sites and one proton per bond. This implies that each proton site $i\alpha$ (indices i and α label oxygen sublattice sites and hydrogen bonds, respectively) can be either occupied or vacant. Their occupation is described here in terms of pseudospin variables $\sigma_{i\alpha} = \pm 1$, where plus and minus correspond to occupied and vacant sites. The Hamiltonian of proton–proton interaction can be written in the nearest-neighbor approximation commonly used in lattice-gas models [1]:

$$H = \frac{J_1}{2}\sum_{i\alpha\beta}\sigma_{i\alpha}\sigma_{i\beta} = \frac{J_2}{2}\sum_{ij}\sigma_{i\alpha}\sigma_{j\alpha}, \quad (1)$$

where summation is performed over all pairs of nearest-neighbor sites in the proton sublattice and the coupling constants J_1 and J_2 correspond to the proton–proton interaction near an oxygen site and in a hydrogen bond, respectively. The index α in the second term varies from 1 to 4, being determined by the combination ij; it is not a summation index.

The order parameter $\mu_{i\alpha}$ is defined by the equation

$$\langle\sigma_{i\alpha}\rangle = \lambda_{i\alpha}\mu_{i\alpha}, \quad (2)$$

where the set of $\lambda_{i\alpha} = \pm 1$ must be such that the ice rules are satisfied. Note that Eq. (2) is a simple generalization of the order parameter for ferromagnets or normal antiferromagnets. Recall that $\lambda_{i\alpha} = 1$ in the case of a ferromagnet. In the case of an antiferromagnet, its sign alternates between the interpenetrating lattices. For an infinite homogeneous medium, the order parameter is independent of coordinates an can be determined by minimizing the Helmholtz free energy or solving the self-consistent equations

$$\langle\sigma_{i\alpha}\rangle = -\tanh\left(a_1\sum_{\beta\neq\alpha}\langle\sigma_{i\beta}\rangle + a_2\langle\sigma_{j\alpha}\rangle\right), \quad (3)$$

where $a_i = J_i/kT$. At extremely low temperatures, the order parameter is close to unity, which implies that the occupancy is close to unity and zero for sites with $\lambda_{i\alpha} = +1$ and –1, respectively. The order parameter decreases with increasing temperature and vanishes at a certain critical point T_c. In the critical state, the proton site occupancy is 0.5; i.e., the ice rules are disobeyed. Since bulk ice with a totally frustrated proton sublattice does not exist under normal conditions (at zero external pressure), it is highly plausible that this transition in the proton sublattice destabilizes the oxygen sublattice, which leads to the actual melting of ice. For this reason, we assume that $T_c \approx T_m$, as in [2]. The presence of an interface makes ice an inhomogeneous medium, and the corresponding order parameter depends on coordinates. In the phase transition region, it is a slowly varying function that can be found by solving the Ginzburg–Landau equation [1, 2]

$$-c\nabla^2 m(\mathbf{r}) + (1 - a_1 - a_2)m(\mathbf{r}) + \frac{1}{3}m^3(\mathbf{r}) = 0, \quad (4)$$

where $c = a_2 r_{HH}(r_{OH} + r_{HH})/3$ and $m(\mathbf{r})$ is the macroscopic order parameter (averaged over interatomic distances).

Equation (4) must be supplemented with boundary conditions. In the case of a semi-infinite crystal with free boundary, the conditions on the surface and in the bulk ice are written as follows [2]:

$$r_{HH}\frac{dm(0)}{dz} = m(0), \quad m(\infty) = \sqrt{\frac{3(1-\tau)}{\tau}}, \quad (5)$$

where $z > 0$ is the domain occupied by ice and $\tau = T/T_c$. To derive the boundary condition on the ice–metal interface, we recall the reasoning behind the derivation of the first equation in (5). Since surface pseudospins do not have any neighbors belonging to other water molecules, they satisfy the self-consistent equations

$$\langle\sigma_{i\alpha}\rangle = -\tanh\left(\tilde{a}_1\sum_{\beta\neq\alpha}\langle\sigma_{i\beta}\rangle\right). \quad (6)$$

This means that the Ginzburg–Landau equation that follows from (3) is not valid on the surface. However, we can assume that the Ginzburg–Landau equation is valid everywhere (even on surface) if the order parameter satisfies an additional condition entailed by equivalence between Eqs. (3) and (6). Note that a similar condition leads to the first equation in (5) in the case of free surface. For an ice–metal interface, the following equation is obtained instead of (3) or (6):

$$\langle\sigma_{i\alpha}\rangle = -\tanh\left(\tilde{a}_1\sum_{\beta\neq\alpha}\langle\sigma_{i\beta}\rangle - \tilde{a}_2\langle\sigma_{i\alpha}\rangle\right). \quad (7)$$

The tildes in $\tilde{a}_1$ and $\tilde{a}_2$ indicate the difference between these constants and their analogs for bulk ice due to the surface relaxation of the oxygen sublattice. Whereas the second term vanishes in the case of free surface, its finite value in (7) represents the interaction between a proton and the electron gas in metal. It is treated here as the interaction between a proton and a corresponding mirror image charge. Indeed, the occu-

pancy of the site $i\alpha$ can be expressed in terms of the pseudospin $\sigma_{i\alpha}$ as $n_{i\alpha} = (\sigma_{i\alpha} + 1)/2$. The energy $-\tilde{J}_2 n_{i\alpha}$ of interaction between an occupied site and the corresponding image charge can easily be represented as a function of $\sigma_{i\alpha}$. Then, assuming that the dependence on coordinates is weak and using relation (2), we obtain

$$r_{\mathrm{HH}}\frac{dm(0)}{dz} = \left(1 - \frac{\tilde{J}_2}{J_2}\right)m(0), \tag{8}$$

which differs from the first equation in (5) only by the factor $1 - \tilde{J}_2/J_2$. It should be noted that this factor can be negative in the case of good contact between ice and metal.

The Ginzburg–Landau equation, boundary condition (8), and the condition in the bulk ice are rewritten in dimensionless form as

$$\begin{aligned} &\frac{d^2 f}{dt^2} + f(t) - f^3(t) = 0, \\ &\frac{df(0)}{dt} = \beta f(0), \quad f(\infty) = 1. \end{aligned} \tag{9}$$

Here,

$$f(t) = \frac{m}{m(\infty)}, \quad t = \frac{z}{\xi}, \tag{10}$$

and the constants are defined as

$$\begin{aligned} \alpha_0 &= \sqrt{\frac{J_2(r_{\mathrm{OH}}/r_{\mathrm{HH}} + 1/2)}{3(J_1 + J_2)}}, \\ \beta &= \frac{\alpha_0(1 - \tilde{J}_2/J_2)}{\sqrt{1-\tau}}, \quad \xi = \frac{\alpha_0 r_{\mathrm{HH}}}{\sqrt{1-\tau}}. \end{aligned} \tag{11}$$

The solution to (9), (10) can be represented as

$$f(t) = \tanh(t/\sqrt{2} + c), \quad \beta \geq 0, \tag{12}$$

$$f(t) = \cosh(t/\sqrt{2} - c), \quad \beta < 0, \tag{13}$$

where $\sinh(2c) = \sqrt{2}/\beta$.

Curve *3* in Fig. 1 shows a solution given by (12) when the interaction between ice and metal is weak ($\tilde{J}_2 < J_2$). In this case, the ice–metal interaction effectively reduces the constant β, i.e., disorder in the proton system. In other words, the ice rules are weaker at the interface as compared to the bulk ice, but to a lesser extent than at a free crystal surface (curve *4* in Fig. 1). As the ice–substrate interaction increases, a state with $\tilde{J}_2 = J_2$ can be reached. The corresponding solution has a trivial form: the order parameter is independent of the spatial coordinate (curve *2* in Fig. 1). The ice rules are uniform throughout the crystal in this state; i.e., there is

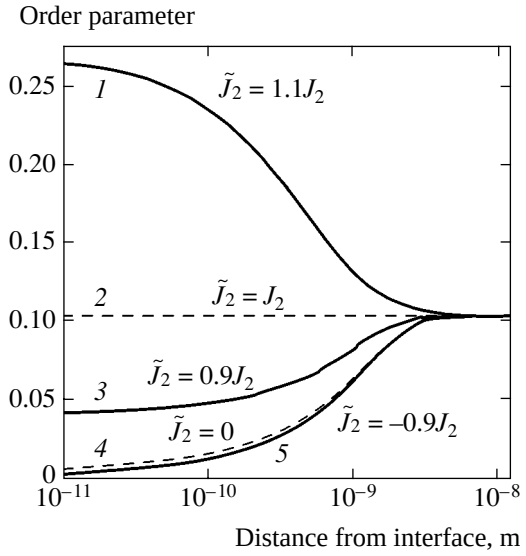

Fig. 1. Order parameter in surface layer vs. distance from interface for several values of the proton–substrate interaction constant and $J_2 = 0.4$ eV at $T = 272$ K and $T_c = T_m = 273$ K.

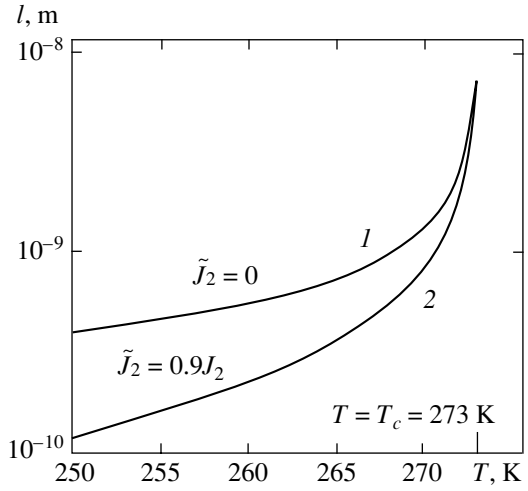

Fig. 2. Quasi-liquid layer thickness vs. temperature for free surface and interface with weak proton–substrate interaction at $T = 272$ K. The critical point where the quasi-liquid layer thickness is divergent ($T_c = 273$) is indicated on the temperature axis.

no quasi-liquid layer. When the interaction is so strong that $\tilde{J}_2 > J_2$, solution (13) is valid, which predicts that the order parameter is greater (i.e., the ice rules are stronger) in a surface layer as compared to the bulk crystal (curve *1* in Fig. 1). Finally, curve *5* corresponds to a negative constant of proton–substrate interaction (hydrophobic coating). The curves presented here correspond to $T_c = T_m = 272$ K; $\alpha_0 = 0.6$; $\tilde{J}_2/J_2 = 0.9$, 1.1, and –0.9 for weak and strong interactions and hydrophobic coating, respectively; and the absolute value of

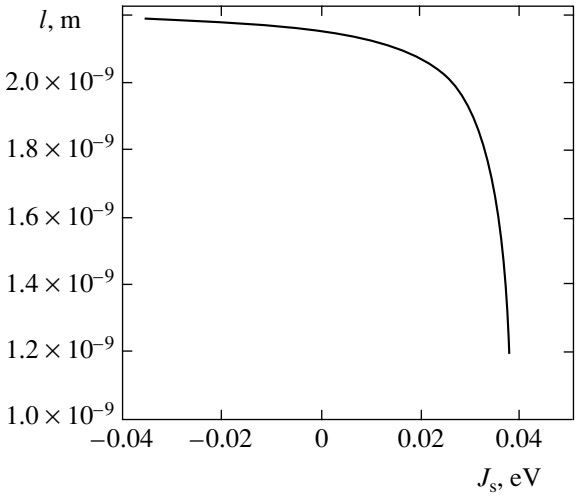

Fig. 3. Quasi-liquid layer thickness vs. proton–substrate interaction constant for $-J_2 < J_s < J_2$ ($J_2 = 0.4$ eV) at $T = 272$ K.

J_2 equal to 0.4 eV. Figure 2 shows the quasi-liquid layer thickness l as a function of temperature. The boundary of the quasi-liquid layer is defined as the point where the order parameter decreases to 0.9 times its bulk value. Curves *2* and *1* correspond to $\tilde{J}_2/J_2 = 0.9$ and $\tilde{J}_2 = 0$ (i.e., to curves *3* and *4* in Fig. 1, respectively). Figure 3 shows the quasi-liquid layer thickness plotted versus the constant of proton–substrate interaction J_s ($J_s \equiv \tilde{J}_2$). It should be noted that the theory predicts continuous curves only for $l \geq a$. In the case of free ice surface (curve *1* in Fig. 2), this condition corresponds to temperatures ranging from 248 to 273 K; in the case of weak interaction with the substrate (curve *2* in Fig. 2), the corresponding temperature interval lies between 260 and 273 K. This behavior appears to be analogous to macroscopic electrodynamics, in which the field is averaged over regions larger than interatomic distances, but the macroscopic permittivity can be used to obtain adequate results even when applied to regions comparable to interatomic distances.

It follows from the results obtained here that the existence and characteristics of a quasi-liquid layer depend on the single parameter $\tilde{J}_2/J_2$ determined by the properties of the substrate, which is very difficult to calculate form the first principles. However, some qualitative results can be obtained by using the physical interpretation of $\tilde{J}_2$ as a constant of interaction with an image charge. According to the laws of electrostatics, both image charge and interaction energy are proportional to $(\epsilon_m - \epsilon_i)/(\epsilon_m + \epsilon_i)$, where ϵ_i and ϵ_m denote the permittivities of ice and substrate material, respectively. In the case of a perfect metal, we have $\epsilon_m \longrightarrow \infty$, and both interaction energy and parameter $\tilde{J}_2/J_2$ have maximum values for a given distance between ice and substrate.

When $\epsilon_i = \epsilon_m$, the image charge vanishes, and so does the interaction constant, as in the case of free ice surface (curve *4* in Fig. 1). Note that the following argumentation seems plausible when the permittivities are equal. Consider a plane in bulk ice. Since it can be viewed as an ice–ice interface (i.e., an interface separating two materials characterized by equal permittivities), it may seem that a quasi-liquid layer must exist on this boundary by analogy with free crystal surface considered above. Indeed, there is no image charge in the case of an ice–ice interface. However, we must take into account the long-range interaction between real protons, whereas neither image charges nor neighboring protons exist in the substrate when $\epsilon_i = \epsilon_m$. Thus, the argumentation above does not lead to any physical paradox. It should also be noted that we allow only for direct electrostatic interaction between protons and substrate electrons, which is generally stronger than both short-range interactions responsible for chemical bonding and van der Waals interactions [13].

Finally, we discuss the physical realizability of a system with $\tilde{J}_2 < 0$. It is obvious that a constant $\tilde{J}_2$ corresponding to interaction with an image charge cannot change sign, since the sign of the image charge must be opposite to that of the proton charge. However, we can consider a system with positive charges on the substrate surface. For instance, it may be an ionic material that has been subjected to special treatment or a molecular material with outward-pointing dipole moments of surface molecules. In this case, the interaction with positive charges is stronger than the interaction with image charges and the constant $\tilde{J}_2$ can be negative. The proton disorder at an ice–substrate interface of this kind must be higher than at the free ice surface (curve *5* in Fig. 1). Moreover, the quasi-liquid layer at such an interface must persist at lower temperatures.

Proton disorder enhances the self-diffusion of water molecules and dislocation motion and can even cause ice to melt. Therefore, materials with permittivity comparable to the high-frequency permittivity of ice or with positive surface charges (i.e., those characterized by tendency toward formation of a quasi-liquid layer) must be least adhesive to ice and have the lowest friction coefficients. The presence of a quasi-liquid layer must manifest itself not only in mechanical phenomena, but also in measurements of low-frequency conductivity. Indeed, proton disorder can be interpreted as an elevated concentration of ionic defects, i.e., charge carriers in ice. As noted in the Introduction, the conductivity calculated by rescaling the extra conductance to the quasi-liquid layer thickness is $10^{-1}\ \Omega^{-1}\ \mathrm{m}^{-1}$. We believe that this result calls for a careful quantitative verification. Indeed, as the critical temperature is approached, disorder increases not only in the surface layer, but also in bulk ice (see second equation in (5)).

When this effect is ignored, the surface conductivity may be overpredicted. This may explain the fact that the quasi-liquid layer thickness inferred from electrical measurements is always higher than that determined by other experimental methods. The contribution of bulk ice to the extra conductivity can be separated from the surface contribution, since it must depend on sample thickness. Both studies of the quasi-liquid layer and verification of their results can be advanced by using electrical measurements with live or grounded intermediate electrodes (guard rings). Guard rings made from materials characterized by strong electrostatic interaction with ice would reduce its surface conductivity to zero even without applied voltage. These techniques can also be used to test materials for low adhesion and low friction coefficients.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 01-02-16459 and by ARO under the Earth and Environmental Sciences Program.

REFERENCES

1. I. A. Ryzhkin and V. F. Petrenko, Phys. Rev. B **62**, 11280 (2000).
2. I. A. Ryzhkin and V. F. Petrenko, Phys. Rev. B **65**, 12205 (2002).
3. V. F. Petrenko and R. W. Whitworth, *Physics of Ice* (Oxford Univ. Press, Oxford, 1999).
4. M. Faraday, Philos. Mag. **17**, 162 (1859).
5. J. Thomson, Proc. R. Soc. London, Ser. A **11**, 198 (1861).
6. I. Golecki and C. Jaccard, Phys. Lett. A **63A**, 374 (1977).
7. A. Kouchi, Y. Furukawa, and T. Kuroda, J. Phys. C1 **48**, 675 (1987).
8. V. I. Kvlividze, V. F. Kiselev, and L. A. Ushakova, Dokl. Akad. Nauk SSSR **191**, 1088 (1970).
9. Y. Mizuno and N. Hanafuza, J. Phys. C1 **48**, 511 (1987).
10. N. Maeno and H. Nishimura, J. Glaciol. **21**, 193 (1978).
11. N. Maeno, in *Physics and Chemistry of Ice*, Ed. by E. Whalley, S. J. Jones, and L. W. Gold (Royal Society of Canada, Ottawa, 1973), p. 140.
12. M. A. Maidique, A. von Hippel, and W. B. Westphal, J. Chem. Phys. **54**, 150 (1971).
13. V. F. Petrenko and I. A. Ryzhkin, J. Phys. Chem. B **101**, 6285 (1997).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 322–330.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 370–380.

ORDER, DISORDER, AND PHASE TRANSITIONS IN CONDENSED SYSTEMS

The Possibility of the "Giant" Isotope Effect for Ultrasound Absorption in Crystals

I. G. Kuleev[a,*], I. I. Kuleev[a], A. V. Inyushkin[b], and V. I. Ozhogin[b]

[a] *Institute of Metal Physics, Ural Division, Russian Academy of Sciences, Yekaterinburg, 620219 Russia*
[b] *Institute of Molecular Physics, Russian Research Centre Kurchatov Institute, pl. Kurchatova 1, Moscow, 123182 Russia*
**e-mail: kuleev@imp.uran.ru*

Received December 21, 2004

Abstract—The effect of isotopic disorder on ultrasound absorption in germanium, silicon, and diamond crystals is considered. The "giant" isotope effect is predicted in the ultrasound absorption coefficient (in contrast to the isotope effect in the thermal conductivity and thermopower) of these crystals. The parameters determining the ultrasound absorption coefficients for the crystals under study with different degrees of isotopic disorder are determined from the known values of elastic moduli of the second and third order. The ultrasound absorption coefficients are analyzed as functions of temperature and wavevector for isotopically modified crystals. The possibility of experimental observation of this effect is considered. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

The properties of isotopically enriched germanium, silicon, and diamond crystals have been attracting the attention of researches over the last decade [1–11]. Like many other crystals, these crystals consist of two or more stable isotopes distributed at random over crystal lattice sites. The parameters of the force interactions between atoms in isotopically disordered crystals are virtually independent of the isotopic composition since the electron shell configuration of an atom weakly depends on the nuclear mass (the scale of the effect is on the order of the ratio of the electron and nuclear masses, i.e., $\sim 10^{-4}$–10^{-5}). The dependence of atomic vibrational modes on the isotopic composition emerges due to the fact that the motion of an atom in the potential produced by neighboring atoms is determined, among other things, by its mass M_i. Atomic vibrations at lattice sites can be treated as motion in a harmonic potential whose parameters depend on the unit cell volume of the crystal; in the quasi-harmonic approximation, the frequency and the squared amplitude of atomic vibrations are proportional to $M_i^{-1/2}$. As a rule, the effect of isotopic disorder on the vibrational spectra of the crystals is weak in view of the smallness of the ratio $|\Delta M/\overline{M}|$, where $\Delta M = M_i - \overline{M}$, M_i is the mass of the ith isotope, $\overline{M} = \sum_i C_i M_i$ is the average mass of the isotopic composition, and C_i is the concentration of the ith isotope. However, some kinetic coefficients may change significantly depending on the isotopic composition since the isotopic disorder violates translational invariance of the lattice and leads to phonon scattering. The effect of isotopic disorder on phonons can be described in terms of temperature-independent isotope scattering of phonons. The relaxation rate of this process is proportional to the fourth power of the phonon wavevector and the isotopic disorder factor

$$g = \sum_i C_i \left(\frac{\Delta M_i}{\overline{M}}\right)^2. \quad (1)$$

Upon a transition from germanium crystals with the natural isotopic composition ^{nat}Ge ($g = 5.87 \times 10^{-4}$) to crystals enriched to 99.99% in the ^{70}Ge isotope ($g = 0.816 \times 10^{-8}$), the value of g decreases almost by four orders of magnitude. However, according to [2, 3], the maximal (in T) values of thermal conductivity increases by an order of magnitude, while the absolute values of thermopower [7] increase by more than two times. For silicon crystals enriched to 99.983% in the ^{28}Si isotope ($g = 3.2 \times 10^{-7}$), the maximal values of thermal conductivity increase, according to [4, 5], by a factor of 7.5 as compared to its value for ^{nat}Si with the natural isotopic composition ($g = 2.01 \times 10^{-4}$). It follows from these results that the observed increase in the maximal values of thermal conductivity and thermopower is much smaller than the ratio of the isotopic disorder factors for natural and enriched compositions of both germanium and silicon. This is due to the fact that thermal conductivity is determined by the total time of phonon relaxation averaged over all thermally excited phonon states. The contribution to this quantity comes not only from isotope scattering, but also from the phonon–phonon scattering processes associated with anharmonism in lattice vibrations, phonon scattering at the sample surface, etc.

Here, we pay attention to another kinetic effect, namely, ultrasound absorption which can be affected by

scattering of acoustic waves from isotopic disorder more strongly than thermal conductivity and thermopower.

The absorption coefficient $\alpha_\lambda(\mathbf{q})$ of ultrasound waves with wavevector $\mathbf{q}$ and polarization λ is proportional to the phonon relaxation rate $\nu_{ph}^{\lambda}(\mathbf{q}, T)$ (see, for example, [12, 13]),

$$\alpha_\lambda(\mathbf{q}, T) = \frac{\nu_{ph}^{\lambda}(\mathbf{q}, T)}{2s_\lambda}, \quad (2)$$

where s_λ is the velocity of phonons and T is the temperature. This expression implies that the procedure of averaging over phonon modes is not involved in determination of absorption coefficient $\alpha_\lambda(\mathbf{q}, T)$. Expression (2) is valid when the inequality $\omega_{q\lambda}\tau_{ph}^{\lambda}(\mathbf{q}, T) \gg 1$ holds, where $\tau_{ph}^{\lambda}(\mathbf{q}, T) = 1/\nu_{ph}^{\lambda}(\mathbf{q}, T)$. For this reason, we will confine subsequent analysis to the range of temperatures and wavevectors $\mathbf{q}$, in which this inequality holds. The dominant contribution to attenuation of acoustic waves comes from the boundary scattering of phonons, scattering from defects (including isotope scattering) and normal phonon–phonon scattering processes (see, for example, [12]). We will confine our analysis to these relaxation processes. Experimental studies of ultrasound absorption [12, 13] show that the inequality $\omega_{q\lambda}\tau_{ph}^{\lambda}(\mathbf{q}, T) \gg 1$ holds at low temperatures (namely, at temperatures lower than 50, 100, and 400 K for germanium, silicon, and diamond crystals, respectively). Obviously, phonon–phonon Umklapp scattering processes are ineffective in these temperature regions. In the opposite limiting case $\omega_{q\lambda}\tau_{ph}^{\lambda}(\mathbf{q}, T) \ll 1$ corresponding to the Akhiezer absorption mechanism, it is more appropriate to treat phonons as quasiparticles moving in a smoothly varying field of an acoustic wave. In this case, we must take into account the phonon–phonon Umklapp processes [12].

It should be noted that isotope scattering is independent of temperature and makes a constant contribution to the ultrasound absorption coefficient. The relaxation rates of phonon modes in anharmonic scattering processes rapidly decrease upon cooling; consequently isotope scattering at low temperatures can make the main contribution to attenuation of acoustic waves. In this case, for perfect crystal we have

$$\alpha_\lambda(g_i) \approx \alpha_{0\lambda} g_i. \quad (3)$$

Since the isotopic disorder factor decreases by more than 7000 (!) times upon a transition from ^{nat}Ge to ^{70}Ge (99.99%), the magnitude of the isotope effect for the ultrasound absorption coefficient may reach giant values in contrast to isotope effects for thermal conductivity and thermopower.

One more circumstance that considerably increases the value of experimental study of ultrasound absorption in isotopically modified crystals is also worth noting. Namely, such structurally perfect and chemically pure crystals may exhibit fine features of phonon mode relaxation associated with anharmonism of lattice vibrations.

2. ISOTOPE EFFECT IN TRANSVERSE ULTRASOUND ABSORPTION IN GERMANIUM, SILICON, AND DIAMOND CRYSTALS

Let us consider by way of example the transverse sound absorption coefficient. In accordance with the prevailing idea [1–11, 14–16], the main mechanism of transverse ultrasound relaxation in normal three-phonon scattering processes is the Landau–Rumer mechanism, in which a longitudinal phonon is generated as a result of merging of a transverse and a longitudinal phonon ($T + L \longrightarrow L$) [17]. This relaxation mechanism is regarded as the main mechanism for transverse thermal phonons as well [1–11, 14–16]. It should be noted that the matrix element of three-phonon scattering processes for the relaxation mechanism of transverse phonons ($T \longleftrightarrow T + T$) is shown to vanish in the model [18] of an isotropic medium.

For this reason, we will henceforth assume that isotope scattering and normal phonon scattering processes in the Landau–Rumer mechanism make the dominant contribution to relaxation of transverse sound. In this case, in accordance with the Matthiesen rule, we have

$$\nu_{ph}^{T} \approx \nu_{phN}^{TLL} + \nu_{phi}^{T}, \quad (4)$$

where the transverse phonon relaxation rate for scattering upon isotopic disorder is given by

$$\nu_{phi}^{T} \approx B_i T^4 z_1^4, \quad z_1 = z_{qT} = \frac{\hbar\omega_q T}{k_B T},$$
$$B_i = g\left(\frac{\hbar}{k_B}\right)^4 \frac{V_0}{4\pi s_T^3}. \quad (5)$$

Here, V_0 is the volume per atom and s_T is the velocity of transverse photons.

The transverse phonon relaxation rate in the Landau–Rumer mechanism for cubic crystals has the form

$$\nu_{phN}^{TLL} = T^5 z_1 B^T(z_1, T),$$
$$B^T(z_1, T) = \frac{k_B^5 S^{*2}(1 - S^{*2})}{16\pi\hbar^4\rho^3 s_T s_L^8} \frac{\sinh(z_1/2)}{(z_1/2)} J_z(z_1, T). \quad (6)$$

Coefficient $B^T(z_1, T)$ was defined in [19]. In the long-wave limit ($z_1/z \ll 1$), at temperatures much lower than the Debye temperature, expression (6) is transformed into the Landau–Rumer formula [17] with coefficient

Table 1. Thermodynamic elastic moduli for Ge, Si, and diamond crystals according to [19, 20], in normalization adopted in [21] in units of 10^{12} dyne/cm^2

	c_{11}	c_{12}	c_{44}	c_{111}	c_{112}	c_{123}	c_{144}	c_{155}	c_{456}
Germanium	1.289	0.483	0.671	–7.10	–3.89	–0.18	–0.23	–2.92	–0.53
Silicon	1.657	0.638	0.796	–8.25	–4.51	–0.64	0.12	–3.10	–0.64
Diamond	10.76	1.25	5.758	–62.6	–22.6	1.12	–6.74	–28.6	–8.23

Table 2. Parameters ΔC, $\tilde{c}_{112}$, $\tilde{c}_{111}$, $\tilde{c}_{155}$, A, and B_{T0} for Ge, Si, and diamond crystals in units of 10^{12} dyne/cm^2

	ΔC	$\tilde{c}_{112}$	$\tilde{c}_{111}$	$\tilde{c}_{155}$	A	B_{T0} (s^{-1} K^{-5}) [100]
Germanium	–0.54	–3.25	28.01	–1.63	–0.08	1.2
Silicon	–0.57	–4.1	32.4	–1.9	0.71	0.099
Diamond	–2.01	–10.24	138.1	–5.4	–27.9	0.0063

$B^T = B_{T0}$, which is independent of the phonon wavevector and temperature and has the following form for cubic crystals:

$$B_{T0} = \pi^3 k_B^5 S^{*2}(1-S^{*2})[15\hbar^4\rho^3 s_T s_L^8]^{-1}$$
$$\times\{A^2 + A[2S^{*2}(2\tilde{c}_{155}+\Delta C) + 3\tilde{c}_{155}(1-S^{*2})] + (2\tilde{c}_{155}+\Delta C)^2(S^{*4}) \quad (7)$$
$$+ 3(2\tilde{c}_{155}+\Delta C)\tilde{c}_{155}S^{*2}(1-S^{*2}) + 2.5\tilde{c}_{155}^2(1-S^{*2})^2\}.$$

Here,

$$\Delta C = c_{11} - c_{12} - 2c_{44}, \quad \tilde{c}_{155} = c_{155} - c_{144} - 2c_{456},$$
$$S^* = \frac{s_T}{s_L}, \quad A = c_{12} + 3c_{44} + 2c_{144} + 4c_{456},$$

s_L and s_T are the velocities of longitudinal and transverse phonons and c_{ij} and c_{ijk} are thermodynamic elastic moduli of the second and third orders. In contrast to [19], where the Tucker and Rampton normalization [13] was used for the elastic moduli, we use the normalization adopted in [20–22]. A transition to the Tucker and Rampton normalization [13] is made by substituting $c_{ijk}^{TR} = (1/6)c_{ijk}$. It should be noted that elastic moduli ΔC and $\tilde{c}_{155}$ are typical of cubic crystals and distinguish these crystals from an isotropic medium. The equality to zero of these moduli indicates a transition to the model of an isotropic medium. Consequently, the first term in formula (7) corresponds to isotropic scattering; the second term corresponds to interference between isotropic and anisotropic scattering, while the three last terms correspond to the contribution associated with cubic anisotropy.

The values of third-order thermodynamic moduli were determined experimentally in [20, 21] in the normalization adopted by Brugger [22] and are given in Table 1. A transition to the Birch normalization c_{ijk}^B [23] was carried out using the substitution

$$c_{111} = 6c_{111}^B, \quad c_{112} = 2c_{112}^B, \quad c_{123} = 2c_{123}^B,$$
$$c_{144} = \frac{1}{2}c_{144}^B, \quad c_{155} = \frac{1}{2}c_{155}^B, \quad c_{456} = \frac{1}{4}c_{456}^B. \quad (8)$$

It should be noted that the coefficient of transitions to the Birch normalization for the thermodynamic elastic modulus c_{456} given in review [20] is erroneous (1/8 instead of 1/4; see [22, 24]). This coefficient was used in [19] for calculating the phonon relaxation rates in the Landau–Rumer mechanism. To correct this inaccuracy, it is necessary to double the value of modulus c_{456} given in [19] (Table 1) in the Tucker–Rampton normalization [13]. This correction does not qualitatively affect the results obtained in [19] for relaxation rates in Ge, Si, and diamond crystals (cf. Figs. 3a, 3b, and 5 for $g = 0$ in the present paper and Fig. 4 from [19]). The quantitative relations between the values of parameters A and $\tilde{c}_{155}$ characterizing the contributions of isotropic and anisotropic scattering change in this case, as well as the relations between different contributions to the relaxation frequencies for the crystals under study. The corrected values for calculated parameters ΔC, A, and $\tilde{c}_{155}$ are given in Table 2. Our estimates show that the contribution from isotropic scattering in the crystallographic [100] direction amounts approximately to 0.1% for germanium, 6% for silicon, and 52% for diamond; the interference term gives 5% for germanium and 40% for diamond, while the contribution from anisotropic scattering amounts to approximately 95% for germa-

nium and 8% for diamond. Thus, germanium is the most anisotropic crystal among those considered here. The interference term of Ge and diamond crystals is positive since quantities A and $\tilde{c}_{155}$ have the same sign (see Table 2), while quantities A_{cub} and $\tilde{c}_{155}$ for Si crystals have opposite signs. Interference scattering makes a negative contribution to the transverse phonon relaxation rate, its absolute value being one third of the contribution from anisotropic scattering.

Table 2 contains the corrected values for parameters B_{T0}. It turns out that the values of coefficients B_{T0} predicted by the theory of elasticity for Ge and Si crystals are approximately 40 times the values of the fitting parameters B_{T0} used in [3–7, 25, 26] for analyzing the thermal conductivity of these crystals with various isotopic compositions (see Table 2 in [26]). It should be noted that the relaxation rate $\nu_{\text{ph}N}^{TLL}(z_1, T)$ for thermal phonons ($z_1 > 1$) turns out to be much smaller than that predicted by the Landau–Rumer theory for Ge, Si, and diamond crystals in the temperature range $1 < T < 100$ K. Consequently, the effective value of parameter B_{T0} for thermal phonons is substantially smaller than predicted by the theory of elasticity in the long-wave approximation (see [19]). However, even if we take this effect into account, the effective relaxation rates $\nu_{\text{ph}N}^{TLL}$ averaged over thermal phonons and calculated in the framework of the theory of elasticity for Ge and Si crystals with different isotopic compositions were found to be an order of magnitude higher than those obtained using fitting parameters B_{T0} [3–5, 25, 26]. This discrepancy between the results obtained using the theory of elasticity and the results of meticulous fitting of experimental data on the thermal conductivity of Ge crystals with different isotopic compositions [3–5, 25, 26] called for additional experimental investigations of ultrasound absorption coefficients in isotopically modified crystals of germanium, silicon, and diamond.

It should be noted that, when the temperature dependence of thermal conductivity is approximated in the framework of the relaxation method [3–16], expressions for the phonon relaxation rates for anharmonic scattering processes derived in the long-wave approximation for the isotropic medium model are conventionally used. It was shown in [19] that both these approximations are inapplicable for calculating the thermal conductivity of Ge, Si, and diamond crystals of cubic symmetry with noticeable elastic anisotropy of both second and third orders. It is hardly appropriate to speak of unambiguity of the procedure for finding the parameters that determine phonon relaxation rates in anharmonic scattering processes, which are obtained using incorrect approximations with variation of four (!) fitting parameters of the theory. A more reliable estimate of relaxation frequencies can be obtained by calculating the phonon relaxation rates in anharmonic scattering processes from experimentally determined second- and third-order elastic moduli for Ge and Si crystals since cubic anisotropy of the crystal is taken into account both for harmonic and anharmonic energies. Such calculations make it possible to find the absorption coefficient of both long-wave and short-wave ultrasound. In addition, effective processes of relaxation of thermal phonons of various polarizations can be determined and the number of fitting parameters of the theory required for calculating the thermal conductivity of these crystals can be noticeably reduced. In this respect, experimental investigations of the absorption coefficients of longitudinal and transverse ultrasound in isotopically modified crystals of germanium, silicon, and diamond are of vital importance, as well as calculations of the relaxation rates for thermal and high-frequency phonons of various polarizations in anharmonic scattering processes proceeding from experimentally determined values of elastic moduli of the second and third orders. It should be noted that the model of anisotropic continuum is confined to phonon wavevectors much smaller than the Debye wavevectors The energies $\hbar\omega_{qt}$ of such phonons in k_B units are lower than 50, 100, and 400 K for Ge, Si, and diamond, respectively.

To analyze the dependence of relaxation rate $\nu_{phN}^{TLL}(z_1, T)$ on the reduced wavevector z_1 and temperature, it is convenient to introduce the quantities $\nu_{TLL}^{*}(z_1, T) = \nu_{\text{ph}N}^{TLL}(z_1, T)/(B_{T0}T^5)$; then, in the Landau–Rumer approximation, we obtain

$$\nu_{TLL}^{*LR}(z_1) = z_1. \tag{9}$$

The relaxation rates calculated in [19] for Ge, Si, and diamond crystals in the [100] directions showed that, in contrast to isotropic media, their dependences on the phonon wavevector are nonmonotonic with two peaks, the second of which lies in the range of high-frequency phonons, $\hbar\omega_{qT} \gg k_B T$. These features of relaxation rates $\nu_{TLL}^{*}(z_1, T)$ are associated with the angular dependence of the anharmonic scattering probability and anisotropy in the elastic properties of the crystals under investigation. However, we disregarded in [19] the effect of isotope scattering as well as scattering from defects, which can mask the features associated with anharmonic scattering processes at low temperatures. In contrast to [19], we will analyze here the effect of isotope scattering on the transverse ultrasound absorption coefficient. We will prove that the above-mentioned features of phonon relaxation rate in the Landau–Rumer mechanism cannot be observed in Ge, Si, and diamond crystals with a natural isotopic composition. However, these features can be manifested in isotopically enriched crystals; accordingly, the dependences of the ultrasound absorption coefficients on the phonon wavevector for such crystals have a qualitatively different form.

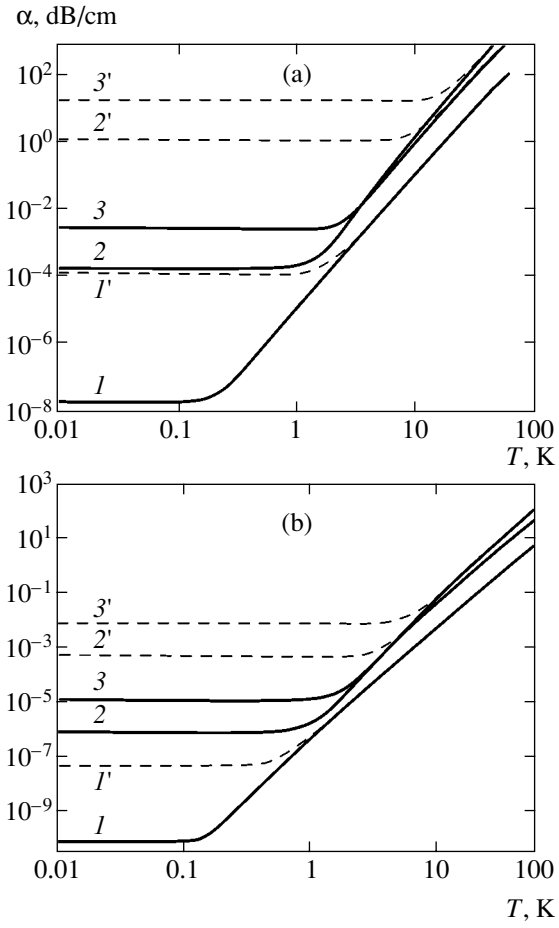

Fig. 1. Temperature dependence of the ultrasound absorption coefficient for (a) Ge and (b) Si crystals in the [100] crystallographic direction for a fixed energy of ultrasonic quantum: T_q = 1 K (curves *1* and *1*'), 10 K (curves *2* and *2*'), and 20 K (curves *3* and *3*'). Dashed curves correspond to ^{nat}Ge and ^{nat}Si crystals and solid curves, to ^{70}Ge (99.99%) and ^{28}Si (99.983%) crystals.

Let us analyze the temperature dependence of the absorption coefficient of transverse ultrasound for a fixed wavevector in crystals with different isotopic compositions. Formulas (2) and (4)–(6) imply that

$$\alpha_T(z_1, T) = \frac{8.68(B_{T0}\nu_{TLL}^{*}T^5 + B_i T_q^4)}{2s_T}\ \text{dB/cm}, \qquad T_q \equiv \frac{\hbar\omega_{qT}}{k_B}, \tag{10}$$

where T_q is the energy of an ultrasonic quantum in kelvins. Figure 1 shows the temperature dependences of the transverse ultrasound absorption coefficient for crystals ^{nat}Ge and ^{70}Ge (99.99%) as well as ^{nat}Si and ^{28}Si (99.983%) at a fixed energy of an ultrasonic quantum. It can be seen from the figure that the isotope effect attains high values only for high-energy phonons since the Rayleigh scattering from isotopic disorder dominates precisely for such phonons. Different heights of plateaus at the origin indicate the role of isotope scattering. It should be noted that coefficient α tends to a constant typical of isotope scattering for $T < T_q$ and follows the T^4 dependence typical of the Landau–Rumer mechanism for $T \gg T_q$. The maximal isotope effect is observed in the temperature range $T < T_q$ ($T_q/T = z_1 > 1$). In this case, the ratio $\alpha(g_{nat})/\alpha(g_{99.99\%})$ of the ultrasound absorption coefficients for Ge crystals at $T < T_q$ tends to the ratio of the isotopic disorder factors $g_{nat}/g_{99.99\%}$ (which is approximately 7000 for Ge). This rough estimate shows that the isotope effect in the ultrasound absorption coefficient may reach "giant" values for high frequencies and low temperatures.

In experiments on ultrasound absorption in crystals, the following question arises above all: in what crystallographic direction is the magnitude of the isotope effect in sound absorption maximal in the presence of competition between isotopic and anharmonic scattering processes? For this reason, we will consider the effect of cubic anisotropy in elastic properties on the isotope effect in absorption of transverse ultrasound using Ge crystals as example. It follows from formula (10) that the magnitude of the isotope effect in the sound absorption coefficient in the presence of competition between isotopic and anharmonic scattering processes is maximal in the crystallographic direction, for which coefficient B_i has the maximal value, while coefficient B_{T0} has the minimal value. In accordance with formula (5), coefficient B_i has the largest value for the direction in which the velocity of transverse phonons is minimal. For example, for Ge crystals, we have $s_T([100]) = 3.55 \times 10^5$ cm/s, $s_T([111]) = 3.04 \times 10^5$ cm/s, and $s_{T1}([110]) = 2.75 \times 10^5$ cm/s for a slow transverse wave and $s_{T2}([110]) = 3.55 \times 10^5$ cm/s for a fast transverse wave [11]. If we take the velocity of sound for the slow transverse mode in the [110] direction, the ratio of coefficients $b_i = B_i(^{nat}\text{Ge})/B_i(^{70}\text{Ge}(99.99\%))$ for Ge in different crystallographic directions has the form $b_i([100]) : b_i([110]) : b_i([111]) = 1.7 : 3.7 : 2.7$. According to estimates, coefficient B_{T0} is minimal in the [111] direction, while the ratio of coefficients B_{T0} for various crystallographic directions has the form $B_{T0}([100]) : B_{T0}([110]) : B_{T0}([111]) = 1.2 : 0.46 : 0.37$.

Numerical analysis of the ratio of the ultrasound absorption coefficients $b_\alpha = \alpha(^{nat}\text{Ge})/\alpha(^{70}\text{Ge}(99.99\%))$, in accordance with formula (10), in the presence of competition between isotopic and anharmonic scattering processes shows that (see Fig.2) the values of ratio b_α for the [111] and [110] directions are close (for the slow transverse mode), while for the [100] and [110] directions (for fast transverse mode) the values of b_α are much smaller. It can be seen from Fig. 2 that coefficient $b_\alpha(T)$ tends to unity at $T \gg T_q$; as the temperature decreases ($T < T_q$), it rapidly approaches the ratio g_{nat}/g_{enrich} of the isotopic disorder factors. To illustrate

the low-temperature behavior of coefficient $b_\alpha(T)$ in various crystallographic directions, Table 3 gives its values for various temperatures and values of parameter T_q (an ultrasound frequency of 10 GHz corresponds to $T_q = 0.48$ K). It follows from the table that the measurements of the ultrasound absorption coefficient in Ge crystals at a frequency of 10 GHz should be performed at temperatures much lower than 4 K to obtain an appreciable isotope effect. Although coefficient b_α in the [110] direction for the slow transverse mode is slightly higher than in the [111] direction, transverse modes are degenerate in the [111] direction and no additional efforts are required to ensure the preset polarization of the transverse mode. Thus, the [111] direction for Ge and Si crystal is optimal for obtaining maximal values of the isotope effect in sound absorption in the presence of competition between isotopic and anharmonic scattering processes.

The manifestation of the isotope effect in the dependence of the transverse ultrasound absorption coefficient on reduced wavevector z_1 at a fixed temperature is more interesting. To analyze this dependence, we introduce a dimensionless coefficient $\alpha_T^*(z_1, T)$:

$$\alpha_T^*(z_1, T) = \frac{\alpha_T(z_1, T)2s_T}{B_{T0}T^5} = \nu_{TLL}^* + \frac{B_i}{B_{T0}T}z_1^4. \quad (11)$$

It can be seen from Fig. 3 that, in the long-wave limit ($z_1 \ll 1$), these dependences for crystals ^{nat}Ge and ^{70}Ge (99.99%), as well as for ^{nat}Si and ^{28}Si (99.983%), are virtually identical since the Landau–Rumer mechanism dominates in this region (see formulas (9) and (11)). However, for thermal and high-frequency transverse modes for $z_1 \gg 1$, the dependences of the ultrasound absorption coefficients for natural and enriched germanium differ qualitatively. The absorption coefficient $\alpha_T^*(z_1, T)$ for ^{nat}Ge is a monotonically increasing function of reduced wavevector z_1. However, for highly enriched crystals ^{70}Ge (99.99%) and ^{28}Si (99.983%), these dependences are nonmonotonic with a peak at $z_1 \approx 4.5$ and a minimum at $z_1 \approx 7.5$. It is only for values of $z_1 > 15$ that the absorption coefficient of ultrasound attains the z_1^4 dependence typical of isotope scattering. However, an interesting situation

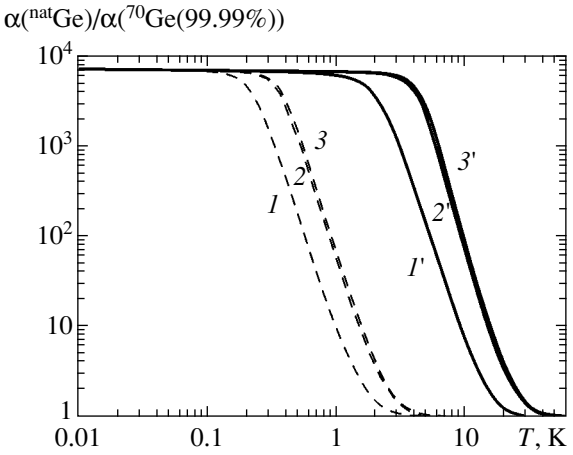

Fig. 2. Temperature dependences of the ratio $\alpha(^{nat}Ge)/(^{70}Ge(99.99\%))$ of the ultrasound absorption coefficients for Ge crystals at fixed energies T_q of an ultrasonic quantum of 1 K (curves *1*, *2*, and *3*) and 20 K (curves *1*', *2*', and *3*') in different crystallographic directions: curves *1* and *1*' correspond to [100], *2* and *2*' correspond to [111], and *3* and *3*' correspond to the slow transverse mode in the [110] direction.

emerges for a higher degree of isotopic enrichment of Ge and Si crystals: the ultrasound absorption coefficient as a function of z_1 acquires a second peak and a second minimum (see, for example, curves *3*, *3*' and *4*, *4*' for silicon in Fig. 3b). These new features are due to the dominant role of anharmonic scattering processes as compared to isotope scattering for short-wave phonons. Figure 4 illustrates the effect of isotopic disorder on the ultrasound absorption coefficient as a function of reduced wavevector z_1 for isotopically modified Ge crystals investigated in [2, 3]. It can be seen from the figure that the ultrasound absorption coefficient for isotopically disordered crystals as well as crystals of ^{70}Ge (96.3%, $g = 7.57 \times 10^{-5}$) with a comparatively low degree of isotopic enrichment is a monotonically increasing function of reduced wavevector z_1; singularities associated with anharmonic scattering processes are not manifested in these crystals.

An analogous type of dependence is observed for diamond crystal with the natural and isotopically enriched compositions (see Fig. 5). For diamond crystals ^{12}C with the natural isotopic composition (98.9%)

Table 3. Coefficients $b_\alpha(T)$ for Ge crystals in various crystallographic directions for various temperatures and values of parameter T_q

	T_q = 0.48 K			T_q = 1 K			T_q = 10 K		
	1 K	4 K	10 K	1 K	4 K	10 K	1 K	4 K	10 K
b_α [100]	1.92	1.003	1.00012	9.53	1.031	1.001	6791	367.1	7.79
b_α [110]	4.1	1.011	1.0003	59.59	1.18	1.004	7194	4966	86.65
b_α [111]	3.57	1.0093	1.00029	51.66	1.16	1.005	7193	4419	69.11

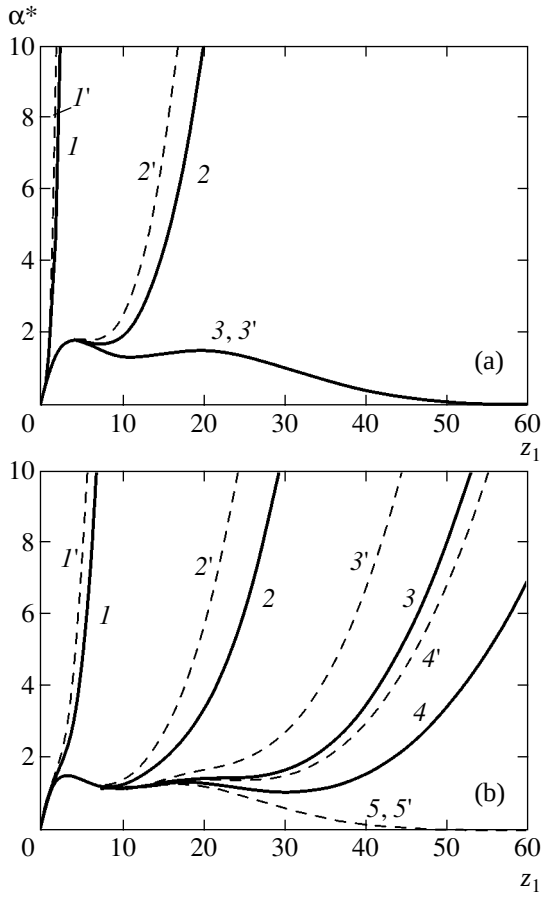

Fig. 3. Transverse ultrasound absorption coefficient as a function of reduced wavevector z_1 for (a) Ge and (b) Si crystals in the [100] crystallographic direction at fixed temperatures (dashed and solid curves correspond to 10 and 20 K, respectively): (a) curves *1* and *1*' correspond to ^{nat}Ge crystals, *2* and *2*'—^{70}Ge (99.99%) crystals, *3* and *3*'—monoisotopic ^{70}Ge (g = 0); (b) curves *1* and *1*' correspond to ^{nat}Si crystals; *2* and *2*'—^{28}Si (99.983%) crystals, *3* and *3*'—^{28}Si (99.994%) crystals, $g = 7 \times 10^{-8}$; *4* and *4*'—^{28}Si (99.998%) crystals, $g = 3 \times 10^{-8}$; *5* and *5*'— monoisotopic ^{28}Si (g = 0).

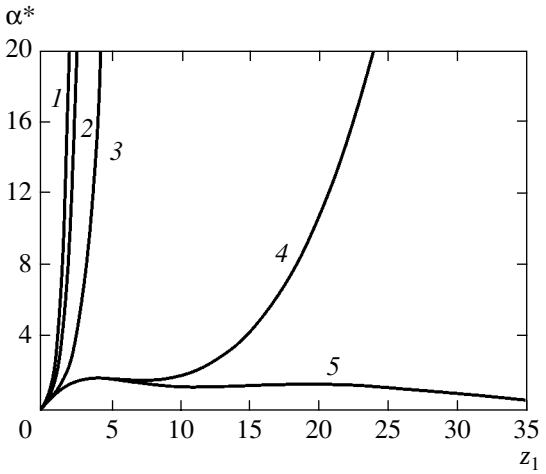

Fig. 4. Transverse ultrasound absorption coefficient as a function of reduced wavevector z_1 at 20 K for Ge crystals with various degrees of isotopic enrichment: curve *1*—$^{70/76}$Ge crystals (50/50, $g = 1.53 \times 10^{3}$); *2*—^{nat}Ge ($g = 5.87 \times 10^{-4}$); *3*—^{70}Ge (96.3%, $g = 7.57 \times 10^{-5}$); *4*—^{70}Ge (99.99%, $g = 0.816 \times 10^{-7}$); *5*—monoisotopic ^{70}Ge (g = 0).

and for enriched ^{12}C (99.93%), the ultrasound absorption coefficient shows no singularities associated with anharmonic scattering processes. Nonmonotonicity typical of anharmonic scattering processes in cubic crystals may appear in the dependence of the absorption coefficient ultrasound only when the concentration of the ^{13}C isotope is lower than 0.01%. It was noted above that the nonmonotonicity in the dependence of the ultrasound absorption coefficient as a function of the phonon wavevector is associated with cubic anisotropy of the crystals under investigation and the angular dependence of the phonon scattering probability in anharmonic scattering processes.

Let us consider the possibility of experimental observation of a peculiar dependence of the transverse ultrasound absorption coefficient on the wavevector in highly enriched crystals of Ge, Si, and diamond taking isotope scattering into account (see [19]). At low temperatures, the value of $\nu^*(z_1, T)$ (see Fig. 5 in [19]) weakly depends on temperature and is in fact a function of z_1 alone. The positions of maxima and minima and their values change by less than 1% in the temperature range 1–20 K for Ge, 1–60 K for Si, and 1–100 K for diamond. It can be seen from Figs. 2 and 4 that isotope scattering in isotopically enriched crystals of Ge, Si, and diamond for $z_1 < 6$ makes a small contribution to the ultrasound absorption coefficient. This circumstance makes it possible to determine the dependence of the ultrasound absorption coefficient on the wavevector (at a fixed temperature) from the measurement of the temperature dependence of coefficient $\alpha_T(T)$ at a fixed phonon energy. For this purpose, we must determine the value of $\alpha_T^*(z_1, T)$ from the data on $\alpha_T(T)$ and reconstruct $\alpha_T^*(z_1, T)$ as a function of z_1 in accordance with expression (11), assuming that the temperature is constant. Thus, by measuring the temperature dependence $\alpha_T(T)$ for ultrasound with $\hbar\omega_q \approx$ 10 K in the temperature range 1–50 K, we can determine $\alpha_T^*(z_1)$ in the range of reduced wavevectors $0.2 < z_1 < 10$. This circumstance might be interesting for experimenters.

3. DISCUSSION OF THE ROLE OF BOUNDARY SCATTERING

Up to now, we have considered the volume absorption coefficient of transverse ultrasound. However, finite-size samples always exhibit scattering of acoustic quanta at the boundary, which leads not only to

momentum relaxation of acoustic waves, but also to a change in their polarization. The important role of the change in polarization of scattered waves upon their reflection from the boundary surface in cubic crystals was indicated in [27]. Unfortunately, the role of boundary scattering in ultrasound absorption has not been studied comprehensively (see [12, 13]). For this reason, we will estimate this effect proceeding from the results obtained for the lattice thermal conductivity of Ge and Si crystals [2–5, 25, 26]. The phonon relaxation rate for scattering at sample boundaries, as well as for scattering from isotopic disorder, is independent of temperature and is defined by the formula

$$\nu_{\mathrm{ph}B}^{\lambda} = \frac{s_\lambda}{L_c}\left\{\frac{1-P}{1+P} + \frac{L_c}{l}\right\} = C_{B\lambda}\times 10^6\ \mathrm{s}^{-1}, \qquad (12)$$

$$C_{BL} = C_{Bt}S_*, \quad L_c = 1.12\sqrt{S},$$

where l is the sample length, S is the cross-sectional area, P is the probability of specular reflection of phonons, and L_c is the Casimir length. The length of our samples [2, 3] was 20–40 mm and the cross-sectional area was 2–6 mm^2. The value of $\nu_{\mathrm{ph}B}$ in Ge and Si crystals with various isotopic compositions was (1–2) $\times$ 10^6 s^{-1}. The necessary condition for observing the isotope effect in the ultrasound absorption coefficient is the fulfillment of the inequality $\nu_{\mathrm{ph}i}^{\lambda} > \nu_{\mathrm{ph}B}^{\lambda}$.

For Ge and Si crystals, we can easily derive the following expressions for rates $\nu_{\mathrm{ph}i}^{T}$:

$$\nu_{\mathrm{ph}i}^{T} \approx 9.8\frac{g}{g_{\mathrm{nat}}}T_q^4, \quad \nu_{\mathrm{ph}i}^{T} \approx 7.2\frac{g}{g_{\mathrm{nat}}}T_q^4. \qquad (13)$$

Order-of-magnitude estimates show that the relaxation rate of transverse phonons scattered from isotopic disorder for high-frequency ultrasound with energies $T_q \approx$ (10, 30, 50) K is $\nu_{\mathrm{ph}i}^{T} \approx (10^{-1}, 10, 10^2)\times 10^6$ s^{-1} both for germanium and for silicon. Thus, the isotope effect in the ultrasound absorption coefficient measured on the same Ge samples as in [2, 3] is strong only for hypersonic quanta with energies $T_q > 20$ K. However, by increasing the cross-sectional area of the samples and polishing the sample surface more thoroughly, it is possible to reduce boundary scattering by at least an order of magnitude. This will permit observation of the isotope effect closer to its bulk value for highly enriched Ge and Si crystals (see above estimates). In this respect, it would be interesting to use another experimental geometry [12, 13], namely, to measure the ultrasound absorption coefficient on disk-shaped samples with $l \ll d$ (where d is the disk diameter) for reflection or transmission. If the size of the emitter and detector is much smaller than the disk diameter, we can get rid of scattering from lateral faces of the sample, while optical polishing of end faces may substantially reduce the role of boundary scattering. In this way, it is possible to lower the ultrasound frequency on which an appreciable isotope effect can be observed and enhance the effect at a fixed ultrasound frequency, since boundary scattering for highly enriched crystals will apparently determine the lower boundary of the absorption coefficient. However, the giant isotope effect in ultrasound absorption can be observed only in the terahertz frequency range (1 THz $\approx$ 50 K). It should be noted that the conditions for observation of the giant isotope effect in Si and diamond crystals are more favorable in view of higher Debye temperatures as compared to germanium. On account of advances made in studying the generation, propagation, and absorption of hypersonic quanta in the terahertz range [28–30], the technical possibility of such investigations appears as quite realistic.

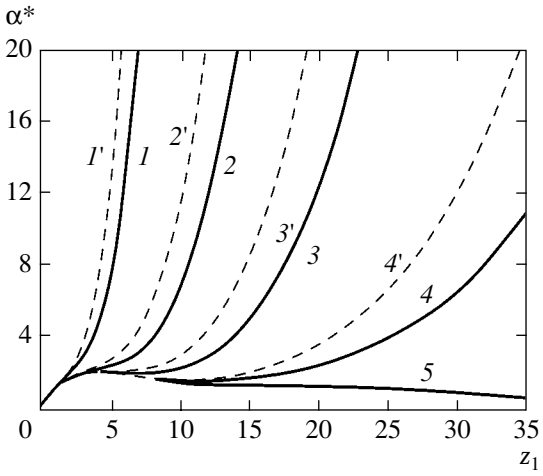

Fig. 5. Transverse ultrasound absorption coefficient as a function of reduced wavevector z_1 for diamond crystals in the [100] crystallographic direction at fixed temperatures (dashed and solid curves correspond to 50 and 100 K, respectively): (a) curves *1* and *1*' correspond to natural diamond with 1.1% of isotope ^{13}C ($g = 7.5 \times 10^5$); *2* and *2*'—diamond with 0.07% of ^{13}C ($g = 4.9 \times 10^{-6}$); *3* and *3*'—diamond with 0.01% of ^{13}C ($g = 6.9 \times 10^{-7}$); *4* and *4*'—diamond with 0.001% of ^{13}C ($g = 6.9\times 10^{-8}$); *5* — monoisotopic diamond ^{12}C ($g = 0$).

4. CONCLUSIONS

Thus, we have obtained a number of estimates of the isotope effect in the ultrasound absorption coefficient in Ge, Si, and diamond crystals. This effect may reach "giant" values in contrast to the isotope effect in thermal conductivity and thermopower. It is shown that investigation of the ultrasound absorption coefficient in isotopically enriched crystals is undoubtedly of considerable interest as a tool for studying the features of anharmonic scattering processes in such popular crystals in microelectronics as germanium, silicon, and diamond.

ACKNOWLEDGMENTS

The authors are grateful to S.N. Ivanov and A.P. Tankeev for fruitful discussions and critical remarks.

This study was supported financially by the Russian Foundation for Basic Research (project no. 05-02-16921), NSh grant no. 1380.2003.2 from the President of the Russian Federation, as well as the Foundation for Supporting National Science.

REFERENCES

1. T. H. Geballe and G. W. Hull, Phys. Rev. **110**, 773 (1958).
2. V. I. Ozhogin, A. V. Inyushkin, A. N. Taldenkov, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **63**, 463 (1996) [JETP Lett. **63**, 490 (1996)].
3. M. Asen-Palmer, K. Bartkowski, E. Gmelin, *et al.*, Phys. Rev. B **56**, 9431 (1997).
4. R. K. Kremer, K. Graf, M. Cardona, *et al.*, Solid State Commun. **131**, 499 (2004).
5. A. V. Inyushkin, A. N. Taldenkov, A. M. Gibin, *et al.*, Phys. Status Solidi C **1**, 11, 2995 (2004).
6. A. P. Zhernov and A. V. Inyushkin, Usp. Fiz. Nauk **171**, 827 (2001) [Phys. Usp. **44**, 785 (2001)]; **172**, 573 (2002) [**45**, 527 (2002)].
7. I. G. Kuleev, I. I. Kuleev, A. N. Taldenkov, *et al.*, Zh. Éksp. Teor. Fiz. **123**, 1227 (2003) [JETP **96**, 1078 (2003)].
8. R. Berman, Phys. Rev. B **45**, 5726 (1992).
9. W. S. Capinski, H. J. Maris, E. Bauser, *et al.*, Appl. Phys. Lett. **71**, 2109 (1997).
10. J. E. Graebner, M. E. Reiss, L. Seibles, *et al.*, Phys. Rev. B **50**, 3702 (1994).
11. J. R. Olson, R. O. Phol, J. W. Vandersande, *et al.*, Phys. Rev. B **47**, 14 850 (1993).
12. B. Truel, C. Elbaum, and B. B. Chick, *Ultrasonic Methods in Solid State Physics* (Acad., New York, 1969; Mir, Moscow, 1972).
13. J. W. Tucker and V. W. Rampton, *Microwave Ultrasonics in Solid State Physics* (North-Holland, Amsterdam, 1972; Mir, Moscow, 1975).
14. V. L. Gurevich, *Kinetics of Photon Systems* (Nauka, Moscow, 1980) [in Russian].
15. R. Berman, *Thermal Conduction in Solids* (Clarendon, Oxford, 1976; Mir, Moscow, 1979).
16. B. M. Mogilevskiĭ and A. F. Chudnovskiĭ, *Thermal Conductivity of Semiconductors* (Nauka, Moscow, 1972) [in Russian].
17. L. Landau and J. Rumer, Sov. Phys. J. **11**, 18 (1937).
18. R. A. H. Hamilton and J. E. Parrot, Phys. Rev. **178**, 1284 (1969).
19. I. G. Kuleev and I. I. Kuleev, Zh. Éksp. Teor. Fiz. **126**, 129 (2004) [JETP **99**, 109 (2004)].
20. W. P. Mason, in *Physical Acoustics. Principles and Methods*, Ed. by W. P. Mason (Academic, New York, 1965; Mir, Moscow, 1968), Vol. 3, Part B, p. 235.
21. I. N. Frantsevich, F. F. Voronov, and S. A. Bakuta, *Elastic Constants and the Moduli of Elasticity of Metals and Nonmetals* (Naukova Dumka, Kiev, 1982) [in Russian].
22. K. Brugger, Phys. Rev. **133**, A1611 (1964).
23. F. Birch, Phys. Rev. **71** (11), 809 (1947).
24. R. N. Thurston, in *Physical Acoustics. Principles and Methods*, Ed. by W. P. Mason (Academic, New York, 1964; Mir, Moscow, 1966), Vol. 1, Part A, p. 235.
25. I. G. Kuleev and I. I. Kuleev, Zh. Éksp. Teor. Fiz. **120**, 649 (2001) [JETP **93**, 568 (2001)].
26. I. G. Kuleev and I. I. Kuleev, Zh. Éksp. Teor. Fiz. **122**, 558 (2002) [JETP **95**, 480 (2002)].
27. P. Carruthers, Rev. Mod. Phys. **33**, 92 (1961).
28. C. Thomsen, H. T. Grahn, H. J. Maris, and J. Tauc, Phys. Rev. B **34**, 4129 (1986).
29. S. Kojima, N. Tsumura, M. Takeda, and S. Nishizawa, Phys. Rev. B **67**, 035102 (2003).
30. N. M. Stanton, R. N. Kini, A. J. Kent, *et al.*, Phys. Rev. B **68**, 113302 (2003).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 331–340.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 381–391.

ORDER, DISORDER, AND PHASE TRANSITIONS IN CONDENSED SYSTEMS

On the Need for a Phenomenological Theory of *P*-Vortices, or Does the Spaghetti Confinement Pattern Admit Condensed-Matter Analogues?¶

A. D. Mironov[a,b], A. Morozov[b], and T. N. Tomaras[c]

[a]*Tamm Department, Lebedev Physical Institute, Russian Academy of Sciences, Moscow, 119991 Russia*
[b]*Alikhanov Institute for Theoretical and Experimental Physics, Moscow, 117218 Russia*
[c]*Department of Physics and Institute of Plasma Physics, University of Crete, and Fo.R.T.H., Greece*
e-mail: mironov@itep.ru, mironov@lpi.ac.ru, morozov@itep.ru, tomaras@physics.uoc.gr

Received January 20, 2005

Abstract—The intuition from condensed-matter physics is commonly used to generate ideas for possible confinement mechanisms in gauge theories. Today, with a clear but puzzling "spaghetti" confinement pattern arising from a decade of lattice computer experiments and implying the formation of a fluctuating net of peculiar magnetic vortices rather than condensation of homogeneously distributed magnetic monopoles, the time has come to reverse logic and search for similar patterns in condensed-matter systems. The main effect to be sought in a condensed-matter setup is the simultaneous existence of narrow tubes (*P*-vortices or 1-branes) of the direction-changing electric field and broader tubes (Abrikosov lines) of the magnetic field, a pattern dual to the one presumably underlying the confinement in gluodynamics. As one possible place to search, we suggest systems with coexisting charge-density waves and superconductivity. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

A possible solution of the confinement problem [1–12] should answer questions at two related but somewhat different levels.[1]

(i) It should allow a reliable evaluation of various quantities, such as the gap in the spectrum of perturbations around a true vacuum, the string tensions in the area laws for the Wilson loops in different representations, as well as the masses of glueballs and other hadrons (when light quarks are taken into consideration).

(ii) It should provide a simple qualitative picture of how a vacuum is formed, how the linear potential arises between remote sources with nonvanishing *N*-alities in the absence of light quarks, and how massive colorless hadrons are formed in the absence as well as in the presence of light quarks.

Of principal importance for the development of theoretical (not computer-experimental) quantitative methods at level (i) would be identification of a true vacuum $|vac\rangle$—a functional of fields at a given moment of time, which is the lowest eigenstate of the nonperturbative Yang–Mills Hamiltonian—with all the other eigenstates presumably separated from $|vac\rangle$ by a nonvanishing gap.

The relevant approach to (ii) would instead identify a relatively small subspace in the space of all field configurations (labeled by a sort of collective coordinates) and substitute the original problem of Yang–Mills dynamics by that of a more or less familiar medium, the QCD ether (like a gas of monopoles or *P*-vortices, a dual superconductor, or something else). The underlying belief here is that the original functional integral at low energies receives a dominant contribution from a restricted set of field configurations and can therefore be substituted by some more familiar effective theory, describing (at least qualitatively) the low-energy quantities as averages over this auxiliary medium and expressing the problems of low-energy quantum Yang–Mills theory via those of the medium dynamics.

Understanding confinement requires certain achievements at both levels (i) and (ii): the existence of a picture is what distinguishes understanding from just calculability, while the possibility of making calcula-

¶ The text was submitted by the authors in English.

[1] We discuss confinement as a problem in pure gluodynamics and ignore all issues related to fermion condensates and chiral symmetry breaking. In real-world QCD, effects related to light quarks can be more important for a large part of hadron physics, and even the dominant confinement mechanism may be different [9]. Therefore, in the study of confinement in gluodynamics, one should rely more upon computer than accelerator experiments. We also do not dwell upon the promising "holistic" approaches to confinement, exploiting various general properties of gluodynamics [10] or building a particular kind of self-consistent approximation to correlation functions [11, 12]. Instead, we discuss the lattice experiment results, providing a microscopic description of relevant field configurations and their common properties, and address the question of whether this mysterious pattern has ever been observed in other types of physical systems.

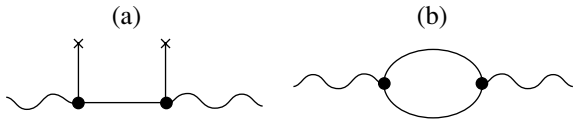

Fig. 1. The origin of the gauge field mass in the Debye screening mechanism. (a) The case where charged particles are originally in the medium. The entire diagram is proportional to the concentration n_0 of these particles in the medium. For nonvanishing temperatures (unavoidable in any lattice calculations), n_0 is never zero (but can be exponentially small). (b) The case where the charged pairs are created in the medium (including the physical vacuum) by the gauge field itself. In this case, the screening is usually much softer and can result in a slow running of the coupling constant rather than in exponential screening.

tions or at least estimates is a criterion for selection of the correct picture among alternative ones.

The problem of confinement consists of two parts: one should explain why

(i) all gauge fields are screened (i.e., all gluons, electric and magnetic, acquire effective masses $\sim\Lambda_{\mathrm{QCD}}$) and

(ii) there nevertheless exists a peculiar long-range color–electric interaction described by a narrow tube where electric force lines (carrying a flux with nonvanishing N-ality, i.e., in a representation that cannot be obtained in a product of adjoints, such that the tube is stable against string-breaking caused by creation of a set of gluons) are collimated and give rise to the linear interaction potential $V(R) \sim \sigma R$ at $R \gg \Lambda_{\mathrm{QCD}}^{-1}$ with the string tension $\sigma \sim \Lambda_{\mathrm{QCD}}^2$ and the string width $r_e \sim \Lambda_{\mathrm{QCD}}^{-1}\log(R\Lambda_{\mathrm{QCD}})$.

We call this double-face situation the dual Meissner–Abrikosov (MA) effect.

The spaghetti vacuum pattern [6], discussed below, implies that, in addition to (i) and (ii),

(iii) one more long-range interaction survives, described by a very narrow tube (P-vortex or 1-brane), with collimated color–magnetic force lines, populated by 0-branes, looking in certain aspects like magnetic monopoles and antimonopoles, with the direction of the magnetic field reversed at the locations of the 0-branes;

(iv) the P-vortices can merge, and when split they form a dense net percolating through the entire volume.

Thus, in some sense, the dual MA effect is complemented by a kind of ordinary MA effect, although the magnetic Abrikosov tubes carry an essential additional structure (moreover, as we discuss below, the oversimplified description of this structure given in (iii) is not gauge-invariant and, hence, is not fully adequate).

2. SCREENING IN THE ABELIAN THEORY

It is well known that the MA effect per se does not require a non-Abelian gauge theory for its manifestation. It can already be discussed at the Abelian level.

There are many ways to obtain a particular kind of the screening effect (α), and many of them allow a particular kind of long-range interactions to survive.

Massive photon. Complete screening with no long-range interactions is described by the effective Lagrangian of the type

$$\frac{1}{e^2}F_{\mu\nu}^2 + m^2A_\mu^2. \tag{1}$$

It explicitly breaks gauge invariance and contains nonpropagating degrees of freedom A_0, giving rise to an instantaneous, but still screened, interaction.

Debye screening. This occurs in ordinary conductors, electrolytes, and some phases of plasma and is described by the effective Lagrangian

$$\frac{1}{e^2}F_{\mu\nu}^2 - E_i\frac{m^2}{\partial^2}E_i. \tag{2}$$

It explicitly breaks the Lorentz invariance and completely screens static electric fields, while magnetic and time-oscillating electric fields remain long-range. The massive term is usually produced by the process shown in Fig. 1, with m^2 being proportional to the concentration n_0 of the electric charges in the medium. If these charges are not originally present, then $m^2 \propto n_0$ either due to nonvanishing temperature, or, if the temperature is zero, to the probability of charge–anticharge creation by an imposed external electric field. This probability, and hence, m^2, normally contains extra powers of 4-momenta, such that the screening mechanism becomes essentially softened and leads, for example, to the slow running coupling phenomenon in QED and QCD, described (in these Lorentz-invariant cases) by the effective Lagrangian

$$F_{\mu\nu}\frac{1}{e^2(\Delta)}F_{\mu\nu}. \tag{3}$$

In 3 + 1 dimensions, the Δ-dependence is just logarithmic, at least in the leading approximation, and, hence, no real screening occurs; gauge fields remain massless. In non-Abelian theories, magnetic interactions also enter the game, producing the antiscreening effect in (3), outweighing the screening one [13]. It is not quite clear whether just this antiscreening could lead to the confinement effect beyond the leading-logarithm approximation (see, e.g., [11]).

To be more precise, in realistic systems, the effective Lagrangian (in the case of linear response, i.e.,

weak fields) is expressed in terms of the dielectric constant[2]

$$\epsilon_{ij} \equiv \left(\delta_{ij} - \frac{k_i k_j}{\mathbf{k}^2}\right)\epsilon_\perp(\omega, \mathbf{k}) + \frac{k_i k_j}{\mathbf{k}^2}\epsilon_\parallel(\omega, \mathbf{k}), \tag{4}$$

$$\mathscr{L} = \frac{1}{e^2}\left[F_{\mu\nu}^2 + (\epsilon_\perp - 1)\mathbf{E}^2 + (\epsilon_\perp + \epsilon_\parallel)\left(\mathrm{div}\mathbf{E}\frac{1}{\boldsymbol{\partial}^2}\mathrm{div}\mathbf{E}\right)\right],$$

and is not universal, because the frequency and momentum dependence of $\epsilon_\parallel$ and $\epsilon_\perp$ can be very different in different regimes. Important for the Debye screening (long-distance exponential decay of the field correlator) is the presence of a singularity in the longitudinal dielectric constant at large distances (small $\mathbf{k}^2$) [15]:

$$\epsilon_\parallel = 1 + \frac{e^2 m^2}{\mathbf{k}^2} + O(\omega),$$

where the omitted terms describe a highly nontrivial frequency dependence. Indeed, the static correlator is

$$\langle E_i E_j\rangle \sim \frac{k_i k_j}{\epsilon_\parallel \mathbf{k}^2} = \frac{k_i k_j}{\mathbf{k}^2 + P_{00}}, \tag{5}$$

where $P_{00} = (\epsilon_\parallel - 1)\mathbf{k}^2$ is the static value of the component of the photon polarization operator $P_{\mu\nu}$ (the "electric" mass [16]).

Dual Debye screening. This is described by a dual effective Lagrangian of the type

$$\frac{1}{e^2}F_{\mu\nu}^2 + H_i\frac{m^2}{\partial^2}H_i \tag{6}$$

and implies screening of static magnetic fields. It is unclear whether any condensed-matter systems with this type of behavior have already been discovered. In ordinary electrodynamics without magnetic charges, we have a counterpart of (5),

$$\langle H_i H_j\rangle = \frac{\mathbf{k}^2\delta_{ij} - k_i k_j}{\mathbf{k}^2 + P}, \tag{7}$$

where the "magnetic" mass P is given by the static value of the spatial components of the photon polarization operator ($P_{ij} \overset{\omega=0}{=} (\delta_{ij} - k_i k_j/\mathbf{k}^2)\,P$ due to the gauge invariance). In a gas of magnetic monopoles, it becomes (see Polyakov's book in [1])

$$\langle H_i H_j\rangle = \delta_{ij} - \frac{k_i k_j}{k^2 + M^2}. \tag{8}$$

[2] We note that the formulation in terms of the dielectric constant and magnetic permeability μ can be useful in the search for solid-state counterparts of the confinement phenomenon (see, e.g., [14]): electric confinement (similarly to that in QCD) can be described by $\epsilon = 0$, while magnetic confinement (similarly to the Meissner effect in superconductors) is attributed to $\mu = 0$.

Chern–Simons screening. It is described by the peculiar gauge-invariant Lagrangian,

$$\frac{1}{e^2}F_{\mu\nu}^2 + m^{\alpha\ldots\beta}\epsilon_{\mu\nu\lambda\alpha\ldots\beta}A_\lambda F_{\mu\nu}. \tag{9}$$

It describes aspects of the Hall effect and related phenomena, is Lorentz-invariant (m is a scalar) only in $2 + 1$ dimensions, and—only in this dimension—makes the photon massive, but the long-range Aharonov–Bohm interaction still survives [17].

Abelian Higgs model. The ordinary (not the dual) Meissner–Abrikosov effect is modeled by the Abelian Higgs (Landau–Ginzburg) effective Lagrangian

$$\frac{1}{e^2}F_{\mu\nu}^2 + |D_\mu\phi|^2 + \lambda(|\phi|^2 - m^2)^2. \tag{10}$$

After ϕ condenses, $\langle\phi\rangle = me^{i\theta}$, the gauge fields become massive, thus giving rise to effect (i): the Meissner effect for magnetic and electric fields. However, the mass is actually acquired not by the A_μ field but by the gauge-invariant combination $\hat{A}_\mu = A_\mu - \partial_\mu\theta$, and hence the mode $\hat{A}_\mu = 0$ can still propagate over large distances, which explains effect (ii): emergence of Abrikosov tubes. $\hat{A}_\mu = 0$ does not imply that $A_\mu = \partial_\mu\theta$ is a pure gauge if $\theta(x)$ is singular and $\oint_C A_\mu\,dx^\mu \neq 0$ for some contours C. In an Abrikosov tube stretched along the z axis, $\theta = \arctan(y/x)$ is the angle in the xy plane and C is any contour in this plane encircling the origin. Because θ is the phase of the smooth field ϕ, the modulus $|\phi|$ should vanish on the z axis, where θ is not well defined; i.e., the condition $|\langle\phi\rangle| = m$ is destroyed in the vicinity of z axis, in a tube with the cross section $\Sigma = \pi r_m^2$. This leads to an energy of $\lambda m^4\Sigma$ per unit length of the tube, while the energy of the magnetic flux Φ in the tube is $\sim(\Phi/\Sigma)^2\Sigma = \Phi^2/\Sigma$. Minimization of the sum of these terms with respect to Σ defines the characteristic width of the tube:

$$\Sigma_m = \pi r_m^2 \sim \frac{\Phi}{\sqrt{\lambda}a^2}.$$

If electric charges q smaller than that of the Higgs field ϕ are present in the theory, then $q\Phi$ can be smaller than 1 and the Aharonov–Bohm effect is observed when such charges travel around the Abrikosov tube at any distance: thus, even though all gauge fields are massive, the Aharonov–Bohm interaction also remains long-range (unscreened) [18].

The technical reason allowing magnetic Abrikosov lines to exist is that the equation $F_{xy} = \delta(x)\delta(y)$ can be

easily solved:

$$A_x = \partial_x \arctan\frac{y}{x}, \quad A_y = \partial_y \arctan\frac{y}{x},$$

and the Higgs field just provides a source of the needed form, with the electric current

$$J_x = \partial_y F_{xy} = \delta(x)\delta'(y), \quad J_y = -\partial_x F_{xy} = -\delta'(x)\delta(y)$$

rotating around the z axis.

To obtain an electric Abrikosov line, we need to solve the equation $F_{0z} = \delta(x)\delta(y)$, which violates Bianchi identity and requires the existence of a magnetic current (rotating around the z axis) and, hence, in a Lorentz-invariant setting, of magnetic charges (monopoles).[3] Therefore, in order to describe confinement with properties (i) and (ii), where the dual MA effect is needed, the dual Abelian Higgs model (the dual superconductor model) is often used where the Higgs field $\tilde{\phi}$ is magnetically charged; i.e., it interacts with the dual field $\tilde{A}_\mu$, such that

$$\tilde{F}_{\mu\nu} = \partial_\mu \tilde{A}_\nu - \partial_\nu \tilde{A}_\mu = \frac{1}{2}\epsilon_{\mu\nu\alpha\beta}F_{\alpha\beta} = \epsilon_{\mu\nu\alpha\beta}\partial_\alpha A_\beta.$$

In this type of scenario, the role of non-Abelian degrees of freedom is thought to be the imitation of Higgs degrees of freedom (see, e.g., $W^\pm$ in Eq. (20) below and [19]), and the problem is to find a mechanism leading to their appropriate condensation.

As already mentioned, the lattice experiments (see Section 4 below) imply that the real pattern (and, perhaps, the mechanism) of confinement can be more sophisticated and may imply the coexistence of (β) electric and (γ) structured magnetic tubes. Therefore, it is important to note that no Abelian model is known that allows the coexistence of magnetic and electric MA effects, e.g., no effective Lagrangian of the form

$$\frac{1}{e^2}F_{\mu\nu}^2 + m_m^2\hat{A}_\mu^2 + m_e^2\hat{\tilde{A}}_\mu^2 \tag{11}$$

is allowed. Therefore, if such coexistence is not an artifact of lattice experiments (which is not considered very probable nowadays), it requires construction of more sophisticated models. A natural hope is that such models can be straightforwardly built in modern string theory (involving branes) and realized in condensed-matter systems.

[3] Similarly, in order to have a magnetic tube, where the field is not constant along the line (in particular, changes direction at some points z_a), we must solve the equation $F_{xy} = \frac{1}{2}\delta(x)\delta(y)\Pi_a \mathrm{sgn}(z - z_a)$, which violates Bianchi identity at $x = y = 0$, $z = z_a$ and, therefore, requires magnetic charges (monopoles) at these points.

We note that some kind of restoration, at least partial, of electromagnetic duality present in Abelian photodynamics is needed. This duality is usually broken by all known relevant modifications: by the introduction of electric charges (without adding their magnetic counterparts), by embedding into a non-Abelian theory (where electric and magnetic interactions of gluons are different), by the addition of a Chern–Simons term, or by coupling to Higgs scalars and going to a superconducting phase. Lattice experiments strongly suggest the need for some—yet unstudied (topological, i.e., with the field content of a field, not string, theory)—stringy phases with both "fundamental" and $D1$ strings present, where screening and MA phenomena do not contradict the electromagnetic duality.

3. 3*d* COMPACT QED

The sample example [2] of the confinement proof in the Abelian (2 + 1)-dimensional compact electrodynamics (embedded into the non-Abelian Georgi–Glashow model to justify compactness and provide ultraviolet regularization, rendering the instanton action finite) actually deals with random confinement [6, 12] and with Wilson's confinement criterion [1]: not fluxes but their squares acquire vacuum averages, $\langle\Phi\rangle = 0$, $\langle\Phi^2\rangle \neq 0$, and this suffices to provide the area-law behavior for the Wilson loop averages. In this example, the relevant medium in two space dimensions is obtained as a time slice of an instanton gas with Debye screening. Instantons in Abelian (2 + 1)-dimensional theory are just ordinary three-dimensional monopoles and antimonopoles with magnetic fields

$$H_\mu = \epsilon_{\mu\nu\lambda}F^{\nu\lambda} = \pm g\frac{r_\mu}{r^3}, \tag{12}$$

or

$$H_\mu = \pm g\frac{r_\mu}{(r^2 + \epsilon^2)}e^{-r/\xi}, \tag{13}$$

where ϵ and ξ provide the respective ultraviolet (from the underlying non-Abelian theory) and infrared (from the Debye screening in the monopole–antimonopole gas) regularizations; g is the monopole charge, normalized such that $2eg$ = integer. Thus, the medium looks like a set of appearing and disappearing vortex–antivortex pairs with the pseudoscalar $2d$ magnetic and vector $2d$ electric fields

$$B = \epsilon_{ij}F^{ij} = \pm\frac{gt}{(\mathbf{x}^2 + t^2)^{3/2}},$$
$$E_i = F_{0i} = \pm g\frac{\epsilon_{ij}x^j}{(\mathbf{x}^2 + t^2)^{3/2}}. \tag{14}$$

The field E_i produced by the time variation of B has nontrivial vorticity and, hence, contributes to the rectangular Wilson average over this medium,

$$\left\langle \exp\left(ie\oint_C (A_0 dt + A_i dx^i) \right) \right\rangle = \left\langle \exp\left(ie\int_S \mathbf{E}\cdot d\mathbf{x}dt \right) \right\rangle, \quad (15)$$

where the contour C lies in the xt plane and S intersects the xy plane by a segment $\tilde{C}$. The contribution of a vortex to the integral $\int_C \mathbf{E}\cdot d\mathbf{x}$ is

$$\pm\int_{-L}^{L} \frac{y dx}{(x^2+y^2+t^2)^{3/2}} \sim \pm\frac{2y}{y^2+t^2} \quad (16)$$

for $L \gg \sqrt{y^2+t^2}$ (with the distance $\sqrt{y^2+t^2}$ actually bounded from above by the Debye radius ξ), and further integration over t gives

$$\pm 4\pi g\Phi = \pm 2\pi g\frac{y}{|y|} = \pm 2\pi g \operatorname{sgn} y \quad (17)$$

for the contribution of a vortex provided that the vortex lies in a slice of width $\xi \ll L$ around the surface S. This flux is one-half of the full flux $4\pi g$ of the charge-g monopole. The factor 1/2 appears here because only half of the vorticity of $\mathbf{E}$ contributes to the integral. Because contributions of vortices and antivortices have opposite signs, the average of $\int \mathbf{E}d\mathbf{x}dt$ itself is, of course, vanishing, but the even powers of this integral, and hence the Wilson exponential, can have nonvanishing averages. The simplest estimate with the help of Poisson distributions gives [20]

$$\left\langle \exp\left(ie\oint_C (A_0 dt + A_i dx^i) \right) \right\rangle = \sum_{n_+,\,n_- = 0}^{\infty} \left[e^{-\bar{n}}\frac{\bar{n}^{n_+}}{n_+!} \right]\left[e^{-\bar{n}}\frac{\bar{n}^{n_-}}{n_-!} \right] \times \exp(4\pi i e g (n_+ - n_-)\Phi) = \exp(-2\bar{n}(1-\cos(4\pi e g \Phi))). \quad (18)$$

Because the average number of contributing vortices and antivortices is $\bar{n} = \xi A_S n_0$, where A_S is the area of the surface S and n_0 is the concentration of vortices (depending primarily on the instanton action, which is in turn defined by the ultraviolet regularization), we obtain the area law for the Wilson loop, at least for the minimal value $eg = 1/2$ allowed by the Dirac quantization condition.[4] The average

$$\left\langle \exp\left(ie\oint_C (A_0 dt + A_i dx^i) \right) \right\rangle = \left\langle \exp\left(ie\int_S B dx dy \right) \right\rangle \quad (19)$$

of a spacelike Wilson loop with S lying in the xy plane and bounded by the curve C can be calculated similarly. This average is given by the same formula (18).

Another interpretation of the same calculation [3] implies that the distribution of vortices is affected by the presence of the loop, such that the vortices and antivortices are concentrated around the surface S and screen it.

4. CONFINEMENT IN 4D

In 3 + 1 dimensions, no such simple calculation from first principles is yet known. The main difference is that ordinary instantons in 3 + 1 dimensions are no longer charged: their field vanishes too fast at infinity, and, therefore, the confinement mechanism should involve an additional dissociation of instantons into something like magnetically charged merons [3, 22]. Time slices of instantons are now 3D objects, namely monopole–antimonopole pairs (if viewed with a special gauge), and the instanton describes the process of their spontaneous creation and annihilation.

The expectation is that, in a dense instanton gas (or liquid), recombination takes place between monopoles and antimonopoles from different pairs, thus picking up a chain of instantons from the liquid (see Fig. 2).

The spaghetti vacuum pattern implies that such chains are actually spread out through the entire volume and form a "percolating cluster" [20, 23].

As in the (2 + 1)-dimensional case, electric fields with nonvanishing vorticities, caused by the moving monopoles and antimonopoles, contribute to the Wilson averages in 3 + 1 dimensions and give rise to the area laws.

At present, there is no absolutely convincing theoretical argument in favor of these ideas; instead, they received considerable support from computer experiments.

"Experimental" lattice results. Lattice computer simulations are primarily targeted at producing qualitative results in the spirit of (i) and, thus, at providing a proof that the Yang–Mills functional integral indeed describes a theory with a mass gap, a linear potential, a realistic hadronic spectrum, and realistic hadron interactions. Remarkably enough, these experiments could

[4] There are corrections to this oversimplified calculation [20, 21], which can, in particular, destroy the prediction in (18) that confinement disappears for even magnetic charges (when the relevant flux Φ is integer).

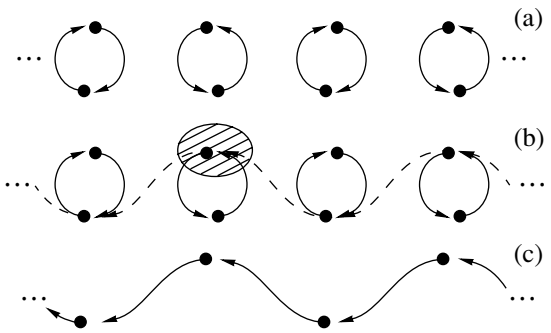

Fig. 2. Possible phases of the recombinant plasma of the instanton gas. (a) Recombinant phase (ordinary instanton gas in 3 + 1 dimensions): each instanton is the process of creation and annihilation of a monopole–antimonopole pair. (b) Transition to the jumping recombinant phase (instantons dissociated into merons): created pairs do not coincide with annihilating pairs. The hatched domain corresponds to a meron. (c) Jumping recombinant phase: a chain is naturally formed.

also be used for research in direction (ii), and they indeed produced very inspiring results. However, up to now, the simulations are not very detailed and the functional integral is actually replaced by a sum over a rather small random subset of field configurations that are believed to make the dominant contribution. According to (ii), one can hope that most of these dominant configurations have something in common—and this is what actually happens—providing a clear description of the medium required in (ii).

This experimentally discovered [24, 25] medium appears somewhat unexpected (see [6] for the original suggestion of this "Copenhagen spaghetti vacuum" and [26] for comprehensive modern reviews and references): it turns out to be filled with peculiar one-dimensional objects (with two-dimensional world surfaces)—*P*-vortices—which, in a certain Abelian approximation (see the next subsection), look like narrow (with a width of $r_m \ll \Lambda_{\mathrm{QCD}}^{-1}$) tubes of magnetic field, directed along the tube and changing direction to the opposite at locations of monopoles and antimonopoles, which form a one-dimensional gas inside the tube.[5] Such objects are obviously stable against the creation of monopole–antimonopole pairs: such processes cannot break the tube in two, because the magnetic flux through any section outside the monopole cores is 1/2.[6] The net of these direction-changing color–magnetic tubes fills the entire space [20] (forming a "percolating cluster" [23]),[7] and in this medium, the force lines of color–electric fields (emitted by sources of nonvanishing *N*-ality) also form tubes (of width $r_e \sim \Lambda_{\mathrm{QCD}}^{-1}$), thus giving rise to the confinement phenomenon. In lattice experiments, the area laws for appropriate Wilson-loop averages are explicitly verified, and the *P*-vortices from a percolating cluster are shown to make the dominant contribution to string tensions. Theoretically, the contribution of *P*-vortices to the string tension depends on their abundance, and one of the tasks of the theory is to explain the origin of the medium of *P*-vortices and ensure its consistency with Lorentz invariance.

So far, there is no clear theoretical explanation of why and how such a medium is formed in non-Abelian gauge theories and why—once formed—it can give rise to a dual Meissner effect and lead to confinement, although the (lattice) experimental evidence in favor of this pattern is rapidly growing.

A serious drawback to the published results of lattice experiments is that they do not provide essential information about instanton-like and meronlike configurations and their probable association with the localized *P*-vortex clusters; furthermore, they have not explicitly studied the configurations of collimated color–electric force lines between sources with nonvanishing *N*-ality (which do not need to be fermions). Information about these color–electric tubes has been extracted indirectly from the study of Wilson averages. This is not enough to understand what happens to these tubes, for example, after the maximal Abelian projection, and whether their content indeed looks like an Abelian electric field exactly in the same projection where the *P*-vortices look like the tubes of a direction-changing Abelian magnetic field. Any data touching upon this issue would be extremely useful for further clarification.

Maximal Abelian projection. The "*P*" in "*P*-vortices" comes from the word "projection" [29]. It is inspired from the way they are often sought and studied, which is not gauge-invariant, even though the *P*-vortices themselves are in fact gauge-invariant (see Fig. 3).

[5] In contrast to the *P*-vortices themselves, the monopoles and antimonopoles inside them are difficult to define in a gauge-invariant way. Even the direction of the tentative Abelian magnetic field and, hence, the exact positions of monopoles and antimonopoles inside the *P*-vortex are unphysical: they can be changed by gauge transformations. Indeed, to change the direction of an Abelian field strength $F_{\mu\nu}^3$ at a given point, it suffices to perform a singular gauge transformation conjugating the fields by a unitary matrix like σ^1 at this point (although it is not absolutely clear how to perform such an operation consistent with the maximal Abelian projection, described in the next subsection). There is still a controversy in the literature (see, e.g., [27] for different points of view) about the actual internal structure of the *P*-vortices and the (dis)advantages of visualizing them in terms of monopoles and antimonopoles.

[6] This does not contradict the possibility that isolated monopoles are screened [28].

[7] In addition to the percolating cluster, there also exists a variety of nonpercolating ones also populated by monopoles. There is no agreement in the literature on whether these nonpercolating clusters are lattice UV artifacts or if they actually contribute in the continuum limit.

A procedure called maximal Abelian projection (MAP) is commonly used.[8] It splits into two steps. First, for every configuration of the fields $A^a_\mu(x)$, taken with the weight dictated by the true non-Abelian action, the maximal Abelian gauge is chosen by minimizing the lattice counterpart of $\int W^+_\mu W^-_\mu(x)d^4x$ along the gauge orbit. This first step is absolutely justified (although technically it suffers from ambiguities caused by the existence of Gribov copies).

This makes it possible to introduce the induced effective action $S(A)$, obtained after integration over the other components ($W^\pm_\mu \equiv A^1_\mu \pm iA^2_\mu$, $D_\mu(A) \equiv \partial_\mu + ieA^3_\mu$):

$$\begin{aligned}\exp(-\tilde{S}(A)) &= \int DW^+DW^-\delta(|D_\mu(A)W^+_\mu|^2)\\ &\times \det{}^2_{FP}(\partial_\mu D_\mu(A))\\ \times \exp\Bigg(-\frac{1}{g^2}\Big[&(F_{\mu\nu} + (W^+_\mu W^-_\nu - W^-_\mu W^+_\nu))^2\\ &+ |D_\mu(A)W^+_\nu|^2\Big]\Bigg).\end{aligned} \tag{20}$$

At the second step, $\tilde{S}(A)$ is used to define Abelian correlation functions

$$\begin{aligned}\left\langle \prod_i O(A^a_\mu)\right\rangle_{MAP} &\equiv \left\langle \prod_i O(W^\pm_\mu = 0, A^3_\mu)\right\rangle\\ &= \int DA^3_\mu e^{-\tilde{S}(A^3_\mu)}\prod_i O(W^\pm_\mu = 0, A^3_\mu).\end{aligned} \tag{21}$$

This step implies that the true non-Abelian action is used; i.e., contributions from the virtual $W^\pm$-bosons in loops are included, although omitted from external lines. Therefore, the second step—the projection itself—is an approximation:

$$\begin{aligned}&\left\langle \prod_i O(W^\pm_\mu, A^3_\mu)\right\rangle\\ &\neq \int DA^3_\mu e^{-\tilde{S}(A^3_\mu)}\prod_i O(W^\pm_\mu = 0, A^3_\mu).\end{aligned} \tag{22}$$

Its experimentally discovered [31], surprising efficiency (as compared to the complete answer including non-Abelian fields) is often called the hypothesis of Abelian dominance. Although so far theoretically unjustified and uncontrollable, it provides a convenient language for description (visualization) of the confinement phase: it is at this level that monopoles and antimonopoles appear. Figure 3 can serve as an illustration of how MAP works.

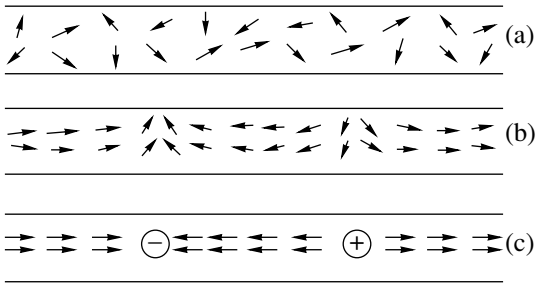

Fig. 3. This figure, taken from seminal paper [25], is the best existing illustration of what *P*-vortices are and what the maximal Abelian projection does. (a) A fragment of the distribution on field strength in an original configuration of fields $A^a_\mu(x)$, from the set of those fields that give a dominant contribution to the non-Abelian functional integral. The strengths are nonvanishing within a narrow tube, the *P*-vortex. Actually, the entire configuration looks like a net of *P*-vortices, containing the "percolating cluster," which has proper scaling properties and survives in the continuum limit. The arrows indicate directions in color space. (b) The maximal Abelian gauge is chosen, which minimizes $\int W^+_\mu W^-_\mu(x)d^4x$. It is just a choice of gauge (field strengths are rotated), no approximation is involved. Certain structures are clearly seen in the distribution of field strengths inside the tube. (c) Maximal Abelian projection is performed: $W^\pm_\mu(x)$ are set equal to zero. The structures in (b) turn into a clear (but approximate) pattern of collimated magnetic force lines, changing direction at the location of monopoles and antimonopoles. No peaks of magnetic energy occur at these locations.

The theoretical problem of evaluating $\tilde{S}(A)$ remains open. We refer to [19] for interesting attempts to identify condensing modes and vortexlike structures in functional integral (20) and to [32] for a supersymmetric model with BPS configurations that look like magnetic *P*-vortices populated by monopoles.

5. ARE THERE CONDENSED-MATTER ANALOGUES OF CONFINEMENT?

Returning to the lattice results above, a natural question to ask is whether anything similar can be found in other avatars of gauge theories, for example, in condensed-matter or plasma physics. There, one would rather expect to encounter a dual type of medium: electric *P*-vortices formed by chains of positive and negative electric charges, connected by narrow tubes of electric fields with fluxes ±1/2, and the ordinary (magnetic) MA effect, implying formation of magnetic-field tubes with a constant unit flux (and confinement of hypothetical magnetic charges), caused by or at least

[8] Comparison with the results of lattice experiments in other Abelian approximations usually demonstrates that the (gauge-noninvariant and necessarily approximate) language of monopoles is most reliable in MAP, the use of this language in other calculational schemes can often be misleading [30].

consistent with the existence of such electric *P*-vortices. In condensed-matter analogues, the underlying non-Abelian Yang–Mills dynamics responsible for the formation of *P*-vortices should presumably be replaced by some other dynamics (additional forces) allowed in condensed-matter systems. The whole situation (the coexistence and even mutual influence of electric *P*-vortices and magnetic MA effect) is already exotic enough to make one wonder if anything like this can occur in any kind of natural matter systems.

The main effect to be sought in a condensed-matter setup is the simultaneous existence of narrow tubes (*P*-vortices) of direction-changing electric field and broader tubes (Abrikosov lines) of magnetic field—a dual pattern to the one underlying the spaghetti confinement mechanism of gluodynamics. This clearly implies that superconductivity (from the dual superconductor scenario), if relevant at all, should be of a more sophisticated nature than just the single-field condensation (monopole condensation), and the superconducting order should be caused by, or at least coexist with, an order of some other type (responsible for the formation of *P*-vortices). This looks almost like the requirement that the Meissner–Abrikosov effect (for the magnetic field) coexists with (or, perhaps, is even implied by) the dual Meissner–Abrikosov effect (for electric field), but actually, the tubes of the electric field should be different: they should have an internal structure, namely, a one-dimensional gas of positive and negative electric charges, and an electric field along the tube that changes direction at the locations of these charges and that is stable against possible "string breaking" caused by creation or annihilation of charge–hole pairs. Moreover, the width of electric tubes should/can be different (much smaller?) than that of magnetic tubes.

The main goal of this paper is to bring these issues to the attention of experts in other fields, such as condensed-matter and plasma physics, and to emphasize the fact that the discovery of a similar picture arising under any circumstances would be of great help for the development of the confinement theory and, in particular, for the understanding of possible 2D vortex theories living on the world sheets of the relevant branes, as well as of the phase structure of these theories.[9] If, on the contrary, no such pattern exists in condensed-matter physics, this would once again emphasize the peculiarities of non-Abelian gauge theories (where elementary quanta carry more structure than just pointlike charges, and, thus, the naive screening behavior is from the very beginning substituted by antiscreening and further nonnaive phenomena are naturally expected to occur).

The rest of this paper is purely speculative, added for encouragement: in order to demonstrate that superconductivity (probably responsible for the magnetic Meissner–Abrikosov effect) can indeed coexist with some kind of dual order at least (although the example below falls short of exhibiting narrow tubes of the direction-changing electric field).

Charge density waves. As a possible (but, by no means, unique) candidate analog of the electric *P*-vortices, we suggest the charge density waves (CDWs); the questions that then arise are:

(a) Are there any tubelike CDWs with a charge density similar to $\rho(x, y, z) \sim \delta^{(2)}(x, y)\sin z$ and (perhaps, direction-changing) electric force lines collimated along the z axis?

(b) Can CDWs coexist with superconductivity (SC), which would be a natural reason for the Meissner–Abrikosov effect?

(c) Can CDWs cause or at least enhance superconductivity?

(d) Can the widths of CDW-like *P*-vortices be much smaller than those of Abrikosov lines (where the Cooper–Higgs-like condensate is broken)?

Remarkably, a very similar set of questions is currently under intense investigation in connection with high-T_c superconductivity (where an adequate theoretical pattern also remains unknown), and it looks like the above possibilities are indeed open, as can be seen in [36] and references therein. Of course, the real media appearing in condensed-matter examples have a lot of additional structure (primarily, the highly anisotropic crystal lattice in the background, playing a key role in the formation of realistic CDWs), which one does not expect to find in gluodynamics. For closer analogues to gluodynamics, one can also look for phenomena in liquid He [33], dense relativistic plasma, segnetoelectrics [14], or even biological membranes [37]. Still, we want to emphasize once again that, today, when the formulation of a phenomenological theory of *P*-vortices is so important, one needs to consider all examples where objects of this kind are presumably present, irrespective of the underlying microscopic structure, and the solid-state systems with the coexisting CDW and SC orders should not be neglected—especially because, along with the confinement in gauge theories, they are now under close scrutiny and considerable progress can result rather quickly from comparison of ideas from the two fields.

The simplest facts and ideas about the CDW–SC systems, although not immediately coinciding with (a)–(d), do not seem to be in obvious contradiction. The relevant properties seem to include the following list:

[9] Among other things, it would be interesting to exploit the idea of topological confinement, which, in different versions, often works in condensed-matter physics. A characteristic feature of topological confinement is that it depends on the dynamics of the theory only via the properties of particular excitations (quasiparticles), while their interactions do not matter. For example, one-dimensional objects can be tied and, therefore, be unseparable, and this can work for real one-dimensional excitations, like Abrikosov tubes, and for pointlike magnetic monopoles and/or hedgehogs that have Dirac strings attached. In practice, topological confinement can look very similar to the mechanism we discuss throughout the paper (see [33, 34] for some examples; see also [35]).

—CDW formation causes transition to an insulator phase (Peierls–Fröhlich–Mott transition), while the SC transition gives rise to a (super)conductor;

—thus, CDW and SC orders compete with each other, with CDW usually a stronger competitor than SC [38];

—the CDW and SC orders can nevertheless coexist [39, 40];

—even if both CDW and SC orders are not established simultaneously at long distances, they interfere locally, one phase appearing in the regions where the other is broken: SC appears in the vicinity of CDW vortices, and CDWs appear in the vicinity of Abrikosov lines [40]. This can be sufficient, for example, to obtain the SC phase when CDW dislocations percolate through the entire volume.

The phenomenological description of CDWs is in terms of electron–phonon interactions [41]. We note that the vector nature of phonons makes them closer to *W*-fields in (20) than to the scalar fields used in the Abelian–Higgs model (10).

6. CONCLUSIONS

The theory of the Copenhagen spaghetti vacuum should, of course, be developed in the context of string theory. The appropriate name for *P*-vortices is 1-branes. Monopoles living on these 1-branes are, naturally, 0-branes. The coexistence of electric and magnetic Abrikosov tubes should be modeled by that of coexisting "fundamental strings" and *D*1 branes. The problems raised in this paper are related to the lack of any "underlying model" for which the theory of strings and branes would be an effective model, a lack which seriously undermines progress in modern string theory. We emphasize that the spaghetti vacuum in gluodynamics can by itself provide such a model, and we also suggest starting a more extensive search for possible underlying models in modern condensed-matter physics.

ACKNOWLEDGMENTS

We are grateful to T. Mironova for help in composing the figures. This work was supported in part by the EU under the RTN contract no. MRTN-CT-2004-512194. A.D.M. and A.M. acknowledge the support of two NATO travel grants and the hospitality of the Department of Physics of the University of Crete, where this work was performed. This work was also partially supported by the Federal Program of the Russian Ministry of Industry, Science, and Technology (grant no. 40.052.1.1.1112), Volkswagen Stiftung, INTAS (grant no. 00-561), the Russian Foundation for Basic Research (project nos. 04-02-16538a and 01-02-17488), and under the State program for Support of Leading Science Schools (grant no. 96-15-96798 (Mironov)).

REFERENCES

1. K. Wilson, Phys. Rev. D **10**, 2445 (1974); Y. Nambu, Phys. Rev. D **10**, 4262 (1974); A. M. Polyakov, *Gauge Fields and Strings* (Harwood Academic, New York, 1987)Ê hep-th/0407209; S. Mandelstam, Phys. Rep. C **23**, 245 (1976); G. 't Hooft, Nucl. Phys. B **190**, 455 (1981); J. Kogut and L. Susskind, Phys. Rev. D **11**, 395 (1975).
2. A. M. Polyakov, Phys. Lett. B **59B**, 82 (1975); Nucl. Phys. B **120**, 429 (1977); T. Banks, J. Kogut, and R. Myerson, Nucl. Phys. B **129**, 493 (1977).
3. C. Callan, R. Dashen, and D. Gross, Phys. Rev. D **17**, 2717 (1978).
4. S. G. Matinyan and G. K. Savvidy, Nucl. Phys. B **134**, 539 (1978).
5. G. 't Hooft, Nucl. Phys. B **153**, 141 (1979); G. Mack and V. Petkova, Z. Phys. C **12**, 177 (1982); E. Tomboulis, Phys. Rev. D **23**, 2371 (1981); T. Kovács and E. Tomboulis, Phys. Rev. D **57**, 4054 (1998); hep-lat/0108017.
6. H. Nielsen and P. Olesen, Nucl. Phys. B **61**, 45 (1973); **160**, 380 (1979); J. Ambjorn and P. Olesen, Nucl. Phys. B **170**, 60 (1980); **170**, 265 (1980); J. Ambjorn, B. Felsager, and P. Olesen, Nucl. Phys. B **175**, 349 (1980).
7. R. P. Feynman, Nucl. Phys. **188**, 479 (1981).
8. P. Vinciarelli, Phys. Lett. **78B**, 485 (1978); J. M. Cornwall, Nucl. Phys. B **157**, 392 (1979); Phys. Rev. D **65**, 085045 (2002); hep-th/0112230.
9. V. N. Gribov, hep-ph/9905285; Eur. Phys. J. C **10**, 91 (1999); hep-ph/9902279; hep-ph/9512352; Yu. Dokshitzer, hep-ph/0404216.
10. G. 't Hooft, Nucl. Phys. B **138**, 1 (1978); C. Bachas, Phys. Rev. D **33**, 2723 (1986); P. van Baal, hep-ph/0008206; I. Kogan and A. Kovner, hep-th/0205026; C. P. Korthals Altes, hep-ph/0308229; A. Smilga, Ann. Phys. (N.Y.) **234**, 1 (1994); Acta Phys. Pol. B **25**, 73 (1994).
11. J. Greensite and C. Thorn, J. High Energy Phys. **02**, 014 (2002); hep-ph/0112326; V. N. Gribov, Nucl. Phys. B **139**, 1 (1978); Eur. Phys. J. C **10**, 71 (1999); hep-ph/9807224; Yu. Dokshitzer, hep-ph/0404216; G. 't Hooft, hep-th/0207179; D. Zwanziger, Nucl. Phys. B **518**, 237 (1998); J. Greensite, S. Olejnik, and D. Zwanziger, Phys. Rev. D **69**, 074506 (2004); hep-lat/0401003; hep-lat/0407032.
12. M. Shifman, A. Vainstein, and V. Zakharov, Nucl. Phys. B **147**, 385 (1979); **147**, 448 (1979); **147**, 519 (1979); V. Novikov, L. Okun, M. Shifman, *et al.*, Phys. Rep. **41**, 1 (1978); Yu. Simonov, Phys. At. Nucl. **58**, 107 (1995); hep-ph/9311247; Phys. Usp. **39**, 313 (1996); hep-ph/9709344.
13. R. J. Hughes, Phys. Lett. **97B**, 246 (1980); J. Iliopoulos, D. V. Nanopoulos, and T. N. Tomaras, Phys. Mag. **12**, 111 (1990).
14. D. Kirzhnitz and M. Mikaelyan, Pis'ma Zh. Éksp. Teor. Fiz. **39**, 571 (1984) [JETP Lett. **39**, 701 (1984)]; Zh. Éksp. Teor. Fiz. **97**, 795 (1990) [Sov. Phys. JETP **70**, 444 (1990)].
15. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 10: *Physical Kinetics* (Nauka, Moscow, 2002; Pergamon, Oxford, 1981).

16. D. Gross, R. Pisarski, and L. Yaffe, Rev. Mod. Phys. **53**, 43 (1981).
17. I. Kogan and A. Morozov, Zh. Éksp. Teor. Fiz. **88**, 3 (1985) [Sov. Phys. JETP **61**, 1 (1985)].
18. F. A. Bais, A. Morozov, and M. de Wild Propitius, Phys. Rev. Lett. **71**, 2383 (1993); hep-th/9303150.
19. L. D. Faddeev and A. Niemi, Phys. Lett. B **525**, 195 (2002); hep-th/0101078; A. Niemi, hep-th/0403175.
20. A. Hart and M. Teper, Phys. Rev. D **58**, 014504 (1998); hep-lat/9712003.
21. J. Ambjorn and J. Greensite, J. High Energy Phys. **9805**, 004 (1998); hep-lat/9804022.
22. J. W. Negele, F. Lenz, and M. Thies, hep-lat/0409083.
23. T. Ivanenko, A. Pochinskii, and M. Polykarpov, Phys. Lett. B **252**, 631 (1990); **302**, 458 (1993); A. Hart and M. Teper, Phys. Rev. D **60**, 114 506 (1999); hep-lat/9902031.
24. L. Del Debbio, M. Faber, J. Greensite, and S. Olejník, hep-lat/9708023; Ph. de Forcrand and M. Pepe, Nucl. Phys. B **598**, 557 (2001); hep-lat/0008016; C. Alexandrou, Ph. de Forcrand, and M. D'Elia, Nucl. Phys. A **663**, 1031 (2000); hep-lat/9909005.
25. J. Ambjorn, J. Giedlt, and J. Greensite, J. High Energy Phys. **02**, 033 (2000); hep-lat/9907021.
26. J. Greensite, Prog. Part. Nucl. Phys. **51**, 1 (2003); hep-lat/0301023; V. G. Bornyakov, M. I. Polikarpov, M. N. Chernodub, *et al.*, Usp. Fiz. Nauk **174**, 19 (2004) [Phys. Usp. **47**, 17 (2004)]; V. E. Zakharov, Usp. Fiz. Nauk **174**, 39 (2004) [Phys. Usp. **47**, 37 (2004)]; hep-ph/0410034; R. W. Haymaker, Phys. Rep. **315**, 153 (1999); hep-lat/9809094.
27. J. M. Carmona, M. D'Elia, A. Di Giacomo, *et al.*, Phys. Rev. D **64**, 114507 (2001); hep-lat/0103005; F. Gubarev and V. Zakharov, hep-lat/0204017; F. Gubarev, hep-lat/0204018; A. Kovner, M. Lavelle, and D. McMillan, hep-lat/0211005.
28. Ch. Hoebling, C. Rebbi, and V. A. Rubakov, Phys. Rev. D **63**, 034506 (2001); hep-lat/0003010.
29. L. Del Debbio, M. Faber, J. Greensite, and S. Olejník, Phys. Rev. D **55**, 2298 (1997); hep-lat/9801027.
30. M. N. Chernodub, Phys. Rev. D **69**, 094504 (2004); hep-lat/0308031; V. A. Belavin, M. N. Chernodub, and M. I. Polikarpov, hep-lat/0403013.
31. A. S. Kronfeld, M. L. Laursen, G. Schierholz, and U. J. Wiese, Phys. Lett. B **198**, 516 (1987); T. Suzuki and I. Yotsuyanagi, Phys. Rev. D **42**, 4257 (1990).
32. D. Tong, Phys. Rev. D **69**, 065003 (2004); hep-th/0307302.
33. G. Volovik, *The Universe in a Helium Droplet* (Clarendon, Oxford, 2003); Proc. Natl. Acad. Sci. USA **97**, 2431 (2000); cond-mat/9911486.
34. M. M. Salomaa and G. E. Volovik, Rev. Mod. Phys. **59**, 533 (1987); Y. Kondo, J. S. Korhonen, M. Krusius, *et al.*, Phys. Rev. Lett. **68**, 3331 (1992); G. E. Volovik and T. Vachaspati, Int. J. Mod. Phys. B **10**, 471 (1996); cond-mat/9510065; G. E. Volovik, V. B. Eltsov, and M. Krusius, cond-mat/0012350.
35. B. Julia and G. Toulouse, J. Phys. Lett. **40**, L395 (1979); F. Quevedo and G. A. Trugenberger, Nucl. Phys. **501**, 143 (1997).
36. G. Grüner, Rev. Mod. Phys. **60**, 1129 (1988); A. Gabovich, A. Voitenko, J. Annett, and M. Ausloos, Supercond. Sci. Technol. **14**, R1 (2001).
37. P. Nelson and T. Powers, Phys. Rev. Lett. **69**, 3409 (1992); cond-mat/9211008.
38. E. Arrigoni, E. Fradkin, and S. A. Kivelson, cond-mat/0409693.
39. S. A. Kivelson, G. Aeppli, and V. J. Emery, cond-mat/0105200; J.-P. Hu and S.-C. Zhang, cond-mat/0108273.
40. D.-H. Lee, Phys. Rev. Lett. **88**, 227003 (2002); cond-mat/0111393.
41. A. B. Migdal, Zh. Éksp. Teor. Fiz. **34**, 1438 (1958) [Sov. Phys. JETP **7**, 999 (1958)]; G. M. Eliashberg, Zh. Éksp. Teor. Fiz. **38**, 966 (1960) [Sov. Phys. JETP **11**, 696 (1960)].

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 341–349.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 392–402.

ORDER, DISORDER, AND PHASE TRANSITIONS IN CONDENSED SYSTEMS

Calculation of Critical States of Superconducting Multilayers Based on Numerical Solution of the Ginzburg–Landau Equations for Superconducting Plates

A. N. Lykov, A. Yu. Tsvetkov, and G. F. Zharkov†

Lebedev Physical Institute, Russian Academy of Sciences, Moscow, 119991 Russia
e-mail: lykov@mail1.lebedev.ru

Received February 1, 2005

Abstract—Numerical methods are used to analyze the Ginzburg–Landau equations for a superconducting plate carrying transport current in a magnetic field. Critical current is calculated as a function of the applied magnetic field strength for superconducting plates with different thicknesses. The relations between the field dependence of critical current and the distributions of order parameter, magnetic field, and supercurrent in a plate are analyzed. The field-dependent critical currents computed for plates are used to determine the critical current as a function of the applied magnetic field strength and local magnetic field and current distributions for multilayers in parallel magnetic fields. The constituent superconducting layers are assumed to interact only via magnetic field. A simple method is proposed for analyzing the critical states of multilayers in magnetic fields of arbitrary strength, based on elementary transformations of the critical current-density distribution over individual layers in zero applied magnetic field. The method can be used to analyze experimental results. *© 2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

Analyses of the behavior of bulk superconductors in magnetic fields based on the Ginzburg–Landau theory [1] have been presented in numerous studies [2, 3]. Recent publications were focused on mesoscopic superconductors of various geometries [4–6]. In this paper, we consider a vortex-free state, in which critical current density is equivalent to depairing critical current density, as a basis for analyzing the critical states of superconducting plates and multilayers carrying transport currents perpendicular to applied magnetic fields parallel to their surfaces. Most studies of the critical states of superconductors rely on models of the interaction between the vortex system and crystalline defects [7]. Analysis of this problem is complicated by the diversity of quantum properties of the superconductor vortex lattice as an elastic medium described by nonlinear electrodynamics. For this reason, simplifying assumptions are invoked, such as the London approximation for the vortex system or model distributions of magnetic field in superconductors [7]. This frequently leads to poor agreement between theoretical calculations and experimental results [8]. Even in the simplest case of ordered defects, as in a multilayer embedded in parallel magnetic fields, the calculation of critical current density is a difficult task. The most interesting results in this area were obtained in [9, 10]. In [9], the Ginzburg–Landau equations were solved to find the field-dependent critical current density in a multilayer for magnetic field strength close to the upper critical field. In [10], critical current was found in the London approximation by representing the vortex lattice as a set of linear chains and analyzing their interaction with the layers making up the multilayer. The scope of both studies was substantially limited by assuming that the vortex lattice matches the multilayer structure in the limit of weak order-parameter modulation. For superconductors of this kind, this condition implies that the critical current density is low, whereas more interesting for practical applications are superconductors with strong pinning centers, i.e., superconducting multilayers characterized by large amplitudes of order-parameter modulation.

We propose here a new method for analyzing the critical state of a superconducting multilayer based on exact solution of the Ginzburg–Landau equation for a thin film [11]. We consider a multilayer consisting of superconducting layers in the vortex-free Meissner state and assume that their mutual influence is mainly due to their interaction with magnetic field. This approach makes it possible to develop a rigorous analysis of the properties of these superconducting structures. Note that the description of a superconducting plate in a parallel magnetic field based on the model of vortex-free state has a limited scope. It was shown in [12] that vortices begin to penetrate into a film when the field strength reaches $H_s(D) \propto \phi_0/D^2$ (ϕ_0 is the magnetic flux quantum, D is the film thickness), which is substantially stronger than the lower critical field H_{c1} characteristic of conventional bulk type II supercon-

† Deceased.

ductors. Furthermore, in the limit of $D < \lambda$ (magnetic field penetration depth), the highest superheating field for the Meissner state corresponds to an even stronger applied field strength [13]:

$$H_s \approx \phi_0/2\pi\xi D,$$

where ξ is the coherence length. This field strength restricts the applicability of our approach in the limit of $\kappa \gg 1$, where κ is the Ginzburg–Landau parameter. Thus, the present approach is valid in sufficiently wide ranges of magnetic field strength and film thickness. In recent analyses of properties of superconductors with $\kappa \sim 1$ [14, 15], it was shown that some of their characteristics must be inconsistent with current theories of both type I and type II superconductors. In this paper, we consider type II superconductors with $\kappa > 1$, which are of primary practical interest.

This area of research is very important, because the Ginzburg–Landau equations are of fundamental importance and their exact solutions can be used to deal with various problems in superconductivity, including validation of these equations as applied to high-temperature superconductors. Moreover, the results of this study lead to better understanding of the processes taking place in real superconducting structures.

2. STATEMENT OF THE PROBLEM

We consider a stack of long and wide superconducting plates of thickness D in a parallel magnetic field H. Each plate carries transport current perpendicular to the applied field. The transport current I_t is defined as the current density multiplied by the plate thickness, i.e., the current per unit plate width. The calculation of critical current for this structure is divided into two steps. First, a self-consistent solution to the Ginzburg–Landau equations is used to find the dependence of the critical current I_c on the applied magnetic field strength H for an individual plate, which is assumed to be in the vortex-free state. Second, the critical current is determined for a multilayer by finding an optimal distribution of transport current over individual plates.

We start from the Ginzburg–Landau equations [1] combined with Maxwell's equation:

$$\frac{4\pi}{c}\mathbf{j}_s = \frac{\psi^2}{\lambda^2}\left(\frac{\phi_0}{2\pi}\nabla\Theta - \mathbf{A}\right), \tag{1}$$

$$\nabla^2\psi\left(\nabla\Theta - \frac{2\pi}{\phi_0}\mathbf{A}\right)^2\psi + \frac{1}{\xi^2}(\psi - \psi^3) = 0, \tag{2}$$

$$\mathrm{curl}\,\mathrm{curl}\mathbf{A} = \frac{4\pi}{c}\mathbf{j}_s, \tag{3}$$

where $\mathbf{A}$ is the magnetic vector potential ($\mathbf{B} = \mathrm{curl}\mathbf{A}$), $\mathbf{j}_s$ is the supercurrent density, and c is the speed of light in free space. In the general case, the order parameter is expressed as $\Psi = \psi e^{i\Theta}$, where ψ and Θ are the corresponding magnitude and phase.

We write the Ginzburg–Landau equations in a Cartesian coordinate system (x, y, z) with y and z axes parallel to the plate surface and z axis parallel to the magnetic field, assuming that transport current flows along the y axis. Using the Landau gauge for $\mathbf{A}$, which entails $\mathbf{A} = \mathbf{e}_y A(x)$, we can rewrite the Ginzburg–Landau equations as

$$\frac{d^2U}{dx_\lambda^2} - \psi^2 U = 0, \tag{4}$$

$$\frac{d^2\psi}{dx_\lambda^2} + \kappa^2(\psi - \psi^3) - U^2\psi = 0. \tag{5}$$

We introduce dimensionless quantities $U(x_\lambda)$, $b(x_\lambda)$, and $j(x_\lambda)$ instead of the dimensional potential A, magnetic induction B, and current density j_s:

$$\begin{aligned} A &= \frac{\phi_0}{2\pi\lambda}U, \quad B = \frac{\phi_0}{2\pi\lambda^2}b, \quad b = \frac{dU}{dx_\lambda}, \\ j(x_\lambda) &= j_s\left(\frac{c\phi_0}{8\pi^2\lambda^3}\right)^{-1} = -\psi^2 U, \quad x_\lambda = \frac{x}{\lambda}. \end{aligned} \tag{6}$$

Since the transport current I_t carried by the plate generates the magnetic field

$$H_I = \frac{2\pi}{c}I_t, \tag{7}$$

the total field strengths at the plate surfaces are $H \pm H_I$. Accordingly, Eq. (4) is supplemented with the following boundary conditions:

$$b|_{x_\lambda = 0} = h - h_I, \quad b|_{x_\lambda = d} = h + h_I, \tag{8}$$

where

$$h = \frac{H}{H_\lambda}, \quad h_I = \frac{H_I}{H_\lambda}, \quad d = \frac{D}{\lambda}, \quad H_\lambda = \frac{\phi_0}{2\pi\lambda^2}.$$

Equation (5) is subject to standard boundary conditions on the plate surfaces [1]:

$$\left.\frac{d\psi}{dx_\lambda}\right|_{x_\lambda = 0} = 0, \quad \left.\frac{d\psi}{dx_\lambda}\right|_{x_\lambda = d} = 0. \tag{9}$$

Recalling that both London length λ and coherence length ξ are functions of temperature, we note that the relations written out above implicitly depend on temperature, being formally valid at any T. However, the

Ginzburg–Landau equations are applicable only at $T \longrightarrow T_c$.

To find a self-consistent solution to Eqs. (4) and (5), we use the following iterative procedure. Introducing a trial function $\psi(x_\lambda)$, we solve Eq. (4) for $U(x_\lambda)$. We substitute the resulting $U(x_\lambda)$ into Eq. (5) and use boundary conditions (9) to find a new $\psi(x_\lambda)$. Then, we solve Eq. (4) and repeat the procedure until both $\psi(x_\lambda)$ and $U(x_\lambda)$ become invariant and can therefore be adopted as a self-consistent solution to the system. It is obvious that the solution obtained by this method is stable with respect to small perturbations (see [16] for details). The critical current I_{ci} carried by the ith plate is set equal to the value of I_t corresponding to $\psi(x_\lambda) \equiv 0$. Thus, we find the critical current per unit width of a superconducting plate as a function of the applied magnetic field strength h. A more detailed description of this method was presented in [11], where we found the temperature dependence of depairing critical current density for several values of applied magnetic field strength and plate thickness. Moreover, it was shown in [11] that the expression for the Ginzburg–Landau critical current corresponding to zero applied magnetic field is valid for films of thickness comparable to coherence length and magnetic field penetration depth. Its value can be estimated by using the Ginzburg–Landau theory with a constant order parameter [17]:

$$I_c = \frac{1}{3\sqrt{6}}\frac{c}{\pi}H_{cm}\frac{D}{\lambda(T)}, \tag{10}$$

where H_{cm} is the thermodynamic critical field. This expression is obtained for $D \ll \lambda, \xi$.

Proceeding to the second step, we seek the critical current for the multilayer. We assume that adjacent superconducting layers are separated by relatively thick insulating layers, i.e., the Josephson coupling between the layers is negligible. To allow for electrical coupling between the superconducting layers, we assume that they are connected by superconducting links at $y = \pm\infty$. We seek such a distribution of transport current over the layers that transition to the normal state occurs in all layers simultaneously. If h_i is the magnetic field corresponding to the ith layer, then the current per unit width of the film in the critical state equals the critical current $I_c(h_i)$, which is determined by the numerical solution of the Ginzburg–Landau equations obtained at the first step. Under this condition, each layer in the structure carries a corresponding critical current. The current flowing through the ith plate generates the magnetic field given by (7), which is independent of the distance from the plate and has opposite directions on its opposite sides. According to the field superposition principle, we must add up the contributions of all layers to find the magnetic field that acts on the ith superconducting layer:

$$h_i = h + \sum_{j=1}^{i-1} h_{ij} - \sum_{j=i+1}^{N} h_{tj}, \tag{11}$$

where h_{tj} is the dimensionless magnetic field generated by the transport current carried by the jth layer. The magnetic field distribution of over the layers that corresponds to their simultaneous transition to the normal state is found by successive approximation. First, we set some initial conditions. For example, we assume that the magnetic field that acts on each layer is equal to the applied field, and the corresponding critical current per unit width of the film is $I_c(h)$. Then, we combine relations (7) and (11) to find the magnetic field for the ith layer. Using the previously calculated function $I_c(h)$, we determine the critical currents for the layers in the respective magnetic fields h_i and substitute them into (7) and (11) to find new values of h_i. The iterative process is terminated when the change in the critical currents from cycle to cycle becomes negligible. Note that this method can also be applied to analyze the critical states of multilayers consisting of different layers.

The magnetic field distribution over a multilayer consisting of similar layers in zero applied magnetic field can be found by a simpler method. Suppose that the number of layers is odd. First, consider a three-layer structure. By virtue of its symmetry, it is obvious that the central layer is in zero magnetic field and the corresponding critical current $I_c(0)$ is determined by the numerical solution of the Ginzburg–Landau equations obtained at the first step. The magnetic field H_{I3} acting on each outer layer is generated by the other two, and the corresponding critical current I_{c3} is

$$I_{c3} = I_c(H_{I3}) = I_c\left(\frac{2\pi}{c}(I_c(0) + I_{c3})\right). \tag{12}$$

This quantity can be found by fitting. It is obvious that there exists a unique value of $I_{c3}(h = 0)$ satisfying (12) if the initial $I_c(h)$ is a monotonically decreasing function. In a five-layer structure, the three central layers exhibit similar behavior in the critical state, because the magnetic fields generated by the outer layers compensate each other, and the corresponding critical current

$$I_{c5} = I_c\left(\frac{2\pi}{c}(I_c(0) + 2I_{c3} + I_{c5})\right) \tag{13}$$

can also be found by fitting. Adding two outer layers at a time and calculating the corresponding critical current by the method described above, we can find the critical current for a multilayer consisting of any num-

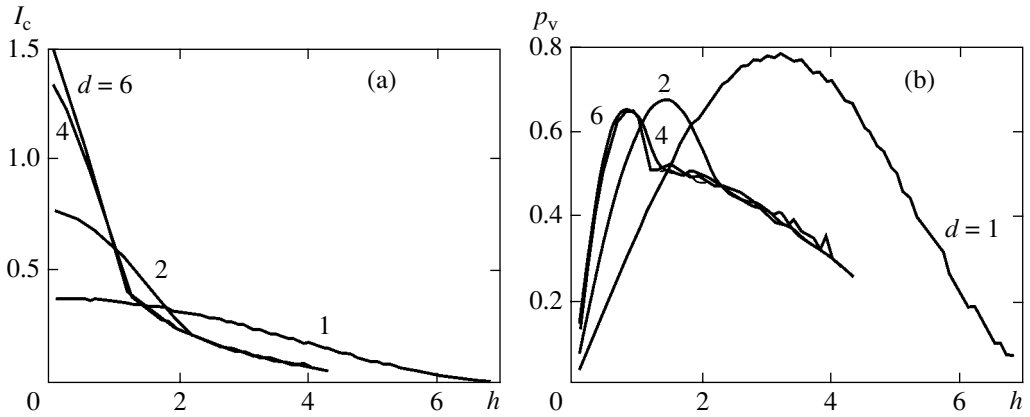

Fig. 1. Critical current (a) and $p_v = I_c h$ (b) vs. applied magnetic field h for $\kappa = 2$ and several values of dimensionless plate thickness d (indicated at curves). Noise in curves, particularly in (b), is due to limited numerical accuracy.

ber of layers. The critical current in the added outer layers can be expressed as

$$I_{cN} = I_c\left(\frac{2\pi}{c}\left[I_c(0) + I_{cN} + 2\sum_{n=1}^{L-1} I_{c2n+1}\right]\right), \quad (14a)$$

where $L = (N-1)/2$, and

$$I_{cN} = I_c\left(\frac{2\pi}{c}\left[I_{cN} + 2\sum_{n=1}^{L-1} I_{c2n}\right]\right), \quad (14b)$$

where $L = N/2$, for an even and odd number of layers with $N > 2$ and $N > 3$, respectively. The present analysis shows that there exists a unique distribution of I_{ci} over the layers in zero magnetic field in the case of a monotonically decreasing $I_c(h)$. Thus, if the field-dependent critical current $I_c(h)$ for a single layer is known (e.g., from experiment), the distribution of transport current over the layers in zero applied magnetic field can readily be found for the critical state of a multilayer consisting of an arbitrary number of layers.

3. NUMERICAL SOLUTION OF THE GINZBURG–LANDAU EQUATIONS FOR A SUPERCONDUCTING PLATE IN AN APPLIED MAGNETIC FIELD

Figure 1a shows $I_c(h)$ calculated for $\kappa = 2$ and several values of the superconducting plate thickness d at the first step. The curves demonstrate that critical current decreases with the layer thickness for small h, whereas $I_c(h)$ is higher for thin plates as compared to relatively thick ones at a moderate field strength. Moreover, thin superconducting plates can carry relatively low transport currents without dissipation in much stronger fields, as compared to thick ones. Note also that, for layers of thickness $d = 6$, the curves of $I_c(h)$ are indistinguishable; i.e., critical current becomes independent of the plate thickness.

We introduce the function $p_v(h) = I_c(h)/h$ as a measure of pinning force in a vortex-lattice model. In the vortex-free state considered here, the plate boundary is the only inhomogeneity, and $p_v(h)$ characterizes the effect of the boundary of a thin superconducting plate on the supercurrent that can be carried by the plate. Examples of this function are shown in Fig. 1b. Note that

$$p_v(h) \propto h(h_{c2} - h) \quad (15)$$

for $d \le 1$, where h_{c2} is the upper critical field for a plate, whereas there exists a value of applied magnetic field strength at which the derivative $\partial p_v(h)/\partial h$ sharply changes and a slow decrease in $p_v(h)$ is observed in stronger fields. Thus, the curve of $p_v(h)$ for a type II superconductor has three portions corresponding to domains of different characteristic behavior: an increasing $p_v(h)$ (domain I) and fast and slowly decreasing $p_v(h)$ (domains II and III). Note that the curve of $p_v(h)$ for a type I superconductor consists of only two portions described by (15) for any d.

The approach employed here makes it possible to obtain detailed information about the distributions of order parameter $\psi(x_\lambda)$, magnetic induction $b(x_\lambda)$, and current density $j(x_\lambda)$ for different values of transport current and applied magnetic field strength h. To understand the behavior of $p_v(h)$ in the case of a thick plate, we analyze these distributions for $d = 6$. Even though the vortex-free state can hardly be reached in such a plate, this case is discussed here to expose the phenomena in question. Figure 2 shows $\psi(x_\lambda)$, $b(x_\lambda)$, and $j(x_\lambda)$ calculated for several values of h in each domain of $p_v(h)$ and values of transport current close to the critical current. Curves *1* represent distributions corresponding to the domain of increasing $p_v(h)$ ($h = 0.5$). Curves *2* and *3* are obtained for values of h ($h = 1.09$ and 1.1,

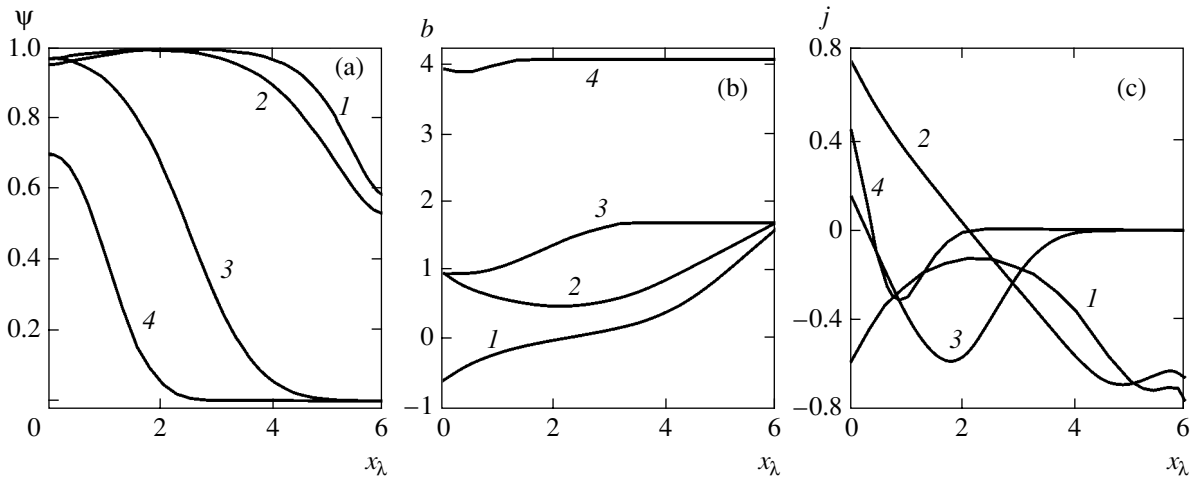

Fig. 2. Distributions of (a) order parameter, (b) magnetic induction, and (c) supercurrent across a superconducting plate for $d = 6$, $\kappa = 2$, and $I_t/I_c = 0.9$ at $h = 0.5$ (*1*), 1.09 (*2*), 1.1 (*3*), and 4.0 (*4*).

respectively) close to the field strength at which the slopes of $I_c(h)$ and $p_v(h)$ sharply change ($h \approx 1.095$). Curves *4* correspond to $h = 4.0$, i.e., the extreme right point of the $I_c(h)$ curve. The shapes of the curves representing domains I and II (curves *1* and *2* in Fig. 2) completely agree with the behavior of the Meissner state. In these domains, the order parameter is suppressed at the boundaries of the superconducting multilayer and slowly varies inside it, and magnetic field penetrates into the multilayer to a depth on the order of λ. The distributions obtained for domain III, where $p_v(h)$ slowly varies with increasing magnetic field (Fig. 1b), exhibit a totally different behavior (see curves *3* and *4* in Fig. 2). In particular, the order parameter $\psi(x_\lambda)$ strongly deviates from zero only at the left boundary of the plate and exponentially decreases toward its right boundary (curves *3* and *4* in Fig. 2a); i.e., the state of a considerable part of the plate is close to normal. Note that the distributions of all macroscopic characteristics of the superconducting plate drastically change at the point $h \approx 1.095$ separating the domains of fast and slow decrease in $p_v(h)$, as illustrated by curves *2* and *3* in Fig. 2. Now, consider the transition to the normal state caused by variation of transport current at $h \approx 4.0$ (curve *4* in Fig. 2). It is obvious that the size of the region of nonzero order parameter is $x_\lambda \approx 1$. Further increase in applied magnetic field strength leads to complete suppression of superconductivity even at zero transport current. Note that the superconducting order parameter vanishes abruptly at $I_t = I_c$, i.e., a first-order phase transition is observed. Thus, the transport properties of a thick plate in a strong magnetic field (see Fig. 1) should be attributed to surface superconductivity.

Next, we consider the effect of I_t on the order-parameter value ψ_0 at the plate boundaries as illustrated by the curves of $\psi_0(I_t)$ plotted in Fig. 3 for $\kappa = 0.5$ and 1.5 (Fig. 3). These curves demonstrate behavior characteristic of both type I and type II superconductors as the shape of $\psi_0(I_t)$ changes with increasing magnetic field strength. Note that the order-parameter values $\psi_0 \equiv \psi(x_\lambda = 0)$ and $\psi_1 \equiv \psi(x_\lambda = d)$ at the plate boundaries satisfy the relation $\psi_1(I_t) = \psi_0(-I_t)$; i.e., the curve of $\psi_1(I_t)$ is the mirror reflection of the curve of $\psi_0(-I_t)$ with respect to the line $I_t = 0$ in Fig. 3. It is clear that $\psi_0(I_t)$ is symmetric about the line $I_t = 0$ at $h = 0$ (see Fig. 3a). This is explained by the fact that suppression of superconductivity by transport current at the plate boundaries at zero applied magnetic field strength is independent of the direction of transport current. According to Fig. 3b, the symmetry of $\psi_0(I_t)$ is lost in applied magnetic field, and I_c decreases with increasing h. The absence of curves of $\psi_0(I_t)$ for $\kappa = 0.5$ in Fig. 3c is explained by the increase in the upper critical field with κ: superconductivity is suppressed by magnetic field in the plate with $\kappa = 0.5$, but persists in the plate with $\kappa = 1.5$. Note that ψ_0 drops from a finite value to zero as superconductivity disappears with increase in magnetic field strength.

Now, consider the behavior of $\psi_0(I_t)$ in strong magnetic fields for type II superconductors as illustrated by Fig. 3d ($h = 1.25$ and $\kappa = 1.5$). This curve substantially differs from those discussed above in that it has two new distinct portions: $\psi_0(I_t) \neq 0$ at low transport currents, whereas $\psi_0(I_t) \approx 0$ and $\psi_1(I_t) \approx 1$ at high currents (when $|I_t|$ is close to I_c). Thus, the low-current state is similar to a previously considered one and is characteristic of type I superconductors. The high-current state, in which the superconducting properties are due to surface superconductivity, is characterized by suppression of the order parameter at one boundary and its increase at the other. This state emerges abruptly. Note that the current state corresponding to the distributions of mac-

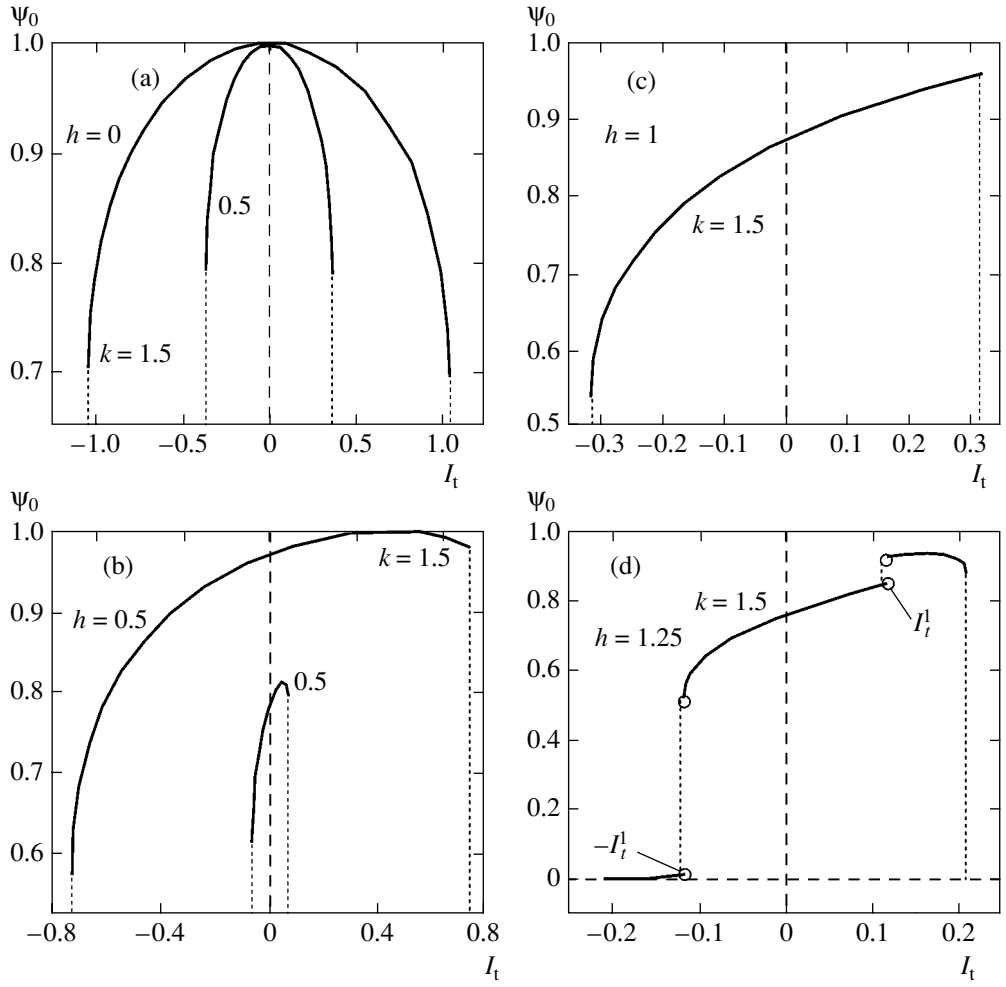

Fig. 3. Dependence of the order parameter at the boundary of a plate of thickness $D = 4\lambda$ on the magnitude and direction of transport current for several values of Ginzburg–Landau parameter and applied magnetic field strength.

roscopic characteristics illustrated by curves *3* and *4* in Fig. 2 is due to the strong asymmetry of boundary conditions (8). This state exists only in type II superconductors in strong magnetic fields (domain III in Fig. 1b) and at transport currents close to the critical current (Fig. 3d). Therefore, the domain of slowly decreasing $p_v(h)$ in Fig. 1b corresponds to this state.

The functions $I_c(h)$ obtained here can be interpreted as the upper bounds for the critical currents carried by superconducting plates with different thicknesses D in magnetic fields of different strength h. Numerical solution of the Ginzburg–Landau equations provides detailed information about macroscopic characteristics of superconductors in this state.

4. CALCULATION OF THE CRITICAL STATES OF SUPERCONDUCTING MULTILAYERS

Figure 4a illustrates the behavior of the average dimensionless current calculated by the method described above as a function of magnetic field,

$$\langle I_c \rangle = \frac{1}{N}\sum_{i=1}^{N} I_{ci},$$

where N is the number of layers in a multilayer (indicated at each curve), for $\kappa = 10$ and $d = 0.3$, i.e., $D = 3\xi$. When the number of layers is relatively small (several tens for layered structures of the type analyzed here), the critical current is close to the Ginzburg–Landau depairing current given by (10). The magnetic field generated by the layers increases with their number, and their increasing role manifests itself by decrease in critical current. This effect weakens with increasing magnetic field, and the average critical current approaches the value of $I_c(h)$ for a single layer. Note that $\langle I_c \rangle$ is a monotonically decreasing function of magnetic field strength in the model considered here for multilayers consisting of any number of layers.

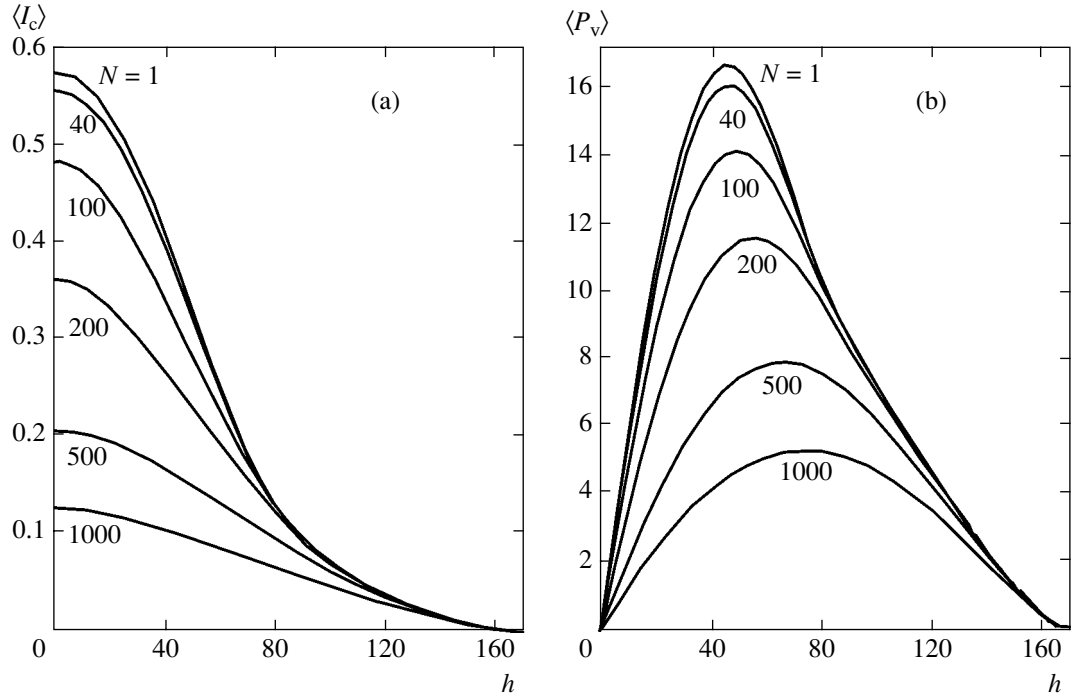

Fig. 4. Dependence of critical current density (a) and bulk pinning force (b) on applied magnetic field for several structures with different number of layers (indicated at curves) for $D = 3\xi$ and $\kappa = 10$.

We define a quantity analogous to the bulk pinning force: $P_v(h) = \langle I_c \rangle h$. Note that the individual layers considered here are in a vortex-free state; i.e., they are inhomogeneous only at their boundaries. This model provides a good approximation of the real state of SIS junctions, because the formation of the strongest pinning centers is due to vortex–boundary interaction. Figure 4b shows the bulk pinning force as a function of magnetic field strength calculated by this method. The most important result demonstrated here is the deviation from similarity. As the role played by the layers increases with their number, the slope $\partial P_v(h)/\partial h$ in the weak-field limit decreases, and the maximum of the curve of $P_v(h)$ moves toward higher field strengths, while its shape tends to that described by (15). Recall that the weak-field limit for an individual layer corresponds to the Meissner behavior (see curves *1* and *2* in Fig. 2), whereas transport properties are due to surface superconductivity (curves *3* and *4* in Fig. 2).

Figure 5 shows distributions of transport current and magnetic field over individual layers calculated by the present method for analyzing the critical states of superconducting multilayers. Here, $\langle b \rangle$ denotes either the magnetic induction averaged over the thickness of a layer or the strength of the magnetic field acting on the layer. Our calculations can also be used to obtain detailed magnetic field, current, and order-parameter distributions across each layer analogous to those presented in Fig. 2. In particular, the critical current density varies from layer to layer. At zero magnetic field strength, its distribution reaches a maximum in the central layers. When a magnetic field is applied, the maximum shifts toward one of the multilayer boundaries. Note that the distribution of transport current predicted for the critical state is similar to its distribution in at zero field strength, but is also shifted toward the boundary. This similarity admits the following simple explanation. It is obvious that the applied magnetic field h is equivalent to the superposition of the magnetic fields generated by l "virtual" layers if two conditions are satisfied. First, the total strength of the magnetic fields generated by these layers in each real layer must be equal to the corresponding applied field strength: $\sum_{j=1}^{l} h_j = h$, where h_j is the magnetic field generated by the jth virtual layer. Second, the transport-current distribution across the multilayer obtained by combining N real layers with l virtual ones in the critical state must be similar to the symmetric critical current-density distribution across the multilayer consisting of $N + l$ real layers at zero magnetic field strength. However, the central maximum of the critical current in the multilayer obtained by adding l virtual layers will not exactly match the shifted critical-current maximum in a real multilayer if the shift is not a multiple of the half-period of the layered structure. When the equivalent model is applied to determine $\langle I_c \rangle(h)$, the error due to this mismatch can be reduced by using the following algorithm: the maximum of the critical-current distribution over individual layers is shifted with a step equal to the multilayer half-period, and then the average critical-current densities and applied magnetic field

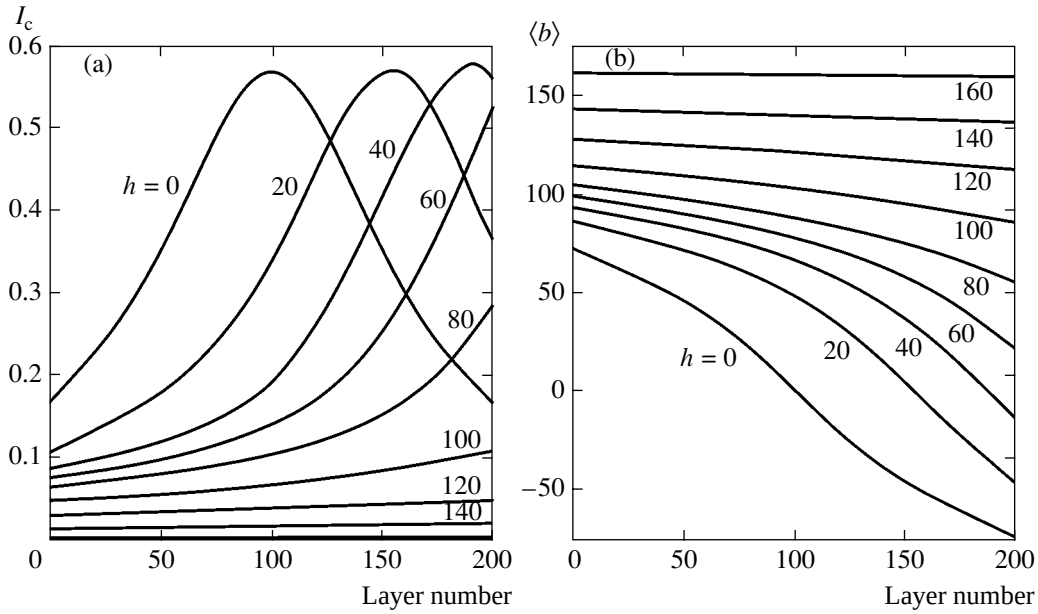

Fig. 5. Dependence of critical current (a) and local magnetic induction (b) on the location of a layer in a structure consisting of 200 layers for $D = 3\xi$ and $\kappa = 10$. Numbers at curves are applied magnetic field strengths.

strengths corresponding to the resulting distributions are calculated. Based on this method for calculating I_{ci}, we conclude that since the distribution of I_{ci} in zero magnetic field is unique, the distribution of I_{ci} in a nonzero magnetic field is also unique. Therefore, a known $I_c(h)$ for a single superconducting layer can be used to calculate the average critical current as a function of magnetic field strength and determine the distribution of transport current over a layered structure in the critical state.

Note that the resulting transport-current and magnetic field distributions over layers are different from those predicted by using the Bean [18] and Anderson–Kim [19] models. In the Bean model, the critical current density is constant across the sample, and the magnetic field varies as a linear function. In the Anderson–Kim model, magnetic field is characterized by a parabolic distribution. At moderate field strengths ($1 \ll h \ll h_{c2}$), the magnetic field distributions shown in Fig. 5b agree with those predicted by the Anderson–Kim model, whereas the distributions obtained for strong fields ($h \leq h_{c2}$) tend to exhibit linear behavior, as in the Bean model. Since the magnetic fields generated by the low critical currents corresponding to strong applied fields are weak, the transport current is uniformly distributed over individual layers in the critical state.

Note that the results presented here will not change significantly if the Josephson coupling between layers separated by a dielectric is taken into account, because the Josephson current density is lower than the depairing current density by several orders of magnitude, and its effect on the order-parameter modulation amplitude and the transport-current distribution is weak. However, if superconducting layers are separated by normal-metal layers, the proximity effect induces superconductivity in the normal conductors and suppresses the order parameter in the superconductors. In the thin-layer limit, this leads to weak order-parameter modulation in a layered structure. In this case, the Ginzburg–Landau equations can be solved for a structure embedded in a parallel magnetic field of strength close to the upper critical field [9]. Note that weak modulation of the order parameter corresponds to weak interaction between Abrikosov vortices and the lattice induced by inhomogeneities, i.e., to low critical current density.

5. CONCLUSIONS

The main results of the present study can be summarized as follows.

In the framework of a vortex-free model, a method based on solution of the Ginzburg–Landau equations is developed for finding the distributions of order parameter $\psi(x_\lambda)$, magnetic induction $b(x_\lambda)$, and current density $j(x_\lambda)$ for various transport currents carried by a superconducting plate and various strengths h of parallel applied magnetic field.

The critical current carried by a superconducting plate is calculated as a function of the applied magnetic field strength h. The effect of surface phenomena on the distribution of transport current across a superconducting plate is examined, and the dependence of processes taking place in the plate on its thickness D and the Ginzburg–Landau parameter κ is analyzed. The existence of a surface superconductivity regime at $h_{c1} < h < h_{c2}$ and a high transport current is revealed.

The nonlinear Ginzburg–Landau equations are solved numerically to calculate critical current as a function of magnetic field strength and both current and magnetic field distributions over layers for superconducting multilayers in parallel applied magnetic fields. The problem is solved for vortex-free layers whose mutual influence is entirely due to their interaction with magnetic field. A simple method is proposed for calculating and analyzing the critical states of layered structures in magnetic fields of arbitrary strength, based on elementary transformations of the critical current-density distribution over individual layers in zero applied magnetic field.

The vortex-free Meissner state of the layers assumed in this study is an important restriction, because vortices do exist in the layers examined in most experiments. Moreover, their interaction with pinning centers is the key factor that determines the critical current density. The dependence $\langle I_c \rangle(h)$ obtained here can be interpreted as an upper bound for the critical current carried by a multilayer in a parallel applied magnetic field. Note that the vortex-free Meissner state cannot exist in an oblique magnetic field, because the perpendicular field component will induce vortices in the layers; i.e., the critical current density in an oblique field will be determined by the pinning of these vortices.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Industry, Science, and Technology of the Russian Federation under State Contract no. 40.012.1.1.1357.

REFERENCES

1. V. L. Ginzburg and L. D. Landau, Zh. Éksp. Teor. Fiz. **10**, 1064 (1950).
2. H. J. Fink and A. G. Presson, Phys. Rev. **151**, 219 (1966).
3. V. V. Moshchalkov, X. G. Qiu, and V. Bruyndoncx, Phys. Rev. B **55**, 11 793 (1997).
4. P. S. Deo, V. A. Shweigert, F. M. Peeters, and A. K. Geim, Phys. Rev. Lett. **79**, 4653 (1997).
5. V. A. Schweigert and F. M. Peeters, Phys. Rev. B **57**, 13817 (1998).
6. G. F. Zharkov, V. G. Zharkov, and A. Yu. Tsvetkov, Phys. Rev. B **61**, 12 293 (2000).
7. G. Blatter, M. V. Feigel'man, V. B. Geshkenbein, *et al.*, Rev. Mod. Phys. **66**, 1125 (1994).
8. S. Takacs, Phys. Status Solidi A **74**, 437 (1982).
9. S. Ami and K. Maki, Prog. Theor. Phys. **53**, 1 (1975).
10. M. Kulic and L. Dobrosavljevic, Phys. Status Solidi B **75**, 677 (1976).
11. A. Yu. Tsvetkov, G. F. Zharkov, and A. N. Lykov, Kratk. Soobshch. Fiz., No. 6, 25 (2004).
12. A. A. Abrikosov, Zh. Éksp. Teor. Fiz. **46**, 1464 (1964) [Sov. Phys. JETP **19**, 988 (1964)].
13. V. V. Shmidt and G. S. Mkrtchyan, Usp. Fiz. Nauk **112**, 459 (1974) [Sov. Phys. Usp. **17**, 170 (1974)].
14. V. I. Marchenko and E. R. Podolyak, Zh. Éksp. Teor. Fiz. **121**, 235 (2002) [JETP **94**, 200 (2002)].
15. V. I. Marchenko and E. R. Podolyak, Zh. Éksp. Teor. Fiz. **124**, 172 (2003) [JETP **97**, 154 (2003)].
16. G. F. Zharkov and V. G. Zharkov, Phys. Scr. **57**, 664 (1998).
17. V. L. Ginzburg, Dokl. Akad. Nauk SSSR **118**, 464 (1958) [Sov. Phys. Dokl. **3**, 102 (1958)].
18. C. P. Bean, Rev. Mod. Phys. **36**, 31 (1964).
19. P. W. Anderson and Y. B. Kim, Rev. Mod. Phys. **36**, 39 (1964).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 350–356.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 403–410.

ELECTRONIC PROPERTIES OF SOLIDS

To the Quantum Theory of Chemical Activity of the Surface of Transition Metals

V. E. Egorushkin, S. E. Kul'kova, N. V. Mel'nikova, and A. N. Ponomarev

Institute of Strength Physics and Materials Science, Siberian Division, Russian Academy of Sciences, Tomsk, 634021 Russia

e-mail: kulkova@ms.tsc.ru

Received January 11, 2005

Abstract—The dissociation of a gas molecule and the formation of a new chemical bond upon adsorption of this molecule on the surface of a transition metal are studied using the method of equations of motion. It is shown that both processes involve the formation of a mixed intermediate state during the adsorbate–substrate interaction. The dissociation is caused by a resonance growth of the vibrational mode, whereby the dissociation barrier is determined by the hybridization energy and by the frequency of electron transitions between molecular levels and the *d* electron energy levels of the metal in the mixed intermediate state. The resonance conditions for the formation of new surface structures are established. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

The chemical activity (reactivity) of the surface of a metal consists in the ability to break chemical bonds in the molecules of a gas or a liquid occurring in contact with the surface, and to form new chemical bonds. The surface activity plays a determining role in various physical phenomena such as adsorption, catalysis, and corrosion. This activity cannot be explained by considering only the behavior of electrons [1]. Nevertheless, the electron structure of solid surfaces has received special attention in investigations of the above phenomena. The results of theoretical calculations [2–4] showed that the activity of the surface of transition metals and a barrier for the dissociation of various molecules on this surface are determined by relations between the energies of the *d* electron states and the adsorbate states, the Fermi level, and the hybridization energy. In addition, the final structure of the adsorbate and the energy of adsorption depend on the atomic geometry and relaxation of the surface [5]. The existing descriptions of various surface processes [5] leave open the questions concerning mechanisms of the surface activity. In order to answer these questions, it is necessary to consider in detail the quantum kinetics of the many-electron system involved in the dissociation of adsorbed molecules and the formation of new chemical bonds.

During the formation of new chemical bonds, atoms or molecules (in the absence of dissociation) exhibit a kind of ordering and form islands with a short-range order, which are composed of hydrides, oxides, etc. Under certain conditions, new structures can form and decay. The adsorbed atoms are not merely kept on the surface, but are capable of forming new molecules. The three processes—dissociation and the formation of new bonds (structures) and new molecules—can be jointly considered in the analysis of relaxation of the electron–ion system of interacting molecules and the surface.

The relaxation of a gas–solid heterogeneous system is determined by the interactions between electrons of the gas and phonons of the metal [6, 7]. However, there is another kind of relaxation in systems such as amorphous metals, where short-range-order islands can be formed [8–10]. In this case, new chemical bonds are formed in the course of the electron–electron interactions, with the corresponding reconstruction of the atomic configuration. The results of such a relaxation strongly depend on the concentrations of components, the degree of the short-range order, the size of clusters, and the thermodynamical conditions. All these aspects are also important for the adsorption processes, which depend on the degree of surface coverage and on the systematic heating and cooling of the system [5]. It is also important to note that hybridization and electron–electron interactions involved in the adsorption lead to the formation of a new, mixed intermediate state of the molecule–surface system, which is essentially a precursor to the final state (oxides, hydrides, etc.) [9]. This intermediate state features the dissociation of adsorbed molecules and the formation of new chemical bonds. The new bonds appear in the course of relaxation of the excited electrons of gas atoms to a partly filled surface *d* electron band. The electron–electron interactions in this system are mediated by the dynamic concentration excitations of a boson type [10], rather than by phonons. This boson-type interaction introduces an additional uncertainty (besides that related to the hybridization) in the electron energies, which can lead to anomalies in the transport properties of a system featuring structural relaxation [8].

Below we will describe such a relaxation by in terms of the equations of motion and consider the mechanisms and conditions controlling the dissociation of gas molecules and the formation of new chemical bonds and compounds—that is, we will analyze the mechanisms and conditions determining the chemical activity of the surface of transition metals.

2. PHYSICAL MODEL AND HAMILTONIAN

Let us consider two stages in the interaction of a gas molecule (X_2) with the surface of a transition metal (Me), which are described by the reaction

$$\mathrm{Me} + \mathrm{X}_2 \longrightarrow \mathrm{Me} + 2\mathrm{X} \longrightarrow \mathrm{MeX}_2.$$

In the initial stage of this interaction, the molecule dissociates and then the two atoms are deposited from the gas phase to form new chemical bonds. Accordingly, the description can be subdivided into two parts: the first, devoted to the hybridization of separated electrons of the dissociated molecule and the metal surface, and the second, devoted to the interaction of common electron states with ions of the adsorbate–metal substrate system.

The first stage will be studied in terms of the standard Hamiltonian in the secondary quantization representation:

$$H_1 = \sum_{l,\sigma}(\varepsilon_0)_l c^+_{l\sigma}c_{l\sigma} + \sum_{m,\sigma}(\varepsilon_d)_m a^+_{m\sigma}a_{m\sigma} + \sum_{l,m,\sigma} V_{lm}(c^+_{l\sigma}a_{m\sigma} + a^+_{m\sigma}c_{l\sigma}), \quad (1)$$

where $c^+_{l\sigma}$ ($c_{l\sigma}$) are the electron creation (annihilation) operators of the gas molecule; $a^+_{m\sigma}$ ($a_{m\sigma}$) are the *d* electron creation (annihilation) operators of the metal surface in the Wannier representation; ε_0 and ε_d are the electron energies in the molecular states and the *d* metal states, respectively (measured relative to the Fermi level); V_{lm} are the hybridization matrix elements; *m* are the coordinates of lattice sites occupied by atoms of the *d* metal; *l* are the positions of adsorbed atoms (top, bridge, fourfold hollow, etc.); and σ is the spin index. In the general case, the surface states with the ε_0 and ε_d energies are decaying.

In considering the second interaction stage, let us define states in the ion subsystem as the dynamic atomic concentration excitations (DCEs) above the ground state corresponding to the metal surface structure. Then, we will consider the interaction of these states with the surface electrons using the method described in [10]. The corresponding Hamiltonian H_2 will include the contributions due to the electron–electron, DCE–DCE, and the DCE–electron interactions. In order to write this Hamiltonian, let us consider the adsorption-induced changes in the ion subsystem and define the set of occupation numbers $C_i(\mathbf{R})$ for positions $\mathbf{R}$ occupied by atoms of the *i*th kind, their average $\langle C_i(\mathbf{R})\rangle$ (equal to the macroscopic concentration C_i), and the concentration fluctuations

$$\delta C_i(\mathbf{R}) = C_i(\mathbf{R}) - C_i.$$

These fluctuations determine the short-range order in the new structure, while the Fourier image $\langle \delta C \delta C\rangle_{\mathbf{q}}$ determines the surface wavevector $\mathbf{q}$ corresponding to a certain type of short-range order [11].

In these terms, the second-stage Hamiltonian can be written as

$$H_2 = \int d\mathbf{r}\psi^+(\mathbf{r})H_e\psi(\mathbf{r}) + \frac{1}{2}\sum_{\substack{\mathbf{R},\mathbf{R}'\\ i,j}} W_{ij}(\mathbf{R}-\mathbf{R}')\delta C_i(\mathbf{R})\delta C_j(\mathbf{R}') + \int d\mathbf{r}\sum_{\mathbf{R}} g(\mathbf{r}-\mathbf{R})\psi^+(\mathbf{r})\psi(\mathbf{r})\delta C(\mathbf{R}), \quad (2)$$

where the first term represents the electron part (H_e is the electron Hamiltonian with the field operators ψ), the second term corresponds to the configurational part, and the third term describes the interaction between the electron and ion configuration subsystems. In Eq. (2), $W(\mathbf{R}-\mathbf{R}')$ is the density of the adsorption energy,

$$g(\mathbf{r}-\mathbf{R}) = \frac{z}{|\mathbf{r}-\mathbf{R}|},$$

and *z* is the charge of the adsorbed ion in the metal–adsorbate "solution." Interactions of the type

$$\int d\mathbf{r}\sum g(\mathbf{r}-\mathbf{R})\psi^+(\mathbf{r})\psi(\mathbf{r})$$

give additive contributions to the spectrum and are not included into the expression for H_2.

The electron operators and the concentration field can be represented as

$$\psi^+(\mathbf{r}) = \sum_{\mathbf{k}} a^+_{\mathbf{k}}e^{-i\mathbf{k}\cdot\mathbf{r}}, \quad \psi(r) = \sum_{\mathbf{k}} a_{\mathbf{k}}e^{i\mathbf{k}\cdot\mathbf{r}}, \quad (3)$$

$$\delta C(\mathbf{R}) = \frac{1}{2}\sqrt{C(1-C)}[b^+(\mathbf{R}) + b(\mathbf{R})], \quad (4)$$

where $b^+(\mathbf{R}) = \alpha^+_m(\mathbf{R})\alpha_l(\mathbf{R})$ and $b(\mathbf{R})$ are the operators of creation and annihilation, respectively, for a pair of metal and adsorbate atoms (i.e., of the creation and annihilation of a new chemical bond); $\alpha^+_m(\mathbf{R})$ corresponds to a metal atom at the *m*th lattice site and $\alpha_l(\mathbf{R})$

corresponds to an adsorbed atom at the *l*th lattice site (the subscripts at b^+ and b are omitted, since metal atoms occur at the fixed lattice sites; $\mathbf{R}_l \equiv \mathbf{R}$).

Now it is convenient to pass from the coordinate to **k** representation in Eq. (4), because the new bonds (atomic pairs) form identical bosons and the system should be symmetric with respect to permutations of these bosons. The **k** representation will automatically take this symmetry into account. In order to pass to this representation, let us introduce the elementary excitations

$$b_{\mathbf{k}} = \frac{1}{\sqrt{N}}\sum_{\mathbf{R}} b(\mathbf{R})e^{i\mathbf{k}\cdot\mathbf{R}},$$
$$b_{\mathbf{k}}^+ = \frac{1}{\sqrt{N}}\sum_{\mathbf{R}} b^+(\mathbf{R})e^{-i\mathbf{k}\cdot\mathbf{R}}, \quad (5)$$

where N is the number of atoms in the lattice on the metal surface. At small degrees of coverage, these excitations represent the aforementioned bosons. Substituting relations (3)–(5) into Eq. (2) and determining the Fourier image of function $g(\mathbf{R})$ as

$$\Delta(\mathbf{q}) = \sqrt{C(1-C)}\sum_{\mathbf{R}} g(\mathbf{R})e^{i\mathbf{q}\cdot\mathbf{R}}, \quad (6)$$

we obtain the Hamiltonian in the secondary quantization representation:

$$H_2 = \sum_{\mathbf{k}} \varepsilon(\mathbf{k})a_{\mathbf{k}}^+ a_{\mathbf{k}} + \frac{1}{8}C(1-C)$$
$$\times \sum_{\mathbf{q}} \{W(\mathbf{q})[2b_{\mathbf{q}}^+ b_{\mathbf{q}} + b_{-\mathbf{q}}b_{\mathbf{q}} + b_{\mathbf{q}}^+ b_{-\mathbf{q}}^+] - \mu b_{\mathbf{q}}^+ b_{\mathbf{q}}\} \quad (7)$$
$$+ \sum_{\mathbf{k},\,\mathbf{q}} \Delta(\mathbf{q})[a_{\mathbf{k}+\mathbf{q}}^+ a_{\mathbf{k}} b_{\mathbf{q}} + a_{\mathbf{k}}^+ a_{\mathbf{k}+\mathbf{q}} b_{\mathbf{q}}^+],$$

where $W(\mathbf{q})$ is the Fourier image of the energy density $W(\mathbf{R})$, $\varepsilon(\mathbf{k})$ is the electron spectrum of the adsorbate–metal substrate system, and μ is the chemical potential of elementary excitations.

3. THE METHOD OF EQUATIONS OF MOTION AND THE DISSOCIATION OF MOLECULES

Let us consider the first interaction stage using the method of equations of motion,

$$\dot{c}_{l\sigma}^+ = i[H_l, c_{l\sigma}^+], \quad \dot{a}_{m\sigma} = i[H_l, a_{m\sigma}^+], \quad (8)$$

where $[\ldots, \ldots]$ is the commutator. Accomplishing simple transformations, we obtain the following system of equations for a single spin subband:

$$\ddot{c}_l^+ + \gamma\dot{c}_l^+ + \sum_n |V_{ln}|^2 c_l^+ = \sum_n V_{ln}\omega_{ln}\exp(-\omega_{ln}t)a_n^+, \quad (9)$$

$$\ddot{a}_m^+ + \sum_p |V_{pm}|^2 a_m^+ = \sum_p V_{pm}(\omega_{pm} - i\gamma)\exp(\omega_{lp}t)c_p^+, \quad (10)$$

where $\omega_{lm} = (\varepsilon_0)_l - (\varepsilon_d)_m$ (in $\hbar$ units) and γ is the c_l^+ mode decay due to hybridization. The decay of *d* states is not taken into account. Equations (9) and (10) represent two interrelated oscillation processes, in which the hybridization V plays the role of the natural frequency of the electron density oscillations, $\omega = \varepsilon_0 - \varepsilon_d$ determines the driving frequency, and the product $V\omega$ enters into the expression for the driving force amplitude. The amplitude of the external force for the adsorbed molecules is determined by the electron density in the metal; for the surface, it is determined by the electron density in the molecule.

As can be seen from Eqs. (9) and (10), the gas molecules and the metal surface at the very beginning of the interaction form a new, mixed intermediate state with common electronic and vibrational modes. At the initial stage of the interaction, these modes differ only slightly from those in the initial non-interacting systems, but the symmetries of these modes and of the new intermediate state on the whole differ from the initial symmetry of both the molecule and the surface. Under certain conditions, the evolution of this state leads to the dissociation of adsorbed molecules. In order to establish these conditions, let us solve system of equations (9) and (10) in a two-mode approximation by assuming that only the terms with $n = m$ and $p = l$ are significant in summing over n and p.

Seeking a solution to Eq. (10) in the form of $a_m^+ = \rho e^{i\omega t}$, we express a_m^+ from Eq. (10) and substitute this expression into Eq. (9), which yields

$$\ddot{c}_l^+ + \gamma\dot{c}_l^+ + V^2\frac{V^2 + i\gamma\omega}{V^2 - \omega^2}c_l^+ = 0, \quad (11)$$

where $V_{lm} = V$ and $\omega_{lm} = \omega$. This equation describes the reactive (vibrational) and relaxation motions in terms of the average values. In both form and meaning, Eq. (11) is identical to the phenomenological equation of

motion for fluctuations in the "order parameter" [12], in this case—in the electron density c_l^+ in the molecule. Therefore, the coefficient at c_l^+ in Eq. (11) is the inverse susceptibility (χ^{-1}) in the intermediate state of the system.

As is known, the χ^{-1} value corresponds to the square frequency Ω^2 of the normal vibrational mode (renormalized due to hybridization) of the molecule [12]:

$$\chi^{-1} = \Omega^2 = V^2\frac{V^2 + i\gamma\omega}{V^2 - \omega^2}. \quad (12)$$

This relation shows that, for $\omega^2 \longrightarrow V^2$ ($\omega^2 < V^2$), the frequency of oscillations exhibits a resonance growth. This resonance accounts for the dissociation of molecules when the $\hbar\Omega$ value becomes equal to the dissociation energy D. By equating $\mathrm{Re}\Omega$ to D, we obtain $D^2 \approx V^2 + \omega^2$ ($D_{\min} \sim \sqrt{2}\,|\omega_{\min}|$). Therefore, the higher the Fermi energy ε_F and the closer ε_d to ε_F, the faster the difference $\varepsilon_0 - \varepsilon_d$ tends to zero and the lower the dissociation barrier: $D_{\min} \longrightarrow 0$ for $\varepsilon_d \leq \varepsilon_F$ (i.e. for a nearly half-filled d band in the metal).

Thus, the values of V, ω, and their ratio determine the activity of the metal surface with respect to dissociation. The dissociation proceeds according to the mechanism of the "rigid" vibrational mode, which appears under the conditions of a resonance between $\omega = \varepsilon_0 - \varepsilon_d$ and V upon the hybridization of electron states of the molecule and the surface d states of the metal. There are two possible situations:

(i) The resonance of small V and ω values: low dissociation barrier. In this case, the molecular states are hybridized with narrow d bands and the ε_0 level occurs in the region of d states below or at the Fermi levels. If the ε_0 level does not coincide with the Fermi level, the resonance conditions are not obeyed.

(ii) The resonance of large V and ω values: high dissociation barrier. In this case, the hybridized involves less localized states of the metal and the ε_0 level occurs far from the center of the d band. The mode with the frequency Ω exhibits substantial decay, which also complicates the dissociation of molecules.

Now let us consider the regime of electron motions in the molecule. Substituting $c_l^+ = \rho e^{i\omega t}$ into Eq. (11), we obtain the following results:

(i) For $\gamma \ll V \ll \omega$, we have $\chi^{-1} \longrightarrow 0$ (the system is far from the resonance) and

$$c_l^+ = \rho\exp\left(\frac{iV^2t}{\omega} + \frac{V^4t}{\gamma\omega^2}\right). \quad (13)$$

This solution to Eq. (13) decays only when $\gamma < 0$ ("negative friction"). In this stage, the metal surface acts as a reservoir of energy for the molecule and excites the molecular states. Thus, in the beginning of the interaction (when a molecule is at a large distance from the surface), the repulsion increases the energy of electron states in the molecule until the ε_0 level approaches close to ε_F so that the system will pass to the other regime. In this regime, the ε_0 level occurs below the Fermi level and approaches ε_d, thus driving the system to resonance.

(ii) For $\gamma \ll V \leq \omega$, we have $\chi^{-1} \longrightarrow \infty$ (the system is close to the resonance) and

$$c_l^+ = \rho\exp\left[-\frac{i\gamma t}{2}\sqrt{\frac{\omega}{2(\omega - V)}} - Vt\sqrt{\frac{\omega}{2(\omega - V)}}\right]. \quad (14)$$

In this case, the decay at any γ determines the dissociation of molecules for any $\omega \longrightarrow V$.

(iii) For $\gamma \ll \omega \leq V$. $\chi^{-1} \longrightarrow \infty$ (far from the resonance) and we have

$$c_l^+ = \rho\exp\left[iVt\sqrt{\frac{\omega}{2(V - \omega)}} - \frac{\gamma t}{2}\sqrt{\frac{\omega}{2(V - \omega)}}\right]. \quad (15)$$

The oscillations decay for $\gamma > 0$, which implies that the surface takes the energy from an adsorbed molecule until it decays at $\omega \longrightarrow V$.

4. THE RELAXATION OF ELECTRON STATES AND THE FORMATION OF NEW COMPOUNDS ON THE SURFACE

In the stage of interaction involving the relaxation of electron states and the formation of new chemical bonds, the ε_0 level falls under the Fermi level and approaches ε_d. The electrons of adsorbed molecules form the states with inversed populations relative to the unoccupied d states of the metal. It should be noted that falling under the Fermi level can be related to the uncertainty $\Delta\varepsilon$ in the energy. This uncertainty in the energy appears because the interaction event is localized in time with an uncertainty of $\Delta t \sim 10^{-15}$ s, whereas the characteristic electron interaction time is on the order of 10^{-14} s. The resulting value of $\Delta\varepsilon$ is on the order of several electronvolts. This implies that electrons of the molecule, which possess an energy of $\varepsilon_0 \sim \varepsilon_d$, may occur in the surface band of the metal owing to the uncertainty principle.

For the deposition of atoms on the substrate surface to be irreversible, it is necessary that the inversely populated states would exhibit relaxation. A change in the population naturally leads to the appearance of polarization which, in turn, leads to the excitation of slower modes in the ion subsystem. All these processes are interrelated. In the system under consideration, a slower mode represents the dynamic concentration wave forming a region with the new short-range order.

In order to describe this process, let us write the equations of motion for the operators $b_{\mathbf{q}}^{+}$, $b_{\mathbf{q}}$ of the new chemical bonds, the operators $d_{\mathbf{kq}}$ of the electron polarization due to electron transitions ($d_{\mathbf{kq}} = a_{\mathbf{k}}^{+} a_{\mathbf{k}+\mathbf{q}}$ and $d_{\mathbf{kq}}^{+} = a_{\mathbf{k}+\mathbf{q}}^{+} a_{\mathbf{k}}$), and the operators $\sigma_{\mathbf{kq}}$ of the inverse population ($\sigma_{\mathbf{kq}} = a_{\mathbf{k}+\mathbf{q}}^{+} a_{\mathbf{k}+\mathbf{q}} - a_{\mathbf{k}}^{+} a_{\mathbf{k}}$):

$$\dot{b}_{\mathbf{q}}^{+} = i[H_2, b_{\mathbf{q}}^{+}], \quad \dot{d}_{\mathbf{kq}}^{+} = i[H_2, d_{\mathbf{kq}}^{+}], \quad \dot{\sigma}_{\mathbf{kq}}^{+} = i[H_2, \sigma_{\mathbf{kq}}]. \tag{16}$$

Accomplishing transformations and using the approximation of exponential relaxation in the absence of decay of the $b_{\mathbf{q}}^{+}$ mode, we obtain

$$\dot{b}_{\mathbf{q}}^{+} = i[2C(1-C)W(\mathbf{q}) - \mu]b_{\mathbf{q}}^{+} + i\Delta(\mathbf{q})\sum_{\mathbf{k}} d_{\mathbf{kq}}^{+}, \tag{17a}$$

$$\dot{d}_{\mathbf{kq}}^{+} = (i\omega_{\mathbf{kq}} - \Gamma_{\mathbf{q}})d_{\mathbf{kq}}^{+} - i\Delta(\mathbf{q})\sigma_{\mathbf{kq}}b_{\mathbf{q}}^{+}, \tag{17b}$$

$$\dot{\sigma}_{\mathbf{kq}} = -\sigma_{\mathbf{kq}}^{0}\exp(-t/\tau_{\sigma}), \tag{17c}$$

where $\omega_{\mathbf{kq}} = \varepsilon_{\mathbf{k}+\mathbf{q}} - \varepsilon_{\mathbf{k}}$, $\Gamma_{\mathbf{q}}$ is the decay of polarization related to the electron–electron interaction, $\sigma_{\mathbf{kq}}^{0}$ is the average population at the beginning of the process, and τ_{σ} is the population relaxation time. According to Eqs. (17), the polarization $d_{\mathbf{q}}$, which excites the $b_{\mathbf{q}}^{+}$ mode, has the frequency $\omega_{\mathbf{kq}}$ and decays over a time on the order of $\Gamma_{\mathbf{q}}^{-1}$. The last nonlinear term in the right-hand side of Eq. (17b) describes a change in the polarization related to the change in the population and the formation of new chemical bonds.

Let us consider a simpler variant of Eqs. (7). The sum $\sum_{\mathbf{k}} d_{\mathbf{kq}}^{+}$ in the right-hand side of Eq. (17a) allows us to assume that the $b_{\mathbf{q}}^{+}$ mode does not decay in the stationary regime and the new phase (at a given coverage) uniformly occupies most of the metal surface. Then, taking into account that

$$\sum_{\mathbf{k}} \exp[i\mathbf{k}\cdot(\mathbf{R}_n - \mathbf{R}_m)] = \delta_{n,m},$$

we obtain

$$\sum_{\mathbf{k}} d_{\mathbf{kq}}^{+} = n_{\mathbf{q}}, \quad \sum_{\mathbf{k}} d_{\mathbf{kq}} = n_{-\mathbf{q}},$$

$$\sum_{\mathbf{k}} \sigma_{\mathbf{kq}} = 0, \quad \sum_{\mathbf{k}} \omega_{\mathbf{kq}} d_{\mathbf{kq}}^{+} = 0,$$

where $n_{\mathbf{q}}$ and $n_{-\mathbf{q}}$ are the numbers of emitted and absorbed electrons. These relations imply that the inversion population no longer exists. Writing the explicit time dependences,

$$b_{\mathbf{q}}^{+} = \exp(i\omega_{\mathbf{q}}t)b_{\mathbf{q}}^{+}, \quad n_{\mathbf{q}} = \exp(-i\omega_{\mathbf{q}}t)n_{\mathbf{q}},$$

considering the exact resonance,

$$\omega_{\mathbf{kq}} = \omega_{\mathbf{q}} = 2C(1-C)W(\mathbf{q}) - \mu, \tag{18}$$

and taking the sum over $\mathbf{k}$ in Eq. (17a), we eventually obtain the relations

$$\dot{b}_{\mathbf{q}}^{+} = i\Delta(\mathbf{q})n_{\mathbf{q}}^{0}\exp(-i\omega_{\mathbf{q}}t - \Gamma_{\mathbf{q}}t), \quad \dot{b}_{\mathbf{q}} = -i\Delta(\mathbf{q})n_{-\mathbf{q}}^{0}\exp(i\omega_{\mathbf{q}}t - \Gamma_{\mathbf{q}}t), \tag{19}$$

where $n_{\mathbf{q}}^{0}$ and $n_{-\mathbf{q}}^{0}$ are the numbers of occupied and vacant states in the energy band of the metal.

Equations (19) show that it is the donor–acceptor interaction that determines the appearance of a new chemical bond in the system under consideration. In addition to the one-electron characteristics (the number of transferred electrons and the relative posit ions of $\varepsilon_{\mathbf{k}}$ and $\varepsilon_{\mathbf{k}+\mathbf{q}}$ terms), a significant role is played by the structural and thermodynamic characteristics such as the electron–bond interactions, the degree of surface coverage, and the energy of adsorption (more exactly, the behavior of the potential surface in the region of wavevectors $\mathbf{q}$ corresponding to a given type of the short-range order).

Thus, Eqs. (7) and (19) interrelate the electron, structural, and thermodynamic aspects of the formation of new chemical bonds during dissociative gas adsorption on a solid surface.

Upon solving Eqs. (19) with the initial conditions

$$b_{\mathbf{q}}^{+}b_{\mathbf{q}}\big|_{t=0} = 0,$$

we obtain the following relation for the values averaged over a nonequilibrium ensemble:

$$\overline{b_{\mathbf{q}}^{+}b_{\mathbf{q}}} = N_{\mathbf{q}} = \frac{2\Delta(\mathbf{q})n_{\mathbf{q}}^{0}n_{-\mathbf{q}}^{0}}{\omega_{\mathbf{q}}^{2} + \Gamma_{\mathbf{q}}^{2}} \times \exp(-\Gamma_{\mathbf{q}}t)(\cosh\Gamma_{\mathbf{q}}t - \cos\omega_{\mathbf{q}}t). \tag{20}$$

The resonance condition (18) relates the energy difference of the donor–acceptor states to the Fourier image of the adsorption energy density. Under the conditions formulated above, the quantity $\omega_{\mathbf{q}} = \varepsilon_{\mathbf{k}+\mathbf{q}} - \varepsilon_{\mathbf{k}}$ weakly depends on C. This implies that the adsorption energy is inversely proportional to the surface coverage L:

$$W_0 = \frac{W}{N} \approx \frac{\text{const}\cdot N}{L(N-L)} \propto \frac{\text{const}}{L}. \tag{21}$$

In Eq. (20), the $N_{\mathbf{q}}$ value is equivalent to the reaction yield and $\eta(t) = N_{\mathbf{q}}(t)/N_0$ is the probability of the deposition reaction, where

$$N_0 = \frac{2\Delta(\mathbf{q})n_{\mathbf{q}}^0 n_{-\mathbf{q}}^0}{\omega_{\mathbf{q}}^2 + \Gamma_{\mathbf{q}}^2}$$

is proportional to a product of the numbers of occupied and vacant electron states in the energy band of the metal. For *d* electrons, this product reaches maximum for the half-filled band. The N_0 value grows with increasing parameter $\Delta(\mathbf{q})$ (determining the adsorbate binding to the substrate) and drops with increasing $\omega_{\mathbf{q}}$. The $\Delta(\mathbf{q})$ value determines only the amplitude (N_0) and does not influence the time dependence of $\eta(t)$.

Let us consider the kinetics of formation of the new chemical bonds, which is described by the relation

$$\eta(t) = \exp(-\Gamma_{\mathbf{q}}t) \times \{\cosh\Gamma_{\mathbf{q}} - \cos[2C(1-C)W(\mathbf{q}) - \mu]t\}. \quad (22)$$

Figure 1 shows the functions $\eta(t)$ and $\dot{\eta}(t)$ (the latter quantity determines the reaction rate) in the cases of $W_{\mathbf{q}} \gg \Gamma_{\mathbf{q}}$, $W_{\mathbf{q}} \sim \Gamma_{\mathbf{q}}$, and $W_{\mathbf{q}} \ll \Gamma_{\mathbf{q}}$. The formation of a new chemical bond—that is, a change in the number of atomic excitations with time—is determined by the electron parameters $\Gamma_{\mathbf{q}}$ and $\omega_{\mathbf{q}}$, provided that the resonance condition (18) is obeyed. If the condition (18) is not fulfilled, the value of $\eta(t)$ only oscillates about zero and the new bond is not formed.

Now let us consider in more detail the resonance condition (18). Taking the derivative with respect to the wavevector from both parts of this equation, we obtain the following relation for the velocities of electrons participating in the formation of the given structure:

$$V_{\mathbf{q}} \sim \left.\frac{\partial W(\mathbf{k})}{\partial \mathbf{k}}\right|_{\mathbf{k}=\mathbf{q}}.$$

For the formation of a stable structure corresponding to vector **q**, it is necessary that

$$\left.\frac{\partial W(\mathbf{k})}{\partial \mathbf{k}}\right|_{\mathbf{k}=\mathbf{q}} = 0,$$

and, therefore,

$$V_{\mathbf{q}} = 0, \quad W(\mathbf{q}) = \min W(\mathbf{k}). \quad (23)$$

This result implies that the formation of the new stable structure involves (via the interaction with DCEs having the wavevector **q**) a certain group of electrons with $V_{\mathbf{q}} = 0$ and $\omega(\mathbf{q}) = \min\omega(\mathbf{k})$, that is, electrons from a flat band with the minimum difference of energies of the donor–acceptor states. This condition is obeyed by the *d* states in the transition metals and their alloys. The effect of the *d* band narrowing on the adsorption energy is illustrated in Fig. 2.

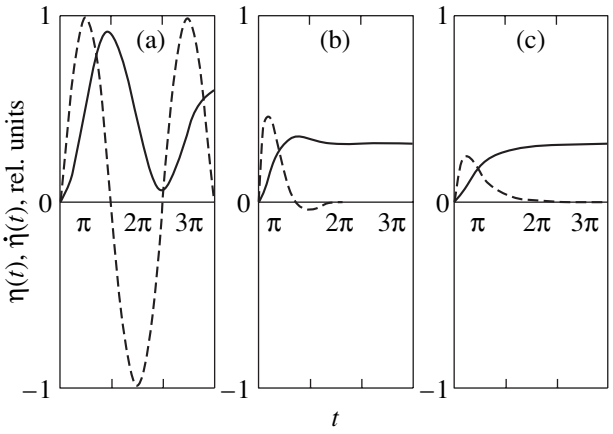

Fig. 1. Kinetics of chemical bond formation in cases of (a) $W_{\mathbf{q}} \gg \Gamma_{\mathbf{q}}$, (b) $W_{\mathbf{q}} \sim \Gamma_{\mathbf{q}}$, and (c) $W_{\mathbf{q}} \ll \Gamma_{\mathbf{q}}$. Solid and dashed curves show $\eta(t)$ and $\dot{\eta}(t)$, respectively.

The value of $\Gamma_{\mathbf{q}}$, which is equal to the inverse relaxation time of the mixed state, is not less important. The relation

$$\omega_{\mathbf{q}} \leq \Gamma_{\mathbf{q}} \quad (24)$$

is a necessary condition for the formation of a coherent standing wave (20), that is, the condition of formation of the new short-range order. Thus, the greater the total width of terms, the more actively the new structure is formed. In the general case, the total width includes (in addition to the broadening due to the interaction) the intrinsic widths of atomic and surface electron states. It is the substantial decay of the surface *d* states that determines the increased activity of the surface of transition metals.

In the case of $W_{\mathbf{q}} \gg \Gamma_{\mathbf{q}}$ (Fig. 1a), the initial stage of the process is characterized by the appearance of beats with a period of $T = 2\pi/\omega_{\mathbf{q}}$. The new structure is formed for a time of $\pi/\omega_{\mathbf{q}}$ and decays for the same time because of beats in the mixed interference state under the conditions of a weak electron–electron interaction via DCEs. In this case, the interaction cannot lead to a coherent alignment in the ion subsystem. (Over long periods of time, by virtue of the necessary appearance of the new phase, the $\eta(t)$ value exhibits saturation.) In the course of deposition, condition (23) is probably not always exactly satisfied even when condition (18) is valid. This may take place at high frequencies $\omega_{\mathbf{q}}$, that is, in wide bands where a small number of electrons can possess the velocity $V_{\mathbf{q}} = 0$. In this case, the Doppler line broadening seem to be possible and the resonance conditions may change. The regime with $W_{\mathbf{q}} \gg \Gamma_{\mathbf{q}}$ is most favorable for the catalytic activity of the surface, since the binding of gas atoms to the surface is established for a short time. In the case, say, of oxygen adsorption, the condition $W_{\mathbf{q}} \gg \Gamma_{\mathbf{q}}$ will not allow the metal surface to be oxidized, thus implying all subsequent events.

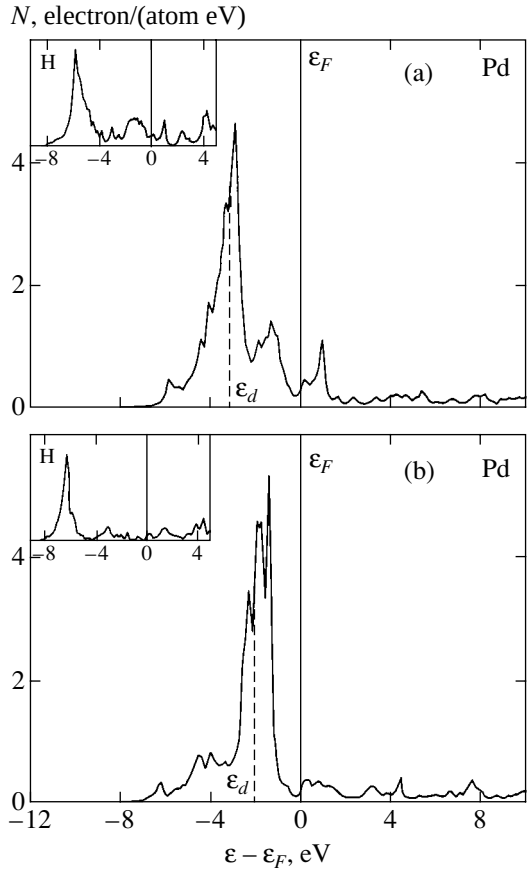

Fig. 2. Local densities of electron states in the surface layer of PdTa(001) in cases when the surface exposes (a) a single monolayer of palladium (E_{ad} = –0.16 eV, ε_d = –3.14 eV) and (b) two monolayers of palladium (E_{ad} = –0.43 eV, ε_d = –2.06 eV). The insets show the analogous curves for the adsorbate (hydrogen occurring in a four-fold hollow at the center of a 2D lattice) on the (001) surface. The calculations were performed using a full-potential linearized augmented-plane-wave method [13].

The situation changes when $W_{\mathbf{q}} \sim \Gamma_{\mathbf{q}}$ (Fig. 1b). In this case, there are no beats and the $\eta(t)$ value reaches a maximum and begins to decrease. However, the decrease is compensated by enhanced interaction and by narrowing of the band, so that the $\eta(t)$ value exhibits saturation for the time $t = 2\pi/\omega_{\mathbf{q}}$. At this point, the derivative $\dot{\eta}(t)$ exhibits a minimum. The compensation leads to an increase in the distance between metal atoms and to a decrease in their mutual binding.

Finally, the condition $W_{\mathbf{q}} \ll \Gamma_{\mathbf{q}}$ (Fig. 1c) is most favorable for the formation of new bonds.

5. CONCLUSIONS

The above analysis shows that the decay of electron states at the initial stage of adsorption does not play a decisive role and the dissociation of molecules takes place as a result of a sharp enhancement of the vibrational mode of molecules in the mixed state. This enhancement is caused by the resonance between the frequency of natural oscillations (the hybridization energy) and the driving frequency equal to the frequency of transitions between the molecular levels and the *d* states of the surface. These frequencies and the Fermi level position determine the dissociation barrier. In the absence of the resonance, the dissociation does not take place and the molecules can adsorb and exist on the surface. In the second stage involving the formation of new chemical bonds, the decay of electron states is among the main factors determining the possibility of this process. In fact, the physical mechanism of adsorption consists in the electron–electron interaction mediated by bosons representing the new bonds (atom pairs) obeying resonance conditions (19), (23) and (24), which determine the conditions of the new structure formation.

ACKNOWLEDGMENTS

This study was supported in part by the Russian Foundation for Basic Research, project nos. 04-02-17221 and 04-02-39009-GFEN2004_a.

REFERENCES

1. G. J. Kroes, A. Gross, E. J. Baerends, *et al.*, Acc. Chem. Res. **35**, 193 (2002).
2. B. Hammer and J. K. Norskov, Surf. Sci. **343**, 211 (1995).
3. Wei-Xue Li, C. Stampfl, and M. Scheffler, Phys. Rev. B **65**, 075 407 (2002).
4. J. Greeley, J. K. Norskov, and M. Mavrikakis, Annu. Rev. Phys. Chem. **53**, 319 (2002).
5. C. Stampfl, M. V. Granduglia-Pirovano, K. Reuter, and M. Scheffler, Surf. Sci. **500**, 368 (2002).
6. S. F. Borisov, N. F. Balakhonov, and V. A. Gubanov, *Interaction of Gases with Solid Surfaces* (Nauka, Moscow, 1988) [in Russian].
7. V. P. Vereshchagin and M. P. Kashchenko, Teor. Mat. Fiz. **42**, 133 (1980).
8. A. V. Vedyaev, V. E. Egorushkin, and N. V. Mel'nikova, Teor. Mat. Fiz. **71**, 259 (1988).
9. V. E. Egorushkin and N. V. Mel'nikova, Zh. Éksp. Teor. Fiz. **103**, 189 (1993) [JETP **76**, 103 (1993)].
10. V. E. Egorushkin and N. V. Mel'nikova, Zh. Éksp. Teor. Fiz. **103**, 214 (1993) [JETP **76**, 280 (1993)].
11. V. I. Iveronova and A. A. Katsnel'son, *Short-Range Order in Solid Solutions* (Nauka, Moscow, 1977) [in Russian].
12. E. M. Lifshitz and L. P. Pitaevskiĭ, *Physical Kinetics* (Nauka, Moscow, 1979; Pergamon, Oxford, 1981).
13. P. Blaha, K. Schwarz, G. K. M. Madsen, *et al.*, in *Proceedings of WIEN2k* (Vienna University of Technology, Vienna, 2001), ISBN 3-9501031-1-2.

Translated by P. Pozdeev

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 357–366.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 411–421.

ELECTRONIC PROPERTIES OF SOLIDS

Prebreakdown Generation of Nonequilibrium Electron–Hole Pairs: The Multiphoton Avalanche Effect

E. Yu. Perlin, A. V. Ivanov, and R. S. Levitskiĭ

All-Russia Research Center Vavilov State Optical Institute, St. Petersburg, 199034 Russia
e-mail: smperlin@online.ru

Received March 5, 2005

Abstract—A new mechanism of prebreakdown generation of electron–hole pairs is considered. It includes a cascade of interband multiphoton transitions as well as Auger-type processes involving two or three photons. A combination of these processes leads to the multiphoton avalanche effect. The threshold pumping radiation intensities required for triggering the avalanche mechanism lie in the range of 10^{11}–10^{12} W/cm^2. The band population balance equations describing the kinetics of pair production are obtained and solved numerically. It is shown that the proposed mechanism of production of nonequilibrium electron–hole pairs is more effective than "conventional" multiphoton absorption for intensities exceeding the threshold values. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

The first publications devoted to the breakdown of transparent insulators and wide-gap semiconductors induced by high-intensity laser radiation appeared more than four decades ago (see, for example, [1–4]). Nevertheless, research in this field remains topical in view of the diversity and complexity of its physical aspects and exceptionally important practical applications of the breakdown effect. Here, we consider only one aspect of this problem, viz., production of nonequilibrium electron–hole (*e–h*) pairs by high-intensity light with a quantum energy $\hbar\omega$ higher than the energies of vibrational excitations, but smaller than the band gap E_g of the material. Apparently, the production of a considerable number of free nonequilibrium carriers under the action of high-intensity light is the main premise for processes ultimately leading to breakdown of the material in most cases.

It was noted even in early publications that two mechanisms of production of nonequilibrium *e–h* pairs, viz., avalanche ionization and multiphoton interband transitions, play the leading role in the breakdown of high-purity transparent materials. Breakdown due to avalanche ionization was considered by many authors (see, for example, [5–13]). However, in the case studied here (for $\hbar\omega \gtrsim 1$ eV and $E_g \gtrsim 5$ eV), the efficiency of impact ionization triggering an avalanche is apparently not high. As a matter of fact, the kinetic energy of an electron in the conduction band required for the production of *e–h* pairs must exceed the threshold value

$$E_{\text{th}} = \frac{2\zeta + 1}{\zeta + 1} E_g, \qquad (1)$$

where $\zeta = m_c/m_v$, m_c and m_v being the effective masses of electrons in the lower conduction band and holes in the upper valence band, respectively. Here and below, we consider electrons for definiteness, although processes of the impact ionization type may also occur due to nonequilibrium holes. Obviously, the allowed band width ΔE_c can be smaller than E_{th}; in this case, impact ionization mechanism does not work (at least in its simple form). If, however, $\Delta E_c > E_{\text{th}}$, an electron can acquire the required kinetic energy either through a multistage cascade of intraband single-photon transitions, or due to multiphoton intraband transitions. For $\hbar\omega \gtrsim 1$ eV, both mechanisms are ineffective. The efficiency of the cascade mechanism is low in view of the smallness of the intraband absorption coefficient in the spectral range of interest (the effective cross sections are on the order of 10^{-20}–10^{-19} cm^2) and a short carrier relaxation time (on the order of 10^{-13} s), which is primarily determined by the interaction of charge carriers with polar optical vibrations of the lattice. The probabilities of five–eight-photon intraband transitions (see, for example, [14]) in the spectral range under investigation are obviously lower than the probabilities of multiphoton interband transitions leading to production of nonequilibrium *e–h* pairs.

Multiphoton interband transitions undoubtedly play an important role in prebreakdown production of *e–h* pairs. The concentration of nonequilibrium charge carriers sufficient for breakdown is either produced directly owing to such transitions, or such transitions generate electrons initiating avalanche ionization. The role of multiphoton interband transitions in the physics

of breakdown and various breakdown mechanisms are considered in [3, 4, 15–19].

In the 1990s, a number of works devoted to breakdown in solids initiated by ultrashort (subpico- and femtosecond) laser pulses were published. This revived the discussion of possible breakdown mechanisms (see [20–28]). In particular, the dependence of the breakdown threshold on the pulse duration and light frequency was studied in detail in [20–26]. Some of these dependences were comprehensively described in [21, 23] in the framework of a model in which avalanche ionization is the main triggering mechanism, and multiphoton processes ensured the initial population of electrons required for initiating the avalanche. However, this model, which is based on the Fokker–Planck equations for electron dynamics in the radiation field, failed to explain the results of measurement of the breakdown threshold as a function of the time delay between two laser pulses [25, 26]. A slightly different model based on the Boltzmann kinetic equations was proposed in [27]. In contrast to [21], it was assumed that multiphoton ionization dominates over avalanche in the case of a short laser pulse with $\tau_p \leq 200$ fs (see also [17]). The model proposed in [21] was modified in [28] by taking into account relaxation processes in the kinetic equations and by describing electron sources. At the same time, this model lays no claim to a comprehensive interpretation of the entire body of available experimental data.

In addition to avalanche ionization and multiphoton interband transitions, intermediate-type processes [29, 30] may contribute to the production of *e–h* pairs. If the relation $(n-1)\hbar\omega < E_g < n\hbar\omega$ holds, the energy deficit $\Delta^{(n-1)} = E_g - (n-1)\hbar\omega$ required for the production of an *e–h* pair upon the absorption of $n-1$ photons can be replenished by the kinetic energy of free electrons, which must exceed the threshold value

$$E'_{\text{th}} = \frac{2\zeta + 1}{\zeta + 1}\Delta^{(n-1)} \qquad (2)$$

in this case. Obviously, E'_{th} is much smaller than E_{th}. If $E'_{\text{th}} \ll k_B T$, only those free electrons that possess a large kinetic energy may take part in the process. This energy must be acquired only due to interband absorption of light. Such processes are vital only for the long-wavelength range of the spectrum, in which the cross sections of intraband light absorption are large enough. In particular, the absorption of radiation from a CO_2 laser ($\hbar\omega \sim 10^{-1}$ eV) in *n*-InAs due to three-photon interband transitions involving free electrons was studied in [29, 30]. Since the rate of production of free carriers in such processes increases with their concentration, an avalanche increase in the number of nonequilibrium *e–h* pairs becomes possible under certain conditions. Although the processes studied in [29, 30] can hardly play a significant role in the problem of breakdown of a broadband material by micrometer-range radiation studied here, modified versions of a process of this type can be quite operative. In particular, such processes play a key role in the model of prebreakdown production of *e–h* pairs proposed here and based on the photon avalanche scheme.

The photon avalanche effect is usually employed for exciting short-wave luminescence with the help of long-wave pumping. In most publications devoted to this effect (see, for example, [31–37]), systems of rare-earth dopant ions are considered. The simplest version of a photon avalanche is realized in a system of three-level impurities. We denote the ground state and two excited states of impurities as 1, 2, and 3. The frequencies of transitions between the *i*th and the *k*th states are denoted as ω_{ik}. In the absence of pumping, only state 1 is occupied. If the pumping frequency $\omega \approx \omega_{32} > \omega_{21}$, no effects are observed in the system for low intensities *j* of light. For large values of *j*, an electron of one of impurities, which for some reason or other is in excited state 2, rapidly absorbs a quantum $\hbar\omega$ and passes into the upper state 3. For a high concentration of impurities, the electron from one of the impurities returns from state 3 to state 2 due to the cross-relaxation process $31 \longrightarrow 22$, while an electron from an adjacent impurity passes from state 1 to state 2. Then electrons from two impurities are in state 2. Each of these electrons may in turn participate in such processes. As a result, many electrons from different impurities are accumulated in state 2; strong absorption takes place on the $2 \longrightarrow 3$ transitions and, accordingly, level 3 becomes highly populated. This may result in luminescence on the $3 \longrightarrow 1$ transition at a frequency $\Omega > \omega$. A photon avalanche is characterized by a clearly manifested threshold nature of the effect: the populations of excited states and light absorption increase jumpwise at $j \approx j_{\text{th}}$; at the same time, the time τ_{eq} of establishment of a quasi-equilibrium distribution of the electron population also increases sharply. The threshold nature of the emergence of a photon avalanche makes it possible to analyze this effect in terms of the Landau theory of second-order phase transitions [34].

The effect of photon avalanche in a semiconductor system with doped quantum wells was considered in [38, 39]. In this case, switching to the photon avalanche mode may occur over a much shorter time and with lower energy expenditures than in a system of rare-earth ions.

A scheme of emergence of a photon avalanche in a structure with type II quantum wells was proposed in [40]. It was shown that this scheme makes it possible to excite a considerable number of nonequilibrium *e–h* pairs for moderate pumping intensities ($j \sim 10^5$–10^6 W/cm^2) and to obtain luminescence at a wavelength 3–5 times larger than that of the exciting light. The energy corresponding to switching between weakly and strongly absorbing states in this case is

$E_{sw} \sim 1$–10 pJ/μm^2, which is four to five orders of magnitude lower than in a system of rare-earth ions.

Here, we propose a new mechanism of production of nonequilibrium e–h pairs by high-intensity light ($j \sim 10^{11}$–10^{12} W/cm^2). This mechanism, which is based on a photon-avalanche-type effect, includes both "ordinary" multiphoton processes and interband transitions involving free charge carriers.

2. MULTIPHOTON AVALANCHE MODEL

We assume that the material is free of impurities whose light-induced ionization could lead to a noticeable concentration of nonequilibrium charge carriers. We also assume that inhomogeneities whose local heating by radiation could lead to breakdown are absent.

Let us consider the model of a crystal, which includes two conduction bands (c and c_1) as well as the valence band v. For definiteness, we assume that the energy gap E_g between the valence band and the lower conduction band c is slightly smaller than $n\hbar\omega$, while the gap E_g' between the conduction bands is smaller than $l\hbar\omega$ ($l < n$; Fig. 1). We will consider two versions of the process: $n = 5$, $l = 3$ and $n = 5$, $l = 2$. Direct n-photon $v \longrightarrow c$ transitions lead to the appearance of a certain number of free electrons in the lower conduction band c. These electrons arrive at the bottom of the band over time periods on the order of 10^{-13} s. Owing to l-photon transitions between conduction bands c and c_1, electrons pass to the upper conduction band. The Auger-type process $c_1 + (n - l)\hbar\omega \longrightarrow ccv$ plays the key role in the proposed mechanism; in other words, an electron in the c_1 band generates (with the help of $n - l$ photons) a pair consisting of a hole in the v band and an electron in the c band, passing thereby to the c band, where two new electrons appear as a result of this process. If we compare this with the "classical" scheme of the photon avalanche effect described in the Introduction, the n-photon $v \longrightarrow c$ transitions play the role of low-intensity nonresonant $0 \longrightarrow 1$ transitions; l-photon $c \longrightarrow c_1$ transitions correspond to rapid resonant $2 \longrightarrow 3$ transitions, while transitions $c_1 + (n-l)\hbar\omega \longrightarrow ccv$ play the same role as cross-relaxation $31 \longrightarrow 22$ transitions.

The mechanism of production of e–h pairs proposed here will be referred to as a multiphoton avalanche.

Naturally, the model under consideration also includes the $c_1 \longrightarrow c$ relaxation process involving phonons as well as processes of recombination of nonequilibrium photoexcited electrons and holes.

Obviously, the proposed model gives correct results only in the range of very strong electromagnetic fields with an electric field strength of the light wave of $F_\omega \gtrsim 10^7$ V/cm. For $\hbar\omega = 1.17$ eV, $m_r = 0.4m_0$ ($m_r^{-1} = m_c^{-1} + m_v^{-1}$, where m_0 is the free electron mass), $E_g = 5.7$ eV, and $F_\omega = 10^7$ V/cm, the Keldysh parameter [41, 42]

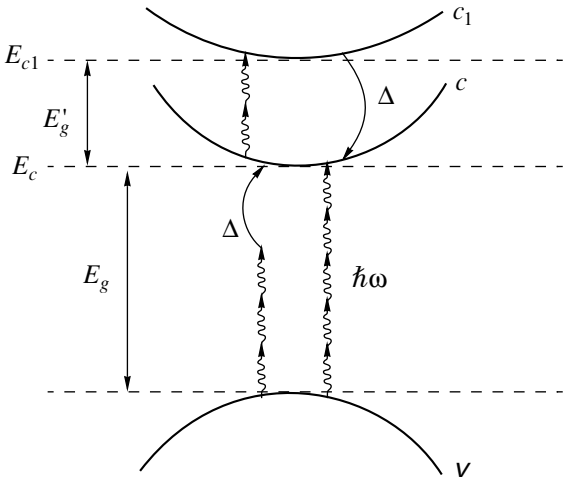

Fig. 1. Diagram of transitions in a system consisting of valence band v and two conduction bands c and c_1 ($l = 3$). Undulated lines with arrows denote photons (see text for explanation).

$$\gamma = \frac{\omega\sqrt{m_r E_g}}{eF_\omega} \tag{3}$$

assumes a value of $\gamma \approx 6.4$. The large value of parameter γ leads to the conclusion that the pattern of multiphoton interband transitions rather than interband tunneling in the field of a strong electromagnetic wave is adequate in the situation considered here for intensities lower than 10^{13} W/cm^2. In this case, the inequality $1/(4\gamma)^2 \ll 1$ suggests that transitions with the smallest number of photons permitted by conservation laws occur in fact between any two bands.[1] To estimate the probabilities of multiphoton interband transitions, we will henceforth use the results obtained in [44, 45], where the following formulas were derived for the probabilities of n-photon interband transitions for arbitrary n

$$W_{vc}^{(n)} \equiv \sigma_{vc}^{(n)} j^n.$$

In the region of small values of j, these formulas for $n = 3$ coincide with those derived in the framework of standard perturbation theory. For $n \gg 1$, formulas from [44, 45] give an asymptotic form close to that

[1] Such a situation is typical, but exceptions are also observed sometimes. For example, it was shown in [43] that the probabilities of n-photon interband transitions with $n = 3$ and $n = 4$ in the indirect gap material AgBr are large than the probabilities of transitions with $n = 2$ for $j \gtrsim 10^{10}$ W/cm^2 ($\lambda = 560$ nm) owing to specific features of the electron band structure and manifestations of the optical Stark effect under the conditions of double interband resonance.

obtained in [41, 42]. The corresponding formulas are given in the Appendix.

It was noted above that Auger-type transitions $c_1 + s\hbar\omega \longrightarrow ccv$, $s = n - l$ play a key role in the development of a multiphoton avalanche. The probabilities of these transitions,

$$W_a^{(s)} \equiv \gamma^{(s)} j^s n_{c_1},$$

were calculated in the $(s + 1)$th order of perturbation theory (s orders for the interaction of the electron subsystem with the field of an electromagnetic wave and one order for the electron–electron Coulomb interaction). The criteria of applicability of the Born approximation in the Coulomb interaction and the principles for selecting Feynman diagrams giving the major contribution to the cross section are described in detail in [46]. The results are presented in the Appendix.

3. BALANCE EQUATIONS FOR BAND POPULATIONS

The balance equations for electron concentrations n_c and n_{c_1} in the two conduction bands have the form

$$\begin{aligned}
\dot{n}_{c_1} &= -W_{c_1c}n_{c_1} + \sigma_{cc_1}^{(l)} j^l f_{cc_1} - \gamma^{(l)} j^{n-l} n_{c_1} \\
&\quad - d_{c_1}[(p_0 + n_c - n_0 + n_{c_1})n_{c_1} - n_0 p_0], \\
\dot{n}_c &= W_{c_1c}n_{c_1} - \sigma_{cc_1}^{(l)} j^l f_{cc_1} + 2\gamma^{(l)} j^{n-l} n_{c_1} \\
&\quad + \sigma_{vc}^{(n)} j^n (1 - f_c)(1 - f_p) \\
&\quad - d_c[(p_0 + n_c - n_0 + n_{c_1})n_c - n_0 p_0]
\end{aligned} \tag{4}$$

with the initial conditions $n_{c_1} = 0$ and $n_c = n_0$ for $t = 0$. In formulas (4), W_{c_1c} is the rate of relaxation electron transitions from the upper conduction band to the lower conduction band and n_0 and p_0 are the equilibrium concentrations of electrons in the lower conduction band and holes in the valence band, respectively. The rates $w_{cc_1}^{(l)} = \sigma_{cc_1}^{(l)} j^l$ of l-photon transitions between the conduction bands c and c_1 are described by the formulas, which can easily be derived using standard perturbation theory. These formulas are also given in the Appendix. The terms proportional to d_{c_1} and d_c on the right-hand sides of Eqs. (4) describe the conventional bimolecular recombination of electrons in the conduction bands with holes in the valence band. The hole concentration is defined by the relation

$$p = n_c - n_0 + n_{c_1} + p_0, \tag{5}$$

where $p = p_0$ for $t = 0$.

At high light intensities, the states near the bottom of the conduction bands c and c_1 are filled, while the states near the top of the valence band v are depleted (i.e., a high concentration p of holes appears). This affects the rates of multiphoton interband transitions and multiphoton Auger-type transitions. These effects are not critical for the problem under investigation. Nevertheless, they should be taken into account at least in a rough approximation. For this purpose, real distributions of nonequilibrium electrons and holes are approximated by the Fermi distribution functions corresponding to instantaneous electron and hole concentrations $n_c(t)$, $n_{c_1}(t)$, and $p(t)$. For example, formulas (4) contain the distribution function f_{cc_1} for electrons in the c band, which corresponds to energy $\varepsilon_c(\mathbf{k}_{cc_1}^{(l)})$ at point $\mathbf{k}_{cc_1}^{(l)}$ of the l-photon resonance between bands c and c_1:

$$f_{cc_1} = \left\{ \exp\left[\frac{\varepsilon_c(\mathbf{k}_{cc_1}^{(l)}) - \mu_c(n_c)}{k_B T} \right] + 1 \right\}^{-1}. \tag{6}$$

Analogously, we can introduce the distributions functions f_c and f_p of electrons in the c band and holes in the v band, which correspond to energies $\varepsilon_c(\mathbf{k}_{vc}^{(n)})$ and $\varepsilon_v(\mathbf{k}_{vc}^{(n)})$ at point $\mathbf{k}_{vc}^{(n)}$ of the n-photon resonance between the v and c bands (see also formulas (A.2)–(A.4) in the Appendix). In the case of relatively large quanta of exciting light ($\hbar\omega \gtrsim 1$ eV) and not very small durations of light pulses ($\tau_p \gtrsim 1$ ps), this approximation turns out to be satisfactory on account of small cross sections of intraband absorption of light and short times of intraband relaxation of electrons and holes. Naturally, this approximation is inapplicable for longer wavelengths ($\hbar\omega \lesssim 0.1$ eV) (see [29, 30]).

To calculate chemical potentials μ_c and $\mu_p (\mu_p > 0)$ for electrons and holes, which appear in the formulas for distribution functions, we will use the transcendental equations

$$\frac{n_c(t)}{N_c(T)} - \Phi_{1/2}(\mu_c^*) = 0, \quad \frac{p(t)}{N_p(T)} - \Phi_{1/2}(\mu_p^*) = 0, \tag{7}$$

where $\mu_{c,p}^* = m_{c,p}/k_B T$; N_c and N_p are the effective densities of states in the conduction and valence bands, respectively,

$$\begin{aligned}
N_c &= 2\left[\frac{2\pi m_c k_B T}{(2\pi\hbar)^2} \right]^{3/2}, \\
N_p &= 2\left[\frac{2\pi m_v k_B T}{(2\pi\hbar)^2} \right]^{3/2},
\end{aligned} \tag{8}$$

and $\Phi_{1/2}(x)$ is the Fermi–Dirac integral,

$$\Phi_{1/2}(x) = \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{z^{1/2} dz}{1 + \exp(z - x)}. \tag{9}$$

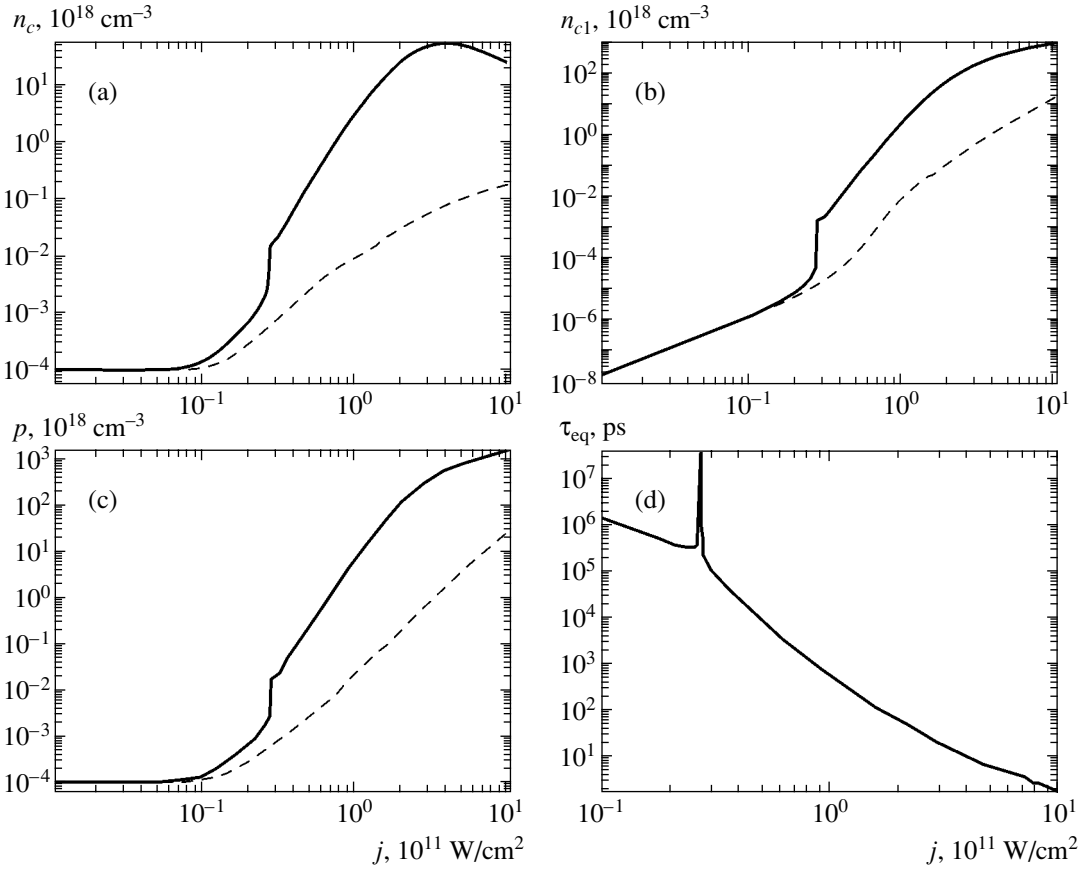

Fig. 2. Solid curves describe the dependences of quasi-equilibrium concentrations of electrons in bands c (a), c_1 (b), and holes in band v (c), as well as the time of establishment of quasi-equilibrium concentrations (d) on the light intensity for $l = 2$. The dashed curves are the same dependences calculated disregarding the process $c_1 + 3\hbar\omega \longrightarrow ccv$. Calculations were made using the following values of parameters: $E_g = 5.5$ eV, $E_g' = 2.28$ eV, $\hbar\omega = 1.17$ eV, $m_c = 0.9m_0$, $m_{c_1} = 0.03m_0$, $m_v = 0.9m_0$, $d_{c_1} = 3 \times 10^{-4}$ cm^3/ps, $d_c = 0.01$ cm^3/ps, $W_{c_1c} = 1.0$ ps^{-1}, $T = 10^3$ K, $\varepsilon_\infty = 4.5$, $\varepsilon_0 = 12.4$, $p_0 = 10^{14}$ cm^{-3}, and $n_0 = 10^{14}$ cm^{-3}.

Thus, we must obtain a self-consistent solution to the system of nonlinear differential equations (4) and transcendental equations (6). The complexity of this problem rules out a qualitative analysis of the system dynamics of the type carried out in [34, 35, 39], as well as the derivation on its basis of simple formulas for the threshold light intensities and the times of establishment of quasi-equilibrium populations in the electron system. For this reason, we will use the results of numerical calculations that will be given in the next section.

4. RESULTS OF NUMERICAL SOLUTION OF THE BALANCE EQUATIONS

Figures 2 ad 3 show the results of numerical solution of Eqs. (4)–(7) for $l = 2$. Solid curves in Figs. 2a–2c represent the quasi-equilibrium concentrations of electrons in both conduction bands (n_c and n_{c_1}) and holes in the valence band (p) as functions of light intensity j. Dashed curves in these figures show the same dependences calculated for the same parameters, but disregarding three-photon Auger-type processes. Figure 2d shows the time τ_{eq} of establishment of quasi-equilibrium populations vs. intensity j. The following features of these dependences are worth noting. A well-pronounced stepwise increase in the concentration of nonequilibrium carriers with increasing j occurs in the vicinity of the threshold value ($j_{th} \approx 2.7 \times 10^{10}$ W/cm^2 for the chosen values of parameters). At this value of intensity, time τ_{eq} increases abruptly. Such features are typical of the photon avalanche effect (a multiphoton avalanche takes place in our case). It can also be seen that the concentration of nonequilibrium carriers for $j > j_{th}$ turns out to be one or two orders of magnitude higher than for ordinary multiphoton interband transitions with the probability calculated for the same values of parameters.

Figure 2d shows that time τ_{eq} is ~1 ps for $j \sim 10^{12}$ W/cm^2. With decreasing j, times τ_{eq} increase rapidly and fall in the nanosecond range for $j \sim$

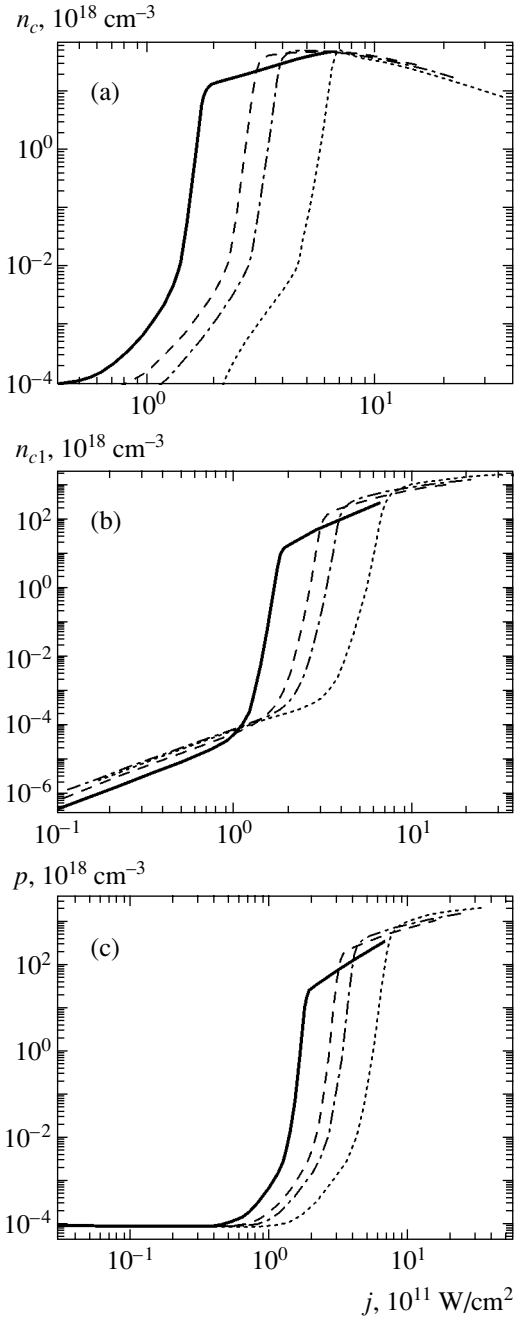

Fig. 3. Concentrations of electrons in bands c (a) and c_1 (b) and holes in band v (c), emerging as a result of action of pulses of a duration of 1 ps (solid curve), 5 ps (dashed curve), 10 ps (dot-and-dash curve), and 50 ps (dotted curve) as functions of light intensity j.

10^{11} W/cm^2, while the interaction of shorter pulses with solids is of utmost practical importance. For this reason, we must also consider the production of nonequilibrium e–h pairs for laser pulse durations τ_p smaller than τ_{eq}. Figure 3 shows the concentrations n_c, n_{c_1}, and p emerging as a result of action of pulses of durations of 1, 5, 10, and 50 ps as functions of the light intensity. These dependences were plotted for Gaussian-shaped pulses. It can be seen that a region of extremely sharp increase in concentrations with the light intensity is present in this case as well (when the value of j increases by 30–40%, the concentrations increase by four or five orders of magnitude). A situation is possible when a slight increase in the light intensity is sufficient for transition from the reversible prebreakdown excitation of the material to its destruction, which takes place for nonequilibrium carrier concentrations exceeding 10^{19} cm^{-3}. If we denote by j'_{th} the intensity near which the rate of variation of, say, quantity $p(j)$ is maximal, we can write the following approximate dependence: $j'_{th} \propto \tau_p^{-1/3}$.

For $l = 3$, the multiphoton avalanche effect is clearly manifested only for relatively long laser pulses ($t_p \gtrsim 100$ ps). Solid curves in Figs. 4a–4c represent the dependences of the quasi-equilibrium concentrations of electrons and holes on the light intensity. The dashed curves in the same figures show the analogous dependences disregarding two-photon Auger-type processes. Figure 4d shows the time τ_{eq} of establishment of quasi-equilibrium populations as a function of light intensity. It can be seen that, in this case, an increase in the nonequilibrium carrier concentration due to formation of a multiphoton avalanche is weaker than in the case of $l = 2$.

5. DISCUSSION

The exceptional complexity of the problem of a multiphoton avalanche necessitated a simplified approach to its solution. We will mention some of the factors that may to a certain extent affect the results of analysis.

1. Although, in almost all wide-gap materials, there exist conduction bands (or valence bands) with band gaps between them and the lowest conduction band (or upper valence band) that are smaller than E_g, the three-band model of the energy spectrum considered above is obviously idealized. Analysis of real band structures of specific materials will probably require that additional photoinduced transitions (in particular, intervalley intraband transitions and indirect multiphoton interband transitions) be included in the model.

2. In spite of the fact that light with $\hbar\omega \sim 1$ eV experiences relatively weak intraband absorption and intraband relaxation times of carriers are short (see Section 1), the intra- and interband dynamics of carriers in the high-intensity radiation field should be generally considered simultaneously. Strictly speaking, the system of balance equations for carrier concentrations in the bands in this case is not quite adequate. At the same time, a rigorous analysis of the carrier dynamics taking into account all transitions significant for the problem of a multiphoton avalanche is hardly feasible.

3. The rearrangement of the energy band spectrum of a crystal in the field of a high-intensity light wave in

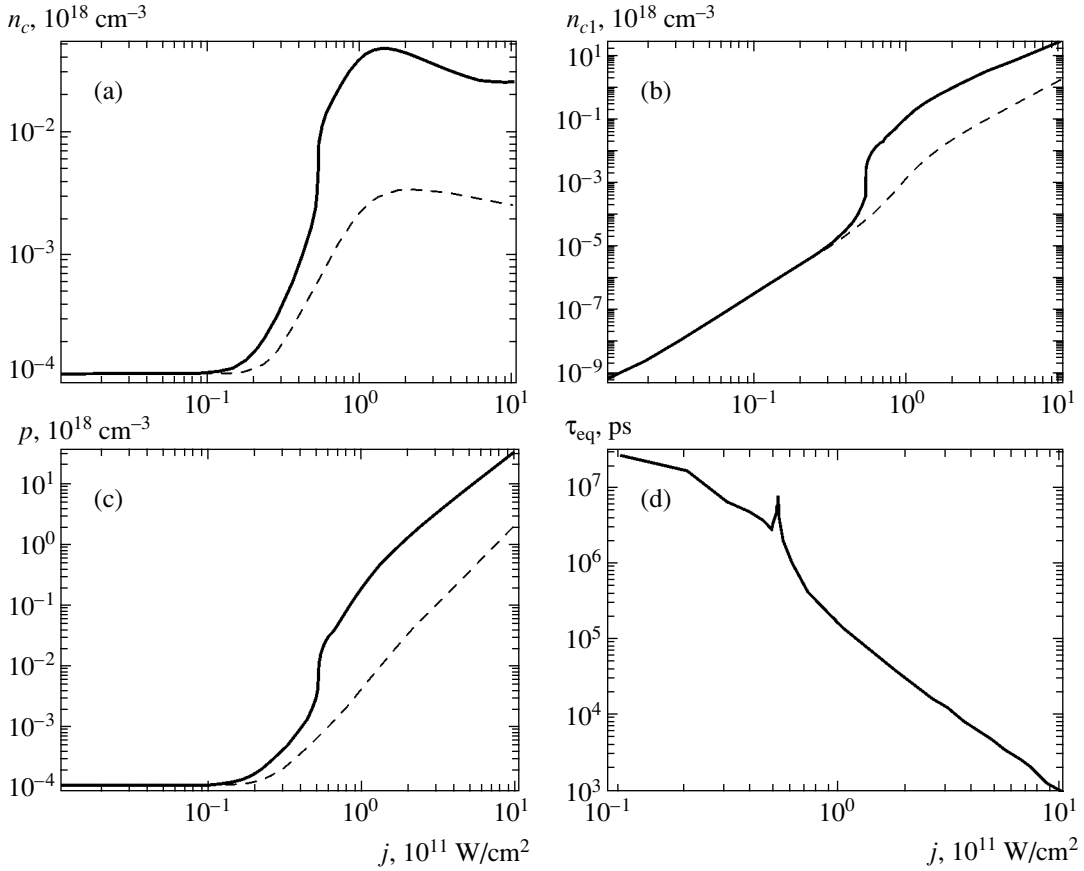

Fig. 4. Solid curves describe the dependences of quasi-equilibrium concentrations of electrons in bands c (a), c_1 (b), and holes in band v (c), as well as the time of establishment of quasi-equilibrium concentrations (d) on the light intensity for $l = 3$. The dashed curves are the same dependences calculated disregarding the process $c_1 + 2\hbar\omega \longrightarrow ccv$. Calculations were made using the following values of parameters: $E_g' = 3.45$ eV, $W_{c_1c} = 0.01$ ps^{-1}, and $d_c = 1.0 \times 10^{-3}$ cm^3/ps. The values of the remaining parameters are the same as in Fig. 2.

the presence of multiphoton resonances at various points of the **k** space may in principle play an important role in the breakdown problem (see, for example, [47, 48]). In this case, two types of effects may take place. First, the band spectrum may change in such a way that three-photon interband transitions, whose probabilities are usually much higher, may occur in strong fields in the spectral regions where, say, four-photon transitions were possible in weak fields. Second, according to [49–51], new critical points (Van Hove singularities) may appear in the electron density of states in the field of a strong electromagnetic wave. Since the positions of these additional critical points depend on light intensity j, intensity regions in which the rate of production of the e–h pairs abruptly increases upon a small increase in j may appear under certain conditions; this may lead to an abrupt increase in the concentration of free carriers and to breakdown of the material. However, in spite of the obvious importance of the effects of band spectrum rearrangement, their simultaneous analysis with the processes of development of a multiphoton avalanche appears as premature in view of the complexity of the corresponding calculations and insufficient clarity of results. Here, we confined our analysis to multiphoton avalanches, assuming that spectral rearrangement effects are insignificant.

4. Processes of photon reemission may noticeably affect transition probabilities for light intensities $j \sim 10^{10}$–10^{12} W/cm^2. Reemission effects can be accounted for in the probabilities of direct multiphoton interband transitions comparatively easily (see [44, 45, 52]); however, it is impossible to make allowance for these effects adequately and obtain a foreseeable result in calculating the cross sections of Auger-type multiphoton processes. Nevertheless, there are no grounds to assume that reemission processes may critically affect the development of a multiphoton avalanche.

5. It should be borne in mind that the above analysis corresponds to pulses with durations exceeding the characteristic momentum relaxation times for nonequilibrium carriers. To describe the production of nonequilibrium e–h pairs by light pulses of duration $t_p \lesssim 100$ fs, a fundamentally different approach is required (see, for

example, [52]),which cannot be based on balance equations of type (4).

Taking into account the above arguments, we can refine and supplement the pattern of prebreakdown generation of nonequilibrium *e–h* pairs. Nevertheless, the results obtained here indicate that a multiphoton avalanche might be much more effective than an "ordinary" multiphoton interband transition for high intensities of light.

6. CONCLUSIONS

We have proposed a new mechanism of prebreakdown production of nonequilibrium *e–h* pairs in wide-gap insulators and semiconductors by high-intensity light in the micrometer range. This mechanism is based on the multiphoton avalanche effect. The probabilities of direct interband multiphoton transitions, as well as multiphoton interband transitions involving free carriers, have been calculated in the framework of the three-band model of the electron energy spectrum of the crystal. We derived a system of nonlinear balance equations for the populations of carriers in the bands. Numerical solution of this system proves that the dependence of populations on the light intensity is of the threshold nature. The threshold intensities for various materials and durations of pumping pulses may assume values ranging from 10^{10} to 10^{12} W/cm^2. The times of establishment of quasi-equilibrium band populations sharply increase for pumping intensities close to threshold values. For intensities exceeding the threshold values, the proposed mechanism makes it possible to obtain a larger number of *e–h* pairs than for "ordinary" multiphoton absorption. In the case when the material is exposed to high-intensity picosecond light pulses, the photon avalanche mechanism can yield a concentration of nonequilibrium electrons and holes, which is sufficient for breakdown of the material.

ACKNOWLEDGMENTS

This study was supported by the Russian Foundation for Basic Research (project no. 04-02-16175-a).

APPENDIX

Probabilities $W_a^{(s)}$ of s-photon (s = 2, 3) transitions between bands v and c involving free electrons in the c_1 band are defined by the formulas

$$W_a^{(s)} = \frac{2^{10} e^{2(s+2)} |p_{cv}|^2 |p_{cc_1}|^2 m_c^{(s-1)}}{\varepsilon_l^2 \varepsilon_t^{1/2} (\hbar\omega m)^{2^s} c^s \zeta (1+\zeta)^2} n_{c_1} j^s I_s, \quad \text{(A.1)}$$

$$I_2 = \pi\hbar \int_{a_1}^{b_1}\int_{a_2}^{b_2} dx dy \frac{(\zeta'-1)^2 f_2}{\delta_c^2 (y-x)} \left\{ \frac{(1+\zeta)y - x}{y[(1-\zeta')x + \delta_c \zeta']} + \frac{y - (1+\zeta)x + (\delta_c + \delta_v)\zeta}{[\delta_c - (1-\zeta')(y-\delta_v)](\delta_c + \delta_v - x)} \right\}^2, \quad \text{(A.2)}$$

$$I_3 = 2\pi^2 \hbar^3 \int_{c_1}^{d_1}\int_{c_2}^{d_2} dx dy \frac{f_3}{(\delta_c - \hbar\omega)^2 (\delta_c - x)(y - \delta_v)} \times \left\{ \frac{\sqrt{y-\delta_v}}{[(\zeta'-1)(y-\delta_v) + \delta_c]\left(y - \frac{2}{1+\zeta}\hbar\omega\right)} + \frac{\sqrt{\delta_c - x}}{[(1-\zeta')x + \zeta'\delta_c]\left(\delta_c + \delta_v - x - \frac{2}{1+\zeta}\hbar\omega\right)} - \frac{4\zeta^2 \hbar\omega(\sqrt{y-\delta_v} + \sqrt{\delta_c - x})}{(1+\zeta) z \left(z + \frac{2\zeta}{1+\zeta}\hbar\omega\right)\left[\left(1 + \frac{\zeta'}{\zeta}\right)z + (\zeta'-1)\delta_v - \delta_c + 2\hbar\omega\right]} \right\}^2. \quad \text{(A.3)}$$

Here,

$$f_2 = \{1 + \exp[(y - x + \delta_v\zeta + \mu_p)/k_B T]\}^{-1},$$

$$f_3 = \{1 + \exp[(z + \delta_v\zeta + \mu_p)/k_BT]\}^{-1},$$

$$a_1 = \frac{(1+\zeta)^2}{1+2\zeta}\delta_v + \frac{\zeta}{1+2\zeta}\delta_c, \quad b_1 = \delta_c, \quad z = y - x - \hbar\omega,$$

$$\left.\begin{matrix} a_2 \\ b_2 \end{matrix}\right\} = \left(\frac{\zeta}{1+\zeta}\right)^2\left\{\sqrt{\delta_c - x} \mp \sqrt{\frac{1+2\zeta}{\zeta^2}x - \frac{\delta_c}{\zeta} - \left(\frac{1+\zeta}{\zeta}\right)^2\delta_v}\right\}^2 + \delta_v,$$

$$c_1 = \frac{(1+\zeta)^2}{1+2\zeta}\delta_v + \frac{\zeta}{1+2\zeta}\delta_c - \frac{1+\zeta}{1+2\zeta}\hbar\omega, \quad d_1 = \delta_c, \tag{A.4}$$

$$\left.\begin{matrix} c_2 \\ d_2 \end{matrix}\right\} = \left(\frac{\zeta}{1+\zeta}\right)^2\left\{\sqrt{\delta_c - x} \mp \sqrt{\frac{1+2\zeta}{\zeta^2}x - \frac{\delta_c}{\zeta} - \left(\frac{1+\zeta}{\zeta}\right)^2\delta_v + \frac{1+\zeta}{\zeta^2}\hbar\omega}\right\}^2 + \delta_v,$$

where p_{cv} and p_{c_1c} are the interband matrix elements of the momentum operator, c is the velocity of light in vacuum, $\zeta' = m_c/m_{c_1}$, $\delta_c = \varepsilon_{c_1}(0) - \varepsilon_c(0) + \hbar\omega$, $\delta_v = (E_g - \hbar\omega)/(1+\zeta)$, and ε_l and ε_t are the longitudinal and transverse permittivities.

Probabilities $W_{vc}^{(n)}$ of n-photon transitions from the filled valence band v to the empty conduction band c can be estimated with the help of the formulas derived in [44, 45]. In particular, for $n = 5$, we have

$$W_{vc}^{(5)} = \frac{\sqrt{2E_{0c}}m_c^{3/2}}{\pi\hbar^4(\zeta+1)}\left(\frac{\rho}{2}\right)^2|g|^2, \tag{A.5}$$

where

$$\rho = \frac{1}{4\hbar\omega}\left\{\frac{10|V_{vc}|^2}{24\hbar\omega} - \frac{|V_{cc_1}|^2\Delta_{c_1c}}{\Delta_{c_1c}^2 - (\hbar\omega)^2}\right\},$$

$$g = -\frac{V_{vc}}{144\hbar\omega}\left[\frac{5|V_{vc}|}{6\hbar\omega} + \frac{|V_{cc_1}|^2(5\hbar\omega - 3\Delta_{c_1c})}{\Delta_{c_1c}^2 - (\hbar\omega)^2}\right], \tag{A.6}$$

$$E_{0c} = \frac{5\hbar\omega - E_g}{\zeta+1}, \quad \Delta_{c_1c} = \varepsilon_{c_1}(0) - \varepsilon_c(0),$$

$$V_{ij} = \sqrt{\frac{8\pi j}{c\sqrt{\varepsilon_t(\omega)}}}\frac{ep_{ij}}{m_0\omega}.$$

The rates of three- and two-photon transitions between the conduction bands c and c_1 for $l = 3$ and $l = 2$ can be described by the formulas, which can easily be derived using the standard perturbation theory:

$$w_{cc_1}^{(3)} = \sigma_{cc_1}^{(3)}j^3 = \frac{4\pi^2e^6\sqrt{2m_{cr}\Delta_{c_1c}^{(3)}}|p_{cc_1}|^6m_{cr}}{\pi m^6\hbar^8\omega^{10}c^3\varepsilon_t^{3/2}(\omega)},$$

$$w_{cc_1}^{(2)} = \sigma_{cc_1}^{(2)}j^2 = \frac{64\pi e^4|p_{cc_1}|^2[-2m_{cr}\Delta_{c_1c}^{(2)}]^{3/2}}{m^2m_{cr}c^2\varepsilon_t(\omega)(\hbar\omega)^6}j^2, \tag{A.7}$$

$$\Delta_{c_1c}^{(l)} = l\hbar\omega - \Delta_{c_1c}, \quad m_{cr}^{-1} = m_c^{-1} - m_{c_1}^{-1}.$$

REFERENCES

1. B. M. Ashkinadze, V. I. Vladimirov, V. A. Likhachev, *et al.*, Zh. Éksp. Teor. Fiz. **50**, 1187 (1966) [Sov. Phys. JETP **23**, 788 (1966)].
2. V. A. Likhachev, S. M. Ryvkin, V. M. Salmanov, and I. D. Yaroshetskiĭ, Fiz. Tverd. Tela (Leningrad) **8**, 3432 (1966) [Sov. Phys. Solid State **8**, 2754 (1966)].
3. B. S. Sharma and K. E. Riekhof, Can. J. Phys. **45**, 3781 (1967).
4. B. S. Sharma and K. E. Riekhof, Can. J. Phys. **48**, 11 781 (1970).
5. A. G. Molchanov, Fiz. Tverd. Tela (Leningrad) **12**, 954 (1970) [Sov. Phys. Solid State **12**, 749 (1970)].
6. D. W. Fradin, E. Yablonovitch, and M. Bass, Appl. Opt. **12**, 700 (1973).
7. E. Yablonovitch and N. Bloembergen, Phys. Rev. Lett. **29**, 907 (1972).
8. D. W. Fradin and M. Bass, Appl. Phys. Lett. **22**, 206 (1973).
9. A. S. Epifanov, Zh. Éksp. Teor. Fiz. **67**, 1805 (1974) [Sov. Phys. JETP **40**, 897 (1975)].
10. A. S. Epifanov, A. A. Manenkov, and A. M. Prokhorov, Pis'ma Zh. Éksp. Teor. Fiz. **21**, 483 (1975) [JETP Lett. **21**, 223 (1975)].
11. L. H. Holway and D. W. Fradin, J. Appl. Phys. **46**, 279 (1975).
12. A. S. Epifanov, A. A. Manenkov, and A. M. Prokhorov, Zh. Éksp. Teor. Fiz. **70**, 728 (1976) [Sov. Phys. JETP **43**, 377 (1976)].
13. A. S. Epifanov, A. A. Manenkov, and A. M. Prokhorov, Tr. Fiz. Inst. im. P. N. Lebedeva, Akad. Nauk SSSR **101**, 87 (1977).
14. E. P. Pokatilov and V. M. Fomin, Phys. Status Solidi B **73**, 553 (1976).
15. A. Schmid, P. Kelly, and P. Braunlich, Phys. Rev. B **16**, 4569 (1977).

16. S. C. Jones, X. A. Shen, R. F. Braunlich, *et al.*, Phys. Rev. B **35**, 894 (1987).
17. X. A. Shen, S. C. Jones, P. F. Braunlich, and P. Kelly, Phys. Rev. B **36**, 2831 (1987).
18. S. C. Jones, P. Braunlich, R. T. Casper, *et al.*, Opt. Eng. **28**, 1039 (1989).
19. E. Cartier, D. Arnold, D. J. Dimaria, *et al.*, Rev. Solid State Sci. **5**, 531 (1991).
20. D. Du, X. Liu, G. Korn, *et al.*, Appl. Phys. Lett. **64**, 3071 (1994).
21. B. C. Stuart, D. Feit, S. Herman, *et al.*, Phys. Rev. B **53**, 1749 (1996).
22. H. Varel, D. Ashkenasi, A. Rosenfeld, *et al.*, Appl. Phys. A **62**, 1749 (1996).
23. M. Lenzner, J. Kruger, S. Sartania, *et al.*, Phys. Rev. Lett. **80**, 4076 (1998).
24. A. C. Tien, S. Backus, H. Kapteyn, *et al.*, Phys. Rev. Lett. **82**, 3883 (1999).
25. M. Li, S. Menon, J. P. Nibarger, and G. N. Gibson, Phys. Rev. Lett. **82**, 2394 (1999).
26. G. Petite, S. Guizard, Ph. Martin, and F. Quere, Phys. Rev. Lett. **83**, 5182 (1999).
27. A. Kaiser, B. Rethfeld, M. Vicanek, and G. Simon, Phys. Rev. B **61**, 11437 (2000).
28. T. Apostolova and Y. Hahn, J. Appl. Phys. **88**, 1024 (2000).
29. E. Yu. Perlin, A. V. Fedorov, and M. B. Kashevnik, Zh. Éksp. Teor. Fiz. **85**, 1357 (1983) [Sov. Phys. JETP **58**, 787 (1983)].
30. A. M. Danishevskiĭ, E. Yu. Perlin, and A. V. Fedorov, Zh. Éksp. Teor. Fiz. **93**, 1319 (1987) [Sov. Phys. JETP **66**, 747 (1987)].
31. J. S. Chivian, W. E. Case, and D. D. Eden, Appl. Phys. Lett. **35**, 124 (1979).
32. A. W. Kueny, W. E. Case, and M. E. Koch, J. Opt. Soc. Am. B **6**, 639 (1989).
33. A. W. Kueny, W. E. Case, and M. E. Koch, J. Opt. Soc. Am. B **10**, 1834 (1993).
34. S. Guy, M.-F. Joubert, and B. Jacquier, Phys. Rev. B **55**, 8240 (1997).
35. M.-F. Joubert, Opt. Mater. **11**, 181 (1999).
36. M. P. Hehlen, A. Kuditcher, A. L. Lenef, *et al.*, Phys. Rev. B **61**, 1116 (2000).
37. E. Yu. Perlin, A. M. Tkachuk, M.-F. Joubert, *et al.*, Opt. Spektrosk. **90**, 772 (2001) [Opt. Spectrosc. **90**, 691 (2001)].
38. E. Yu. Perlin, Opt. Spektrosk. **91**, 777 (2001) [Opt. Spectrosc. **91**, 729 (2001)].
39. E. Yu. Perlin, J. Lumin. **94–95**, 249 (2001).
40. E. Yu. Perlin, A. V. Ivanov, and R. S. Levitskiĭ, Zh. Éksp. Teor. Fiz. **123**, 612 (2003) [JETP **96**, 543 (2003)].
41. L. V. Keldysh, Zh. Éksp. Teor. Fiz. **47**, 1945 (1964) [Sov. Phys. JETP **20**, 1307 (1964)].
42. Yu. A. Bychkov and A. M. Dykhne, Zh. Éksp. Teor. Fiz. **58**, 1734 (1970) [Sov. Phys. JETP **31**, 928 (1970)].
43. E. Yu. Perlin and D. I. Stasel'ko, Opt. Spektrosk. **98**, 944 (2005) [Opt. Spectrosc. **98**, 844 (2005)].
44. V. A. Kovarskiĭ and E. Yu. Perlin, Fiz. Tverd. Tela (Leningrad) **13**, 1217 (1971) [Sov. Phys. Solid State **13**, 1013 (1971)].
45. V. A. Kovarskii and E. Yu. Perlin, Phys. Status Solidi B **45**, 47 (1971).
46. A. V. Ivanov and E. Yu. Perlin, Opt. Spektrosk. **99**, 790 (2005).
47. Yu. I. Balkareĭ and É. M. Épshteĭn, Fiz. Tverd. Tela (Leningrad) **15**, 925 (1973) [Sov. Phys. Solid State **15**, 641 (1973)].
48. N. Tzoar and J. I. Gersten, Phys. Rev. B **12**, 1132 (1975).
49. E. Yu. Perlin and A. V. Fedorov, Fiz. Tverd. Tela (St. Petersburg) **37**, 1463 (1995) [Phys. Solid State **37**, 792 (1995)].
50. E. Yu. Perlin and A. V. Fedorov, Opt. Spektrosk. **78**, 445 (1995) [Opt. Spectrosc. **78**, 400 (1995)].
51. E. Yu. Perlin and D. I. Stasel'ko, Opt. Spektrosk. **88**, 57 (2000) [Opt. Spectrosc. **88**, 50 (2000)].
52. S. D. Ganichev, E. L. Ivchenko, E. Yu. Perlin, *et al.*, Zh. Éksp. Teor. Fiz. **91**, 1233 (1986) [Sov. Phys. JETP **64**, 729 (1986)].
53. E. Yu. Perlin, Zh. Éksp. Teor. Fiz. **105**, 186 (1994) [JETP **78**, 98 (1994)].

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 367–371.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 422–427.

ELECTRONIC PROPERTIES OF SOLIDS

Negative Magnetoresistance in Binary Distorted Perovskites $Ca(Cu_xMn_{3-x})Mn_4O_{12}$

A. N. Vasil'ev*, O. S. Volkova, and E. A. Goodilin

Moscow State University, Vorob'evy gory, Moscow, 119899 Russia
**e-mail: vasil@mig.phys.msu.ru*

Received March 15, 2005

Abstract—The substitution of copper for manganese in one position of the double distorted perovskite $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ leads to a change in the valence of manganese atoms in the other position. This is accompanied by a sharp increase in the temperature of magnetic ordering and by a change in the conductivity type from semiconductor to metallic. These perovskites also exhibit a negative magnetoresistance in a broad temperature range, with the $[\rho(H) - \rho(0)] \times 100\%/\rho(0)$ ratio (at liquid nitrogen temperature) reaching 10% in a field of 1 T. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

The phenomenon of colossal negative magnetoresistance (MR) in complex oxides and chalcogenides of 3*d* metals is usually observed in the vicinity of the transition to a magnetically ordered state [1–4]. In this state, the ferromagnetic (FM) component is related in many cases to the presence of transition metal ions in different valence states. In this situation, the magnetic ordering is due to a double exchange interaction, which also accounts for high electric conductivity [5]. The magnetic transitions in substances exhibiting colossal MR are frequently closely related to structural phase transitions, since the structures of these compounds contain Jahn–Teller ions [6].

The double distorted perovskite $CaMn_3Mn_4O_{12}$ contains manganese ions in various valence states (one Mn^{4+} and six Mn^{3+} per unit cell), but the magnetic ordering is observed only at low temperatures (T_C = 49 K) and the corresponding FM component is rather small [7]. The main factor determining the magnetic behavior of this compound is probably the superexchange interaction, which possesses an antiferromagnetic character. A structural phase transition in $CaMn_3Mn_4O_{12}$ takes place at T = 440 K and does not influence the magnetic properties of this compound [8].

As Mn^{3+} ions in one position of the parent $CaMn_3Mn_4O_{12}$ structure are replaced by copper ions (Cu^{2+}), the valence of manganese ions in the other position increases from Mn^{3+} to Mn^{4+}. The end member $CaCu_3Mn_4O_{12}$ in this series of copper-substituted compounds no longer contains transition metal ions in different valence states, but it has a high temperature of magnetic ordering (T_C = 340 K) and is characterized by a large FM component of magnetization [9, 10]. All copper-substituted double distorted perovskites of this series exhibit a negative MR, the magnitude of which increases when the sample temperature moves away from the Curie point.

The properties of double distorted perovskites of the $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ series are significantly different from those of other compounds with perovskitelike structures [11–15]. This fact spurred us to undertake a complex investigation of the physical properties of $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ compounds.

2. SAMPLE PREPARATION

The metal oxide compound $CaMn_3Mn_4O_{12}$ belongs to double distorted perovskites of the $AC_3B_4O_{12}$ type and crystallizes in a rhombohedral lattice of the $NaMn_7O_{12}$ type (space group $R\bar{3}$) [16]. Figure 1 shows the structure of this compound in a polyhedral representation. In this structure, MnO_4 squares (position C) contain only Mn^{3+} ions, while MnO_6 (position B) octahedra involve both Mn^{3+} and Mn^{4+} ions. The MnO_4 squares do not come in contact with each other, while the MnO_6 octahedra have common vertices and form a framework of the structure under consideration. When Mn^{3+} in position C are replaced by Cu^{2+} ions so that $x > 0.5$, the crystal lattice symmetry increases to the cubic symmetry of the $CaCu_3Ti_4O_{12}$ type (space group $Im\bar{3}$) [17].

The samples of $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ with x = 0.5, 1, and 2 were synthesized by pyrolysis of nitrate aerosols. The initial reactants $CaCO_3$, Mn_2O_3, and CuO were taken in the stoichiometric ratio and dissolved on heating in excess nitric acid. The obtained solution was dispersed in an ultrasonic bath and transported with air flow to a reactor, where the nitrate aerosol mixture was decomposed at 700°C. Then, submicron particles of the synthesized powder were pressed into tablets and

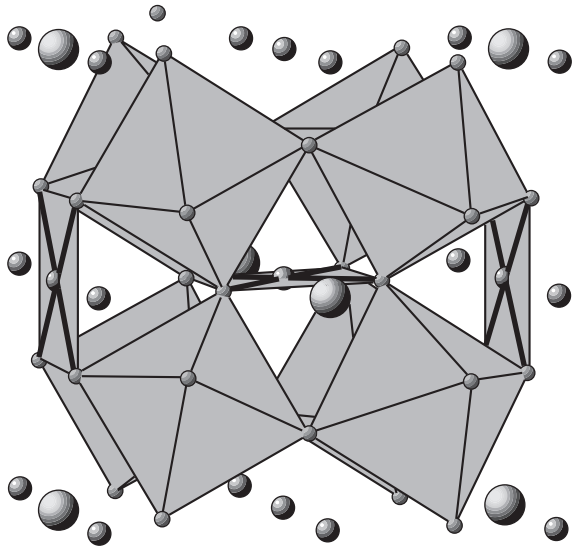

Fig. 1. Crystal structure of $Ca(Cu_xMn_{3-x})Mn_4O_{12}$: Cu^{2+} and Mn^{3+} ions in position C occur in the square environment of O^{2-} ions; Mn^{3+} and Mn^{4+} in position B occur in the octahedral environment of O^{2-} ions; the size of Ca^{2+} cations is much greater than that of Mn^{3+} cations.

annealed in oxygen flow for 48 h at T = 850°C. The samples with x = 2 were annealed in oxygen at a pressure of about 30 bar (48 h, 850°C). The single-phase character of the samples and their structural type ($Im\bar{3}$) were checked by X-ray diffraction. Figure 2 shows the typical morphology of a $CaCuMn_6O_{12}$ sample, which represented a three-dimensional sponge composed of submicron grains.

The physical properties of $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ with x = 0.5, 1, and 2 were studied in a range of temperatures from 5 to 300 K and in applied magnetic fields up to 1 T. The low-field magnetic susceptibility was measured at a frequency of 2.5 kHz using an ac magnetometer (Thermis). The heat capacity was determined with the aid of an quasi-adiabatic calorimeter (Thermis). The electric resistivity and MR measurements were performed using the standard four-point-probe technique.

3. THERMODYNAMIC PROPERTIES

Figure 3 shows the temperature dependences of the low-field magnetic susceptibility χ of $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ samples with x = 0.5, 1, and 2. The shapes of these curves are typical of disordered ferrimagnets featuring competitive magnetic interactions. The initial increase in the χ value with the temperature, accompanying the transition to a magnetically ordered state is followed by a decrease related to spin glass effects. The partial substitution of copper for manganese leads to a growth in the temperature of magnetic ordering, whereby T_C increases from about 90 K for x = 0.5 to ~200 K for x =1 and reaches ~290 K for x = 2. Note that the T_C values observed for the sample studied virtually coincide with the parameter Θ entering into the Curie–Weiss law according to which $\chi = N_A g^2 \mu_{eff}^2 / 3k_B(T - \Theta)$, which is indicative of the absence of frustrations in the structure studied. The effective magnetic moments μ_{eff} = 10–12μ_B in the compounds studied, which were calculated from the temperature dependences of χ in the paramagnetic region assuming that g = 2, are in good agreement with theoretical values calculated for the corresponding combinations of Cu^{2+} (S = 1/2), Mn^{3+} (S = 2), and Mn^{4+} (S = 3/2) ions.

Figure 4 shows the temperature dependences of the heat capacity C for the pressed tablets of

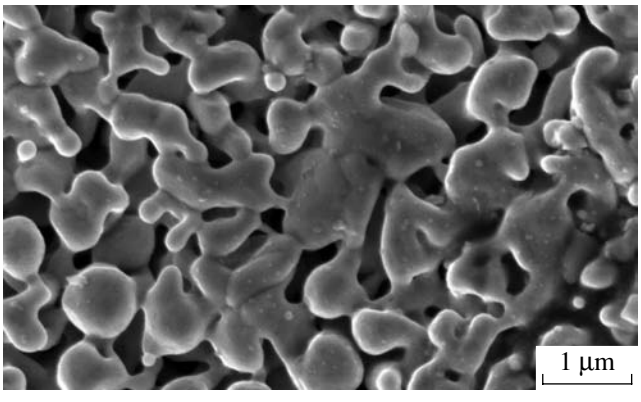

Fig. 2. A micrograph of the $CaCuMn_6O_{12}$ sample structure imaged in the Leo Supra 50 VP scanning electron microscope.

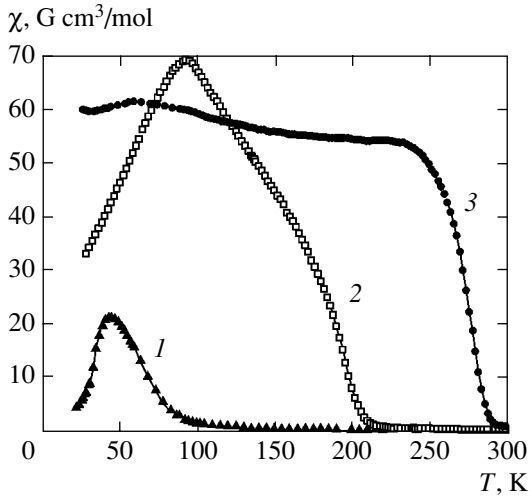

Fig. 3. The temperature dependences of the low-field magnetic susceptibility of $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ with x = 0.5 (*1*), 1 (*2*), and 2 (*3*).

$Ca(Cu_xMn_{3-x})Mn_4O_{12}$ with $x = 0.5$, 1, and 2. The $C(T)$ curves for the samples with $x = 0.5$ and 1 exhibit no clearly pronounced anomalous features upon the transition to a magnetically ordered state, whereas the curve for the sample with $x = 2$ shows an anomaly in the vicinity of T_C. The absence of clearly distinguishable features on the $C(T)$ curves is indicative of a magnetic inhomogeneity of $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ on a molecular level. Indeed, Cu^{2+} cations replace Mn^{3+} cations in position C in a random way, and both Mn^{3+} and Mn^{4+} cations are also randomly distributed in position B. This inhomogeneity leads to smearing of the phase transitions and to spreading of the magnetic specific heat in a broad temperature range. Nevertheless, it is seen that the heat capacity $C(T)$ of the samples with $x = 0.5$ and 1 noticeably exceeds that of the sample with $x = 2$ at low temperatures. At high temperatures, the opposite situation is observed. This is related to the fact that the magnetic specific heat is liberated predominantly below the Curie point, which is lower in $CaCu_{0.5}Mn_{6.5}O_{12}$ ($T_C \approx 90$ K) and $CaCuMn_6O_{12}$ ($T_C \approx 200$ K) than in $CaCu_2Mn_5O_{12}$ ($T_C \approx 290$ K). These features in the heat capacity variations depending on the degree of copper substitution are more clearly manifested in the reduced heat capacity representation of C/T versus T (see the inset to Fig. 4).

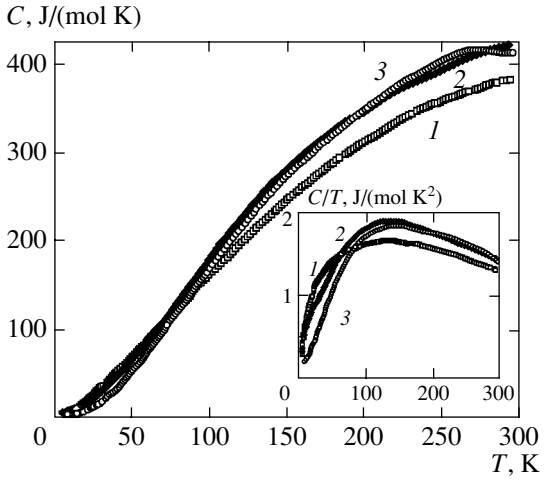

Fig. 4. The temperature dependences of the heat capacity of $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ with $x = 0.5$ (*1*), 1 (*2*), and 2 (*3*). The inset shows the same data plotted as C/T versus T.

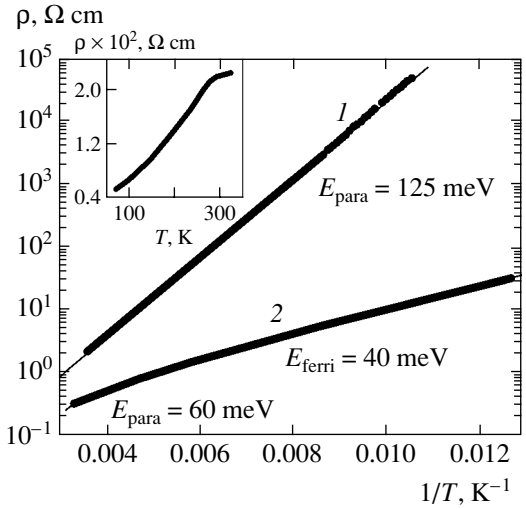

Fig. 5. The temperature dependence of the semiconductor type resistivity in (*1*) $CaCu_{0.5}Mn_{6.5}O_{12}$ and (*2*) $CaCuMn_6O_{12}$. The inset shows the temperature dependence of the metal type resistivity in $CaCu_2Mn_5O_{12}$.

4. KINETIC PROPERTIES

Figure 5 shows the temperature dependence of the resistivity ρ measured in the absence of an external magnetic field. An increase in the content of Cu^{2+} ions in the sample structure leads to a sharp decrease in the resistivity. Moreover, the character of the $\rho(T)$ behavior changes from the semiconductor type in $CaCu_{0.5}Mn_{6.5}O_{12}$ and $CaCuMn_6O_{12}$ to the metallic type in $CaCu_2Mn_5O_{12}$. In the sample with $x = 0.5$, which had a rather large resistance, the conductivity activation energy E_a could be determined from the temperature dependence $\rho = \rho_0 \exp(E_a/k_BT)$ only in the paramagnetic region ($E_{para} = 125$ meV). For the sample with $x = 1$, the $\ln\rho$ versus $1/T$ curve exhibits a bending point at the temperature of magnetic ordering, which corresponds to a decrease in the conductivity activation energy on passage from the paramagnetic state ($E_{para} = 60$ meV) to the magnetically ordered ferrimagnetic state ($E_{ferri} = 40$ meV). In the sample with $x = 2$, the $\ln\rho$ versus $1/T$ curve also exhibits bending, which reflects an increase in the mobility of charge carriers.

All $Ca(Cu_xMn_{3-x})Mn_4O_{12}$ samples ($x = 0.5$, 1, and 2) exhibited a negative MR in the entire range of existence of the magnetically ordered phase, which qualitatively exhibited the same character of evolution depending on the temperature. Figure 6 shows plots of the $[\rho(H) - \rho(0)] \times 100\%/\rho(0)$ ratio versus the magnetic field strength H for the sample of $CaCuMn_6O_{12}$. The shape of the curve exhibits a qualitative change in the vicinity of the Curie temperature ($T_C \approx 200$ K). This behavior is related to the fact that the MR at high temperatures is proportional to the square of the field ($\rho \propto \chi H^2$), whereas at low temperatures it is proportional to the square of the magnetization ($\rho \propto M^2$). In the temperature range of magnetic ordering, the magnetization rapidly grows in weak magnetic fields. As a result, the $\rho(H)$ curves are superlinear at $T > T_C$ and sublinear at $T < T_C$.

In a three-dimensional sponge consisting of sintered submicron particles, the transport properties depend not only on the charge transport inside the grains, but also on the tunneling of carriers between grains. In the case when the dimensions of magnetic domains are

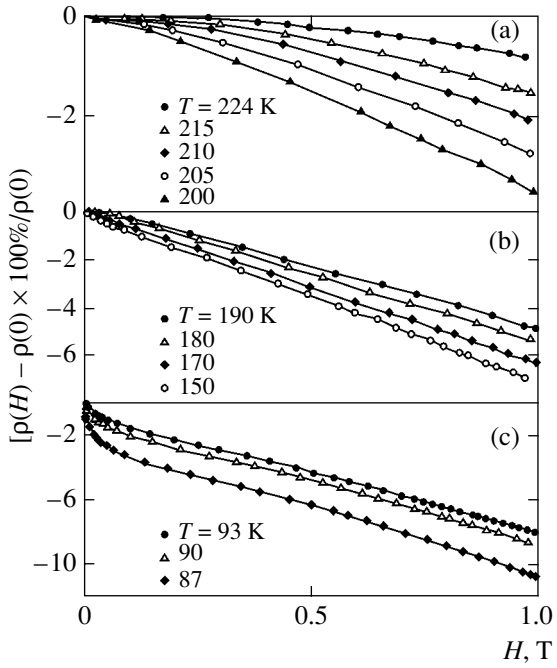

Fig. 6. The magnetoresistance of $CaCuMn_6O_{12}$ at various temperatures (a) above, (b) near, and (c) below the temperature of magnetic ordering ($T_C \sim 200$ K).

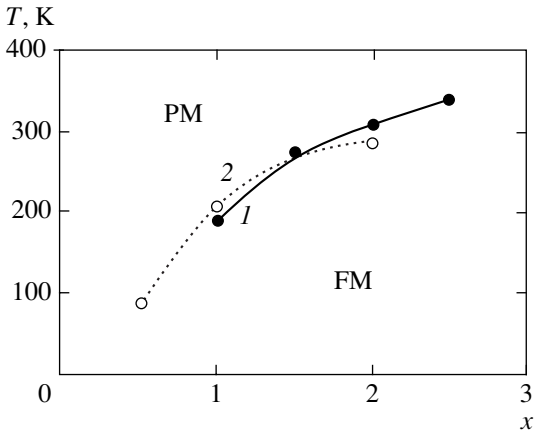

Fig. 7. Plots of the temperature of magnetic ordering versus copper content x in $Ca(Cu_xMn_{3-x})Mn_4O_{12}$: (*1*) data from [9]; (*2*) this study. PM and FM are the regions of the paramagnetic and ferromagnetic state, respectively.

comparable to the grain size, application of the magnetic field leads to parallel alignment of the magnetic moments. This mesoscopic effect is also manifested in the MR behavior.

5. DISCUSSION OF RESULTS

A key role in the formation of a magnetic order in compounds of the $AC_3B_4O_{12}$ type is apparently played by the cations in position B. The BO_6 octahedra share vertices both with each other and with cations in the CO_4 squares, whereas sublattice C has no such short paths of the magnetic interaction between metal cations. The presence of manganese ions with different valences (Mn^{4+} and $3Mn^{3+}$) in positions B of the structure of the parent compound $CaMn_3Mn_4O_{12}$ makes possible the double magnetic interaction in sublattice B. This double interaction is realized via e_g orbitals of manganese and is related to the virtual hopping of electrons between Mn^{3+} and Mn^{4+} ions. In addition, the double magnetic interaction is also possible between Mn^{4+} ions in position B and Mn^{3+} ions in position C, but this interaction is strongly decreased because the schemes of *d* shell splitting in the square and octahedral environment are significantly different. The superexchange between manganese cations via t_{2g} orbitals apparently has an antiferromagnetic character, although this interaction is strongly suppressed due to a particular local geometry of the Mn–O–Mn bonds. In the $CaMn_3Mn_4O_{12}$ structure, the B–O–B angle is 137° and the C–O–B angle varies within 108–113° [16]. As a result, the FM component in the magnetization of $CaMn_7O_{12}$ is small and the magnetic order is established only at a sufficiently low temperature ($T_C = 49$ K).

The substitution of Cu^{2+} for Mn^{3+} in position C leads to a change in the ratio of Mn^{3+} and Mn^{4+} ions in sublattice B. Indeed, the Mn^{3+}/Mn^{4+} ratio is 5 : 3 in $CaCu_{0.5}Mn_{6.5}O_{12}$, 1 : 1 in $CaCuMn_6O_{12}$, and 1 : 3 in $CaCu_2Mn_5O_{12}$. The Mn–O–Mn bond angles vary only slightly as compared to those in the parent $CaMn_7O_{12}$ structure. Thus, in explaining the observed monotonic increase in the Curie temperature and in the FM component of magnetization with increasing x, it is necessary also to take into account the interaction between Cu^{2+} in position C with Mn^{4+} in position B. Only this interaction is retained in the compound $CaCu_3Mn_4O_{12}$, which has the maximum Curie temperature ($T_C \approx 355$ K) in the series of copper-substituted compounds under consideration [9].

Figure 7 shows a plot of the temperature of magnetic ordering versus the content of copper in $Ca(Cu_xMn_{3-x})Mn_4O_{12}$. According to the analysis of magnetic interactions [9, 10, 18], the magnetic structure of these compounds for $0 < x < 3$ corresponds to the ferrimagnetic ordering. In [18], it was postulated that the superexchange between Mn^{4+} ions via t_{2g} orbitals in position B has an FM character. However, we believe that a more important circumstance is that, at an C–O–B angle of about 109°, the electron superexchange interaction between t_{2g} orbitals of Mn^{4+} and e_g orbitals of Cu^{2+} has an antiferromagnetic character.

ACKNOWLEDGMENTS

The authors are grateful to A.G. Veresov, K.V. Klimov, A.V. Knot'ko, and E.A. Popova for their help in the characterization of samples.

This study was supported in part by the Russian Foundation for Basic Research, project nos. 03-02-16108, 04-03-08078, and 04-03-32183a.

REFERENCES

1. *Colossal Magnetoresistance, Charge Ordering and Related Properties of Manganese Oxides*, Ed. by C. N. R. Rao and B. Raveau (World Sci., Singapore, 1998).
2. *Colossal Magnetoresistive Oxides*, Ed. by Y. Tokura (Gordon and Breach, New York, 1999).
3. M. B. Salamon and M. Jaime, Rev. Mod. Phys. **73**, 583 (2001).
4. N. A. Babushkina, L. M. Belova, D. I. Khomskii, *et al.*, Phys. Rev. B **59**, 6994 (1999).
5. A. Rozenzwaig, Phys. Rev. **181**, 946 (1969).
6. I. Bersuker, *The Jahn–Teller Effect* (Cambridge Univ. Press, Cambridge, 2005).
7. E. A. Pomerantseva, D. M. Itkis, E. A. Goodilin, *et al.*, J. Mater. Chem. **14**, 1150 (2004).
8. I. O. Troyanchuk, L. S. Lobanovsky, N. V. Kasper, *et al.*, Phys. Rev. B **58**, 14 903 (1998).
9. Z. Zeng, M. Greenblatt, M. A. Subramanian, and M. Croft, Phys. Rev. Lett. **82**, 3164 (1999).
10. Z. Zeng, M. Greenblatt, J. E. Sustrom IV, *et al.*, J. Solid State Chem. **147**, 185 (1999).
11. H. Hwang, S.-W. Cheong, P. Radaelli, *et al.*, Phys. Rev. Lett. **75**, 914 (1995).
12. P. Schiffer, A. Ramirez, W. Bao, and S.-W. Cheong, Phys. Rev. Lett. **75**, 3336 (1995).
13. A. N. Vasil'ev, T. N. Voloshok, and R. Suriyanarayanan, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 392 (2001) [JETP Lett. **73**, 349 (2001)].
14. D. A. Filippov, K. V. Klimov, R. Z. Levitin, *et al.*, J. Phys.: Condens. Matter **15**, 8351 (2003).
15. I. O. Troyanchuk, V. A. Khomchenko, G. M. Chobot, *et al.*, J. Phys.: Condens. Matter **15**, 8865 (2003).
16. B. Bochu, J. L. Buevoz, J. Chenavas, *et al.*, Solid State Commun. **36**, 133 (1980).
17. J. Chenavas, J. C. Joubert, M. Marezio, and B. Bochu, J. Solid State Chem. **14**, 25 (1975).
18. R. Weht and W. E. Pickett, cond-mat/0011316.

Translated by P. Pozdeev

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 372–379.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 428–435.

STATISTICAL, NONLINEAR, AND SOFT MATTER PHYSICS

Three-Wave Interaction between Interstrand Modes of the DNA¶

V. L. Golo
Moscow State University, Vorob'evy gory, Moscow, 119992 Russia
e-mail: golo@mech.math.msu.su
Received February 21, 2005

Abstract—We consider the regime in which the bands of the torsional acoustic (TA) and hydrogen-bond-stretch (HBS) modes of DNA interpenetrate each other. We propose a simple model accommodating the helix structure of DNA and, within its framework, find a three-wave interaction between the TA and HBS modes. The phenomenon could be useful for studying the action of microwave radiation on a DNA molecule. Thus, using Zhang's mechanism of the interaction between the system of electric dipoles of a DNA molecule and microwave radiation, we show that the latter could bring about torsional vibrations that maintaining HBS vibrations. We show an estimate of the microwave power density necessary for generating the HBS mode, which significantly depends on the viscous properties of the ambient medium. *© 2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

It is generally accepted that the conformational dynamics of DNA relies significantly on elastic vibrations of the DNA molecule in the region of 10^9–10^{12} Hz [1]. According to Kim and Prohofsky [2], the region comprises two domains, which correspond with different degrees of freedom of the molecule: (1) acoustic modes, which do not involve hydrogen bonds; (2) modes that stretch the hydrogen bonds between the base pairs (HBS modes). A local minimum of the frequency is characteristic of HBS modes [2], its position depending on the choice of the band. Vibrations of DNA were observed in the low-frequency Raman scattering [3, 4] and the Fourier-transform infrared absorption experiments [5]. Globus *et al.* [6] reported the existence of internal modes generated by the interaction of artificial DNA-type molecules with electromagnetic radiation in the submillimeter range. It should be noted that the type of modes observed depends on the kind of DNA samples, i.e., in aqueous solutions, or films and filaments [6, 7]. The experimental data [1] is not conclusive as to the relative positions of the acoustic and HBS modes.

This study is based on the observation that if TA bands penetrate the frequency region of HBS modes, the torsional vibrations of the double helix could periodically change the elastic constants of interstrand motions and thus provide a energy supply for HBS modes. If the attenuation is small enough, the TA band, which has double the frequency with respect to that of the HBS mode, could maintain an HBS mode through parametric resonance. Thus, one could obtain a means for generating an HBS mode and studying the interhelical dynamics of DNA.

2. THE ELASTIC DYNAMICS OF TORSIONAL AND INTERSTRAND MODES

In considering the dynamics of DNA, one has to take into account (1) that DNA has two strands; (2) the base pairs are linked by hydrogen bonds; and (3) there is helical symmetry. We utilize a quasi-one-dimensional lattice model for the elastic properties of DNA, which accommodates these requirements.

El Hasan and Calladine [8] set up a scheme for the internal geometry of the double helix of DNA that describes the relative position of one base with respect to the other in a Watson–Crick base pair and the positions of two base pairs. This is achieved by introducing local frames for the bases and the base pairs and translation slides along their long axes. We follow the guidelines of [8], but, in an attempt at qualitative description of the DNA dynamics, we use a simplified set of variables. We describe the relative position of the bases of a base pair by means of the vector **Y** directed along the long axis (the *y* axis in [8]; see also [9]); **Y** is equal to zero when the base pair is at equilibrium. The relative position of the base pairs is described by the torsional angles ϕ_n, which give deviations from the standard equilibrium twist of the double helix. Thus, a twist of the DNA molecule, which does not involve interstrand motion or mutual displacements of the bases inside the pairs, is determined by the torsional angles ϕ_n, which are the angles of rotation of the base pairs about the axis

¶ The text was submitted by the author in English.

of the double helix. The twist energy of the molecule is given by

$$\sum_n \left[\frac{I}{2}\dot{\phi}_n^2 + \frac{\tau}{2a^2}(\phi_{n+1} - \phi_n)^2\right],$$

where I is the moment of inertia and τ is the twist coefficient, which are assumed the same for all base pairs for simplicity and due to the qualitative picture which we are trying to obtain. Interstrand motions should correspond to the relative motion of the bases inside the base pairs, and, therefore, the kinetic energy due to this degree of freedom may be written as

$$\sum_n \frac{M}{2}\dot{\mathbf{Y}}_n^2,$$

where M is the effective mass of a couple.

For each base pair, we have the reference frame in which the z axis corresponds to the axis of the double helix, the y axis to the long axis of the base pair, and the x axis is perpendicular to the z and y axes (see Fig. 1 in [8]). At equilibrium, the change in position of adjacent base pairs is determined only by the twist angle Ω of the double helix. We assume that $\Omega = 2\pi/10$ as for the B-form of DNA. To determine the energy due to the interstrand displacements, we need to find the strain taking into account the constraint imposed by the helical structure of our system. For this, one may utilize the method employed by Kirchhoff for the twisted rod, that is, the covariant derivative, as was done in [10] for the DNA molecule. But a more simple and straightforward approach is possible.

We confine ourselves only to the torsional degrees of freedom of the double lattice and assume the vectors $\mathbf{Y}_n$ to be parallel to the xy plane, or two-dimensional. Consider the displacements $\mathbf{Y}_n$ and $\mathbf{Y}_{n+1}$ determined within the frames of the two consecutive base pairs, n and $n + 1$. Since we must compare the two vectors in the same frame, we rotate the vector $\mathbf{Y}_{n+1}$ to the frame of the nth base pair,

$$\mathbf{Y}_{n+1}^{\text{back}} = R^{-1}(\phi)\mathbf{Y}_{n+1}.$$

Here, $R^{-1}(\phi)$ is the inverse matrix of the rotation of the nth frame to the $(n + 1)$th one given by the equation

$$R(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \quad (1)$$

The matrix R is 2×2, since the vectors $\mathbf{Y}_n$ are effectively two-dimensional. Then, the strain caused by the displacements of the base pairs is determined by the difference $\mathbf{Y}_{n+1}^{\text{back}} - \mathbf{Y}_n$.[1]

It is important that the angle ϕ is given by the twist angle Ω describing the double helix, in conjunction with the torsional angles ϕ_n, so that

$$\phi = \Omega + \phi_{n+1} - \phi_n.$$

Therefore, the energy due to the interstrand stress is given by

$$\sum_n \left\{\frac{M}{2}\dot{\mathbf{Y}}_n^2 + \frac{K}{2a^2}[R^{-1}(\Omega + \phi_{n+1} - \phi_n)\mathbf{Y}_{n+1} - \mathbf{Y}_n]^2\right\}.$$

It corresponds to the fact that the equilibrium position of the double helix is the twisted one determined by Ω and all ϕ_n being equal to zero. We suppose that the size of the DNA molecule is small enough to be visualized as a straight double helix that is not larger than the persistence length. Hence, the number of base pairs $N \le 150$, approximately. Combining the formulas given above, we can write the total energy of the DNA molecule as

$$\mathcal{H} = \sum_n \left[\frac{I}{2}\dot{\phi}_n^2 + \frac{\tau}{2a^2}(\phi_{n+1} - \phi_n)^2\right] + \sum_n \left\{\frac{M}{2}\dot{\mathbf{Y}}_n^2 + \frac{K}{2a^2}[R^{-1}(\Omega + \phi_{n+1} - \phi_n)\mathbf{Y}_{n+1} - \mathbf{Y}_n]^2 + \frac{\epsilon}{2}\mathbf{Y}_n^2\right\}, \quad (2)$$

where K and a are the torsional elastic constant and the interpair distance, respectively. In the summations given above, n is the number of a site corresponding to the nth base pair, $n = 1, 2, \ldots, N$; N being the number of pairs in the segment of the DNA under consideration. The last term $(\epsilon/2)\mathbf{Y}_n^2$ accommodates the energy of the interstrand separation due to the slides of the bases inside the base pairs.

It should be noted that the dynamical variables ϕ_n and $\mathbf{Y}_n$ are of the same order of magnitude, that is, the first. Consequently, preserving only terms up to the

[1] For this argument, I am indebted to D.I. Tchertov.

third order, we can transform Eq. (2), so that it takes the form

$$\mathcal{H} = \sum_n \left[\frac{I}{2}\dot{\phi}_n^2 + \frac{\tau}{2a^2}(\phi_{n+1}-\phi_n)^2\right]$$

$$+ \sum_n \left\{\frac{M}{2}\dot{\mathbf{Y}}_n^2 + \frac{K}{2a^2}[R^{-1}(\Omega)\mathbf{Y}_{n+1}-\mathbf{Y}_n]^2 + \frac{\epsilon}{2}\mathbf{Y}_n^2\right\} \quad (3)$$

$$+ \frac{K}{a^2}\sum_n (\phi_{n+1}-\phi_n)[R^{-1}(\Omega)\mathbf{Y}_{n+1}\times\mathbf{Y}_n]_3.$$

We have used the fact that the axis of the double helix is directed along the z axis.

Simplifying Eq. (3) by diagonalizing it with the help of the unitary transformation

$$\mathbf{Y}_n = S\mathbf{u}_n, \quad S = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},$$

which is a 2×2 matrix, for the vectors $\mathbf{Y}_n$ and $\mathbf{u}_n$ are effectively two-dimensional, their third coordinates being equal to zero. The equation for energy (3) becomes

$$\mathcal{H} = \sum_n \left[\frac{I}{2}\dot{\phi}_n^2 + \frac{\tau}{2a^2}(\phi_{n+1}-\phi_n)^2\right]$$

$$+ \sum_n \left[\frac{M}{2}\dot{\mathbf{u}}_n\cdot\dot{\mathbf{u}}_n^* + \frac{\epsilon}{2}\mathbf{u}_n\cdot\mathbf{u}_n^*\right.$$

$$\left. + \frac{K}{2a^2}\left(\left|e^{i\Omega}u_{n+1}^1 - u_n^1\right|^2 + \left|e^{-i\Omega}u_{n+1}^2 - u_n^2\right|^2\right)\right]$$

$$- \frac{K}{a^2}\sum_n (\phi_{n+1}-\phi_n)[-ie^{i\Omega}u_{n+1}^1 u_n^{*1} + ie^{-i\Omega}u_{n+1}^2 u_n^{*2}],$$

where $*$ signifies complex conjugation.

We can further simplify the equation for the energy by applying the Fourier transformation given by the equations

$$f_n = \frac{1}{\sqrt{N}}\sum_q e^{-inaq} f_q,$$

$$f_q = \frac{1}{\sqrt{N}}\sum_{n=-N/2}^{n=+N/2} e^{inaq} f_n, \quad q = \frac{2\pi}{N_a}m,$$

$$m = 0, \pm 1, \ldots, \pm\frac{N}{2}.$$

It is important that, after the Fourier transformation, the variables $\mathbf{u}_n$ satisfy the following equations for their complex conjugates:

$$u_q^{*1} = iu_{-q}^2, \quad u_q^{*2} = iu_{-q}^1. \quad (4)$$

The equation for energy can be written as

$$\mathcal{H} = \sum_q \left[\frac{I}{2}\dot{\phi}_q\dot{\phi}_q^* + \frac{\tau}{2a^2}\sin^2\frac{aq}{2}\phi_q\phi_q^*\right]$$

$$+ \sum_q \left[\frac{M}{2}\dot{\mathbf{u}}_q\cdot\dot{\mathbf{u}}_q^* + \frac{\epsilon}{2}\mathbf{u}_q\cdot\mathbf{u}_q^*\right.$$

$$\left. + \frac{2K}{a^2}\left(\sin^2\frac{\Omega-aq}{2}u_q^1 u_q^{*1} + \sin^2\frac{\Omega+aq}{2}u_q^2 u_q^{*2}\right)\right] \quad (5)$$

$$+ \frac{K}{a^2}\sum_{q'q''} i\frac{e^{-iaq}}{\sqrt{N}}$$

$$\times\phi_{q'}[-e^{i\Omega}u_{q''}^1 u_{q'+q''}^{*1} + e^{-i\Omega}u_{q'}^2 u_{q'+q''}^{*2}].$$

The interaction term in Eq. (5) corresponds to the three-wave process and may result in resonance. We use this fact for deriving the parametric maintenance of the u_q modes, i.e., the HBS modes (see below).

In the usual way, one can obtain the equations of motion for u_q^α, $\alpha = 1, 2$, and ϕ_q from the equation for energy given above. The essential point is the effects of dissipation, which are due to ions in the close neighborhood of the molecule and water effects (see [11]). The dissipation can be accommodated by writing terms linear in $\dot{u}_q^\alpha$ and $\dot{\phi}_q$. We take the external force, or torque $\mathcal{T}_q$ into account, only in the equation for ϕ_q, because it corresponds to the external degrees of freedom of our model. Thus, the equations of motion can be written as

$$\ddot{u}_q^\alpha + \omega_{\alpha q}^2 u_q^\alpha + \gamma_u\dot{u}_q^\alpha + \frac{4K\sin\Omega}{Ma^2\sqrt{N}}\sum_{q'} e^{-iaq'}\phi_{q'}u_{q-q'}^\alpha = 0, \quad (6)$$

$$\ddot{\phi}_q + \omega_q^2\phi_q + \gamma_\phi\dot{\phi}_q + i\frac{4K\sin\Omega e^{iaq}}{Ia^2\sqrt{N}}\sum_{q'} u_{q'}^1 u_{q-q'}^2 = \mathcal{T}_q, \quad (7)$$

where

$$\omega_{\alpha q}^2 = \frac{4K}{Ma^2}\sin^2\frac{\Omega+(-1)^\alpha aq}{2} + \frac{\epsilon}{M},$$
$$\omega_q^2 = \frac{4\tau}{Ia^2}\sin^2\frac{aq}{2} \quad (8)$$

are the dispersion laws for the fields u_q^α, $\alpha = 1, 2$, and ϕ_q. We see that the spectrum of ϕ_q has a typical acoustic character, whereas that for u_q^α has a local minimum determined by the helical twist Ω. Thus, the spectrum of our model is in qualitative agreement with the conclusions in [2]. The specific nature of the torque is to be specified elsewhere (see Section 3). For the moment, we consider the general dynamical phenomena to which the torque may be conducive.

Suppose that, on the one hand, the amplitudes of the HBS modes given by u_q^α are so small that the quadratic term in Eq. (7) can be neglected, and on the other hand, the external torque $\mathcal{T}_q$ is appreciable enough to maintain the vibration of the torsional mode ϕ_q. Thus, we can visualize the torsional mode as a pump mode that interacts with the HBS mode u_q^α through the nonlinearity in Eq. (6). We confine ourselves to the case of the torque $\mathcal{T}_q$ being nonzero only at $q = q_*$ and having a frequency of 2ω. Therefore, the forced wave, or the pump wave for the HBS mode, has the form

$$\phi_{q^*} = e^{i2\omega t}\Phi\delta_{qq^*}, \quad \phi_{-q^*} = e^{-i2\omega t}\Phi^*\delta_{-qq^*}. \qquad (9)$$

To obtain larger values for the pump wave ϕ_q, the resonance condition

$$\omega_{q^*} = 2\omega, \quad q_* = \Omega/a$$

should be satisfied, even though the resonance behavior of the torsional ϕ_q-mode itself could be attenuated by dissipation; i.e., it may be a mode of small amplitude.

The equations of motion for u_q^α in the pumping regime are

$$\ddot{u}_q^a + \omega_{\alpha q}^2 u_q^\alpha + \gamma_u \dot{u}_q^\alpha + \frac{2K}{Ma^2}\frac{\sin\Omega}{\sqrt{N}}$$
$$\times (Ae^{i2\omega t}u_{q-q^*}^\alpha + A^*e^{-i2\omega t}u_{q+q^*}^\alpha) = 0,$$

where

$$A = e^{-i\Omega}\Phi.$$

Note that the momentum conservation in the q-values is preserved, as required by the three-wave interaction. The equations given above can be rewritten in the matrix form as

$$\ddot{\mathbf{u}}_\alpha + \hat{\omega}_\alpha^2\mathbf{u}_\alpha + \gamma_u\dot{\mathbf{u}}_\alpha = (e^{i2\omega t}\mathcal{H} + e^{-i2\omega t}\mathcal{H}^+)\mathbf{u}_\alpha, \qquad (10)$$

where $\mathcal{H}$ and $\mathcal{H}^+$ are Hermitian conjugate, and

$$\mathcal{H}^+\mathcal{H} = \mathcal{I}\left(\frac{2K}{Ma^2}\frac{\sin\Omega}{\sqrt{N}}\right)^2|A|^2, \quad \mathcal{I}_{ij} = \delta_{ij}.$$

It is worth noting that Eq. (10) is a kind of the matrix Mathieu equation. In fact, we can apply Rayleigh's method to it for studying parametric resonance [12]. For this, we seek the solution to Eq. (10) in the form of a series:

$$\mathbf{u}(t) = \mathbf{A}_1e^{i\omega t} + \mathbf{B}_1e^{-i\omega t} + \mathbf{A}_3e^{i3\omega t} + \mathbf{B}_3e^{-i3\omega t} + \ldots.$$

Substituting the expression given above into Eq. (10) and preserving only the terms corresponding to $e^{\pm i\omega t}$, we obtain the equations

$$[(-\omega^2 + i\gamma_u\omega)\mathcal{I} + \hat{\omega}_\alpha^2]\mathbf{A}_1 + \mathcal{H}\mathbf{B}_1 = 0,$$

$$[(-\omega^2 - i\gamma_u\omega)\mathcal{I} + \hat{\omega}_\alpha^2]\mathbf{B}_1 + \mathcal{H}^+\mathbf{A}_1 = 0.$$

The compatibility condition of the equations given above can be cast in the form of a determinant for the block matrix,

$$\det\begin{bmatrix} \hat{\omega}_\alpha^2 - \omega^2 + i\gamma_u\omega & \mathcal{H} \\ \mathcal{H}^+ & \hat{\omega}_\alpha^2 - \omega^2 - i\gamma_u\omega \end{bmatrix} = 0, \qquad (11)$$

where $\hat{\omega}^2$ is the matrix of frequencies given by Eq. (8), and ω^2 and $\gamma_u\omega$ are the scalar frequencies. We can transform Eq. (11) into a more amenable form. Note that it is equivalent to the equation

$$\det\left\{\begin{bmatrix} \hat{\omega}_\alpha^2 - \omega^2 + i\gamma_u\omega & \mathcal{H} \\ \mathcal{H}^+ & \hat{\omega}_\alpha^2 - \omega^2 - i\gamma_u\omega \end{bmatrix}\right.$$
$$\left.\times\begin{bmatrix} \mathcal{I} & -(\hat{\omega}_\alpha^2 - \omega^2 + i\gamma_u\omega)^{-1} \\ 0 & \rho^2\mathcal{H}^+ \end{bmatrix}\right\} = 0,$$

where

$$\rho = \frac{Ma^2}{2K}\frac{\sqrt{N}}{\sin\Omega}, \qquad (12)$$

and the matrices $\mathcal{H}^+$ and $\mathcal{H}$ satisfy the equation

$$-\mathcal{I} + \rho^2\mathcal{H}\mathcal{H}^+ = 0.$$

We have used the fact that, for the range of frequencies under consideration, the matrix

$$\hat{\omega}_\alpha^2 - \omega^2 + i\gamma_u\omega$$

is nondegenerate. Therefore, the equation given above is equivalent to

$$\det[-\mathcal{I} + \rho^2(\hat{\omega}_\alpha^2 - \omega^2 + i\gamma_u\omega)$$
$$\times\mathcal{J}(\hat{\omega}_\alpha^2 - \omega^2 - i\gamma_u\omega)\mathcal{J}^+] = 0,$$

where the matrix $\mathcal{J}$ is given by $\mathcal{J}_{qq'} = \delta_{q'q-q_*}$. We may

rewrite the last equation as

$$(\omega_{\alpha q}^2 - \omega^2 - i\gamma_u\omega)(\omega_{\alpha q - q_*}^2 - \omega^2 + i\gamma_u\omega) - \left(\frac{2K}{Ma^2}\frac{\sin\Omega}{\sqrt{N}}\right)^2 |A|^2 = 0, \quad (13)$$

and it is quite similar to the usual condition for parametric resonance. Solutions of Eq. (13) are generally complex and, therefore, correspond to attenuated regimes. But there is a specific wavenumber, q_{res}, for which the solution gives the real frequency ω, and it is easy to see that it should satisfy the constraint

$$\omega_{\alpha q - q_*}^2 = \omega_{\alpha q}^2, \quad q = q_{\text{res}}. \quad (14)$$

Thus, we may cast the condition for parametric resonance in the familiar form [12]

$$(\omega^2 - \omega_{\alpha q_{\text{res}}}^2)^2 + \gamma^2\omega^2 - \left(\frac{2K}{Ma^2}\frac{\sin\Omega}{\sqrt{N}}\right)^2 |A|^2 = 0. \quad (15)$$

3. MICROWAVE IRRADIATION AND THE HBS MODES

We may use the results of the previous section for assessing the action of microwave (mw) radiation on a DNA molecule. The key point is accommodating the fact that the wavelength of radiation is by many orders of magnitudes larger than the characteristic size of the region of the molecule involved in the process. It was Zhang who suggested a mechanism to overcome this difficulty [13]. The main point of Zhang's argument is that the helical configuration of the electric dipoles corresponding to the base pairs makes the interaction of the dipole **P** and the field **E**

$$U = -\mathbf{P}\cdot\mathbf{E}$$

angle-dependent. Therefore, different torsional momenta are applied at the base pairs. The equation for the energy of interaction between the DNA dipoles and an incident microwave is given by

$$-\sum_n \mathbf{E}\cdot R(n\Omega + \phi_n)\mathbf{P}_0,$$

where $R(n\Omega + \phi_n)$ is the rotation matrix given by Eq. (1) and $\mathbf{P}_0$ is the dipole at the site $n = 0$. Consequently, even though the radiation has a plane wave configuration at the molecular scale, it still twists the DNA molecule about the double helix axis. Since the momenta change periodically in time with the incident wave, the irradiation results in a periodic stress that may produce elastic vibrations in the DNA molecule. Zhang suggested that the force may generate resonance vibrations, resulting in a crossover mechanism that takes up initial torsion excitations and transforms them into longitudinal acoustic vibrations.

In the present paper, we try to combine Zhang's mechanism [13] and the excitations of the double helix studied by Prohofsky and Kim [2] with the view of generating interstrand waves in the DNA by mw radiation. In contrast to the original idea by Zhang, we do not utilize a crossover into longitudinal acoustic vibrations but employ the interaction between torsional oscillations and the interstrand ones, i.e., three-wave interaction, given by Eq. (5).

The main point is that, by expanding the rotation matrix $R(n\Omega + \phi_n)$ in the angles ϕ_n and keeping only the first-order terms, we can write Zhang's interaction as

$$\mathcal{H}_Z = -\sum_n \phi_n(\mathbf{E}\times\mathbf{P}_n)_3 + \text{const}, \quad \mathbf{P}_n = R(n\Omega)\mathbf{P}_0, \quad (16)$$

where $\mathbf{P}_0$ is the dipole vector at site $n = 0$. Next, using Eq. (1) for the matrix $R(n\Omega)$ and neglecting the constant term, we rewrite Eq. (16) as

$$\mathcal{H}_Z = \frac{1}{2}\sum_n \phi_n\{e^{in\Omega}[(\mathbf{E}\times\mathbf{P}_0)_3 - i(\mathbf{E}\cdot\mathbf{P}_0)] + e^{-in\Omega}[(\mathbf{E}\times\mathbf{P}_0)_3 + i(\mathbf{E}\cdot\mathbf{P}_0)]\}.$$

Applying the Fourier transform for the ϕ_n and using the equation

$$\frac{1}{N}\sum_n e^{i(\Omega \pm aq)n} = \delta_{\Omega, \pm aq},$$

we obtain the following expression for Zhang's interaction:

$$\mathcal{H}_Z = \frac{N}{2}\{\phi_{q=\Omega/a}[(\mathbf{E}\times\mathbf{P}_0)_3 - i(\mathbf{E}\cdot\mathbf{P}_0)] + \phi_{q=-\Omega/a}[(\mathbf{E}\times\mathbf{P}_0)_3 + i(\mathbf{E}\cdot\mathbf{P}_0)]\}.$$

Hence, the torque $\mathcal{T}_q$ in Eq. (7) corresponding to $\mathcal{H}_Z$ is given by

$$\mathcal{T} = \frac{\mathcal{L}}{I}\delta_{q,-q^*} + \frac{\mathcal{L}^*}{I}\delta_{q,q^*}, \quad q_* = \frac{\Omega}{a}, \quad (17)$$

where

$$\mathcal{L} = \frac{N}{2}[(\mathbf{E}\times\mathbf{P}_0)_3 + i(\mathbf{E}\cdot\mathbf{P}_0)].$$

It should be noted that $\pm q_*$ are the local minima of the HBS modes. From Eq. (8), we infer that q satisfies the constraint given by Eq. (14),

$$q_{\text{res}} = \frac{3}{2}q_*. \quad (18)$$

It is worth noting that the wavenumbers q_* and q_{res} correspond to wavelengths of one and two-thirds turns of the double helix.

The equations given above allows us to make numerical, order-of-magnitude estimates, which enable us to assess the effect of mw radiation on the HBS modes. From Eq. (17), we infer that the torque $\mathcal{T}$ has a size

$$\mathcal{T} \propto e^{2i\omega t}EP,$$

where E and P are the external field and the dipole moment of the base pair, respectively. Next, suppose that the resonance condition

$$\omega_q = 2\omega, \quad q = q_* = \frac{\Omega}{a}$$

is true, so that the action of the radiation on the torsional modes should be the largest possible. Then, the amplitude of the pumping wave, ϕ_{q_*}, according to Eq. (7), is on the order of

$$\Phi \propto \frac{\sqrt{N}}{I}\frac{EP}{2\omega\gamma_\phi}. \tag{19}$$

Next, we turn to Rayleigh's condition for the parametric resonance of the HBS mode given by Eq. (15). For a pumping wave corresponding to Eq. (19), it gives

$$(\omega^2 - \omega_{\alpha q_*}^2)^2 + \gamma_u^2\omega^2 \approx 4\left(\frac{K\sin\Omega}{Ma^2}\frac{EP}{I\gamma_\phi}\right)^2.$$

Hence, we have the threshold

$$\gamma_u\gamma_\phi \le \frac{2K\sin\Omega}{Ma^2\omega^2}\frac{EP}{I}, \tag{20}$$

which is the condition that the energy supplied to a DNA molecule is greater than that dissipated, such that the maintaining of the HBS mode can take place. We suppose that the frequency of the HBS modes, given by Eq. (8), is generally determined by the gap term ϵ/M in the equation for $u_{\alpha q}$ and that the first factor in Eq. (20) does not differ much from unity. This signifies that the energies of the interstrand separation per base pair and the twist of the relative positions of the two adjacent base pairs should be comparable. At any rate, the hypothesis appears not to contradict the data reproduced in [2]. If so, we could have the estimate for the dissipative constants, at least by orders of magnitude,

$$\gamma_u\gamma_\phi \le \frac{EP}{I}. \tag{21}$$

Using the relation

$$E \propto 2\sqrt{\frac{\pi S}{c}},$$

which follows from the expression for the Poynting vector

$$\mathbf{S} = \frac{c}{4\pi}\mathbf{E}\times\mathbf{H},$$

where c is the velocity of light, we rewrite Eq. (21) as

$$\gamma_u\gamma_\phi \le 2\frac{P}{I}\sqrt{\frac{\pi S}{c}}, \tag{22}$$

where S is the power density of the interaction. If we assume

$$P \sim 1\ \text{D} \ \text{ or } \ 10^{-18}\ \text{CGS}$$

and the inertia coefficient $I \sim 10^{-36}$ g cm^2, corresponding to the mass of the base pair ~10^{-22} g and the size ~10 Å, then, for a power density of $S \sim 100$ mW/cm^2, we have $\gamma_u\gamma_\phi \le 10^{16}$ Hz^2 or $\gamma_u, \gamma_\phi \le 10^8$ Hz. The estimate suggests that the effect produced by mw radiation should be sought for at the edge of the gigahertz zone, because, in this case, the requirement on the linewidth is less stringent. It should be noted that the crucial point in assessing the feasibility of experiments on mw radiation of DNA and its possible influence is the part played by the ambient solvent and ions contained in it. In fact, irradiation may result just in heating of the solvent, such that the dissipation due to the ions takes up all effects on the molecules of DNA. Generally, the thin boundary layer of water and ions close to the DNA molecule may have an important bearing on the dynamics initiated by the incident mw radiation and result in overdamping of the molecule's torsional oscillations.

Davis and VanZandt [11] put forward arguments that the ions contained in a layer close to the DNA molecule should have an influence small enough to allow the survival of the effect due to mw irradiation. The part played by the dissipation caused by water is more subtle.

The current arguments [14] about the overdamping of the DNA elastic modes rely on the Stokes law for the friction force, $F = 6\pi\eta R v$, for a sphere of radius R moving in a fluid of viscosity η at a speed v; and in the specific case of the DNA, it should involve the gigahertz frequency region. However, classical hydrodynamics, that is, Navier–Stokes theory, breaks down in the region, as can be inferred from the phenomenon of light scattering in liquids, which is characterized by triplet structure: the central Rayleigh line ν due to the elastic scattering and the Mandelstam–Brillouin doublet $\nu \pm f$ of the inelastic one, with f being the frequency of elastic waves in the liquid. Classical hydrodynamics gives a linewidth in the Mandelstam–Brillouin doublet larger than the distance between this line and the maximum of the central line of the triplet, so that the discrete triplet structure should not be observable; in fact, it is [15]. Mandelstam and Leontovich [15] effected the solution to this problem by using the relaxational theory of hydrodynamics, in which liquid is considered as a viscous elastic medium characterized by a viscosity coef-

ficient η and a shear modulus G, the so-called Maxwell model. In fact, the theory also takes anisotropy effects into account [15]. It predicts that, in the region of hypersound, a few gigahertz or more, the attenuation coefficient for sound waves, α_η, ceases to depend on frequency ω, whereas in the low-frequency classical region, in which the Stokes law is valid, the dependence is $\alpha_\eta \sim \omega^2$. Davis and VanZandt [11] used the Maxwell theory to find estimates for the damping of the DNA elastic modes, taking the first and the second hydration layers and the quasicrystalline structure of water in a neighborhood of DNA into account. They found the attenuation to be two orders of magnitude smaller than that given by the Stokes law.

In fact, there is a further reason for rejecting the approach based on the Stokes law. The water molecules form hydration shells of DNA [16]. The primary hydration shell comprises the water molecules immediately adjacent to the DNA, about 20 molecules per nucleotide pair, which constitute a medium different from bulk water. The secondary hydration shell is generally considered to be similar to bulk water. However, on the spatial scale of the diameter of the DNA molecule, that is, several tens of ångstroms, the water in the second hydration shell is hardly a condensed medium. Indeed, in this case, one should have accommodated its local quasicrystalline structure, described by the icosahedral model [17, 18], which should result in sophisticated dynamical equations. The conclusion is that, presently, it is difficult, if at all possible, to construct accurate theoretical estimates for the attenuation of DNA modes.

From the experimental standpoint, the situation is more advanced. The DNA helical modes were observed in the experiments on Raman [3, 4, 16] and far-infrared [5] scattering. Therefore, one may suggest that the attenuation effects due to viscosity should not preclude elastic modes of DNA. At the same time, small relaxation times for damping between DNA and the first hydration layer, on the order of several tens of picoseconds (see [19]) should result in concerted motion of the double helix of the DNA with the surrounding layer of water. The circumstance could be accommodated within the framework of the semiphenomenological model in the present paper. In fact, the DNA molecule and its first hydration layer still form a helix structure, and the mutual motion of constituent bases of a pair together with hydration water molecules could be described with the field $\mathbf{Y}$. Of course, the values of the model constants, K, τ, should be changed, and for the time being, there is lack of information as to their size.

It is also worth noting that the effects of dissipation in aqueous solutions, where a certain form of the Stokes law could be possible, and in films or fibers, should be quite different. So far, there has been no comprehensive theoretical analysis of the dissipation that would allow comparing the DNA dynamics in solutions and in films. Nonetheless, the interplay of internal vibration modes and submillimeter electromagnetic irradiation was recorded in [6] using Fourier transform spectroscopy and films of poly[A]–poly[U] and poly[C]–poly[G] double-stranded homopolymers. Employing the concept of normal modes, or oscillators, of macromolecules developed earlier for proteins [20] and used later for DNA [21], Globus *et al.* [6] performed a numerical simulation of their experimental results and thus obtained an estimate for the relaxational parameter γ, which has the meaning of oscillator dissipation. It turned out that, in the range of frequencies of several tenths of a centimeter, the best fit for γ is less than a tenth of a centimeter, depending on the conformation of the external electric field and the sample. This value of γ is too large for Eq. (22), but the region of frequencies studied in [6] is far from the edge of the gigahertz region, and therefore one may consider the question of the acceptable dissipation rate still open and suggest that studying the effects of mw radiation on DNA modes may be instrumental in understanding the phenomenon.

4. CONCLUSIONS

We have shown that the elastic dynamics of the double helix could have sufficient structure to ensure stretching of the hydrogen bonds of the base pairs of DNA or generating the HBS modes. If the vibrational modes of DNA are not overdamped by the ambient solvent and if the balance between energies supplied and dissipated is favorable, maintenance of the HBS modes could be expected at the edge of the HBS zone. The best technique for studying hydrogen bond stretching is still Raman spectroscopy, to which certain improvements have been made (see [22] and the references therein). Thus, the HBS modes as well as the breathing modes are well accessible from the experimental standpoint.

The choice of specific means for generating torsional excitations of DNA is important and interesting. In this paper, we have envisaged mw irradiation of DNA. In case the interpenetration of the acoustic and HBS modes takes place, mw radiation would maintain the HBS modes if the power density is sufficiently large, 100 mW/cm^2 or more. It is important that there is no need for long exposures of the sample to radiation. At this point, it is worth noting that our estimate for the critical power density, 100 mW/cm^2, is by orders of magnitude larger than that officially prescribed, i.e., 0.1–0.2 mW/cm^2.

ACKNOWLEDGMENTS

I am thankful to G. Bonnet and T. Globus for useful communications and D.I. Tchertov and Yu.S. Volkov for discussions. This paper was supported by the Russian Foundation for Basic Research (project no. 03-02-16173) and Science School (grant no. 1988.2003.1).

REFERENCES

1. E. B. Starikov, Phys. Rep. **284**, 1 (1997).
2. Y. Kim and E. W. Prohofsky, Phys. Rev. B **36**, 3449 (1987).
3. H. Urabe, Y. Sugawara, M. Tsukakoshi, and T. Kasuya, J. Chem. Phys. **95**, 5519 (1991).
4. H. Urabe and Y. Tominaga, J. Phys. Soc. Jpn. **50**, 3543 (1981).
5. J. W. Powell, C. S. Edwards, L. Genzel, and A. Wittlin, Phys. Rev. A **35**, 3929 (1987).
6. T. Globus, M. Bykhovskaia, D. Woolard, and B. Gelmont, J. Phys. D: Appl. Phys. **36**, 1314 (2003).
7. C. S. Edwards and Changle Liu, Phys. Rev. A **44**, 2709 (1991).
8. M. A. El Hassan and C. R. Calladine, J. Mol. Biol. **251**, 648 (1995).
9. Ch. A. Hunter, J. Mol. Biol. **230**, 1025 (1993).
10. V. L. Golo, E. I. Kats, and M. Peyrard, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 225 (2001) [JETP Lett. **73**, 202 (2001)].
11. M. E. Davis and L. L. VanZandt, Phys. Rev. A **37**, 888 (1988).
12. J. W. Rayleigh, *The Theory of Sound* (Macmillan, London, 1926; Gostekhizdat, Moscow, 1955), Vol. 1, Chap. 3.
13. C. T. Zhang, Phys. Rev. A **40**, 2148 (1989).
14. R. K. Adair, Biophys. J. **82**, 1147 (2002).
15. I. L. Fabelinsky, *Molecular Scattering of Light* (Nauka, Moscow, 1965; Plenum, New York, 1968), Chap. 6.
16. N. J. Tao and S. M. Lindsay, Biopolymers **28**, 1019 (1989).
17. M. Chaplin, Biophys. Chem. **83**, 211 (2000).
18. A. Müller, H. Bögge, and E. Diemann, Inorg. Chem. Commun. **6**, 52 (2003).
19. N. J. Tao, S. M. Lindsay, and A. Rupprecht, Biopolymers **27**, 1655 (1988).
20. N. Go, T. Noguti, and T. Nishikawa, Proc. Natl. Acad. Sci. USA **80**, 3696 (1983).
21. T. H. Duong and K. Zakrzewska, J. Comput. Chem. **18**, 796 (1997).
22. L. Movileanu, J. M. Benevides, and C. J. Thomas, Jr., Nucleic Acids Res. **30**, 3767 (2002).

Journal of Experimental and Theoretical Physics, Vol. 101, No. 2, 2005, pp. 380–387.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 128, No. 2, 2005, pp. 436–448.

STATISTICAL, NONLINEAR, AND SOFT MATTER PHYSICS

Pseudoscalar Particle–Photon Interaction Induced by an External Magnetic Field in an Electron–Positron Plasma

N. V. Mikheev and E. N. Narynskaya

Yaroslavl State University, Sovetskaya ul. 14, Yaroslavl, 150000 Russia
e-mail: mikheev@uniyar.ac.ru

Received December 16, 2004

Abstract—The effective interaction of a pseudoscalar particle with a photon in a magnetized electron–positron plasma is investigated. The plasma and field contributions to the effective coupling between the pseudoscalar particle and the photon are calculated. The effective coupling is shown to be independent of the parameters of the medium and the particle 4-momentum in certain limiting cases. © *2005 Pleiades Publishing, Inc.*

1. INTRODUCTION

At present, investigation of the physics of elementary particles in an external active medium, which can be not only a plasma, but also an external magnetic field, becomes increasingly topical. Since both components of an active medium can be present in astrophysical objects simultaneously, studying the quantum processes and, in particular, the processes involving weakly interacting particles is of indubitable interest from the standpoint of astrophysical applications. In general, the emphasis has been on the physics of neutrinos. The reason is that the processes involving neutrinos play a crucial role in such astrophysical cataclysms as supernovas explosions or neutron-star mergers, since it is these elementary particles that carry away almost all of the energy lost by the star during its explosion while propagating through a dense plasma and a strong magnetic field. However, investigation of the processes involving other particles weakly interacting with matter can also be of practical importance in describing astrophysical cataclysms [1]. The light or even strictly massless Goldstone and pseudo-Goldstone bosons that arise from new symmetry breaking on a certain energy scale could act as these particles. For example, axions, familons, majorons, etc., belong to this type of particles. Since these particles interact weakly with matter, they have a large penetrability; as a result, the various processes with their emission could be additional sources of energy losses by stars [2–9]. Studies of this kind provide an additional means for placing astrophysical constraints on such parameters of weakly interactive particles as the masses and the coupling constants. At present, the astrophysical method is being successfully used for hypothetical particles.

Another effect of interest from the standpoint of astrophysical applications is the possible asymmetry in the escape of weakly interacting particles during supernova explosions attributable to the presence of a strong external magnetic field, which could be a beautiful macroscopic manifestation of mirror symmetry breaking in the microscopic world. This phenomenon could probably solve the problem of high pulsar velocities. An asymmetry in the escape of such particles from a supernova could lead to jet thrust; as a result, a nascent pulsar would acquire a kick velocity.

The extreme physical conditions that exist inside astrophysical objects have a significant effect on the properties of particles and their interaction. A dense plasma significantly affects the processes with the emission, absorption, and scattering of neutrinos, which play an important role in the cooling dynamics of stellar objects. For example, the decay of a plasmon into a neutrino pair, $\gamma^{pl} \longrightarrow \nu\bar{\nu}$, is not only kinematically possible in a dense medium, but can also be a dominant source of neutrinos [10]. Being an active component of the external medium, the external magnetic field, just as plasma, can strongly affect the particle properties. First, the magnetic field is capable of inducing new particle–particle interactions, for example, between a familon and a photon. Second, the external field substantially changes the particle kinematics, which opens the processes that are forbidden in a vacuum by the energy–momentum conservation law (e.g., the photon decay into an electron–positron pair [11] and the photon splitting into two photons, $\gamma \longrightarrow \gamma\gamma$ [12]). In addition, the magnetic field can have a catalyzing effect on the processes suppressed in the absence of an external magnetic field (the decay of a massive neutrino, $\nu_i \longrightarrow \nu_j\gamma$ [13]).

It should be noted that this effect of the external magnetic field is particularly significant when the field strength exceeds a critical field value for the electron, the so-called Schwinger value, $B_e = m_e^2/e = 4.41 \times$

10^{13} G.[1] Magnetic fields of such or larger strengths can be generated in certain astrophysical objects. For example, the magnetic field generated during supernova core collapse can reach 10^{12}–10^{13} G. Such fields emerge during one-dimensional collapse, where the fall of matter is strictly radial. However, there is strong evidence that the physics of supernovas is much more complex; in particular, the rotation of the collapsing core should also be taken into account. Rotation can increase the magnetic field by a factor of 10^3–10^4, as in Bisnovatyi-Kogan's rotational supernova explosion model [14].

In this paper, we investigate the effect of a magnetized electron–positron plasma on the effective interaction between a pseudoscalar particle and a photon induced by an external magnetic field. As the pseudoscalar particle, we consider the familon that arises from the breaking of horizontal symmetry between the fermion generations [15, 16]. It should be noted that the familon–photon interaction is possible only in the presence of an external magnetic field (in contrast, for example, to the axion). The reason is that the familon does not have the anomalies $\Phi(G\tilde{G})$ and $\Phi(F\tilde{F})$ in a vacuum ($G_{\mu\nu}$ and $F_{\mu\nu}$ are the gluon and electromagnetic field tensors, respectively).

The effective familon–photon interaction induced by an external magnetic field is described by the loop diagram shown in Fig. 1 and can be represented as

$$L_{\phi\gamma} = g_{\phi\gamma}\tilde{F}^{\alpha\beta}(\partial_\beta A_\alpha)\Phi. \quad (1)$$

Here, A_α is the 4-potential of the quantized electromagnetic field,

$$\tilde{F}^{\alpha\beta} = \frac{1}{2}\varepsilon^{\alpha\beta\rho\sigma}F_{\rho\sigma}$$

is a tensor that is dually conjugate to the tensor of the external magnetic field $F_{\rho\sigma}$, Φ is the familon field, and $g_{\phi\gamma}$ is the effective familon–photon coupling in a magnetized plasma. The $\phi \longrightarrow \gamma$ transition amplitude corresponding to Lagrangian (1) is

$$M = ig_{\phi\gamma}(\varepsilon^*\tilde{F}q), \quad (2)$$

where $q_\mu = (\omega, \mathbf{k})$ is the photon momentum 4-vector, and ε_μ is the photon polarization 4-vector. In what follows, the tensor indices in the 4-vectors and tensors in parentheses are assumed to be folded sequentially, for example, $(\varepsilon^*\tilde{F}q) = (\varepsilon^*_\mu\tilde{F}^{\mu\nu}q_\nu)$.

Another familon–photon transition channel becomes possible in the presence of an e^-e^+ plasma, more specifically, the Compton-like familon–photon

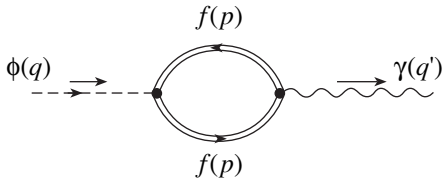

Fig. 1. Diagram describing the loop contribution to the $\phi \longrightarrow \gamma$ transition amplitude in a external magnetic field. The double line corresponds to the exact fermion propagator in magnetic field.

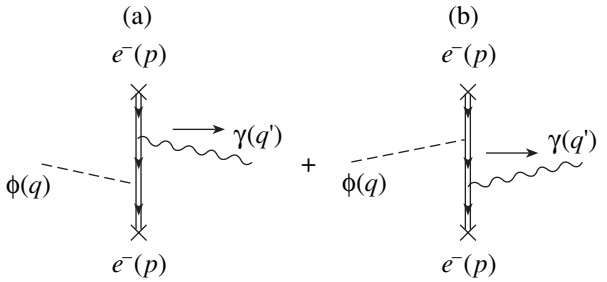

Fig. 2. Compton-like familon forward scattering by plasma electrons. The diagrams describing the familon scattering by plasma positrons can be obtained from the given diagrams by substituting $p \longrightarrow -p$ and by changing the direction of the fermion line.

forward scattering by plasma electrons and positrons indicated by the diagrams in Fig. 2. The contribution of this channel to the effective pseudoscalar particle–photon interaction has previously been disregarded, although, as we show below, it can be much larger than the loop contribution to the effective coupling $g_{\phi\gamma}$ induced by a magnetic field under certain conditions. The total $\phi \longrightarrow \gamma$ conversion amplitude in a magnetized plasma can be represented as the sum of the field and plasma contributions:

$$M = M^F + M^P. \quad (3)$$

The contributions of the external magnetic field and the magnetized plasma to the familon $\longrightarrow$ photon transition amplitude can be determined from the familon–fermion interaction Lagrangian

$$L = \frac{-2im_f c_f}{v_\phi}(\bar{\Psi}_f\gamma_5\Psi_f)\Phi - e_f(\bar{\Psi}_f\hat{A}\Psi_f), \quad (4)$$

where $\hat{A} = (\gamma^\mu A_\mu)$, v_ϕ is the model-dependent energy scale (for the familon, this is the horizontal symmetry breaking scale); c_f is a dimensionless parameter on the order of unity, which is also model-dependent; e_f and m_f are the fermion (f) charge and mass, respectively; and Ψ_f is the fermion field operator. Here, we use the Lagrangian with a pseudoscalar coupling to describe the familon–electron interaction. This does not reduce

[1] The natural system of units in which $c = \hbar = 1$ and the elementary charge $e > 0$ is used.

the generality of the results, since the diagrams under consideration (Figs. 1 and 2) contain only one pseudoscalar particle line.

In this paper, we derive a general expression for the $\phi \longrightarrow \gamma$ conversion amplitude and the effective familon–photon coupling in an electron–positron plasma in the presence of an arbitrary external magnetic field. We consider in detail the strong-field limit commonly realizable in astrophysical conditions where only the ground Landau level is populated.

2. THE FAMILON–PHOTON INTERACTION IN PLASMA IN THE PRESENCE OF AN ARBITRARY CONSTANT MAGNETIC FIELD

2.1. The Contribution to the Effective Familon–Photon Coupling Induced by an External Magnetic Field

The familon–photon interaction in an external magnetic field is indicated by the loop diagram in Fig. 1, where the summation is over the virtual fermions f in the loop. Of all fermions, the electron as the particle with the largest charge-to-mass ratio e/m is most sensitive to the influence of external magnetic field. Therefore, the electron makes a major contribution to the sum; the contribution from other fermions can be ignored.

The $\phi \longrightarrow \gamma$ conversion amplitude induced by an external magnetic field can be taken from [17]. Performing the substitutions

$$j_P \longrightarrow -\frac{2im_e c_e}{v_\phi}, \quad j_{V\mu} \longrightarrow e\varepsilon^*_\mu$$

in Eq. (4.17) from [17] yields

$$\Delta M^F = \frac{-ec_e\beta}{2\pi^2 v_\phi}(\varepsilon^*\tilde{\varphi}q) \times \int_0^1 du \int_0^\infty dt \exp[-i\Omega(u,t)],$$
$$\Omega(u,t) = t\left(1 - \frac{q_\parallel^2}{m_e^2}\frac{1-u^2}{4}\right) + \frac{q_\perp^2}{2m_e^2}\frac{\cos(\eta t u) - \cos(\eta t)}{\eta \sin(\eta t)}, \tag{5}$$

where $\tilde{\varphi}_{\alpha\beta} = \tilde{F}_{\alpha\beta}/B$ is the dual tensor of the external magnetic field reduced to dimensionless form, $\beta = eB$, and $\eta = \beta/m_e^2 = B/B_e$. In what follows, the subscript $\perp$ on the 4-vector means that the vector lies in the plane perpendicular to the magnetic field direction (the field is assumed to be directed along the third axis, $\mathbf{B} = (0, 0, B)$), while the subscript $\parallel$ denotes the longitudinal 4-vector components, i.e.,

$$q_\parallel^\mu = (q_0, 0, 0, q_3),$$
$$q_\perp^\mu = (0, q_1, q_2, 0), \quad q_\parallel^2 = q_0^2 - q_3^2,$$
$$q_\perp^2 = q_1^2 + q_2^2, \quad q^2 = q_\parallel^2 - q_\perp^2.$$

The integration over the variable t in Eq. (5) is performed in the complex plane along the positive direction of the real axis. The integrand on the real axis has a set of poles that should be bypassed below. Let us turn the contour of integration clockwise until it coincides with the imaginary semiaxis, which corresponds to the substitution $t \longrightarrow -i\tau$. After this substitution, Eq. (5) takes a form that is more convenient for analysis:

$$\Delta M^F = \frac{iec_e\beta}{2\pi^2 v_\phi}(\varepsilon^*\tilde{\varphi}q)\int_0^1 du \int_0^\infty d\tau \exp[-\Omega(u,\tau)],$$
$$\Omega(u,\tau) = \tau\left(1 - \frac{q_\parallel^2}{m_e^2}\frac{1-u^2}{4}\right) - \frac{q_\perp^2}{2m_e^2}\frac{\cosh(\eta u\tau) - \cosh(\eta\tau)}{\eta \sinh(\eta\tau)}. \tag{6}$$

It should be noted that, strictly speaking, amplitude (6) is improper, because it contains the Adler anomaly. Since the familon interaction is free from the Adler anomaly, the integral at the tensor structure $(\varepsilon^*\tilde{\varphi}q)$ is zero at $q_\mu = 0$ in the local limit, which is not the case for amplitude (6). Therefore, the triangular Adler anomaly should be subtracted in Eq. (6); in our case, this procedure is reduced to subtracting the field amplitude in the limit of an infinite fermion mass from Eq. (6):

$$M^F = \Delta M^F - \Delta M^F(m_e \longrightarrow \infty).$$

The familon $\longrightarrow$ photon transition amplitude in an external magnetic field can then be written as

$$M^F = \frac{ic_e e^2}{2\pi^2 v_\phi}(\varepsilon^*\tilde{F}q) \times \left(\int_0^1 du \int_0^\infty d\tau \exp[-\Omega(u,\tau)] - 1\right). \tag{7}$$

Comparing (7) with (2), we find the contribution to the effective familon–photon coupling induced by an external magnetic field:

$$g^F_{\phi\gamma} = \frac{c_e e^2}{2\pi^2 v_\phi}\left(\int_0^1 du \int_0^\infty d\tau \exp[-\Omega(u,\tau)] - 1\right). \tag{8}$$

The derived expression is valid for arbitrary magnetic fields.

2.2. The Plasma Contribution to the Effective Familon–Photon Coupling

The contribution of an e^-e^+ plasma to the $\phi \longrightarrow \gamma$ transition amplitude is attributable to the Compton-like familon–photon forward scattering by plasma electrons and positrons (Fig. 2). We will consider in detail the calculation of the transition amplitude M_1^P corresponding to the diagram in Fig. 2b. The amplitude corresponding to the diagram in Fig. 2a and the amplitudes of the scattering processes by plasma positrons can be derived from M_1^P via the corresponding transformations.

Using Lagrangian (4), we obtain the following expression for the S matrix element corresponding to the diagram in Fig. 2b:

$$S_1^P = \frac{iec_e m_e}{v_\phi V\sqrt{\omega\omega'}}\sum_{n=0}^{\infty}\sum_s\int d^4x d^4y dn_{e^-} \times \bar{\psi}_e(p,y)\hat{\varepsilon}^* S(y,x)\gamma_5\psi_e(p,x)e^{-iqx}e^{iq'y}. \quad (9)$$

Here, $q^\alpha = (\omega, \mathbf{k})$ and $q'^\alpha = (\omega', \mathbf{k}')$ are the familon and photon 4-momenta, respectively; $p^\alpha = (E_n, \mathbf{p})$ is the electron 4-momentum;

$$E_n = \sqrt{p_3^2 + m_e^2 + 2\beta n}$$

is the electron energy at Landau level n; ψ_e and $S(y, x)$ are, respectively, the solutions of the Dirac equation and the electron propagator in an external magnetic field; and dn_{e^-} is an element of the phase volume of the plasma electron. In an external magnetic field directed along the z axis in the gauge $\mathbf{A} = (0, Bx, 0)$, the number of plasma electron states is defined as follows:

$$dn_{e^-} = \frac{dp_2 dp_3}{(2\pi)^2}L_2 L_3 f(E_n, \mu).$$

Here, L_2 and L_3 are the auxiliary parameters defining the normalization volume, $L_1L_2L_3 = V$, p_3 is the kinetic momentum along the third axis, p_2 is the generalized momentum defining the position of the center of the Gaussian packet on the first axis in accordance with the equation $x_1 + p_2/\beta = 0$, and $f(E_n, \mu)$ is the equilibrium electron distribution function, which in the rest frame of the plasma is

$$f(E_n, \mu) = \frac{1}{\exp[(E_n - \mu)/T] + 1}, \quad (10)$$

where μ and T are the chemical potential and temperature of the plasma, respectively.

The electron wavefunctions in the chosen gauge can be written as [18]

$$\psi_e(p, x) = \frac{u_s(p,\xi)\exp[-i(E_n x_0 - p_2 x_2 - p_3 x_3)]}{\sqrt{2E_n(E_n + m_e)L_2 L_3}}, \quad (11)$$

where the bispinor amplitudes corresponding to the two spin components along the magnetic field direction $s = \pm 1$ are

$$u_{s=-1}(p,\xi) = \begin{pmatrix} 0 \\ (E_n + m_e)V_n(\xi) \\ -i\sqrt{2\beta n}V_{n-1}(\xi) \\ -p_3 V_n(\xi) \end{pmatrix},$$

$$u_{s=1}(p,\xi) = \begin{pmatrix} (E_n + m_e)V_{n-1}(\xi) \\ 0 \\ p_3 V_{n-1}(\xi) \\ i\sqrt{2\beta n}V_n(\xi) \end{pmatrix}.$$

Here,

$$V_n(\xi) = \frac{\beta^{1/4}}{\sqrt{2^n n!\sqrt{\pi}}}\exp\left(-\frac{\xi^2}{2}\right)H_n(\xi),$$

$H_n(\xi)$ are the Hermitean polynomials, and ξ is a dimensionless coordinate,

$$\xi = \sqrt{\beta}\left(x_1 + \frac{p_2}{\beta}\right).$$

The function $V_n(\xi)$ at negative values of the subscript n is assumed to be zero.

In an arbitrary constant electromagnetic field, the electron propagator has neither translational nor gauge invariance. The noninvariant component is separated out in the form of a phase factor, and the propagator can be represented as [19]

$$S(x,y) = e^{i\Phi(x,y)}S(x-y),$$
$$\Phi(x,y) = e\int_y^x d\xi^\mu\left[A_\mu(\xi) + \frac{1}{2}F_{\mu\nu}(\xi - y)^\nu\right]. \quad (12)$$

Given the explicit form of the electromagnetic field potential in the chosen gauge, the phase in (12) can be calculated in explicit form:

$$\Phi(x,y) = -\frac{\beta}{2}(x_1 + y_1)(x_2 - y_2).$$

The translationally invariant part of the propagator $S(x - y)$ has various representations; in our case, it is

convenient to take this part in the form of a partial Fourier integral expansion:

$$S(z) = -\frac{i}{4\pi}\int_0^\infty \frac{d\tau}{\tanh\tau}\int\frac{d^2p'_\parallel}{(2\pi)^2}$$

$$\times\left\{[p'_\parallel + m_e]\Pi_-(1+\tanh\tau)\right.$$

$$\left. + [\hat{p}'_\parallel + m_e]\Pi_+(1-\tanh\tau) - \frac{i\beta\hat{z}_\perp}{2\tanh\tau}(1-\tanh^2\tau)\right\} \quad (13)$$

$$\times\exp\left(-\frac{\beta z_\perp^2}{4\tanh\tau} - \frac{\tau(m_e^2 - p'^2_\parallel)}{\beta} - i(p'z)_\parallel\right),$$

where $z_\mu = x_\mu - y_\mu$, τ is the imaginary Fock–Schwinger proper time, p'_μ is the virtual electron momentum,

$$d^2p'_\parallel = dp'_0dp'_3, \quad \Pi_\pm = \frac{1}{2}(1 \pm i\gamma_1\gamma_2).$$

After substituting the solutions of the Dirac equation (11) and the electron propagator (12) into the S matrix element (9) and after simple, but slightly cumbersome calculations, the four-dimensional δ function is separated out in the S matrix element as a result of the fact that the initial and final states are formed by neutral particles. In this case, the δ function corresponding to the conservation laws for energy and the second and third photon momentum components is separated out immediately after integration over the 4-coordinates x and y in Minkowski space. The missing δ function corresponding to the law of conservation of the first familon (photon) momentum component is separated out only after integration over the plasma electron momentum p_2. Using the standard definition of the invariant amplitude [20],

$$S = \frac{i(2\pi)^4\delta^4(q-q')}{2\omega V}M,$$

it can be separated from the S matrix element in the form

$$M_1^P = \frac{iec_em_e^2\beta^{3/2}}{8\pi^3 v_\phi}(\varepsilon^*\tilde{\varphi}q)$$

$$\times\sum_{n=0}^\infty\int_{-\infty}^\infty d\rho d^2z_\perp\int_{-\infty}^\infty\frac{dp_3}{E_n}f(E_n,\mu)\int_0^\infty\frac{d\tau}{\tanh(\tau\beta)}$$

$$\times\exp\left[i\left((qz)_\perp + \sqrt{\beta}\rho z_2 + \frac{\beta}{2}z_1z_2\right)\right. \quad (14)$$

$$\left. - \frac{\beta z_\perp^2}{4\tanh(\tau\beta)} - \tau(m_e^2 - (p_\parallel + q_\parallel)^2)\right]$$

$$\times\{V_n(\rho)V_n(\rho+\sqrt{\beta}z_1)(1+\tanh(\tau\beta))$$

$$- V_{n-1}(\rho)V_{n-1}(\rho+\sqrt{\beta}z_1)(1-\tanh(\tau\beta))\}.$$

To calculate the integral over the variable ρ in Eq. (14), we use the relation

$$\int_{-\infty}^\infty d\rho\exp(i\sqrt{\beta}\rho z_2)V_n(\rho)V_n(\rho+\sqrt{\beta}z_1)$$

$$= \sqrt{\beta}\exp\left(-\frac{i\beta z_1z_2}{2}\right)\exp\left(-\frac{u}{2}\right)L_n(u),$$

where $u = \beta z_\perp^2/2$ and $L_n(u)$ are the Laguerre polynomials normalized by the condition

$$\int_0^\infty\exp(-u)L_n^2(u)du = 1.$$

Further integration over $\perp$ space of the 4-coordinate z in polar coordinates $d^2z_\perp = dz_1dz_2 = z_\perp dz_\perp d\varphi$ allows the result to be expressed in terms of the zero-order Bessel function $J_0(x)$:

$$M_1^P = \frac{iec_em_e^2\beta}{4\pi^2 v_\phi}(\varepsilon^*\tilde{\varphi}q)$$

$$\times\sum_{n=0}^\infty\int_0^\infty\frac{d\tau}{\tanh(\tau\beta)}\exp[-\tau(m_e^2 - (p_\parallel+q_\parallel)^2)]$$

$$\times\int_{-\infty}^\infty\frac{dp_3}{E_n}f(E_n,\mu)\int_0^\infty du\exp\left[-\frac{u}{2}\left(\frac{1}{\tanh(\tau\beta)}+1\right)\right] \quad (15)$$

$$\times J_0\left(\sqrt{\frac{2q_\perp^2}{\beta}u}\right)$$

$$\times\{(1+\tanh(\tau\beta))L_n(u) - (1-\tanh(\tau\beta))L_{n-1}(u)\}.$$

Taking the integral over the variable u using the relation

$$\int_0^\infty\exp(-ax)J_0(b\sqrt{x})L_n(x)dx$$

$$= \frac{(a-1)^n}{a^{n+1}}L_n\left(\frac{b^2}{4a(1-a)}\right)\exp\left(-\frac{b^2}{4a}\right)$$

and passing to the new variable $s = -i\tau$, we find the amplitude attributable to the Compton-like familon–

photon forward scattering by plasma electrons corresponding to the diagram in Fig. 2b:

$$M_1^P = \frac{-c_e e m_e^2 \beta}{2\pi^2 v_\phi}(\varepsilon^* \tilde{\varphi} q) \sum_{n=0}^{\infty} \int_{-\infty}^{\infty} \frac{dp_3}{E_n} f(E_n, \mu) \times \int_0^\infty ds \exp[is(q_\parallel^2 + 2(pq)_\parallel)] \times \exp\left(-\frac{iq_\perp^2 \sin(2\beta s)}{2\beta}\right)\lambda_n(x), \quad (16)$$

where

$$\lambda_n(x) = \exp(-x)\{L_n(2x) - L_{n-1}(2x)\}, \quad x = \frac{q_\perp^2}{\beta}\sin^2(\beta s).$$

The calculations of the amplitude M_2^P corresponding to the diagram in Fig. 2a are identical to the calculations of M_1^P presented above. The result for the transition amplitude M_2^P differs from (17) by the common sign and the substitution $-q$ for q. Taking into account the coherent scattering by all plasma electrons and positrons for the familon $\longrightarrow$ photon conversion amplitude in a magnetized plasma, we obtain

$$M^P = \frac{-c_e e^2 m_e^2}{2\pi^2 v_\phi}(\varepsilon^* \tilde{F} q) \times \sum_{n=0}^{\infty} \int_{-\infty}^{\infty} \frac{dp_3}{E_n}\{f(E_n, \mu) + f(E_n, -\mu)\} \times \int_0^\infty ds \{\exp[is(q_\parallel^2 + 2(pq)_\parallel)] + \exp[is(q_\parallel^2 - 2(pq)_\parallel)]\} \times \exp\left[-\frac{iq_\perp^2 \sin(2\beta s)}{2\beta}\right]\lambda_n(x). \quad (17)$$

Comparing the result obtained with amplitude (2), we find the plasma contribution to the effective familon–photon coupling:

$$g_{\phi\gamma}^P = \frac{ic_e e^2 m_e^2}{2\pi^2 v_\phi} \sum_{n=0}^{\infty} \int_{-\infty}^{\infty} \frac{dp_3}{E_n}\{f(E_n, \mu) + f(E_n, -\mu)\} \times \int_0^\infty ds \{\exp[is(q_\parallel^2 + 2(pq)_\parallel)] + \exp[is(q_\parallel^2 - 2(pq)_\parallel)]\} \times \exp\left[-\frac{iq_\perp^2 \sin(2\beta s)}{2\beta}\right]\lambda_n(x). \quad (18)$$

The derived formula, just as the field contribution (8), is valid for arbitrary magnetic fields. As we show below, it is simplified significantly in the strong-field limit.

3. THE EFFECTIVE PSEUDOSCALAR PARTICLE–PHOTON COUPLING IN A STRONGLY MAGNETIZED PLASMA

In this section, we investigate the familon–photon interaction in the limit of a strongly magnetized plasma where the magnetic field strength is the largest parameter of all the physical parameters that characterize the magnetize plasma:

$$\beta \gg \mu^2, T^2, m_e^2. \quad (19)$$

These physical conditions can be realized in certain astrophysical objects. For example, during a supernova explosion, a region with a size on the order of several dozen kilometers exists near the neutrinosphere where the plasma is relatively tenuous, while the magnetic fields can reach 10^{14}–10^{16} G [21].

The field contribution to the effective familon–photon coupling (8) in the strong-field limit ($\eta = B/B_e \gg 1$) is reduced to

$$g_{\phi\gamma}^F = \frac{-c_e e^2}{2\pi^2 v_\phi} H(z), \quad H(z) = \int_0^1 \frac{du}{1 - z(1-u^2) - i\varepsilon} - 1, \quad (20)$$

where $z = q_\parallel^2/4m_e^2$. The function $H(z)$ has various representations, depending on the relationship between $4m_e^2$ and $q_\parallel^2$:

$$H(z) = \begin{cases} \dfrac{1}{2\sqrt{-z(1-z)}} \ln \dfrac{\sqrt{1-z} + \sqrt{-z}}{\sqrt{1-z} - \sqrt{-z}} - 1, & z < 0, \\ \dfrac{1}{\sqrt{z(1-z)}} \arctan\sqrt{\dfrac{z}{1-z}} - 1, & 0 < z < 1, \\ \dfrac{1}{2\sqrt{z(z-1)}} \ln \dfrac{\sqrt{z} - \sqrt{z-1}}{\sqrt{z} + \sqrt{z-1}} - 1 + \dfrac{i\pi}{2\sqrt{z(z-1)}}, & z > 1. \end{cases}$$

The function $H(z)$ at large and small values of the argument have the following asymptotic behavior:

$$H(z) \approx -1 - \frac{1}{2z}\ln(4|z|) + \frac{i\pi}{2z}\Theta(z), \quad |z| \gg 1, \tag{21}$$

$$H(z) \approx \frac{2}{3}z + \frac{8}{15}z^2, \quad |z| \ll 1, \tag{22}$$

where $\Theta(z)$ is the Heaviside function.

The plasma contribution to the effective coupling $g_{\phi\gamma}$ is determined only by the ground Landau level ($n = 0$) under conditions (19), and the integrals over the variables τ and u in Eq. (18) can be easily calculated:

$$g_{\phi\gamma}^P \approx \frac{c_e e^2 m_e^2}{\pi^2 v_\phi} q_\parallel^2 \int_{-\infty}^{\infty} \frac{dp_3}{E} \frac{f(E,\mu) + f(E,-\mu)}{4(pq)_\parallel^2 - q_\parallel^4}, \tag{23}$$

where $E = \sqrt{p_3^2 + m^2}$ is the electron (positron) energy at the ground Landau level.

Using (23) and (20) for the effective familon–photon coupling in a strongly magnetized plasma, we obtain

$$g_{\phi\gamma} = \frac{-c_e e^2}{2\pi^2 v_\phi} \times \left\{ H(z) - 2m_e^2 q_\parallel^2 \int_{-\infty}^{\infty} \frac{dp_3}{E} \frac{f(E,\mu) + f(E,-\mu)}{4(pq)_\parallel^2 - q_\parallel^4} \right\}. \tag{24}$$

Expression (24) for the effective coupling is simplified significantly in certain limiting cases.

(1) The case of relatively high familon energies ($\beta \gg \omega^2 \gg m_e^2$). In this limit, the plasma contribution to the coupling $g_{\phi\gamma}$ includes the suppression related to the electron mass m_e^2, which is the smallest parameter of the problem. Thus, the effective familon–photon coupling for hard familons is attributable only to the field contribution, which is reduced by using (21) to

$$g_{\phi\gamma} \approx g_{\phi\gamma}^F \approx \frac{c_e e^2}{2\pi^2 v_\phi}. \tag{25}$$

(2) The case of soft familons ($\omega^2 \ll m_e^2$). Under these conditions, since the behavior of the function $H(z)$ is asymptotic at small values of argument (22), the effective coupling $g_{\phi\gamma}$ is dominated by the plasma contribution, which can be represented as a single integral,

$$g_{\phi\gamma} \approx g_{\phi\gamma}^P = \frac{-2c_e e^2}{\pi^2 v_\phi} \times \int_0^{\infty} \frac{p_3 dE}{E} \frac{d}{dE}(f(E,\mu) + f(E,-\mu)). \tag{26}$$

The integral in Eq. (26) can be easily calculated in the following cases:

(i) a degenerate plasma ($\mu \gg T$):

$$g_{\phi\gamma} \approx g_{\phi\gamma}^P \approx \frac{2c_e e^2}{\pi^2 v_\phi}\sqrt{1 - \frac{m_e^2}{\mu^2}}; \tag{27}$$

(ii) an ultrarelativistic plasma:

$$g_{\phi\gamma} \approx g_{\phi\gamma}^P \approx \frac{2c_e e^2}{\pi^2 v_\phi}. \tag{28}$$

As we see from Eqs. (25), (27), and (28), the familon–photon coupling at low familon energies ($\omega \ll m_e$) in a degenerate and/or ultrarelativistic plasma and at high energies ($\omega \gg m_e$) does not depend on the parameters of the medium and the familon (photon) 4-momentum.

4. A FAMILON (PHOTON) PROPAGATING ALONG THE MAGNETIC FIELD

Interestingly, irrespective of the external magnetic field strength, a familon (photon) propagating in a magnetized electron–positron plasma along the magnetic field ($q_\perp = 0$) interacts efficiently only with electrons and positrons at the lowest Landau level. In this case, the expression for the effective familon–photon coupling derived in an arbitrary external magnetic field is reduced to Eq. (24), which defines the coupling $g_{\phi\gamma}$ in the strong-field limit.

To prove this assertion, let us first consider the plasma contribution to the coupling $g_{\phi\gamma}$ described by Eq. (18). Since

$$\lambda_n(0) = \begin{cases} 1, & n = 0, \\ 0, & n \neq 0, \end{cases}$$

we immediately find that the plasma contribution to the effective familon–photon coupling for $q_\perp = 0$ is attributable only to the ground Landau level. Substituting this result into (18) and performing integration over the variable s, we reproduce result (23) obtained in the strong-field limit.

As regards the field contribution for a "longitudinal" familon (photon), we can easily remove the integral over the variable τ by setting $q_\perp = 0$ in Eq. (8) and reproduce result (20).

Thus, the results obtained for an arbitrary magnetic field for a longitudinal familon ($q_\perp = 0$) closely reproduce results (23) and (20) obtained in the strong-field limit where only electrons and positrons at the ground Landau level make a contribution.

5. CONCLUSIONS

We investigated the pseudoscalar particle–photon interaction in an electron–positron plasma in the presence of an arbitrary external magnetic field. The familon arising from the breaking of horizontal symmetry between the fermion generations was considered as the pseudoscalar particle. However, our results can be applicable for any particle with a pseudoscalar interaction with fermions.

We derived expressions for the plasma and field contributions to the effective familon–photon coupling. We considered in detail the strong-field limit where the plasma electrons and positrons populate only the ground Landau level. We showed that the effective familon–photon coupling in a strongly magnetized plasma at low familon energies ($\omega^2 \ll m_e^2$) is attributable mainly to the presence of a plasma and is constant. At relatively high familon energies ($\beta \gg \omega^2 \gg m_e^2$), the effective coupling is dominated by the field contribution, which does not depend on the parameters of the medium and the familon 4-momentum in a degenerate and/or ultrarelativistic plasma, as in the case of low familon energies.

Our results can be of use in investigating the physics of pseudoscalar particles in an external active medium, in particular, under extreme stellar conditions where both components of the active medium are present simultaneously: the plasma and the external magnetic field.

ACKNOWLEDGMENTS

This work was supported in part by the Council for Support of Young Russian Scientists and Leading Scientific Schools (grant no. NSh-1916.2003.2), the Russian Foundation for Basic Research (project no. 04-02-16253), and the Ministry of Education of the Russian Federation (project no. E02-11.0-48).

REFERENCES

1. G. G. Raffelt, *Stars as Laboratories for Fundamental Physics* (Chicago Univ. Press, Chicago, 1996).
2. A. V. Averin, A. V. Borisov, and A. I. Studenikin, Yad. Fiz. **50**, 1058 (1989) [Sov. J. Nucl. Phys. **50**, 660 (1989)].
3. A. V. Averin, A. V. Borisov, and A. I. Studenikin, Phys. Lett. B **231**, 280 (1989).
4. A. V. Borisov and V. Yu. Grishina, Zh. Éksp. Teor. Fiz. **106**, 1553 (1994) [JETP **79**, 837 (1994)].
5. M. Kacherliess, C. Wilke, and G. Wunner, Phys. Rev. D **56**, 1313 (1997).
6. N. V. Mikheev, A. Ya. Parkhomenko, and L. A. Vassilevskaya, Mod. Phys. Lett. A **13**, 2593 (1998).
7. N. V. Mikheev, G. Raffelt, and L. A. Vassilevskaya, Phys. Rev. D **58**, 055008 (1998).
8. L. A. Vasilevskaya, N. V. Mikheev, and A. Ya. Parkhomenko, Yad. Fiz. **62**, 1812 (1999) [Phys. At. Nucl. **62**, 1687 (1999)].
9. N. V. Mikheev and E. N. Narynskaya, Yad. Fiz. **66**, 1337 (2003) [Phys. At. Nucl. **66**, 1297 (2003)].
10. J. B. Adams, M. A. Ruderman, and C.-H. Woo, Phys. Rev. **129**, 1383 (1963).
11. N. P. Klepikov, Zh. Éksp. Teor. Fiz. **26**, 19 (1954).
12. S. L. Adler, Ann. Phys. (N.Y.) **67**, 599 (1971).
13. A. A. Gvozdev, N. V. Mikheev, and L. A. Vassilevskaya, Phys. Rev. D **54**, 5674 (1996).
14. G. S. Bisnovatyĭ-Kogan, Astron. Zh. **47**, 813 (1970) [Sov. Astron. **14**, 652 (1971)]; *Physical Problems of Stellar Evolution* (Nauka, Moscow, 1989) [in Russian].
15. Dzh. L. Chkareuli, Pis'ma Zh. Éksp. Teor. Fiz. **32**, 684 (1980) [JETP Lett. **32**, 671 (1980)]; Z. G. Berejiani and Dzh. L. Chkareuli, Pis'ma Zh. Éksp. Teor. Fiz. **35**, 494 (1982) [JETP Lett. **35**, 612 (1982)].
16. A. A. Ansel'm and N. G. Ural'tsev, Zh. Éksp. Teor. Fiz. **84**, 1961 (1983) [Sov. Phys. JETP **57**, 1142 (1983)].
17. A. V. Kuznetsov and N. V. Mikheev, *Electroweak Processes in External Electromagnetic Fields* (Springer, New York, 2003).
18. A. I. Akhiezer and V. B. Berestetskiĭ, *Quantum Electrodynamics*, 4th ed. (Nauka, Moscow, 1981; Wiley, New York, 1965).
19. C. Itzykson and J.-B. Zuber, *Introduction to Quantum Field Theory* (McGraw-Hill, New York, 1980; Mir, Moscow, 1984).
20. V. B. Berestetskiĭ, E. M. Lifshitz, and L. P. Pitaevskiĭ, *Quantum Electrodynamics*, 3rd ed. (Nauka, Moscow, 1989; Pergamon, Oxford, 1982).
21. M. Ruderman, *In Neutron Star: Theory and Observation* (Kluwer Academic, Dordrecht, 1991); G. S. Bisnovatyĭ-Kogan and S. G. Moiseenko, Astron. Zh. **69**, 563 (1992) [Sov. Astron. **36**, 285 (1992)]; G. S. Bisnovatyĭ-Kogan, Astron. Astrophys. Trans. **3**, 287 (1993).

Translated by V. Astakhov