Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 619–628.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 707–716.

GRAVITATION, ASTROPHYSICS

A General Exact Solution of the Einstein–Dirac Equations with the Cosmological Constant in Homogeneous Space

V. A. Zhelnorovich

Institute of Mechanics, Moscow State University, Vorob'evy gory, Moscow, 119992 Russia
e-mail: zhelnor@imec.msu.ru

Received May 29, 2003

Abstract—We have obtained a general exact solution of the system of Einstein–Dirac equations with the cosmological constant in homogeneous Riemannian space of the first type according to the Bianchi classification.

1. INTRODUCTION

We face two problems when integrating the Einstein–Dirac equations.

The first, purely technical problem stems from the fact that the Einstein–Dirac equations constitute a complex system of nonlinear partial differential equations of the second order for 24 unknown functions. Previously, some of the particular exact solutions [1–6] to the Einstein–Dirac equations in homogeneous space have been obtained only for diagonal metrics of the Riemannian event space.

The second problem is fundamental in nature and stems from the fact that the spinor field functions ψ in the Riemannian event space can be determined only in certain nonholonomic orthonormal bases (tetrads) which that must be specified, or a tetrad gauge is said to be needed. A large number of such gauges are known, and different authors have suggested various gauges. All of these are either noninvariant under transformation of the variables of the observer's coordinate system or are written in the form of differential equations, which complicates the initial system of equations.

Physically, all gauges are equivalent, because the Einstein–Dirac equations are invariant under the choice of tetrads. Mathematically, however, using a bad gauge (i.e., additional equations that close the Einstein–Dirac equations) can greatly complicate the equations, while using a good gauge can significantly simplify them.

In many respects, the problem of choosing a reasonable tetrad gauge stems from the fact that the solutions of the Einstein–Dirac equations have previously been obtained only for diagonal metrics. Since the basis vectors of a holonomic coordinate system for such metrics are orthogonal, the tetrads associated with the orthogonal holonomic basis of the Riemannian space can be chosen naturally.

In this paper, we use a new tetrad gauge [7] that is algebraic and, at the same time, is formed in an invariant way. Using this gauge allows the number of unknown functions in the Einstein–Dirac equations to be reduced by six, while keeping the equations invariant under transformation of the observer's coordinate system.

With the tetrad gauge used here, all of the equations can be written as equations of the first order only for two invariants of the spinor field and the Ricci rotation symbols of the proper vector bases determined by the spinor field. The tetrad gauge transforms the Dirac equations to equations for the Ricci rotation symbols and for the spinor field invariants, with the Ricci rotation symbols in the Dirac equations being linear and without derivatives. Therefore, the Dirac equations in homogeneous Riemannian space close the system of Einstein equations for the Ricci rotation symbols without using additional equations. In this case, we can first integrate the first-order equations for the Ricci rotation symbols and the spinor field invariants and then integrate the first-order equations for the tetrad (Lamé) coefficients.

These two factors—the reduction in the number of unknown functions by six and the possibility of integrating the second-order equations in two steps — considerably simplify the Einstein–Dirac equations. As a result, it becomes possible to obtain new exact solutions of these equations.

A general exact solution of the Einstein–Dirac equations in homogeneous Riemannian space of the first type according to the Bianchi classification was obtained in [7–9]. Recently, new papers in which the authors discuss models described by the Einstein equations with the cosmological constant (including those with the spinor fields) have appeared. In this paper, we obtain a general exact solution to the system of Einstein–Dirac equations with the cosmological constant in homogeneous Riemannian space in connection with increasing interest in studying the role of the cosmological constant.

1063-7761/04/9804-0619$26.00 © 2004 MAIK "Nauka/Interperiodica"

2. THE SPINOR FIELDS IN FOUR-DIMENSIONAL RIEMANNIAN SPACE

Let V be the four-dimensional Riemannian space with metric signature (+, +, +, –) referred to a coordinate system with variables x^i and with a holonomic vector basis J_i (i = 1, 2, 3, 4). We specify the metric tensor of space V in basis J_i by the covariant components g_{ij}; the connectivity is defined by the Christoffel symbols Γ^s_{ij}. In space V, we introduce a smooth field of orthonormal bases (tetrads) $\mathbf{e}_a(x^i)$ (a = 1, 2, 3, 4) in the form

$$\mathbf{e}_a = h^i_a J_i, \quad J_i = h^a_i \mathbf{e}_a, \tag{1}$$

where h^a_i and h^i_a are the scale factors. Below, we denote the indices of the tensor components calculated in basis J_i by the Latin letters $i, j, k, \ldots$, and the indices of the tensor components calculated in the orthonormal bases $\mathbf{e}_a$ by the first Latin letters a, b, c, d, e, and f.

The differential of the vectors of the orthonormal basis $\mathbf{e}_a(x^i)$ is defined by the Ricci rotation symbols $d\mathbf{e}_a = \Delta_{i,a}{}^b \mathbf{e}_b dx^i$, which can be expressed in terms of the scale factors

$$\begin{aligned}\Delta_{i,ac} &= \frac{1}{2}[h^j_c(\partial_i h_j a - \partial_j h_{ia}) \\ &- h^j_a(\partial_i h_{jc} - \partial_j h_{ic}) + h^b_i h^j_a h^s_c(\partial_j h_{sb} - \partial_s h_{jb})].\end{aligned} \tag{2}$$

Here, $\partial_i = \partial/\partial x^i$.

Let us define the spinor field of the first rank, $\boldsymbol{\psi}(x^i)$, in Riemannian space V specified by the contravariant components $\psi^A(x^i)$ (A = 1, 2, 3, 4) in the orthonormal bases $\mathbf{e}_a(x^i)$. The spinor indices are juggled by using the formulas $\psi^A = e^{AB}\psi_B$ and $\psi_A = e_{AB}\psi^B$. In these formulas, $E = \|e_{AB}\|$ and $E^{-1} = \|e^{AB}\|$ are the covariant and contravariant components of the metric spinor, respectively, given by the equations

$$\gamma^T_a = -E\gamma_a E^{-1}, \quad E^T = -E. \tag{3}$$

Here, T is the transposition symbol; γ_a are the four-dimensional Dirac matrices, which, by definition, satisfy the equation

$$\gamma_a\gamma_b + \gamma_b\gamma_a = 2g_{ab}I,$$

where I is a unit four-dimensional matrix; and $\|g_{ab}\| = \mathrm{diag}(1, 1, 1, -1)$ are the covariant components of the metric tensor in the orthonormal basis $\mathbf{e}_a$.

Let us also define the conjugate spinor field by the covariant components $\psi^+ = \|\psi^+_A\|$ using the relation $\psi^+ = \dot{\psi}^T\beta$, in which the dot above the letter means complex conjugation; the invariant spinor of the second rank β is given by the equations

$$\dot{\gamma}^T_a = -\beta\gamma_a\beta^{-1}, \quad \dot{\beta}^T = \beta. \tag{4}$$

In general, the four-component spinor field ψ has two real invariants, ρ and η, that can be determined by using the equation

$$\rho\exp(i\eta) = \psi^+\psi + i\psi^+\gamma^5\psi, \tag{5}$$

where

$$\gamma^5 = \frac{1}{24}\varepsilon^{abcd}\gamma_a\gamma_b\gamma_c\gamma_d,$$

ε^{abcd} are the components of the four-dimensional Levi–Civita pseudotensor, $\varepsilon^{1234} = -1$. Using the spinor field ψ and the conjugate spinor field ψ^+ in Riemannian space V, we can determine the proper orthonormal vector basis $\check{\mathbf{e}}_a$ of the spinor field:

$$\check{\mathbf{e}}_1 = \pi^i J_i, \quad \check{\mathbf{e}}_2 = \xi^i J_i, \quad \check{\mathbf{e}}_3 = \sigma^i J_i, \quad \check{\mathbf{e}}_4 = u^i J_i.$$

The vector components π^i, ξ^i, σ^i, and u^i are specified by the relations [7, 10]

$$\begin{aligned}\rho\pi^i = \mathrm{Im}(\psi^T E\gamma^i\psi), \quad \rho\xi^i = \mathrm{Re}(\psi^T E\gamma^i\psi), \\ \rho\sigma^i = \psi^+\gamma^i\gamma^5\psi, \quad \rho u^i = i\psi^+\gamma^i\psi,\end{aligned} \tag{6}$$

in which the spin tensors $\gamma^i = h^i_a\gamma^a$ satisfy the equation

$$\gamma^i\gamma^j + \gamma^j\gamma^i = 2g^{ij}I.$$

Clearly, the scale factors $\check{h}^i_a$ that correspond to the proper basis $\check{\mathbf{e}}_a$ are defined by the matrix

$$\check{h}^i_a = \begin{Vmatrix} \pi^1 & \xi^1 & \sigma^1 & u^1 \\ \pi^2 & \xi^2 & \sigma^2 & u^2 \\ \pi^3 & \xi^3 & \sigma^3 & u^3 \\ \pi^4 & \xi^4 & \sigma^4 & u^4 \end{Vmatrix}. \tag{7}$$

If the spintensors γ_a, E, and β are specified as

$$\gamma_1 = \begin{Vmatrix} 0 & 0 & 0 & i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ -i & 0 & 0 & 0 \end{Vmatrix}, \quad \gamma_2 = \begin{Vmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{Vmatrix},$$

$$\gamma_3 = \begin{Vmatrix} 0 & 0 & i & 0 \\ 0 & 0 & 0 & -i \\ -i & 0 & 0 & 0 \\ 0 & i & 0 & 0 \end{Vmatrix}, \quad \gamma_4 = \begin{Vmatrix} 0 & 0 & i & 0 \\ 0 & 0 & 0 & i \\ i & 0 & 0 & 0 \\ 0 & i & 0 & 0 \end{Vmatrix}, \qquad (8)$$

$$E = \begin{Vmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{Vmatrix}, \quad \beta = \begin{Vmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{Vmatrix},$$

then the components of the spinor ψ calculated in the proper basis $\check{\mathbf{e}}_a$ are defined by the invariants ρ and η as follows [7, 10]:

$$\check{\psi}^1 = 0, \quad \check{\psi}^2 = i\sqrt{\frac{1}{2}\rho}\exp\left(\frac{i}{2}\eta\right),$$
$$\check{\psi}^3 = 0, \quad \check{\psi}^4 = i\sqrt{\frac{1}{2}\rho}\exp\left(-\frac{i}{2}\eta\right). \qquad (9)$$

The covariant derivatives of the spinor fields ψ and ψ^+ in Riemannian space are known to be given by the equations [11]

$$\nabla_s\psi = \partial_s\psi - \frac{1}{4}\Delta_{s,ij}\gamma^i\gamma^j\psi,$$
$$\nabla_s\psi^+ = \partial_s\psi^+ + \frac{1}{4}\psi^+\Delta_{s,ij}\gamma^i\gamma^j. \qquad (10)$$

The following equations are valid for the covariant derivatives of the spinor fields ψ and ψ^+:

$$\nabla_s\psi = \left(\frac{1}{2}I\partial_s\ln\rho - \frac{1}{2}\gamma^5\partial_s\eta - \frac{1}{4}\check{\Delta}_{s,ij}\gamma^i\gamma^j\right)\psi,$$
$$\nabla_s\psi^+ = \psi^+\left(\frac{1}{2}I\partial_s\ln\rho - \frac{1}{2}\gamma^5\partial_s\eta + \frac{1}{4}\check{\Delta}_{s,ij}\gamma^i\gamma^j\right). \qquad (11)$$

Here, the invariants of the spinor field are defined by Eq. (5), and the Ricci rotation symbols, $\check{\Delta}_{i,jk} = \check{h}^b_j\check{h}^c_k\check{\Delta}_{i,bc}$, correspond to the proper bases of the spinor field and can be calculated by using the formula

$$\check{\Delta}_{s,ij} = \frac{1}{2}(\pi_i\nabla_s\pi_j - \pi_j\nabla_s\pi_i + \xi_i\nabla_s\xi_j - \xi_j\nabla_s\xi_i + \sigma_i\nabla_s\sigma_j - \sigma_j\nabla_s\sigma_i - u_i\nabla_s u_j + u_j\nabla_s u_i). \qquad (12)$$

Equations (11) in four-dimensional pseudo-Euclidean space were obtained in [7, 12]. Eqs. (11) in Riemannian space can be derived from the corresponding equations in pseudo-Euclidean space by substituting the covariant derivatives for the partial derivatives. Eqs. (11) are identically valid by the definition of ρ, η, and $\check{\Delta}_{s,ij}$.

3. THE EINSTEIN–DIRAC EQUATIONS WITH THE COSMOLOGICAL CONSTANT

Let us consider the system of equations

$$\gamma^a\nabla_a\psi + m\psi = 0,$$
$$R_{ab} - \frac{1}{2}Rg_{ab} + \lambda g_{ab} = \kappa T_{ab}, \qquad (13)$$
$$T_{ab} = \frac{1}{4}[\psi^+\gamma_a\nabla_b\psi - (\nabla_b\psi^+)\gamma_a\psi + \psi^+\gamma_b\nabla_a\psi - (\nabla_a\psi^+)\gamma_b\psi].$$

Here, ψ is the four-component spinor field in four-dimensional Riemannian event space V specified in the orthonormal basis $\mathbf{e}_a$; m, λ, and κ are constants; $R = g^{ab}R_{ab}$ is the scalar curvature of space V, R^{ab} are the Ricci tensor components calculated in the basis $\mathbf{e}_a$; $g^{ab} = \mathrm{diag}(1, 1, 1, -1)$; and T_{ab} are the energy–momentum tensor components for the spinor field in the basis $\mathbf{e}_a$.

Since Eqs. (13) are invariant under an arbitrary pseudoorthogonal transformation of the tetrads $\mathbf{e}_a$, the equations that define the tetrad $\mathbf{e}_a$ must be added to close Eqs. (13). These additional equations are commonly called gauge conditions. As the gauge conditions for the tetrads $\mathbf{e}_a$, we assume that an arbitrary $\mathbf{e}_a$ tetrad in Eqs. (13) is identical to the proper tetrad of the spinor field ψ, i.e., $\mathbf{e}_a = \check{\mathbf{e}}_a$. For this gauge, the scale factors h^i_a in Eqs. (13) are identical to the coefficients $\check{h}^i_a$ defined by matrix (7).

In this case, the Dirac equations can be written as equations for the Ricci rotation symbols $\check{\Delta}_{a,bc}$ and the invariants of the spinor field [7, 10]:

$$\check{\partial}_a\ln\rho + \check{\Delta}_{b,a}{}^b = 2m\check{\sigma}_a\sin\eta,$$
$$\check{\partial}^a\eta + \frac{1}{2}\varepsilon^{abcd}\check{\Delta}_{b,cd} = 2m\check{\sigma}^a\cos\eta. \qquad (14)$$

Here,

$$\check{\partial}_a = \check{h}^i_a\partial_i = \{\pi^i\partial_i, \xi^i\partial_i, \sigma^i\partial_i, u^i\partial_i\}$$

is the differentiation operator along the vectors of the proper basis, and $\check{\sigma}_a = \check{\sigma}^a = (0, 0, 1, 0)$ are the components of the vector $\check{\mathbf{e}}_3$ in the proper basis.

The symbols $\breve{\Delta}_{a,bc}$ in Eqs. (14) are related to the scale factors $\breve{h}^i_a$ by

$$\breve{\Delta}_{a,bc} = \frac{1}{2}[\breve{h}^j_a(\breve{\partial}_b\breve{h}_{jc} - \breve{\partial}_c\breve{h}_{jb}) + \breve{h}^j_c(\breve{\partial}_a\breve{h}_{jb} + \breve{\partial}_b\breve{h}_{ja}) - \breve{h}^j_b(\breve{\partial}_a\breve{h}_{jc} + \breve{\partial}_c\breve{h}_{ja})]. \quad (15)$$

Equations (14) are identical to the Dirac equations in the proper basis $\breve{\mathbf{e}}_a$. These equations can also be derived from the Dirac equations in system (13) by changing the derivatives in them using formula (11) and by subsequent algebraic transformations.

Substituting $\breve{h}^i_a$ for $\breve{\Delta}_{a,bc}$ in Eqs. (14) leads to the following system of invariant tensor equations [7, 10]:

$$\nabla_i\rho\pi^i = 0, \quad \nabla_i\rho\xi^i = 0,$$
$$\nabla_i\rho\sigma^i = 2m\rho\sin\eta, \quad \nabla_i\rho u^i = 0,$$
$$\nabla^i\eta - \frac{1}{2}\varepsilon^{ijms}(\pi_j\nabla_m\pi_s + \xi_j\nabla_m\xi_s + \sigma_j\nabla_m\sigma_s - u_j\nabla_m u_s) = 2m\sigma^i\cos\eta. \quad (16)$$

It is convenient to write the Einstein equations in the proper basis $\breve{\mathbf{e}}_a$ as

$$\breve{R}_{ab} = \kappa\breve{T}_{ab} + \left(\frac{1}{2}\kappa m\rho\cos\eta + \lambda\right)g_{ab}. \quad (17)$$

To transform the Einstein equations to (17), we should take into account the fact that, in view of (13), the following equation holds:

$$R = -\kappa T^a_a + 4\lambda = \kappa m\rho\cos\eta + 4\lambda. \quad (18)$$

An expression for the tetrad components of the energy–momentum tensor in the gauge $\mathbf{e}_a = \breve{\mathbf{e}}_a$ was obtained in [10]:

$$\breve{T}_{ab} = \frac{1}{4}\rho\Big[-\breve{\sigma}_b\breve{\partial}_a\eta - \breve{\sigma}_a\breve{\partial}_b\eta + \frac{1}{2}\breve{\sigma}_e(\breve{\Delta}_{a,cd}\varepsilon_b{}^{cde} + \breve{\Delta}_{b,cd}\varepsilon_a{}^{cde})\Big]. \quad (19)$$

The tensor components $\breve{R}_{ab}$ can be expressed in terms of the Ricci rotation symbols as

$$\breve{R}_{ab} = \frac{1}{\sqrt{-g}}\partial_j[\sqrt{-g}(\breve{h}^j_c\breve{\Delta}_{b,a}{}^c - \breve{h}^j_b\breve{\Delta}_{c,a}{}^c)] - \breve{\Delta}_{f,b}{}^c\breve{\Delta}_{c,a}{}^f + \breve{\Delta}_{c,a}{}^c\breve{\Delta}_{f,b}{}^f, \quad (20)$$

where $g = \det\|g_{ij}\|$.

Equations (14)–(20) in the given coordinate system x^i form a closed system of equations for the functions $\pi_i(x^j)$, $\xi_i(x^j)$, $\sigma_i(x^j)$, $u_i(x^j)$, $\rho(x^j)$, and $\eta(x^j)$. The metric tensor components for Riemannian space are related to the functions $\pi_i(x^j)$, $\xi_i(x^j)$, $\sigma_i(x^j)$, and $u_i(x^j)$ by

$$g^{ij} = \breve{h}^i_a\breve{h}^j_b g^{ab} = \pi^i\pi^j + \xi^i\xi^j + \sigma^i\sigma^j - u^iu^j. \quad (21)$$

4. THE GENERAL EXACT SOLUTION OF THE ENSTEIN–DIRAC EQUATION IN HOMOGENEOUS SPACE

Let us consider the four-dimensional Riemannian space referred to a synchronous coordinate system with variables x^i in which, by definition, the following equations hold:

$$g_{44} = g^{44} = -1, \quad g_{4\alpha} = g^{4\alpha} = 0, \quad \alpha = 1, 2, 3. \quad (22)$$

We will seek a solution of Eqs. (14)–(20) in the synchronous coordinate system by assuming that all of the sought-for functions depend only on the parameter $x^4 = t$. Thus, the event space is assumed to be a homogeneous space of the first type according to the Bianchi classification.

In this case, Eqs. (16) can be written as

$$\partial_4(\sqrt{-g}\rho\pi^4) = \partial_4(\sqrt{-g}\rho\xi^4) = \partial_4(\sqrt{-g}\rho u^4) = 0,$$
$$\partial_4(\sqrt{-g}\rho\sigma^4) = 2m\sqrt{-g}\rho\sin\eta,$$
$$\partial_4\eta = -2m\sigma^4\cos\eta. \quad (23)$$

Equations (23) and the equation

$$g^{44} \equiv \pi^4\pi^4 + \xi^4\xi^4 + \sigma^4\sigma^4 - u^4u^4 = -1 \quad (24)$$

form a complete system for the functions π^4, ξ^4, σ^4, u^4, η, and $\rho\sqrt{-g}$. The general solution of Eqs. (23) and (24) is [8, 9]

$$\frac{C_\rho}{\rho\sqrt{-g}} = \frac{\pi^4}{C_\pi} = \frac{\xi^4}{C_\xi} = \frac{u^4}{C_u} = \frac{1}{\sqrt{1 + C_\sigma^2\cos^2(2mt + \varphi)}},$$
$$\sigma^4 = \frac{\varepsilon C_\sigma\sin(2mt + \varphi)}{\sqrt{1 + C_\sigma^2\cos^2(2mt + \varphi)}},$$
$$\exp(i\eta) = \varepsilon\frac{1 + iC_\sigma\cos(2mt + \varphi)}{\sqrt{1 + C_\sigma^2\cos^2(2mt + \varphi)}}, \quad (25)$$

where φ, C_π, C_ξ, C_σ, $C_u \geq 1$, and $C_\rho > 0$ are the integration constants; the coefficient ε can take on any of the

two values: +1 or –1. In view of the synchronism condition (24), the constants C are related by

$$C_\pi^2 + C_\xi^2 + C_\sigma^2 - C_u^2 = -1. \quad (26)$$

Let us denote $h_a = \check{h}_a^4 = (\pi^4, \xi^4, \sigma^4, u^4)$. We find from solution (25) that

$$\{h_1, h_2, h_4\} = \frac{1}{\sqrt{1 + C_\sigma^2 \cos^2(2mt + \varphi)}}\{C_\pi, C_\xi, C_u\}. \quad (27)$$

Thus, we see that, in view of solution (25), the direction of the three-dimensional vector with the components h_1, h_2, and h_4 does not depend on the parameter t.

It follows from definition (15) that the Ricci rotation symbols, $\Delta_{a,bc}$, can be represented as

$$\check{\Delta}_{a,bc} = \frac{1}{2}(h_b s_{ac} - h_c s_{ab} - h_a a_{bc}), \quad (28)$$

where, by definition,

$$s_{ab} = s_{ba} = \check{h}_a^i \partial_4 \check{h}_{ib} + \check{h}_b^i \partial_4 \check{h}_{ia},$$
$$a_{ab} = -a_{ba} = \check{h}_a^i \partial_4 \check{h}_{ib} - \check{h}_b^i \partial_4 \check{h}_{ia}. \quad (29)$$

Since the quantities h_a in (28) are defined by solution (25) as functions of the parameter t, Eq. (28) expresses the 24 dependent functions $\check{\Delta}_{a,bc}$ only in terms of the 16 functions s_{ab} and a_{ab}.

It follows from Eqs. (14) that the antisymmetric quantities a_{ab} are defined by the equality

$$a_{ab} = 4m[(\check{\sigma}_a h_b - \check{\sigma}_b h_a)\sin\eta - \varepsilon_{abcd}\check{\sigma}^c h^d \cos\eta]. \quad (30)$$

Using Eqs. (30) and definitions (19) and (20) for $\check{T}_{ab}$ and $\check{R}_{ab}$, we can write the Einstein equations (17) as the equivalent system of equations

$$\partial_4(\sqrt{-g}s_{ab}) - 2m\sqrt{-g}(h_a s_{bc} + h_b s_{ac})\check{\sigma}^c \sin\eta$$
$$-2\sqrt{-g}\left(m\cos\eta + \frac{1}{8}\kappa\rho\right)(\varepsilon_{cefa}s_b^f + \varepsilon_{cefb}s_a^f)h^c\check{\sigma}^e \quad (31)$$
$$= (\kappa m\rho\sqrt{-g}\cos\eta + 2\lambda\sqrt{-g})(g_{ab} + h_a h_b),$$
$$(s_a^a)^2 - s_{ab}s^{ab} = 8(\kappa\rho m\cos\eta + \lambda).$$

In view of solution (25), the quantity $\rho\sqrt{-g}\cos\eta$ on the right-hand side of the first equation in (31) is a constant:

$$\rho\sqrt{-g}\cos\eta = \varepsilon C_\rho. \quad (32)$$

The first equation in (31) can be obtained by contracting the Einstein equations (17) with the components of the tensor $\delta_c^a + h_c h^a$ in index a. The second equation in (31) can be obtained by contracting the Einstein equations (17) with the components of the tensor $g^{ab} + 2h^a h^b$ in indices a and b.

Contracting the first equation in (31) with g^{ab} in indices a and b yields the equation

$$\partial_4\partial_4\sqrt{-g} = \frac{3}{2}\kappa m\varepsilon C_\rho + 3\lambda\sqrt{-g}, \quad (33)$$

that defines the quantity $\sqrt{-g}$.

If the cosmological constant $\lambda > 0$, then the solution of Eq. (33) is

$$\sqrt{-g} = \frac{\varepsilon\kappa m}{2\lambda} \times C_\rho[-1 + f_1\sinh(\sqrt{3\lambda}t) + f_2\cosh(\sqrt{3\lambda}t)], \quad (34)$$

where f_1 and f_2 are arbitrary constants.

For $\lambda < 0$, we obtain

$$\sqrt{-g} = \frac{\varepsilon\kappa m}{2\lambda}C_\rho\{-1 + f\sin[\sqrt{3\lambda}(t - t_0)]\}, \quad (35)$$

where f and t_0 are the integration constants. The case of $\lambda = 0$ was considered in [7–9].

Since the direction of the three-dimensional vector with the components h_1, h_2, and h_4 does not depend on the parameter t, the components h_1 and h_2 can be made equal to zero by a constant Lorentz transformation of the basis vectors $\check{\mathbf{e}}_1$, $\check{\mathbf{e}}_2$, and $\check{\mathbf{e}}_4$. It is easy to see that the initial system of equations (14)–(20) is invariant under an arbitrary Lorentz transformation of the vectors of the basis $\check{\mathbf{e}}_1$, $\check{\mathbf{e}}_2$, and $\check{\mathbf{e}}_4$ that is independent of the variables x^i. Therefore, it will suffice to consider the solution of Eqs. (14)–(20) only for $h_1 = h_2 = 0$. Under this condition, the first equation in (31) can be written in expanded form as

$$\partial_4(s_{33}\sqrt{-g}) - 4m\sigma^4\sin\eta s_{33}\sqrt{-g}$$
$$= (\varepsilon\kappa m C_\rho + 2\lambda\sqrt{-g})(u^4)^2,$$

$$\partial_4(s_{23}\sqrt{-g}) - 2m\sigma^4\sin\eta s_{23}\sqrt{-g} - \left(2m\cos\eta + \frac{1}{4}\kappa\rho\right)u^4 s_{13}\sqrt{-g} = 0,$$

$$\partial_4(s_{13}\sqrt{-g}) - 2m\sigma^4\sin\eta s_{13}\sqrt{-g} + \left(2m\cos\eta + \frac{1}{4}\kappa\rho\right)u^4 s_{23}\sqrt{-g} = 0,$$

$$\partial_4(s_{11}\sqrt{-g}) + 2\left(2m\cos\eta + \frac{1}{4}\kappa\rho\right)u^4 s_{12}\sqrt{-g} = \varepsilon\kappa m C_\rho + 2\lambda\sqrt{-g},$$

$$\partial_4(s_{22}\sqrt{-g}) - 2\left(2m\cos\eta + \frac{1}{4}\kappa\rho\right)u^4 s_{12}\sqrt{-g} = \varepsilon\kappa m C_\rho + 2\lambda\sqrt{-g},$$

$$\partial_4(s_{12}\sqrt{-g}) - \left(2m\cos\eta + \frac{1}{4}\kappa\rho\right) \times u^4(s_{11} - s_{12})\sqrt{-g} = 0, \quad (36)$$

$$\partial_4(s_{14}\sqrt{-g}) + \left(2m\cos\eta + \frac{1}{4}\kappa\rho\right) \times u^4 s_{24}\sqrt{-g} - 2mu^4\sin\eta s_{31}\sqrt{-g} = 0,$$

$$\partial_4(s_{24}\sqrt{-g}) - \left(2m\cos\eta + \frac{1}{4}\kappa\rho\right) \times u^4 s_{14}\sqrt{-g} - 2mu^4\sin\eta s_{23}\sqrt{-g} = 0,$$

$$\partial_4(s_{34}\sqrt{-g}) - 2m\sin\eta(\sigma^4 s_{34} + u^4 s_{33})\sqrt{-g} = (\varepsilon\kappa m C_\rho + 2\lambda\sqrt{-g})\sigma^4 u^4,$$

$$\partial_4(s_{44}\sqrt{-g}) - 4mu^4\sin\eta s_{34}\sqrt{-g} = (\varepsilon\kappa m C_\rho + 2\lambda\sqrt{-g})(\sigma^4)^2.$$

The general solution of Eqs. (36) is

$$s_{11} = \rho u^4\left[-\frac{1}{3}N + \frac{2}{3C_\rho C_u}\partial_4\sqrt{-g} + \frac{1}{2}B\sin(2(\zeta+\beta))\right],$$

$$s_{22} = \rho u^4\left[-\frac{1}{3}N + \frac{2}{3C_\rho C_u}\partial_4\sqrt{-g} - \frac{1}{2}B\sin(2(\zeta+\beta))\right],$$

$$s_{33} = \frac{2}{3}\rho(u^4)^3\left(N + \frac{1}{C_\rho C_u}\partial_4\sqrt{-g}\right),$$

$$s_{44} = \frac{2}{3}\rho u^4(\sigma^4)^2\left(N + \frac{1}{C_\rho C_u}\partial_4\sqrt{-g}\right),$$

$$s_{12} = -\frac{1}{2}\rho u^4 B\cos(2(\zeta+\beta)),$$

$$s_{13} = \frac{1}{2}\rho(u^4)^2 A\cos(\zeta+\alpha), \quad (37)$$

$$s_{23} = \frac{1}{2}\rho(u^4)^2 A\sin(\zeta+\alpha),$$

$$s_{34} = \frac{2}{3}\rho\sigma^4(u^4)^2\left(N + \frac{1}{C_\rho C_u}\partial_4\sqrt{-g}\right),$$

$$s_{14} = \frac{1}{2}\rho u^4\sigma^4 A\cos(\zeta+\alpha),$$

$$s_{24} = \frac{1}{2}\rho u^4\sigma^4 A\sin(\zeta+\alpha).$$

Here, A, B, N, α, and β are arbitrary constants, $\sqrt{-g}$ in solution (37) is defined by equalities (34) and (35), and ζ is given by

$$\zeta = \int\left(2m\cos\eta + \frac{1}{4}\kappa\rho\right)u^4 dt = \varepsilon\arctan\left(\frac{\tan(2mt+\varphi)}{\sqrt{1+C_\sigma^2}}\right) + \frac{1}{4}\kappa\tau. \quad (38)$$

The parameter τ in (38) depends on the cosmological constant λ and is defined by the integral $\tau = \int\rho u^4 dt$.

In view of Eqs. (18) and (32), the scalar curvature R of the Riemannian event space can be expressed in terms of $\sqrt{-g}$:

$$R = \frac{\varepsilon\kappa m C_\rho}{\sqrt{-g}} + 4\lambda. \quad (39)$$

It thus follows that the points at which $\sqrt{-g}$ becomes zero are the singular points of the curvature tensor.

Substituting the components s_{ab} from (37) into the second equation of system (31) yields a relation between the integration constants A, B, N and f_1, f_2. For $\lambda > 0$, we obtain

$$f_2^2 - f_1^2 = 1 - \frac{\lambda}{\kappa^2 m^2}C_u^2\left(\frac{1}{4}A^2 + \frac{1}{4}B^2 + \frac{1}{3}N^2\right) \le 1.$$

If $f_2^2 \ge f_1^2$, then formula (34) for $\sqrt{-g}$ can be represented as

$$\sqrt{-g} = \frac{\varepsilon\kappa m}{2\lambda}C_\rho\{-1 + f\cosh[\sqrt{3\lambda}(t-t_0)]\}, \quad (40)$$

where t_0 and f are arbitrary constants, $f^2 = f_2^2 - f_1^2$.

The singular points of the solution are defined by the equation

$$f\cosh[\sqrt{3\lambda}(t-t_0)] = 1.$$

The solution has one singular point at $f = 1$ and two singular points for $0 < f < 1$. If $f \le 0$ (in this case, $\varepsilon = -1$), then the solution has no singular points.

If $f_2^2 < f_1^2$, then the following equality is valid for $\sqrt{-g}$:

$$\sqrt{-g} = \frac{\varepsilon\kappa m}{2\lambda}C_\rho\{-1 + f\sinh[\sqrt{3\lambda}(t-t_0)]\}, \quad (41)$$

in which $f^2 = f_1^2 - f_2^2$. In this case, the solution has one singular point defined by the equation

$$f\sinh[\sqrt{3\lambda}(t-t_0)] = 1.$$

If $f_2 = \pm f_1 = f$, then

$$\sqrt{-g} = \frac{\varepsilon\kappa m}{2\lambda}C_\rho\{-1 + f\exp(\pm\sqrt{3\lambda}t)\}. \quad (42)$$

For $\lambda < 0$, we obtain the following formula for the integration constant f in Eq. (35):

$$f^2 = 1 - \frac{\lambda}{\kappa^2 m^2}C_u^2\left(\frac{1}{4}A^2 + \frac{1}{4}B^2 + \frac{1}{3}N^2\right) \ge 1.$$

In this case, there is an infinite number of singular points at which $\sqrt{-g}$ becomes zero.

The positivity condition, $\sqrt{-g} > 0$, imposes constraints on the possible values of the integration constants and the domain of existence of the solution.

Using definition (9) and solution (25) for ρ and η, let us write out the solution for the spinor field components in the proper basis:

$$\check{\psi} = \pm\left\|\begin{array}{c} 0 \\ i\sqrt{\varepsilon C_\rho\dfrac{1 + iC_\sigma\cos(2mt+\varphi)}{2\sqrt{-g}}} \\ 0 \\ i\sqrt{\varepsilon C_\rho\dfrac{1 - iC_\sigma\cos(2mt+\varphi)}{2\sqrt{-g}}}\end{array}\right\|, \quad (43)$$

where $\sqrt{-g}$ is given by (35) or (40)–(42), depending on the sign of λ.

Equations (25), (30), and (37) completely define the Ricci rotation symbols $\check{\Delta}_{a,bc}$ by formula (28) and constitute the first integral of Eqs. (14)–(20). To find the general solution of Eqs. (14)–(20), it will now suffice to integrate Eqs. (29), from which it follows that

$$\partial_4\check{h}_{ib} = \frac{1}{2}(s_{ab} + a_{ab})\check{h}_i^a. \quad (44)$$

Given definition (7) of $\check{h}_i^a$ and solutions (30) and (37) for s_{ab} and a_{ab}, Eqs. (44) can be transformed to

$$\begin{aligned}&\frac{d}{d\tau}(u^4\sigma_j - \sigma^4 u_j)\\ &= \frac{1}{4}A[\pi_j\cos(\zeta+\alpha) + \xi_j\sin(\zeta+\alpha)]\\ &+ (u^4\sigma_j - \sigma^4 u_j)\left(\frac{1}{3\sqrt{-g}}\frac{d}{d\tau}\sqrt{-g} + \frac{1}{3}N\right),\\ &\frac{d}{d\tau}(\xi_j + i\pi_j) = (\xi_j + i\pi_j) \quad (45)\\ &\times\left(\frac{1}{3\sqrt{-g}}\frac{d}{d\tau}\sqrt{-g} - \frac{1}{6}N - i\frac{2m}{\rho}\cos\eta\right)\\ &- \frac{i}{4}(\xi_j - i\pi_j)B\exp[-2i(\zeta+\beta)]\\ &+ \frac{i}{4}A(u^4\sigma_j - \sigma^4 u_j)\exp[-i(\zeta+\alpha)].\end{aligned}$$

The system of equations (45) must be complemented with the synchronism condition

$$g_{4\alpha} \equiv \sigma_4\sigma_\alpha - u_4 u_\alpha = 0, \quad \alpha = 1, 2, 3. \quad (46)$$

In view of Eqs. (25), we obtain

$$\tau = \int\rho u^4 dt = C_\rho C_u\int\frac{dt}{\sqrt{-g}}.$$

If the cosmological constant is positive, $\lambda > 0$, and $\kappa m \ne 0$, then using Eqs. (40) and (41), we obtain

$$\tau = \frac{2\lambda C_u}{\varepsilon\kappa m}\int\frac{dt}{-1 + f\sinh[\sqrt{3\lambda}(t-t_0)]} \quad (47)$$

and

$$\tau = \frac{2\lambda C_u}{\varepsilon\kappa m}\int\frac{dt}{-1 + f\cosh[\sqrt{3\lambda}(t-t_0)]}. \quad (48)$$

For $\lambda < 0$ ($f^2 > 1$), using (35), we find that

$$\tau = \frac{2\lambda C_u}{\varepsilon\kappa m}\int\frac{dt}{-1 + f\sin[\sqrt{-3\lambda}(t-t_0)]}. \quad (49)$$

Integrals (47)–(49) can be found in [13].

Let us now change the unknown functions $(\pi_\lambda, \xi_\lambda, \sigma_\lambda, u_\lambda) \longrightarrow (\pi_\lambda^0, \xi_\lambda^0, \theta_\lambda)$ in Eqs. (45):

$$\xi_\lambda + i\pi_\lambda = (\xi_\lambda^0 + i\pi_\lambda^0)(\sqrt{-g})^{1/3}\exp\left(-\frac{1}{6}N\tau - i\zeta\right),$$

$$\sigma_\lambda = \theta_\lambda(\sqrt{-g})^{1/3}u^4\exp\left(-\frac{1}{6}N\tau\right), \quad (50)$$

$$u_\lambda = \theta_\lambda(\sqrt{-g})^{1/3}\sigma^4\exp\left(-\frac{1}{6}N\tau\right),$$

where ζ is given by Eq. (38), σ^4 and u^4 are defined by solution (25), and $\sqrt{-g}$ is defined by equalities (35) or (40)–(42). As a result, the synchronism condition (46) is satisfied identically, and the system of equations (45) transforms to a system of linear equations with constant coefficients:

$$\frac{d}{d\tau}(\xi_\lambda^0 + i\pi_\lambda^0) = \frac{i}{4}\kappa(\xi_\lambda^0 + i\pi_\lambda^0)$$
$$-\frac{i}{4}B\exp(-2i\beta)(\xi_\lambda^0 - i\pi_\lambda^0) + \frac{i}{4}A\exp(-i\alpha)\theta_\lambda, \quad (51)$$
$$\frac{d}{d\tau}\theta_\lambda = \frac{1}{4}A(\pi_\lambda^0\cos\alpha + \xi_\lambda^0\sin\alpha) + \frac{1}{2}N\theta_\lambda.$$

At $j = 4$, Eqs. (45) are satisfied identically in view of the conditions $h_1 = h_2 = 0$.

The following characteristic equation for the eigenvalue q corresponds to Eqs. (51) for each value of the index $\lambda = 1, 2, 3$:

$$2(N-2q)(16q^2 + \kappa^2 - A^2 - B^2) + A^2[2N - B\sin(2(\alpha-\beta))] = 0. \quad (52)$$

In general, the solution of Eq. (52) is given by the Cardano formulas. A simple solution of this equation is obtained, for example, for $A = 0$ or $2N = B\sin(2(\alpha-\beta))$. The solution of Eqs. (51) for $A = 0$ was considered in [8, 9].

Let us consider the case where the equation $2N = B\sin(2(\alpha-\beta))$ holds, $A \neq 0$. In this case, the eigenvalues are

$$\frac{1}{2}N, \quad \frac{1}{4}\sqrt{A^2 + B^2 - \kappa^2}, \quad -\frac{1}{4}\sqrt{A^2 + B^2 - \kappa^2}.$$

If $0 < A^2 + B^2 < \kappa^2$, then the solution of Eqs. (51) can be written as

$$\xi_\lambda^0 + i\pi_\lambda^0 = \exp(-i\alpha)\left\{-AH_\lambda\exp\left(\frac{1}{2}N\tau\right)\right.$$
$$+ [-[\kappa + B\exp(2i(\alpha-\beta))]F_\lambda + 4iQG_\lambda]\cos(Q\tau) \quad (53)$$
$$\left. + [-[\kappa + B\exp(2i(\alpha-\beta))]G_\lambda - 4iQF_\lambda]\sin(Q\tau)\right\},$$

$$\theta_\lambda = [\kappa - B\cos(2(\alpha-\beta))]H_\lambda\exp\left(\frac{1}{2}N\tau\right)$$
$$+ A[F_\lambda\cos(Q\tau) + G_\lambda\sin(Q\tau)].$$

Here,

$$Q = \frac{1}{4}\sqrt{\kappa^2 - A^2 - B^2},$$

F_λ, G_λ, and H_λ are the integration constants.

For the spatial components of the metric tensor, we obtain the following expression using Eqs. (21) and (50):

$$g_{\alpha\beta} = (\sqrt{-g})^{2/3}\exp\left(-\frac{1}{3}N\tau\right)\left\{H_\alpha H_\beta Z^2\exp(N\tau)\right.$$
$$+ F_\alpha F_\beta[S + M\cos(2Q\tau) + 8QN\sin(2Q\tau)]$$
$$+ G_\alpha G_\beta(S - M\cos(2Q\tau) - 8QN\sin(2Q\tau)]$$
$$+ (F_\alpha G_\beta + F_\beta G_\alpha) \quad (54)$$
$$\times[M\sin(2Q\tau) - 8QN\cos(2Q\tau)]$$
$$+ (F_\alpha H_\beta + F_\beta H_\alpha)2\kappa A\exp\left(\frac{1}{2}N\tau\right)\cos(Q\tau)$$
$$\left. + (G_\alpha H_\beta + G_\beta H_\alpha)2\kappa A\exp\left(\frac{1}{2}N\tau\right)\sin(Q\tau)\right\},$$

where $2N = B\sin(\alpha-\beta)$, and we use the following notation for the constants:

$$M = A^2 + B^2 + \kappa B\cos 2(\alpha-\beta),$$

$$S = \kappa[\kappa + B\cos(2(\alpha-\beta))],$$

$$Z^2 = A^2 + [\kappa - B\cos(2(\alpha-\beta))]^2.$$

The quantity $\sqrt{-g}$ on the right-hand side of equality (54) is defined by Eqs. (35) or (40)–(41). Formula (54) defines the oscillatory approach to the singular points of the solution.

The metric is particularly simple when the phases α and β satisfy the condition $\alpha - \beta = k\pi$, $k = 0, \pm 1, \pm 2, \ldots$:

$$g_{\alpha\beta} = \frac{(\sqrt{-g})^{2/3}}{16Q^2} \times \begin{Vmatrix} S + M\cos(2Q\tau) & M\sin(2Q\tau) & 2\kappa A\cos(Q\tau) \\ M\sin(2Q\tau) & S - M\cos(2Q\tau) & 2\kappa A\sin(Q\tau) \\ 2\kappa A\cos(Q\tau) & 2\kappa A\sin(Q\tau) & Z^2 \end{Vmatrix}.$$

In this case, $N = 0$; A and B are arbitrary; and the constants F_λ, G_λ, and H_λ are given by the equalities

$$\begin{aligned} F_\alpha &= \{(\kappa^2 - B^2)^{-1/2}, 0, 0\}, \\ G_\alpha &= \{0, (\kappa^2 - B^2)^{-1/2}, 0\}, \\ H_\alpha &= \{0, 0, (\kappa^2 - B^2)^{-1/2}\}. \end{aligned} \tag{55}$$

If $A^2 + B^2 > \kappa^2$, then the trigonometric functions in (54) are substituted with hyperbolic functions.

The case where the integration constants A and B in Eqs. (51) are equal to zero, $A = B = 0$, corresponds to a diagonal metric g_{ij}. In this case,

$$\begin{aligned} \xi_\lambda^0 + i\pi_\lambda^0 &= \kappa(-F_\lambda + iG_\lambda)\exp\left(\frac{i}{4}\kappa\tau\right), \\ \theta_\lambda &= \kappa H_\lambda \exp\left(\frac{1}{2}N\tau\right), \\ g_{\alpha b} &= \kappa^2(\sqrt{-g})^{2/3}\left[\exp\left(-\frac{1}{3}N\tau\right)(F_\alpha F_\beta + G_\alpha G_\beta)\right. \\ &\quad \left. + \exp\left(\frac{2}{3}N\tau\right)H_\alpha H_\beta\right]. \end{aligned} \tag{56}$$

Here, $\sqrt{-g}$ takes the form of (35) or (40)–(42).

If we define F_λ, G_λ, and H_λ by the relations

$$\begin{aligned} F_\alpha &= \{\kappa^{-1}, 0, 0\}, \quad G_\alpha = \{0, \kappa^{-1}, 0\}, \\ H_\alpha &= \{0, 0, \kappa^{-1}\}, \end{aligned} \tag{57}$$

then metric (56) is diagonal:

$$g_{\alpha\beta} = (\sqrt{-g})^{2/3}\mathrm{diag}\{e^{-N\tau/3}, e^{-N\tau/3}, e^{2N\tau/3}\}. \tag{58}$$

For $A = B = N = 0$, metric (54) corresponds to an isotropic Universe:

$$g_{\alpha\beta} = \kappa^2(\sqrt{-g})^{2/3}(F_\alpha F_\beta + G_\alpha G_\beta + H_\alpha H_\beta). \tag{59}$$

Here, $\sqrt{-g}$ is given by formula (40) in which $f^2 = 1$ for $\lambda > 0$; for $\lambda < 0$, Eq. (35) in which $f^2 = 1$ holds.

Calculating the energy–momentum tensor of the spinor field for the solutions obtained yields the following relation for its component T_{44} in the synchronous coordinate system:

$$T_{44} = m\rho\cos\eta = \frac{\varepsilon m C_\rho}{\sqrt{-g}}. \tag{60}$$

Since, by definition, $C_\rho > 0$, the above solutions with $\varepsilon = -1$ correspond to a negative energy density; for $\varepsilon = 1$, the energy density of the spinor field is positive.

Let us consider the transformation of the variables of the observer's coordinate system $(x^\alpha, t) \longrightarrow (x^\alpha, \tau)$ defined by the equation $d\tau = \rho u^4 dt$. The explicit dependence of the function $\tau(t)$ is given by relations (47), (48), or (49). Our calculation shows that the following equation holds in the coordinate system with variables x^α and τ:

$$[\partial_j(\sqrt{-g}g^{ij})]' = \frac{d}{d\tau}(\sqrt{-g}g^{4i})' = 0. \tag{61}$$

Thus, the coordinate system with variables x^α and τ is harmonic. It is easy to see that the parameters t and τ/C_ρ, where the integration constant C_ρ is defined by solution (25), have the same dimensions.

After the transformation $(x^\alpha, t) \longrightarrow (x^\alpha, \tau)$, the spatial part of the metric does not change. It should be noted that the time dependence of the determinant $\gamma = \det\|g_{\alpha\beta}\|$ of the spatial metric tensor components, which governs the expansion or contraction of the Universe, is distinctly different in the synchronous and harmonic coordinate systems. This difference is responsible for the significant difference between the evolutionary scenarios of the Universe in the synchronous and harmonic coordinate systems.

In conclusion, we note some of the qualitative differences between our solutions of the Einstein equations with the cosmological constant $\lambda \neq 0$ in the synchronous coordinate system and the solutions of these equations without any cosmological constant.

(1) At $\lambda \neq 0$, there are solutions without singular points (solution (40) for $f > 1$, $\varepsilon = -1$ and $f < 0$, $\varepsilon = -1$).

(2) There are solutions with a horizontal asymptote for $\sqrt{-g}$; i.e., the Universe can expand only to a certain limit defined by the integration constants (solution (42)).

(3) There are solutions with an infinite number of singular points. In this case, the Universe passes through a closed cycle of evolution (solution (35)).

REFERENCES

1. C. J. Isham and J. E. Nelson, Phys. Rev. D **10**, 3226 (1974).

2. C. J. Radford and A. N. Klotz, J. Phys. A: Math. Gen. **16**, 317 (1983).
3. P. Wils, J. Math. Phys. **32**, 231 (1991).
4. Yu. P. Rybakov, B. Sakha, and G. N. Shikin, Izv. Vyssh. Uchebn. Zaved. Fiz., No. 7, 40 (1994).
5. V. G. Bagrov, A. D. Istomin, and V. V. Obukhov, Gravit. Cosmol. **2**, 117 (1996).
6. G. Platania and R. Rosania, Europhys. Lett. **37**, 585 (1997).
7. V. A. Zhelnorovich, *Theory of Spinors and Its Applications* (Avgust-Print, Moscow, 2001).
8. V. A. Zhelnorovich, Dokl. Akad. Nauk **346**, 21 (1996).
9. V. A. Zhelnorovich, Gravit. Cosmol. **2**, 109 (1996).
10. V. A. Zhelnorovich, Dokl. Akad. Nauk SSSR **296**, 571 (1987) [Sov. Phys. Dokl. **32**, 726 (1987)].
11. V. Fock, Z. Phys. **57**, 216 (1929).
12. V. A. Zhelnorovich, Dokl. Akad. Nauk SSSR **311**, 590 (1990) [Sov. Phys. Dokl. **35**, 245 (1990)].
13. A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series. Elementary Functions* (Nauka, Moscow, 1981).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 629–642.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 717–733.

GRAVITATION, ASTROPHYSICS

On the Presence of Fictitious Solar Neutrino Flux Variations in Radiochemical Experiments

B. M. Vladimirskiĭ* and A. V. Bruns

Crimean Astrophysical Observatory, pos. Nauchnyĭ, Crimea, 98409 Ukraine

**e-mail: bvlad@crao.crimea.ua*

Received April 4, 2003

Abstract—The currently available data on solar neutrino flux variation in radiochemical experiments and Cherenkov measurements have so far defied a simple interpretation. Some of the results concerning these variations are indicative of their relationship to processes on the solar surface. It may well be that a poorly understood, uncontrollable factor correlating with solar activity indices affects the neutrino flux measurements. This factor is assumed to modulate the detection efficiency on different detectors in different ways. To test this assumption, we have analyzed all available radiochemical measurements obtained with the Brookhaven (1970–1994, 108 runs), GALLEX (1991–1997, 65 runs), and SAGE (1989–2000, 80 runs) detectors for possible instability of the detection efficiency. We consider the heliophysical situation at the final stage of the run, the last 7–27 days, when the products of the neutrino reaction with the target material had already been accumulated. All of the main results obtained previously by other authors were found to be reproduced for chlorine–argon measurements. The neutrino flux anticorrelates with the sunspot numbers only for an odd solar cycle. A similar behavior is observed for the critical frequencies of the *E*-ionosphere. The neutrino flux probably correlates with the A_p magnetic activity index only for an even solar cycle. The predominance of a certain sign of the radial interplanetary magnetic field (IMF) in the last 14 (or 7) days of the run has the strongest effect on the recorded neutrino flux. The effect changes sign when the polarity of the general solar magnetic field is reversed and is most pronounced for the shortest runs (less than 50 days). The dependence of the flux on IMF polarity completely disappears if the corresponding index is taken for the first rather than the last days of the run. The IMF effect on the recorded neutrino flux was also found for short runs in the gallium–germanium experiment, but this effect for a given time interval in the SAGE measurements is opposite in sign to that detected with the Brookhaven and GALLEX detectors. The anticorrelation with the A_p activity index, which is absent in the SAGE measurements, contributes significantly to the flux variations on the GALLEX detector. If a magnetic storm with a sudden commencement occurs in the last 7 days of the run, then an effect of generally the same type, a significant increase in the variance, is observed for all three detectors. In all other indices, the flux variations on the Brookhaven and GALLEX detectors are the opposite of those on the SAGE detector. We have found that the GALLEX and SAGE measurements for the runs that ended simultaneously, to within about 10 days or less, anticorrelate, while the Brookhaven and GALLEX measurements correlate. We conclude that there are fictitious variations in the measurements under consideration that are attributable to the influence of geophysical factors (probably, very-low-frequency electromagnetic fields) controllable by solar activity on the physical–chemical kinetics of the target material. We discuss possible experiments to check whether the detected effects are real. If all of these effects are indeed real, then the neutrino flux was underestimated in the radiochemical measurements. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The question of possible solar neutrino variations in a chlorine–argon experiment that was raised by G.T. Zatsepin back in the 1960s and initiated by Bazilevskaya *et al.* [1] has already been discussed for almost two decades. Many authors have independently analyzed the accumulated experimental data by using various heliogeophysical indices. If we summarize the results of this work, then the following picture emerges (all of the necessary information about the experiment itself and a bibliography of previous publications are given in the monograph [2]).

1.1. The Chlorine–Argon Experiment

(1) There is a significant anticorrelation between the neutrino flux and the global solar activity index, the sunspot numbers *R*, only after 1977 [3, 4]. No convincing interpretation of the absence of such an anticorrelation for 1970–1976 has been offered.

(2) A similar anticorrelation is observed if the photospheric magnetic fields measured with a magnetograph are used as a heliophysical index. The anticorrelation proves to be statistically significant only for low-heliolatitude data (the disk center [5, 6]). It is also more pronounced for 1977–1990, with the spring

months of each year making an important contribution to such variations.

(3) Most of the authors agree that there is an annual period in the intensity variations [7–9]. This period is found when analyzing data irrespective of any heliophysical indices [10, 11] and when correlating the flux Q with the sunspot numbers R [3, 8] and the coronal index (the intensity of the $\lambda = 5303$ Å line [12]). All of the authors interpret the annual period as resulting from the Earth's spatial displacements relative to the helioequator.

(4) There is a weak positive correlation between the neutrino flux and the magnetic activity level (the A_p index [13, 14], the *aa* index [6]). This correlation is usually interpreted as resulting from variations in solar wind parameters. It seems to be more pronounced for the wind density estimated in direct measurements (1973–1993 [15]). It is not quite clear how this correlation matches with the anticorrelation between the neutrino flux and the sunspot numbers.

(5) There is a positive correlation between the neutrino flux and the Galactic cosmic-ray intensity measured with neutron monitors. Most of the authors agree that this correlation is not causal, but results from the anticorrelation between Q and R. The correlation between the neutrino flux and chromospheric flares has not been confirmed.

(6) A significant linear anticorrelation has been observed between the neutrino flux and the frequency shift of the free solar acoustic oscillations (P-modes, 1980–1991 [16]). It is not clear whether this anticorrelation is causal. Since the above shift anticorrelates with the solar radius, the neutrino flux may be assumed to positively correlate with the solar-radius variations.

(7) A set of periods is found in the neutrino flux variations. Various (more than ten) authors are unanimous that a quasi-biennial period (2.17 ± 0.03 years) is observed. As regards other periods, the results slightly disagree. The periods of 8–9 and 4.5–5 years are quoted most commonly (see, e.g., [17–20]). Clearly, the periodograms (power spectra) must exhibit a one-year period; it is actually present. At the same time, it also often exhibits a period of 1.3 years. A big surprise was the detection of one of the solar rotation modes (28.34 days) in the neutrino flux variations [21].

(8) There is compelling evidence suggesting that many authors have overestimated the statistical significance of their results concerning the neutrino flux variations and the corresponding correlations [22–24]. However, we cannot agree with the opinion that all of the above results are attributable to statistical fluctuations. Most of them are consistent with the well-known patterns of solar–terrestrial physics. For example, almost all of the variation periods (listed in (7)) are the well-known cosmophysical periods.

1.2. The Gallium–Germanium Experiments

For obvious reasons, the question of neutrino flux variations in the GALLEX and SAGE experiments is less clear. It seems that no correlation with the sunspot number is found in these measurements [25], although the annual period is probably present. Two specific results deserve rapt attention.

(1) The GALLEX measurements revealed a period close to one of the solar rotation modes (28 days). There may also be a period that coincides with the well-known harmonic of the solar free inertial oscillations (about 157 days) [25]. A period close to 28 days may be present in the Brookhaven experimental data [25].

(2) There is a remarkable feature in the set of data from both gallium–germanium experiments: the distribution of their results exhibits two peaks. The bimodal pattern of the distribution could be indicative of the existence of two discrete intensities [26]. The time scale of the transition between the states was estimated by the authors to lie within the range of 10–60 days.

The situation described above must be complemented by data on the high-energy neutrinos observed with Cherenkov detectors. According to KAMIOKANDE data [27], the neutrino flux shows no anticorrelation with the sunspot numbers for solar cycle 22 (a weak correlation is not ruled out). The absence of intensity variations that correlate with solar activity in these and subsequent measurements calls into question the reality of the neutrino flux variations in the radiochemical experiments.

2. THE EXPERIMENT AND THEORETICAL MODELS

The behavior patterns described above must be compared with the following three major types of models that describe the variations.

(1) The neutrino source in the solar core is stable; variations arise when the neutrino flux propagates through the solar matter due to the transitions between various neutrino states, in particular, due to its interaction with solar subphotospheric magnetic field, as long as the neutrino is postulated to have a large magnetic moment (for the physics of these phenomena, see [28]).

(2) The neutrino source in the core pulsates due to the excitation of special oscillations; these oscillations are synchronized with the solar activity variations [29, 30].

(3) The neutrino flux is stable both in the solar core and near the Earth; variations arise from unnoticed variations in the detection efficiency in the instrument itself; i.e., they are fictitious. These variations are conjugate with the solar activity variations, because the solar activity itself affects the measuring technique through geophysical fields, acting as an uncontrollable factor [31].

Many publications are devoted to the first type of models. A familiarity with them shows that a large

number of various fitting assumptions are made when comparing experimental data with specific computational models. Therefore, the popular belief that the problem of neutrino flux variations has been solved almost completely is a gross exaggeration.

In contrast, the idea of solar core oscillations receives little support. These oscillations, if they actually exist, must seemingly be synchronized with the natural timer—the oscillations attributable to a motion of the Sun with respect to the system's barycenter. The periods found for the parameters of this motion (in years)—2.41, 3.51, 4.26, … [32]—are in poor agreement with the periods found in the chlorine–argon experiment (see Section 1.7).

To most researchers, the third type of models seems to be completely implausible. In this case, the main idea is that the reaction products (radioactive Ar^{37} and Ge^{71}) in the target liquid may prove to be bonded somehow with a certain molecular structure and do not always fall into the counting system. Of course, such bonding is possible in principle at a certain stage of the complex technique of extracting a small number of product atoms from a large number of target atoms. This possibility seems to be ruled out, because the extraction efficiency is controlled in real time by adding a known number of atoms of the corresponding stable isotope to the target material. However, this procedure is not quite correct: neutral Ar^{36} atoms are added for control, while the neutrino reaction product appears in the form of a fast ion. It may well be that the bonding into a structure will be different in these cases. This well-known problem of a hot ion can be solved by adding an appropriate radioactive isotope to the target material, where the product being extracted also appears in hot form. In this case, however, the uncertainty is not completely eliminated either: this experiment only simulates the real situation, without replicating it in all details. In addition, these changes are isolated episodes in nature. Whether the extraction efficiency changes from session to session in normal working conditions is still an open question.

The bonding of Ar^{+37} or Ge^{+71} ions can include various reactions. Until now, only one type of this process has been discussed: falling into a molecular cavity trap. In particular, three possible "containers" for the neutrino reaction product have been proposed for Ar^{+37}: Jacobs [33] assumed that the globules produced by perchloroethylene polymerization could play an important role in the possible argon bonding. Subsequently, argon during the cooling may prove to be bonded in the structure formed in a microbubble (babstone) [34]. If there is water in the form of an admixture in perchloroethylene, then the formation of a gas hydrate of the second type—an $A \cdot 2B \cdot 17H_2O$ structure, where A and B are the cavities populated by various molecules, radicals, and ions—is probable. Cavity B can be a cavitand for Ar^{37} [31].

In the model being described, the correlation with solar activity arises, because the state of the target liquid depends on the parameters of the background geophysical fields, in particular, very-low-frequency electromagnetic fields controllable by solar activity. When Ar^{37} is trapped into a clathrate structure, for example, the number density of these structures in the target depends on the liquid run in the fields mentioned above.

In the past, this kind of reasoning could not be convincingly justified and was perceived as speculation. In the last decade, the situation has changed significantly. The following arguments now force us to treat the idea of modulation of the detection efficiency by solar activity in earnest.

First, the pattern of the dependence of neutrino flux variations on heliogeophysical indices closely follows the patterns attributable to the surface effects of solar activity that are known from solar–terrestrial physics. Significantly, the 28-day period with its ghosts found by Sturrock *et al.* [21] is precisely the period that was observed in the sunspot numbers at the epoch under study [25]. Further, the heliophysical indices taken for the central zone of the solar disk are known to correlate better with the parameters of purely terrestrial processes than the same disk-averaged indices do. Characteristically, the neutrino flux shows a correlation both with the global solar indices (pertaining to the entire disk) and with the solar-wind (magnetic activity) parameters that reflect the processes in a narrow zonal region for a given heliolatitude. In solar–terrestrial physics, this applies to two different connection channels through the short-wavelength radiation (ionosphere) and the solar wind (magnetosphere).

Second, direct evidence for the instability of the detection efficiency in the chlorine–argon experiment has been obtained. For example, for some reason, there is a strange anticorrelation between the flux Q and the run time [4, 24]. If this anticorrelation is real, then it could correspond to continuous bonding of the reaction product, its fictitious destruction, in certain time intervals. The anticorrelation between the flux Q and the background correction [24] appears no less strange.

Third, a number of examples where the effects of solar activity on the processes in condensed phases were found with confidence have been accumulated to date. The directly acting agent—the very-low-frequency electromagnetic background disturbances—acts in all these cases as an uncontrollable factor in the given experiment. The widely known example is a regular increase in the measured biochemical reaction rates during each solar minimum over three 11-year cycles [35]. The above studies (of the so-called macroscopic fluctuations) based on a comparison of histograms have revealed characteristic periods of about 27 days and about a year for normal measurements of the count rate of radioactive standards [36]. A behavior of the same type in an 11-year cycle was also discovered when analyzing the gravitational constant mea-

Table 1

Revealed interval	Runs
Even cycle 20, 1970–1976	18–46
Odd cycle 21, 1977–1986	47–91
Even cycle 22, 1987–1994	92–133
Plus at solar north pole	18–64; 109–133
Minus at solar north pole	65–108

surements using a torsion pendulum [37]. In this case, the point of action of the uncontrollable factor is the torsion pendulum suspension filament. This conclusion was reached after studying the dependence of the measurements with this detector on the index used below—the sign of the radial interplanetary magnetic field (IMF) [38]. Semiannual variations of seemingly the same nature [39] (for an overview of the solar activity effects on the "technosphere," see the monograph [40, Chapter 7]) are now known even in neutrino mass measurements. In general, as long as the solar activity effects, which apparently modulate the detection efficiency in many cases, do not seem absurd, a special search for similar effects in radiochemical experiments is quite justifiable. Although the targets in these experiments are placed at relatively large depths, they are nevertheless within the reach of electromagnetic disturbances. For typical electrical conductivities of rocks, the skin depth for a frequency of 8 Hz is several kilometers. In addition, there is a significant component of lithospheric origin in the variations of these geophysical fields. This component of the electromagnetic background also varies synchronously with solar activity.

If the neutrino flux variations in the radiochemical experiments are attributable not to variations in the number of reaction product atoms, but to variations in the number of these atoms extracted from the target material, then the heliogeophysical situation at the end of the run is of great importance in analyzing the results. In the following sections, we present some of the results of our analysis of data from the chlorine–argon experiment and the GALLEX and SAGE experiments in connection with the solar activity and magnetic disturbance variations when the corresponding indices were calculated for the final stage of the run (less than 25% of the session duration).

3. CORRELATION OF THE NEUTRINO FLUX IN THE RADIOCHEMICAL EXPERIMENTS WITH HELIOGEOPHYSICAL INDICES AT THE END OF THE RUN

3.1. The Chlorine–Argon Detector

The measurements processed by the maximum-likelihood method [41] (a total of 108 runs, 1970–1994) were used as the source data. We took unweighted (best-fit) values. The heliogeophysical indices—the sunspot numbers R and the A_p magnetic activity indices—were taken from the Solar–Geophysical Data. The index I of the mid-latitude E-ionosphere was tabulated in [42]. The IMF signs restored from geophysical data were taken from the catalog of laboratory polar measurements (Institute of Terrestrial Magnetism, Ionosphere and Radio-Wave Propagation, IZMIRAN) [43]. It is clear from general considerations that an optimal time interval for which the effect of heliogeophysical variations on the target material is most pronounced must exist. Since this interval is not known, we used three different intervals in many cases: 27, 14, and 7 days. They were counted backward from the completion date, which was the main reference date of our analysis. As is customary in solar–terrestrial physics, the even–odd solar cycles are considered separately. It is of no less importance to distinguish the epochs at which the general solar magnetic field changes sign. Table 1 presents the distribution of runs (their numbers) for the selected time intervals. The sign of the field in polar solar regions is known to be established relatively slowly. Therefore, referring a certain run to the interval of a particular polarity at such transitional epochs is controversial.

The diagrams presented in the figures were constructed by using the same standard method: the heliophysical index obtained for a given run (its final part) allowed it to be placed in a certain bin of this index. The fluxes Q (in atoms day^{-1}) were averaged within this bin. The indices R (sunspot numbers) and A_p (magnetic activity) were used for comparison with the corresponding data by other authors. The ionospheric and IMF data are physically more meaningful.

The neutrino flux $\bar{Q}$ is plotted against the sunspot numbers in the last 27 days of the run in Fig. 1 (Q was averaged for each R bin). The distinct anticorrelation for the odd solar cycle corresponds to the anticorrelation found for approximately the same time interval by other authors who used the mean R for the entire run (see, e.g., [4]). For the even cycles, we clearly see a weak opposite tendency. The ionospheric index I exhibits a similar pattern (Fig. 2) in the last 27 days of the run; I has the meaning of a corrected daily mean critical frequency. The similarly constructed data for the A_p index are shown in Fig. 3. The index was averaged over the last 7 days of the run. In this case, we see a tendency for $\bar{Q}$ to increase with A_p for the even cycle (this tendency for the entire data set was also found by other authors who used the mean indices for the entire run time).

Finally, Figs. 4–6 show plots of $\bar{Q}$ against IMF sign in the last 14 days of the run (the sum of the numbers of days of positive, negative, and mixed polarities divided by 14). We clearly see a tendency for $\bar{Q}$ to decrease

with increasing dominance of the positive polarity in Fig. 4a and the change in sign of the effect when the polarity of the general solar magnetic field is reversed. In general, this pattern does not change if the IMF signs are taken in the last 7 and 27 days of the run, but for 14 days it is move distinct. Figure 5 shows data convolution for the entire data set (Fig. 4b was inverted relative to the zero line). According to the Mann–Whitney test, the general means for different polarities differ at the 10^{-2} significance level. We can see from Fig. 6 that the effect is systematic in nature; in the means, it manifests itself in a redistribution of extreme values: 80% of the values close to zero belong to one IMF sign, and all of $\overline{Q} > 1.0$ belong to the other IMF sign.

It is easy to verify by direct comparison that the dependence of $\overline{Q}$ on the IMF sign contributes appreciably to the annual variation. According to the standard Coleman–Rosenberg law, the phase of this variation is reversed when the polarity of the general solar magnetic field is reversed.

It is important to emphasize that the dependence $\overline{Q}(R)$, which is distinct for the odd solar cycle (see Fig. 1), holds irrespective of the IMF sign. If we construct such a dependence separately for all cases of different IMF polarities, then it will remain valid. Thus, the dependences $\overline{Q}(R)$ and $\overline{Q}$(IMF) cannot be reduced to each other and represent different effects. It is particularly important to take this into account when analyzing the dependence of $\overline{Q}$ on the run time found in [4]. It is shown in the form adopted here in Fig. 7a and holds irrespective of the solar activity level (R) and the IMF sign. There is no such general anticorrelation for the even cycle. However, it has emerged that this kind of relationship exists for different IMF polarities (Fig. 7b). As we see, the tendencies are opposite for different IMF signs. Note that the dependences in Fig. 7b cannot be explained by the Ar^{37} bonding with a constant rate (the aging of the solution), as should seemingly be done for the case of Fig. 7a. The curves in Fig. 7b were constructed in such a way that, most likely, there is no bonding for one of the IMF signs at short runs. The effect in a weakened form for longer runs proves noticeable because of the inertia of the index used. If the IMF sign is taken for the final 7 days of the run (rather than 14 days, as in Fig. 7b), then this difference in fluxes $\overline{Q}$ disappears abruptly at a run time of L = 60 days. In general, the IMF sign index differs only slightly from one 27-day interval to another. Therefore, when this interval is used, the effect for $L < 80$ days disappears completely. However, when the IMF signs are taken for 7 and 14 days at $L \le 60$ days, the changes in the means are the same: $\overline{Q}(-)/\overline{Q}(+) \approx 3$ (in all cases, when the sign of the general solar magnetic field changed, the IMF polarities also changed in accordance

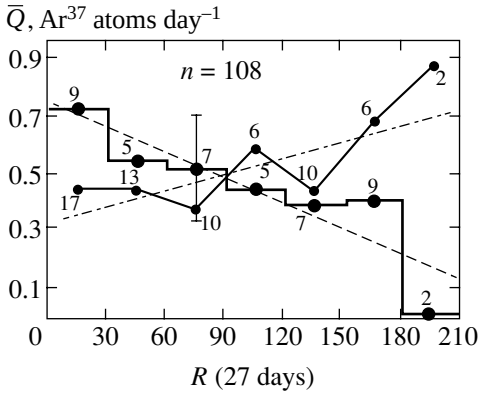


Fig. 1. Mean neutrino flux versus sunspot numbers R averaged over the last 27 days of the run for the even (thin solid and dash-dotted lines) and odd (heavy solid and dashed lines) solar cycles. $\overline{Q}$ was calculated for all of the runs that fell within a given R bin. The numbers near the dots give the number of such runs. The error is a typical rms scatter; the straight lines were drawn by least squares; n is the number of runs.

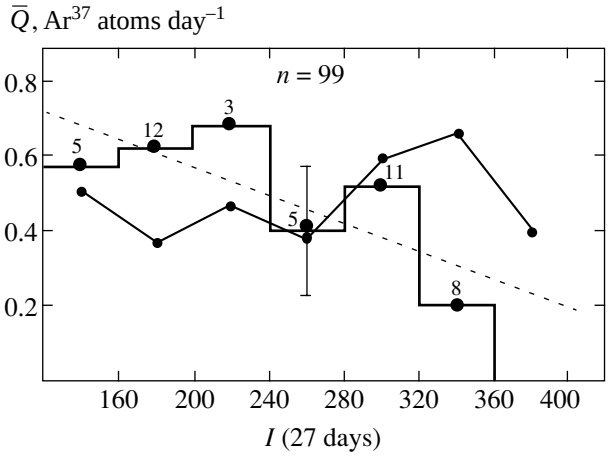


Fig. 2. Same as Fig. 1 for the index I of the mid-latitude E-ionosphere at the Moscow station. The even solar cycle is represented by the thin solid line; the odd solar cycle is represented by the heavy solid and dashed lines.

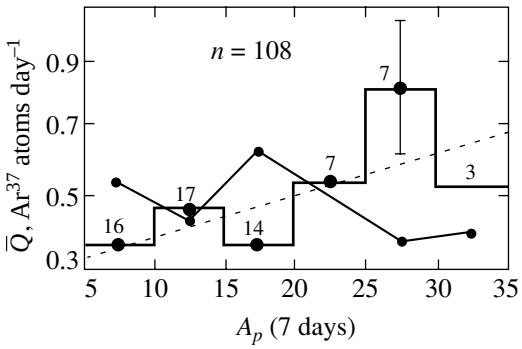


Fig. 3. Same as Fig. 1 for the A_p magnetic activity index in the last 7 days of the run. The even solar cycle is represented by the heavy solid and dashed lines; the odd solar cycle is represented by the thin solid line.

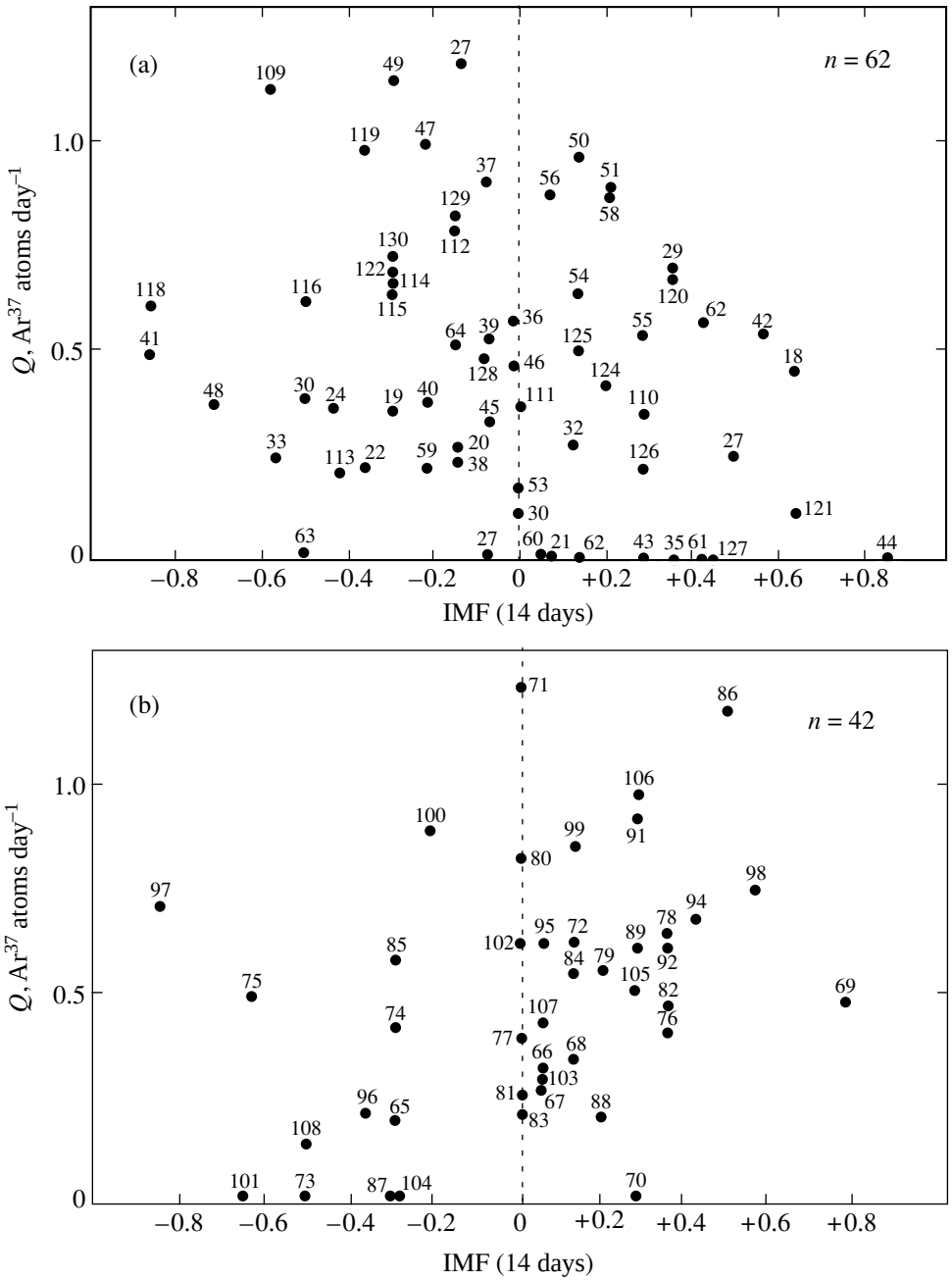


Fig. 4. "Neutrino flux *Q*—mean IMF polarity sign in the last 14 days of the run" scatter diagram: (a) the general solar magnetic field at the north pole is positive; (b) the general solar magnetic field at the north pole is negative. The numbers near the dots give the number of runs.

with the pattern shown in Fig. 4). The $\overline{Q}$ difference is statistically significant at the 1.7×10^{-3} level for an interval of 30–50 days and at the 4.8×10^{-2} level for an interval of 50–60 days.

The effect of the IMF sign in the last days of the run on the recorded neutrino flux $\overline{Q}$, which is important for the subsequent analysis, can be tested independently. For the set of runs used in the experiment under consideration (a median value of about 70 days), the IMF indices for the final and initial run intervals may be shown to be uncorrelated. Therefore, we once again constructed a convolution of the dependence of *Q* on the IMF sign, but for the initial 14 days of the run. We found no correlation between *Q* and the IMF sign: the corresponding values were $\overline{Q}(-) = 0.476 \pm 0.041$ and $\overline{Q}(+) = 0.479 \pm 0.048$ (see Fig. 5, where the corresponding means differ by a factor of 1.6; the standard deviations pertain to the scatter of points with different signs).

3.2 The Gallium–Germanium Detectors—the GALLEX and SAGE Experiments

The well-known published data (GALLEX Coll., 1993, 1994, 1996, and 1999 [44–48], 65 runs in 1991–1997) served as the source data for our analysis of the GALLEX experiment. For a comparison with magnetic activity, we also used 19 runs of the continuation of these measurements (GNO, 1998–1999). The SAGE measurements were received from the authors of the experiment (1989–2000, 80 runs). The method of analysis was similar to that used above. All of the cosmophysical data were counted for the last 7 days of the run. Since the IMF polarity estimate for a 7-day interval may contain a significant error due to the erroneous determination of the field sign for a single day in the IZMIRAN catalog, the data were checked by using independent direct measurements (the OMNI database). The gaps in these data were filled by interpolation. The IMF sign was assumed to have been determined reliably only when it was the same in both types of data. For the intervals of the GALLEX and SAGE measurements, the sign proved to be uncertain for 25% and 10% of the runs, respectively (a trivial factor, geomagnetic disturbances, is responsible for the discrepancy between the data in the catalogs mentioned above in half of the cases). Although, at first glance, the run statistics for the gallium–germanium experiments is comparable to that for the chlorine–argon measurements, analysis of the GALLEX and SAGE data runs into additional difficulties. Since the transition from (even) solar cycle 22 to odd solar cycle 23 occurred in the spring of 1996, these measurements do not completely cover a cycle of this type. The possible variations through the two channels of solar–terrestrial relationships mentioned above are largely independent. Therefore, for a correlation between the neutrino flux and a given cosmophysical index to be found, we must consider only those data that were obtained at constant values of another index. This approach cannot be consistently implemented for the accumulated body of data.

It is all the more surprising that the very important result on the intensity variations of the chlorine–argon experiment (see Fig. 7b) is also reproduced for the gallium–germanium measurements. This result is shown in the same notion in Fig. 8. The corresponding data are presented in Table 2. As we see, the GALLEX data for the measurement interval under consideration behave in exactly the same way as the Brookhaven data: the intensity is higher for the negative IMF polarities for short runs and is close to the mean for longer runs (Fig. 8a). For the SAGE data, the effect is less distinct and is opposite in sign (Fig. 8b).

On average, the intensity is generally slightly higher for short runs for the two experiments under consideration (as in the chlorine–argon experiments, about

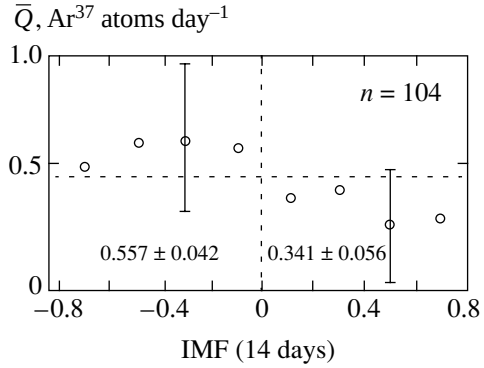


Fig. 5. Convolution of the data shown in Figs. 4a and 4b after the inversion of the data in Fig. 4b relative to the zero IMF line. The difference between the means, 0.557 ± 0.042 and 0.341 ± 0.056, for the negative and positive IMF signs, respectively, is statistically significant at the 10^{-2} level.

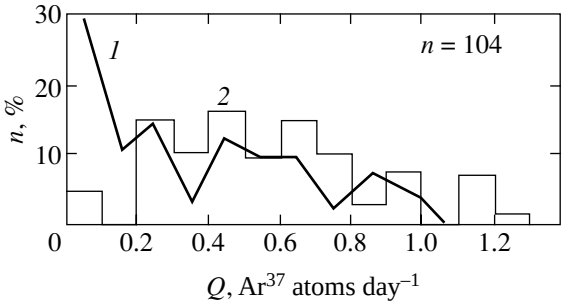


Fig. 6. The distribution (frequency of occurrence) of measured $\bar{Q}$ for the positive (*1*) and negative (*2*) IMF polarities.

Table 2

Experiment	Run	"+" IMF	*n*	"–" IMF	*n*	*P*(*U**)
GALLEX	about 21 days	42.4 ± 69.5	7	94.4 ± 45.7	10	0.13
	about 28 days	73.3 ± 39.9	18	75.6 ± 43.4	10	–
SAGE	35 days	84.6 ± 60.9	20	63.1 ± 49.9	15	0.24
	about 45 days	64.4 ± 35.4	10	70.9 ± 50.3	14	–

Note: *n* is the number of runs, and *P*(*U**) is the significance level.

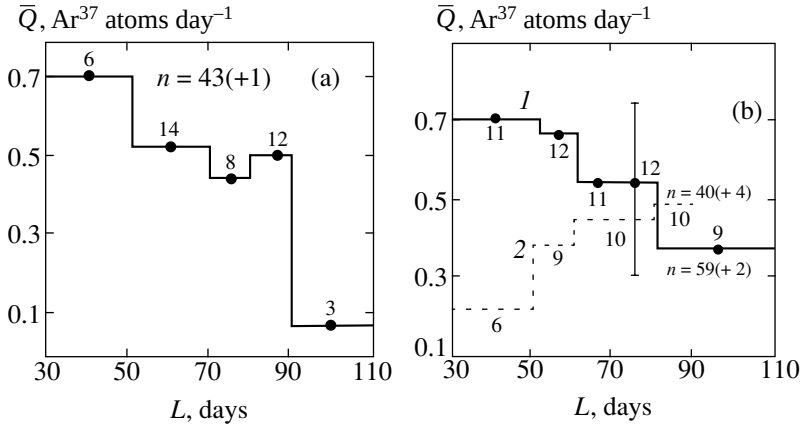


Fig. 7. (a) The relationship between mean neutrino flux $\bar{Q}$ and run time L found in [4]. The data pertain to the odd solar cycle 21. (b) The neutrino flux $\bar{Q}$ for the given run time L for the positive (curve *1*) and negative (curve *2*) IMF signs in the last 14 days of the run. The numbers near the dots give the number of runs in the given interval L. A typical rms scatter is shown for an interval of 70–80 days. Six points from the entire data set were excluded from our analysis (the IMF polarity is equal to zero, $L > 110$). According to the Mann–Whitney test, the difference between the mean $\bar{Q}$ for an interval of 30–50 days is statistically significant at the 1.7×10^{-3} level.

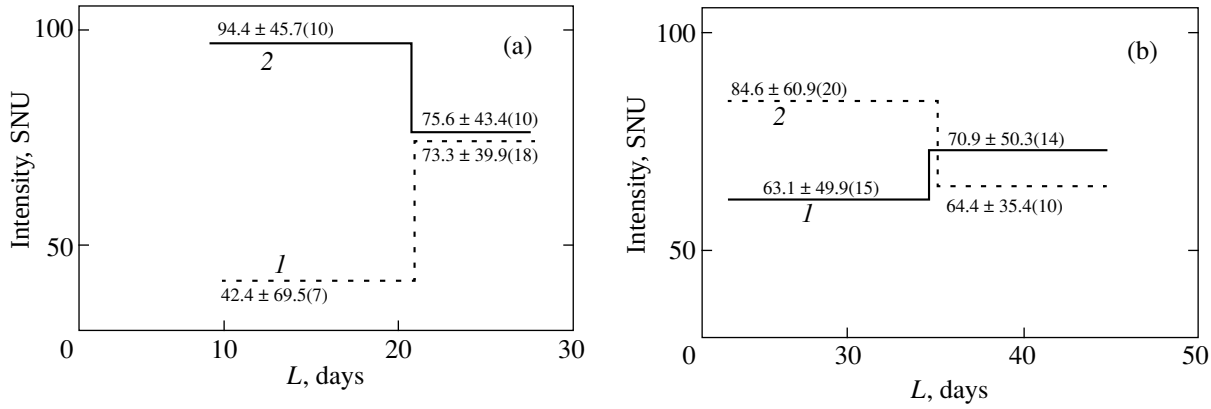


Fig. 8. The neutrino flux (SNU) for the run time under consideration for the positive (curve *1*) and negative (curve *2*) IMF signs in the last 7 days of the run. The numbers are the corresponding means with their standard deviations; the number of runs is given in parentheses. The same as Fig. 7b for GALLEX (a) and SAGE (b).

20%). More importantly, a significant (by a factor of 1.7) increase in the variance is characteristic of short runs. This increase may be considered as evidence for a higher sensitivity of the measurements to the influence of the uncontrollable factor precisely for short runs. Figure 9 shows an "$\bar{A}_p$ (7 days) – intensity" scatter diagram for the GALLEX measurements. There is a weak tendency for an anticorrelation (-0.22 ± 0.09), which is also observed separately for the 19 GNO runs. It is more distinct for short runs and is definitely absent for the SAGE data. This can be seen from Table 3, which gives the means with their rms deviations when the entire data set is broken down into groups for quiet and disturbed conditions. The conditional boundary corresponds to $\bar{A}_p$ (7 days) = 12.5.

The values of $A_p > 25$ correspond with a high probability to the falling of a magnetic storm with a sudden commencement within the seven-day interval in question. This is a special type of global electromagnetic disturbance. Such events should be considered separately. The corresponding counts are summarized in Table 4. (No magnetic storms appear in the right-hand part of Table 3 and Fig. 9).

To facilitate our comparison of the various experiments, the last column in Table 4 gives the data normalized to the corresponding means. In all cases, an effect of the same type is observed (in all cases, the mean $\bar{A}_p$ (7 days) > 30). If the revealed tendencies are real, then it makes sense to return to Table 2 to check the sensitivity of the measurements to the change of IMF sign for the most favorable conditions—short runs and a fixed

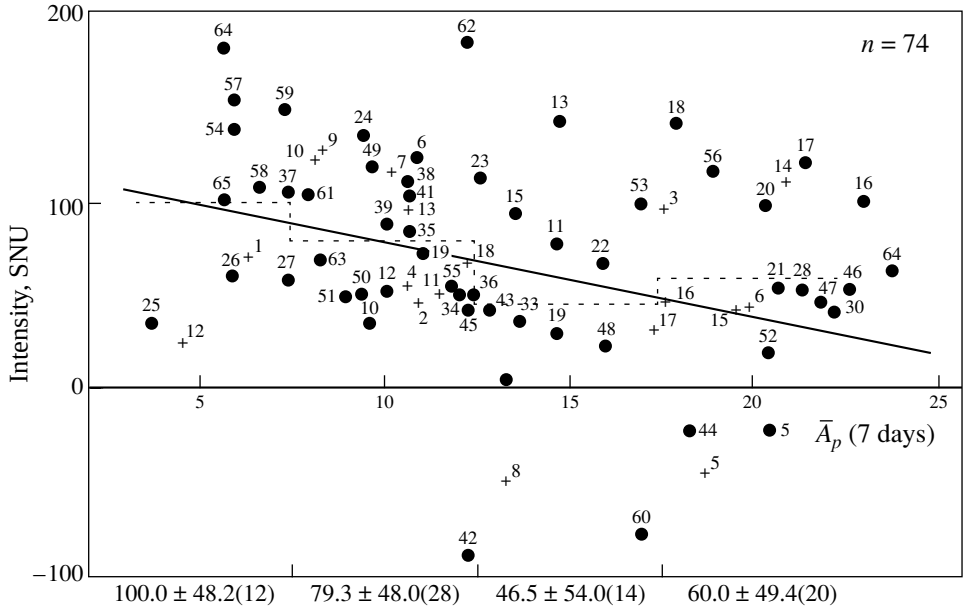


Fig. 9. "Intensity (SNU)—A_p index in the last 7 days of the run" scatter diagram for the GALLEX (circles) and GNO (crosses) measurements. The numbers near the dots give the number of runs. The means shown at the bottom of the figure correspond to the dashed histogram; the errors are standard deviations; the number of runs in the interval under consideration is given in parentheses. The correlation coefficient is $r = 0.22 \pm 0.09$.

level of geomagnetic activity. Of course, the number of corresponding runs decreases. The result can be seen from Table 5.

Note that the variations in the experiments being analyzed show opposite tendencies: in the GALLEX and SAGE measurements, the intensity is enhanced when, respectively, the negative and positive IMF polarities dominate at the end of the run. As the geomagnetic disturbance level increases, the intensity on GALLEX decreases, while the intensity on SAGE increases (the last column in Table 5—the significance of the differences between the means for different signs in quiet conditions). Finally, having selected the runs that ended when the conditions were geomagnetically quiet ($A_p < 12.5$), we may attempt to find the possible influence of sunspot number variations on the results of the measurements. The picture obtained is clear from Table 6. The conditional boundary between the high and low levels of solar activity was adopted for $R = 60.0$. As follows from the data obtained in Section 3.1 for the chlorine–argon measurements, the even and odd solar cycles should be considered separately.

As we see from an examination of Table 6, the differences are marginally significant, but we obtain a coherent picture: the GALLEX–GNO data reveal a correlation for the even solar cycle 22 and an anticorrela-

Table 3

Experiment	Run	All data	Conditions		$P(U^*)$
			quiet, $A_p < 12.5$	disturbed, $A_p > 12.5$	
GALLEX	Less than 21 days	82.5 ± 75.5	86.1 ± 62.0	41.4 ± 75.6	0.18
		$n = 28$	$n = 13$	$n = 7$	
	About 28 days	74.6 ± 44.9	89.1 ± 46.7	67.1 ± 38.6	0.20
		$n = 37$	$n = 16$	$n = 19$	
SAGE	35 days	86.1 ± 75.4	73.1 ± 55.8	85.3 ± 74.6	–
		$n = 51$	$n = 30$	$n = 15$	
	About 45 days	70.4 ± 43.3	72.4 ± 43.4	62.1 ± 40.2	–
		$n = 29$	$n = 17$	$n = 11$	

Table 4

Experiment	*n*	Intensity, SNU	Normalized values
Brookhaven	18	3.09 ± 1.49	1.19 ± 0.58
GALLEX	10	95.9 ± 79.1	1.30 ± 1.08
SAGE	7	136.6 ± 115.7	1.77 ± 1.50

tion for the odd solar cycle 23. This result is in close agreement with the results obtained above for the Brookhaven data. The variations for the SAGE measurements are opposite: an anticorrelation for the even cycle and a correlation for the odd cycle. It follows from these data that the GALLEX and SAGE results show an opposite behavior for all of the cosmophysical indices under consideration: for the IMF sign (Table 2), the A_p magnetic activity indices A_p (Table 5), and the sunspot numbers (Table 6). If the above tendencies are real, then the data obtained with these detectors must generally anticorrelate, while the Brookhaven and GALLEX data must correlate.

3.3. Comparison of the Data from Different Detectors

To test the above prediction, we must choose the runs from the corresponding sets of measurements that would end simultaneously, with a small mismatch. As this mismatch, we chose $|\Delta| \le 5$ days for the gallium–germanium measurements and $|\bar{\Delta}| \le 10$ days for the Brookhaven and GALLEX detectors. For the 12 synchronous runs, the actual value was $|\Delta| = 3.2 \pm 1.6$ days in the former case and $|\bar{\Delta}| = 4.5 \pm 3.1$ days in the latter case. Next, we may choose lower (higher) intensities from the catalogs compiled in this way for the entire data set and calculate the mean of the synchronous runs for another detector. The inverse procedure can serve as a check. The data normalized to the corresponding means are presented in Tables 7 and 8.

The large variance at the first point in the right-hand column in Table 7 is attributable to the presence of two extremely large values. If they are disregarded, then the inequality remains valid. Another check of the GALLEX and SAGE data for anticorrelation is to search for synchronous runs for extreme intensities on a particular detector. The corresponding examination shows that through the extremely large values obtained on SAGE (more than 150 SNU), we can find seven "twins" in the GALLEX–GNO data if the requirement for synchronism is relaxed to $|\Delta| \le 10$ days. The result of our comparison by the epoch-folding technique is shown in Fig. 10. As we see, extremely low values for GALLEX–GNO correspond to anomalously high values for SAGE. The mean mismatch between the runs for this sample is $|\Delta| = 6.7 \pm 2.7$ days, $P(U^*) < 10^{-3}$.

Unfortunately, the number of corresponding synchronous runs is too small for anomalous values of the opposite sign. The aforesaid suggests that the annual neutrino intensity variations must be similar in the GALLEX and Brookhaven measurements. In contrast, the annual variations in the GALLEX and SAGE measurements must be different. Indeed, if we construct the annual variations for the even solar cycle, then we will observe the characteristic minima for the GALLEX data in April and November and the maximum in September that have long been known for the chlorine–argon measurements. The SAGE data do not contain these features and even show a tendency for anticorrelation with the profile mentioned above (it is pertinent to recall that the annual variation in this case was constructed for the completion date of the run).

4. DISCUSSION

Thus, we may summarize the main results presented above as follows: the measured solar neutrino intensity in all three radiochemical experiments depends on the heliophysical indices pertaining to the last run interval. At the same time, all of the previously known main results, with the important refinement that an anticorrelation between the neutrino flux and the sunspot numbers is only observed for the odd solar cycle, while the IMF sign has the strongest effect on the recorded flux, are reproduced in the chlorine–argon measurements. This effect for shorter runs is also found for the gallium–germanium measurements. For all three detectors, we found a magnetic storm effect of the same type: a sharp increase in the mean scatter of results. Evidence suggests that the variations on the Brookhaven and GALLEX detectors are of the same type, while those

Table 5

Experiment	Quiet conditions, $A_p < 12.5$		Disturbances, $A_p > 12.5$		$P(U^*)$
	"+" IMF	"–" IMF	"+" IMF	"–" IMF	
GALLEX	76.3 ± 20.5	103.1 ± 40.0	17.0 ± 81.5	74.0 ± 51.3	0.19
	n = 3	*n* = 7	*n* = 4	*n* = 3	
SAGE	79.6 ± 61.9	48.8 ± 23.1	96.2 ± 56.8	84.7 ± 68.1	0.23
	n = 14	*n* = 9	*n* = 6	*n* = 6	

Table 6

Experiment	Cycle	Solar activity		$P(U^*)$
		low, $R < 60$	high, $R \geq 60$	
GALLEX	22, even	62.4 ± 50.1	82.6 ± 40.9	0.16
		$n = 14$	$n = 5$	
CNO	23, odd	125.8 ± 40.3	74.9 ± 28.7	0.16
		$n = 11$	$n = 10$	
SAGE	22, even	77.4 ± 45.8	60.8 ± 73.4	0.29
		$n = 13$	$n = 10$	
	23, odd	61.1 ± 36.6	84.6 ± 41.7	0.01
		$n = 12$	$n = 14$	

on the SAGE detector are opposite. For the runs that ended simultaneously, to within several days, the GALLEX and SAGE data probably anticorrelate, while the Brookhaven and GALLEX data correlate.

In many cases, the statistical significance of the results obtained above is low (10^{-2}), and, strictly speaking, they need to be tested. Such a test can be made only when additional data will be accumulated. In our case, the statistical reliability is determined by random fluctuations (not by systematic errors). Therefore, circumstantial evidence is also important for assessing whether the tendencies under consideration are real. In particular, it is important to note that the results obtained fit into a self-consistent picture. Note, in particular, that the correlations of the neutrino flux with the IMF sign, R, and A_p necessarily give rise to the already detected about 27-day periodicity. These correlations give a qualitative insight into the bimodal distribution of the results from [26]: the two stated found by the authors of [26] may correspond to the two IMF signs and the corresponding quasi-periodic A_p variations. Some of the periods found in the neutrino flux variations (see point 7 in the Introduction) are equal to the periods found in the IMF variations (1.3, about 3, and about 4.5 years). The annual period is of greatest interest. As was noted above, the profile of the annual variations for the even solar cycle for the Brookhaven and GALLEX detectors share common characteristic features. Interestingly, in the various precision measurements that also reveal the annual period of unknown origin (the frequency drift of the atomic standards, the search for dark matter, etc.), the extreme points are arranged in a similar fashion. Finally, note that an examination of the GALLEX calibration measurements using Cr^{51} (seven runs, June–July 1994 [48]) makes it possible to understand why this procedure does not completely solve the question of flux variations either: the runs mentioned above were taken for various values of the cosmophysical indices used here. What level of the reconstructed activity of the source should be taken—at a low (60.7 ± 3.9) or enhanced (72.2 ± 2.1) geomagnetic disturbance level?

Of course, by no means all of the empirical data presented above can be understood. Even in those cases where a simple explanation can be found for the effect in question, it is most often ambiguous or not comprehensive. We have failed to find any idea that would explain why the measured flux for all three experiments is, on average, higher for shorter runs. The even and odd solar cycles are generally known to differ in various properties, but it is not clear what these differences specifically in parameters of the acting electromagnetic fields are. The anticorrelation between the ionospheric index and the neutrino flux could in principle be an important argument for our hypothesis that very-low-frequency emissions affect the target material. However, the conclusion that the anticorrelation between the ionospheric index and the flux for the odd solar cycle (see Fig. 2) is causal cannot be considered to have been proved. This correlation may just arise from the correlation between I and R.

In contrast, the influence of the sign of the radial IMF on the results of the measurements can be interpreted unambiguously in our case. All of the parameters of geomagnetic micropulsations and low-frequency emissions in the IMF sectors of different signs have long been known to vary significantly. It has been firmly established that some physicochemical systems respond to the change in sign of the IMF sector (in particular, this has been established for the rest reaction that reveals an 11-year solar cycle; see [49]). Clearly, this dependence must change sign with polarity reversal of the general solar magnetic field (although the dynamics of such phenomena has been studied inadequately). It should be emphasized that the change in IMF sign and the accompanying change in the excitation regime of geomagnetic micropulsations and kilohertz magnetospheric emissions is a geophysical effect. Therefore, large variations in neutrino flux Q for differ-

Table 7

GALLEX–GNO, initial sample smaller than $M = 78.0$	SAGE, mean of synchronous runs	$P(U^*)$
$n = 6$; 0.43 ± 0.35	$n = 6$; 1.86 ± 1.79	5×10^{-2}
SAGE, initial sample smaller than $M = 77.0$	GALLEX–GNO, mean of synchronous runs	
$n = 7$; 0.56 ± 0.25	$n = 7$; 1.05 ± 0.41	3×10^{-2}

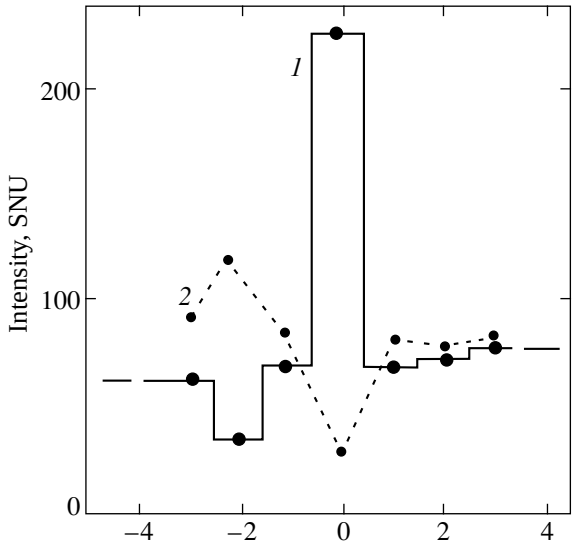


Fig 10. Intensity (SNU) versus run sample on SAGE (*1*) and GALLEX (*2*). The zero interval—synchronous runs (the mismatch between their ends is no more than 10 days). The initial SAGE sample—extremely large values (>150 SNU). The corresponding runs adjacent to the synchronous runs are to the left and to the right of zero (the minus and plus correspond to the preceding and succeeding runs, respectively).

ent IMF signs at the end of the run is a decisive argument that there are fictitious variations in the radiochemical experiments that were not caused by real variations.

The model of the variations under consideration assumes that the action on the physicochemical kinetics through both channels (through the short-wavelength solar radiation, the ionosphere, and through the solar wind, the magnetosphere) always only reduces the measured flux, differently at different epochs and on different detectors. Therefore, the measured fluxes must be lower than the actual fluxes. The actual recorded flux can be estimated by assuming that all of the correlations under consideration are real. For the chlorine–argon detector, these are the correlations (anticorrelations) with the sunspot numbers, the critical ionospheric frequencies, and the A_p index. By analyzing the distribution of these indices, we can obtain the following limiting values for a high (low) activity level:

Table 8

GALLEX–GNO, initial sample larger than $M = 78.0$	Brookhaven, mean of synchronous runs
$n = 7$; 1.76 ± 0.84	1.27 ± 0.46
Brookhaven, initial sample larger than $M = 0.478$	GALLEX, mean of synchronous runs
$n = 9$; 1.40 ± 0.29	1.37 ± 0.96

$R > (<) 90$; $I > (<) 220$; and $A_p > (<) 12.5$. Choosing the intervals of flux-stimulating IMF polarities and requiring the satisfaction of the inequalities written above, we obtain

for the odd cycle,

$$\bar{Q}_{\max} = 0.726 \pm 0.338 \quad (n = 9),$$

$$\bar{Q}_{\min} = 0.431 \pm 0.313 \quad (n = 35);$$

for the even cycle,

$$\bar{Q}_{\max} = 0.690 \pm 0.226 \quad (n = 17),$$

$$\bar{Q}_{\min} = 0.400 \pm 0.267 \quad (n = 47);$$

and for the entire data set,

$$\bar{Q}_{\max} = 0.702 \pm 0.293 \quad (n = 26),$$

$$\bar{Q}_{\min} = 0.413 \pm 0.228 \quad (n = 82)$$

(all of the errors are standard deviations).

Finally, if we take into account the significant effect of the IMF sign (see Fig. 7b) and exclude the long ($\Delta >$ 70 days) runs from the list for $\bar{Q}_{\max}$, then we can obtain the extreme value, $\bar{Q}_{\max} = 0.768 \pm 0.289$ ($n = 17$; $4.5 \pm$ 1.5 in SNU). Clearly, if this flux is taken as the actual value, then the problem of its deficit becomes less acute. Unfortunately, such an estimate for the GALLEX and SAGE data cannot be obtained because of the poor statistics. However, recall the result of analyzing the data distribution (histograms) in the gallium–germanium measurements [26]. The two states found in the above paper may correspond to the two IMF signs, with the higher state approaching the theoretical flux.

Without touching upon the complicated question of the influence of weak electromagnetic fields on the physical–chemical kinetics of solutions (where the most important events develop at the mesolevel, e.g., the drawing of ions or atoms into nanotubes), we will briefly list the possible tests of the ideas presented above. Most of them are simple and quite realizable.

(1) It would be instructive to carry out a series of direct measurements of the variations in electromagnetic fields of low and very low frequencies near the detector, including the fundamental mode frequencies of the Schumann ionospheric resonance and **Pc**3-type geomagnetic pulsations.

(2) It seems of considerable interest to carry out a series of measurements on all of the radiochemical detectors (e.g, for a year) in such a way that the runs end simultaneously.

(3) It would be appropriate to measure the detection efficiency of a suitable radioactive standard near the detector by using standard measuring devices, preferably simultaneously through several channels (scintilla-

tors, Geiger counters, semiconductor detectors). It is quite probable that the count-rate measurements in this case could correlate (anticorrelate) with the recorded neutrino flux variations.

(4) It seems of considerable importance to automatically monitor the parameters of the target liquid (e.g., perchloroethylene) in stable laboratory conditions. The various parameters of the liquid that could be systematically measured, such as the electrical conductivity, the dielectric loss tangent, the heat conductivity, etc., are related. Therefore, if the liquid–liquid phase transition affecting the Ar^{37} (Ge^{71}) extraction efficiency occurs in the target material under external conditions, it can be detected. Such a monitoring is of interest for many reasons, and it would be appropriate to perform it simultaneously at several points on a cooperative basis [Crimean Astrophysical Observatory, Institute of Biophysics of the Russian Academy of Sciences (Pushchino), and Institute of Chemical Physics of the Russian Academy of Sciences (Moscow)].

From the viewpoint of the hypothetical ideas presented here, weak neutrino flux variations could also be detected with water Cherenkov detectors. The effect of solar activity—geomagnetic disturbance—on physicochemical phenomena is universal in nature [40]. On the detectors mentioned above, the effect on the neutrino count rate is possible through variations in the refractive index of water (the intensity of the Cherenkov light is a quadratic function of this parameter) and variations in the quantum efficiency of the photomultiplier cathodes [50]. We think that about one-week and about 27-day periods could be easiest to find at the epochs of the Earth's high heliolatitudes. These variations could be detected and studied in detail in laboratory conditions by measuring the beta activity of a suitable radioactive standard with Cherenkov water detectors.

5. CONCLUSIONS

Our analysis has led us to the following conclusions.

(1) The neutrino flux in all of the radiochemical experiments depends on the heliogeophysical situation at the very end of the run, when the neutrino reaction product has already been accumulated.

(2) This dependence holds for such ionospheric and magnetospheric indices as the critical frequencies of the ionosphere and the sign of the radial interplanetary magnetic field, i.e., geophysical indices.

(3) The correlation of the neutrino flux with cosmophysical indices is mainly or entirely attributable to variations in the detection efficiency. The latter probably arise from the effects of the very-low-frequency background electromagnetic fields controlled by solar activity on the target material.

(4) The proposed model of the variations can only reduce the recorded neutrino flux. Therefore, the mean neutrino flux in the radiochemical measurements has probably been underestimated.

ACKNOWLEDGMENTS

We are grateful to G.S. Ivanov-Kholodnyĭ for providing unpublished ionospheric data and helpful remarks; to V.I. Odintsov for providing IMF data; and to S.E. Shnol and A.A. Konradov for helpful discussion.

REFERENCES

1. G. A. Bazilevskaya, Yu. I. Stozhkov, and T. N. Charakhch'yan, Pis'ma Zh. Éksp. Teor. Fiz. **35**, 273 (1982) [JETP Lett. **35**, 341 (1982)].
2. J. N. Bahcall, *Neutrino Astrophysics* (Cambridge Univ. Press, Cambridge, 1989; Mir, Moscow, 1993).
3. J. W. Bieber, D. Sechel, T. Stanev, and G. Steigman, Nature **348**, 407 (1990).
4. J. N. Bahcall and W. H. Press, Astrophys. J. **370**, 730 (1991).
5. D. S. Oakly, H. B. Snodgrass, R. K. Ulrich, and T. L. Van De Kop, Astrophys. J. **437**, L63 (1994).
6. V. N. Obridko and Yu. R. Rivin, Astron. Astrophys. **308**, 951 (1996).
7. Yu. R. Rivin, *Circles of the Earth and Sun* (Nauka, Moscow, 1989), p. 48.
8. L. I. Dorman, V. L. Dorman, and A. W. Wolffendale, in *Contributed Papers of 23rd ICRC* (Calgary, 1993), Vol. 3, p. 872.
9. Yu. R. Rivin, Astron. Zh. **70**, 392 (1993) [Astron. Rep. **37**, 202 (1993)].
10. P. A. Sturrock, G. Walther, and M. S. Wheatland, Astrophys. J. **507**, 978 (1998).
11. L. I. Dorman, Yad. Fiz. **63**, 1064 (2000) [Phys. At. Nucl. **63**, 984 (2000)].
12. S. Massetti and M. Storini, Astrophys. J. **472**, 827 (1996).
13. D. Basu, Sol. Phys. **81**, 363 (1982).
14. R. Wilson, Sol. Phys. **149**, 391 (1994).
15. R. L. McNutt, Science **270**, 1635 (1995).
16. Ph. Delache, V. Gavryusev, E. Gavryuseva, *et al.*, Astrophys. J. **407**, 801 (1993).
17. H. J. Haubold and E. Gerh, Astron. Nachr. **306**, 203 (1985).
18. H. J. Haubold and E. Gerh, Sol. Phys. **127**, 347 (1990).
19. V. Gavryusev, E. Gavryuseva, and A. Roslyakov, Sol. Phys. **133**, 161 (1991).
20. I. Liritzis, Sol. Phys. **161**, 29 (1995).
21. P. A. Sturrock, G. Watlher, and M. S. Wheatland, Astrophys. J. **491**, 409 (1997).
22. G. Walther, Astrophys. J. **513**, 990 (1999).
23. J. Boger, R. L. Hahu, and J. B. Cumming, Astrophys. J. **537**, 1080 (2000).
24. R. B. Wilson, Astrophys. J. **545**, 532 (2000).
25. P. A. Sturrock, J. D. Scargle, G. Walther, and M. S. Whieatlend, Astrophys. J. **523**, L177 (1999).
26. P. A. Sturrok and J. D. Scargle, Astrophys. J. **550**, L101 (2001).
27. Y. Fukuda, T. Hayakawa, K. Inoe, *et al.*, Phys. Rev. Lett. **77**, 1683 (1996).

28. Yu. V. Kozlov, V. P. Martem'yanov, and K. N. Mukhin, Usp. Fiz. Nauk **167**, 849 (1997) [Phys. Usp. **40**, 807 (1997)].
29. Y. S. Kopysov, in *Proceedings of International Conference Neutrino-82* (Budapest, 1982), Vol. 1, p. 274.
30. O. V. Chumak and V. N. Oraevsky, in *Proceedings of Fourth SOHO Workshop on Helioseismology,* ESA SP-376, Pacific Grove, USA (1995), p. 59.
31. B. M. Vladimirskiĭ and L. D. Kislovskiĭ, Izv. Krym. Astrofiz. Obs. **82**, 153 (1990).
32. A. I. Khlystov, V. P. Dolgachev, and L. M. Domozhilova, Tr. Gos. Aston. Inst., Mosk. Gos. Univ. **64**, 91 (1995).
33. K. S. Jacobs, Nature **256**, 560 (1975).
34. V. P. Vasil'ev and A. I. Kalinichenko, in *Studies on Physics of Cosmic Rays* (Yakutsk, 1985), p. 62.
35. S. É. Shnol', V. A. Kolombet, E. V. Pozharskiĭ, *et al.*, Usp. Fiz. Nauk **168**, 1129 (1998) [Phys. Usp. **41**, 1025 (1998)].
36. S. E. Shnoll, E. U. Pozharskii, T. A. Zenchenko, *et al.*, Phys. Chem. Earth A **24**, 711 (1999).
37. V. P. Izmaĭlov, O. V. Karagioz, and A. G. Parkhomov, in *Atlas of Time Variations of Natural, Anthropogenic, and Social Processes* (Nauchnyĭ Mir, Moscow, 1998), Vol. 2, p. 163.
38. B. M. Vladimirskiĭ and A. V. Bruns, Biofizika **43**, 720 (1998).
39. V. M. Lobashov, V. N. Aseev, A. I. Belesev, *et al.*, Phys. Lett. B **460**, 227 (1999).
40. B. M. Vladimirskiĭ and N. A. Temur'yants, *Solar Activity Impact on the Biosphere–Noosphere* (MNÉPU, Moscow, 2000) [in Russian].
41. B. T. Cleveland, T. Daily, R. Davis, *et al.*, Astrophys. J. **496**, 505 (1998).
42. L. A. Antonova, G. S. Ivanov-Kholodnyĭ, and V. E. Chertoprud, *Aeronomy of the E Layer* (Yanus, Moscow, 1996) [in Russian].
43. http://www.izmiran.rssi.ru/magnetism/SSIMF/index.htm.
44. P. Anselmann *et al.* (GALLEX Collab.), Phys. Lett. B **314**, 445 (1993).
45. P. Anselmann *et al.* (GALLEX Collab.), Phys. Lett. B **327**, 377 (1994).
46. P. Anselmann *et al.* (GALLEX Collab.), Phys. Lett. B **388**, 384 (1996).
47. P. Anselmann *et al.* (GALLEX Collab.), Phys. Lett. B **447**, 127 (1999).
48. P. Anselmann *et al.* (GALLEX Collab.), Phys. Lett. B **447**, 440 (1999).
49. N. V. Udal'tsova, V. A. Kolombet, and S. É. Shnol', *Possible Space Physical Dependence of Macroscopic Fluctuations* (Pushchino, 1987).
50. A. V. Bruns, B. M. Vladimirskiĭ, L. G. Limanskiĭ, and S. M. Shumko, in *Solar Physics: Proceedings of 7th Symposium of Russia and CIS Countries on Solar–Terrestrial Physics* (Troitsk, 1999), p. 179.

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 643–650.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 734–743.

GRAVITATION, ASTROPHYSICS

The Spectrum of Cosmic-Ray Particles and Their Origin

N. L. Grigorov and E. D. Tolstaya*

Skobeltsyn Research Institute of Nuclear Physics, Moscow State University, Vorob'evy gory, Moscow, 119992 Russia
**e-mail: katya@srd.sinp.msu.ru*

Received May 21, 2003

Abstract—An analysis of all the direct measurements of the spectrum for all cosmic-ray particles over the energy range 0.1–10 TeV reveals an anomaly in the spectrum in the form of a step if the spectrum is represented as $E^{\beta}I_0(E)$. The pattern of the anomaly unequivocally implies a proton spectrum with a knee at energy close to 1 TeV. The qualitative difference between the spectra of protons and nuclei with $Z \geq 2$ (the latter have a purely power-law spectrum over a wide energy range) leads us to conclude that the acceleration conditions for protons and nuclei are different. We consider the process characteristic only of protons that may be responsible for the emergence of a knee in the proton spectrum. © *2004 MAIK "Nauka/Interperiodica".*

1. DIRECT MEASUREMENTS OF ALL GALACTIC COSMIC-RAY PARTICLES OVER THE ENERGY RANGE 0.1–100 TeV

Historically, there have been very few direct measurements of the spectrum for all galactic cosmic-ray (GCR) particles. In general, information about the spectrum of all particles, $I_0(E)$, have been obtained by adding the spectra of the individual components.

Since the spectra of the individual components were measured by different methods (using electronic devices at energies $E < 1$ TeV and, as a rule, X-ray emulsion chambers (XEC) with a high detection threshold at $E > 1$ TeV (5–10 TeV)), the proton spectrum contained a range from about 1 to 5–10 TeV in which virtually no direct measurements were carried out. In the spectrum of all particles obtained by adding the individual components, the energy range $1 < E < 5$–10 TeV was not covered by direct measurements. This energy range in the spectrum of all particles is usually drawn by interpolation based on the confidence that the proton spectrum is similar to the spectrum of nuclei. This spectrum is commonly considered as the spectrum of all GCR particles $I_0(E)$ [1].

Direct information about the spectrum of all particles in the energy range 1 to 5–10 TeV can be obtained by using direct measurements of the spectrum for all particles with electronic equipment that covers a wide energy range containing particles both to the left and to the right of the narrow interval 1–5 TeV to be measured. Such information was first obtained in 1972 by measuring the spectrum of all particles with the SEZ-14 instrument onboard the Proton-1,2,3 satellites over the energy range 0.07–17 TeV and with the SEZ-15 instrument onboard the Proton-4 satellite over the energy range 0.19–10^3 TeV [2]. These measurements first revealed an anomaly in the spectrum of all particles in the energy range 1–10 TeV. (Subsequently, the energy spectra taken by the authors of [2] were published in a tabulated form [3].)

The discovered anomaly had been neither confirmed nor disproved over the whole 25 years. The spectrum of all particles was remeasured with a thin ionization calorimeter (TIC) flown on a balloon only in 1997 [4]. The TIC measured the energy release of all the particles that fell on the instrument in any direction. As the authors of [4] showed, the energy release spectrum revealed the same anomaly in the spectrum of all particles that was observed in the Proton satellite measurements [2]. When the energy release was recalculated to the particle energy, as was done in [5], the TIC energy spectrum was quantitatively identical to the spectrum obtained by the authors of [2]. The TIC, SEZ-14, and SEZ-15 measurements are shown in Fig. 1. The solid line in Fig. 1 indicates the best fit $\Phi(E)$ to the experimental spectrum of all particles at $a = 0.4$ TeV:

$$\Phi(E) = E^{2.6}I_0(E) = \frac{1.1}{[1+(E/a)^3]^{0.2}} \times \left\{1 + 0.37\frac{(E/a)^3}{1+(E/a)^3}\right\} + 0.130\ \mathrm{m^{-2}\ s^{-1}\ sr^{-1}\ TeV^{1.6}}. \quad (1)$$

We see from Fig. 1 that the anomaly in the spectrum of all particles, if it is represented as $E^{\beta}I_0(E)$, appears as a step.

Figure 1 shows that the anomaly in the spectrum of all particles is revealed irrespective of the thickness of the ionization calorimeter (IC) in the instrument (it is $\sim 1\lambda_p$ in TIC, $\sim 1.71\lambda_p$ in SEZ-14, and $\sim 3\lambda_p$ in SEZ-15). Therefore, we decided to look for such an anomaly in the spectrum of all particles measured with the BFB-S instrument, in which the IC had a mean thickness of $\sim 0.7\lambda_p$.

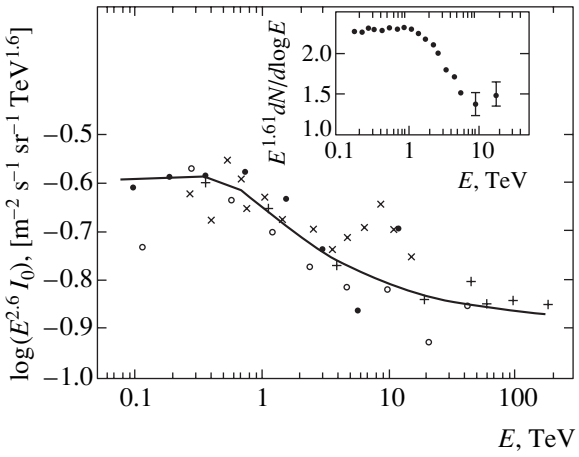


Fig. 1. $E^{2.6}I_0$ versus E, as measured with different instruments: SEZ-14 (●) [3], SEZ-15 (+) [3], TIC (○) [5], and BFB-S (×) (this paper). The ATIC data are shown in the upper right corner; arbitrary units are along the vertical axis, and energy release in the calorimeter is along the horizontal axis.

The BFB-S was installed on the Intercosmos-6 satellite and described in [6]. The ionization calorimeter of this instrument consisted of two identical sections. Each section consisted of eight 1.5-cm-thick lead plates. A 0.5-cm-thick scintillator was placed under each odd plate. All four scintillators were viewed by two photomultipliers from their ends, one from each side. The signals from both photomultipliers were added and fed to a pulse-height analyzer. It recorded the energy release spectrum without being connected with the trigger that controlled the operation of the instrument.

Since the photomultiplier photocathodes were strongly diaphragmed, the particles emerging from the IC affected the signal from the photomultiplier. As a result, the relationship between the measured energy release ε and the particle energy E turned out to be nonlinear, $E = \varepsilon^{\alpha}$ (at $\alpha = 0.78$). Therefore, the spectrum lost its scientific significance and was not published. Now, we are interested in the presence (or absence) of an irregularity in a narrow spectral range that cannot arise from the nonlinear relationship between ε and E. Therefore, we recalculated the measured energy release to the spectrum of all particles $I_0(E)$ and plotted $E^{\beta}I_0(E)$ in Fig. 1 (crosses). We see that the measured BFB-S spectrum also exhibits an irregularity in the form of a step in the spectrum of all particles in the same energy range in which it is recorded by other instruments. The smaller step height is a natural result of the IC thinness (protons—the culprits of the step—make a small contribution to the number of recorded particles).

Table 1

Instrument	β_1	β_2	β_3	Source
SEZ-14	2.59	3.00	–	[3]
SEZ-15*	–	2.94	2.63	[3]
TIC	–	2.80	2.65	[4, 5]
BFB-S	2.59	2.78	2.66	This paper
ATIC	2.61	2.87	–	[8]
Literature	2.62	–	2.67	[1, 3, 9]
Mean	2.60 ± 0.01	2.88 ± 0.04	2.65 ± 0.01	

* In some of the publications, SEZ-15 is called IC-15.

A preliminary result of the ATIC measurement of the spectrum for all particles was published at the 27th International Conference on Cosmic Rays [7]. The spectrum of energy release in the calorimeter of the instrument presented in this paper was thoroughly measured in [8]. As a result, the authors obtained a dependence of $E^{\beta}(dN/d\log E)$ on $\log E$ in different ranges of energy release. This dependence is shown in the upper right corner of Fig. 1. It convincingly demonstrates that there is also the same step as that recorded by previous instruments in the ATIC energy release spectrum for all particles.

Thus, we have measurements of the spectrum for all particles with five different instruments: SEZ-14, SEZ-15, TIC, BFB-S, and ATIC. They all measured the spectrum over a wide energy range that contained the interval 1–10 TeV concerned, and they all revealed a similar anomaly in the spectrum in the form of a step: with different spectral indices in different energy ranges. Accordingly, the values of $E^{\beta}I_0(E)$ before and after the step are also different. The five qualitatively identical results suggest that the step in the spectrum of all particles is an objective reality that has certain quantitative characteristics. These characteristics include the following: the spectral index at energies up to 1 TeV (β_1), the spectral index in the energy range 1–5 TeV (β_2), the spectral index at $E \geq 10$ TeV (β_3), and the mean values of $E^{\beta}I_0(E)$ at $E < 1$ TeV and $E \geq 5$ TeV. We determined all of these characteristics from the results of each experiment and gathered them together in Tables 1 and 2.

The differences between the mean spectral indices in different energy ranges are

$$\langle\beta_2\rangle - \langle\beta_1\rangle = 0.28 \pm 0.04, \quad \langle\beta_2\rangle - \langle\beta_3\rangle = 0.23 \pm 0.04.$$

These values allow us to formulate the first characteristic of the anomaly in the spectrum of all particles: the spectral indices at energies $E < 1$ TeV and $E > 5$ TeV are almost equal and close to 2.6. The spectral index in the energy range 1–5 TeV is larger than that outside this range by 0.2–0.25.

If the spectrum of all particles is represented as

$$\Phi(E) = E^{\beta} I_0(E), \quad \beta = 2.6,$$

then in the energy regions where $I_0(E)$ is described by a power function with a spectral index of 2.6, $\Phi(E)$ will be constant over the entire energy range. This implies that Φ must have some constant values Φ_1 and Φ_2 in the spectrum of all particles at $E < 1$ TeV and $E > 5$ TeV, respectively. The values of Φ_1 and Φ_2 obtained from each experiment are given in Table 2.

The second quantitative characteristic of the anomaly in the spectrum of all particles is the ratio of the step height to the flux of all particles before the step, i.e., $(\Phi_1 - \Phi_2)/\Phi_1$. This parameter is close to the ratio of the proton flux to the total flux of all GCR particles at equal energy per particle.

The following two remarks should be made regarding the results presented in Table 2.

First, if we take data from the All-Union State Standard (GOST) [10] for energies $E < 1$ TeV, then we obtain the sum

$$\sum_{Z=1}^{28} E^{2.6} I_Z = 0.258 \pm 0.005 \text{ m}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ TeV}^{1.6}$$

which is almost the same as that obtained by directly measuring the spectrum of all particles with the instruments listed in Table 2. This implies that there are no significant systematic errors in these measurements.

The above sum consists of the sum of two quantities: one refers to protons and is equal to

$$E^{2.6} I_p = 0.120 \text{ m}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ TeV}^{1.6},$$

and the other refers to all nuclei with $Z \geq 2$ and is equal to

$$E^{\beta} I_Z = 0.138 \pm 0.005 \text{ m}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ TeV}^{1.6}.$$

In other words, the protons at $E < 1$ TeV account for $0.120/0.258 = 0.46$ of the total particle flux at equal energy per particle (this is a well-known result).

Second, it is well known that for the nuclei,

$$E^{2.6} I_Z = \text{const}$$

over a wide energy range of several orders of magnitude. At E < 1 TeV,

$$\sum_{Z=2}^{28} E^{2.6} I_Z = 0.138 \pm 0.005 \text{ m}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ TeV}^{1.6},$$

and at $E > 5$ TeV for the flux of all particles I_0,

$$E^{2.6} I_0 = 0.148 \pm 0.008 \text{ m}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ TeV}^{1.6}.$$

Table 2

Instrument	Φ_1, m^{-2} s^{-1} sr^{-1} TeV$^{1.6}$ ($E < 1$ TeV)	Φ_2, m^{-2} s^{-1} sr^{-1} TeV$^{1.6}$ ($E > 5$ TeV)	$K = \Phi_1/\Phi_2$
SEZ-14	0.247 ± 0.009	–	1.66 ± 0.07
SEZ-15	–	0.149 ± 0.003	–
TIC	0.240 ± 0.018	0.134 ± 0.008	1.79 ± 0.17
BFB-S	0.237 ± 0.012	0.198 ± 0.007	1.20 ± 0.07
ATIC [8]	–	–	1.49 ± 0.08
Literature	0.270 [1]	0.160 ± 0.007 [9]	1.69 ± 0.07
Mean	0.249 ± 0.007	0.148 ± 0.008	1.66 ± 0.06

Note: The mean values do not include the BFB-S data. The errors of the mean are the rms deviations from the mean. The first row in the column $K = \Phi_1/\Phi_2$ was obtained from SEZ-14 and SEZ-15 data.

The fact that these two values are almost equal suggests that there are very few protons in the flux of all particles at $E > 5$ TeV.

Thus, Table 2 and the remarks to it lead us to conclude that the step in the spectrum of all particles is produced by protons.

2. PARAMETERS OF THE PROTON SPECTRUM (FROM THE SPECTRUM OF ALL PARTICLES)

We obtain the proton spectrum $I_p(E)$ from the obvious equality

$$I_0(E) = I_p(E) + I_Z(E),$$

where $I_Z(E)$ is the spectrum of the sum of all nuclear components with $Z \geq 2$. Multiplying all terms of this equality by $E^{2.6}$ and interchanging I_0 and I_p yields

$$E^{2.6} I_p(E) = E^{2.6} I_0(E) - E^{2.6} I_Z(E). \quad (2)$$

Since

$$E^{2.6} I_Z(E) = \text{const} = \Phi_Z,$$

over a wide energy range, equality (2) may be rewritten as

$$E^{2.6} I_p(E) = E^{2.6} I_0 - \Phi_Z.$$

Since

$$E^{2.6} I_0(E) = \text{const} = \Phi_1$$

at $E < 1$ TeV,

$$E^{2.6} I_p = \Phi_1 - \Phi_Z = \text{const}$$
$$= 0.11 \pm 0.01 \text{ m}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ TeV}^{1.6}$$

in this energy range.

Hence, we obtain for $E < 1$ TeV

$$I_p(E) = (0.11 \pm 0.01)E^{-2.6}\ \mathrm{m}^{-2}\ \mathrm{s}^{-1}\ \mathrm{sr}^{-1}\ \mathrm{TeV}^{-1}.$$

The proton spectrum at $E > 1$ TeV must decrease faster than $E^{-2.6}$; i.e., it must have a spectral index $\beta_p > 2.6$. The value of β_p can be determined from the data in Tables 1 and 2. To this end, we represent the proton spectrum in a slightly simplified form:

$$I_p(E) \propto \begin{cases} E^{-2.6}, & E < E_c, \\ E^{-\beta_p}, & E > E_c. \end{cases}$$

For this proton spectrum, the spectrum of all particles is

$$E^{2.6}I_0(E) = BE^{-(\beta_p - 2.6)} + \Phi_Z.$$

At $E = E_c$,

$$E^{2.6}I_0 = \Phi_1, \quad B = (\Phi_1 - \Phi_Z)E_c^{(\beta_p - 2.6)}.$$

Therefore,

$$E^{2.6}I_0(E) = (\Phi_1 - \Phi_Z)(E/E_c)^{-(\beta_p - 2.6)} + \Phi_Z.$$

The sum of two power functions,

$$BE^{-\gamma_1} + CE^{-\gamma_2},$$

may be substituted with a good accuracy by one power function $DE^{-\gamma}$, where

$$\gamma = \frac{B}{C+B}\gamma_1 + \frac{C}{C+B}\gamma_2$$

(see [11]). In our case,

$$\gamma_1 = \beta_p - 2.6, \quad B = \frac{\Phi_1 - \Phi_Z}{\Phi_1},$$

$$\gamma_2 = 0, \quad C = \frac{\Phi_Z}{\Phi_1}.$$

Therefore, the power-law index of the sum of the spectra is

$$\frac{\Phi_1 - \Phi_Z}{\Phi_1}(\beta_p - 2.6).$$

At $E > E_c$, the spectral index of the spectrum for all particles is equal to β. Therefore, $E^{2.6}I_0(E)$ is a power function with an index $\beta - 2.6$. As a result,

$$\frac{\Phi_1 - \Phi_Z}{\Phi_1}(\beta_p - 2.6) = \beta - 2.6.$$

If we use the mean value of $\Phi_1 = 0.249 \pm 0.007$ from Table 2, $\beta - 2.6 = 0.24 \pm 0.04$ from Table 1, and $\Phi_Z = 0.138 \pm 0.005$, then we obtain

$$\beta_p - 2.6 = (0.24 \pm 0.04) \cdot (2.26 \pm 0.18) = 0.54 \pm 0.09,$$

whence $\beta_p = 3.14 \pm 0.09$ at $E > 1$ TeV.

Thus, the step in the spectrum of all particles inevitably leads us to conclude that the proton spectrum has a knee at energy close to 1 TeV. The proton spectral index is $\beta_p = 2.6$ before the knee and $\beta_p = 3.14 \pm 0.09$ after the knee.

The shape of the proton spectrum can be obtained in more detail from the same equality (2) if we subtract the contribution of nuclei Φ_Z from the function $\Phi(E)$ that describes the spectrum of all particles. If expression (1) is taken as $\Phi(E)$, then the proton spectrum takes the form

$$E^{2.6}I_p(E) = \frac{0.11}{[1 + (E/a)^3]^{0.2}} \times \left\{1 + 0.37\frac{(E/a)^3}{1 + (E/a)^3}\right\} \mathrm{m}^{-2}\ \mathrm{s}^{-1}\ \mathrm{sr}^{-1}\ \mathrm{TeV}^{2.6}. \quad (3)$$

The coefficient a should be determined by comparing (3) with the experimental proton spectrum (see below).

It is important to emphasize that we obtained the proton spectrum (3) with a knee at $E = a$ from the spectrum of all particles, i.e., from the experiments to which the reverse particle current from the ionization calorimeter (which frighten many experimenters) bears no relation whatsoever.

Let us now consider the direct measurements of the proton spectra. They all refer to energies $E > 4$–10 TeV, and most of them were carried out by the XEC method without experimental chamber calibration. Therefore, caution should be exercised when dealing with the absolute fluxes determined by this method. However, this remark does not apply to the spectral index β_p concerned.

Table 3 lists the values of β_p obtained by different authors.

Its columns give the following data: (1) the author and the measurement method, (2) the minimum proton energy in the spectrum (in TeV), (3) the value of β_p with its error, and (4) the number of protons N_0 used to construct the spectrum. The number N_0 without an asterisk is given in the paper; the number with an asterisk was estimated by us from the statistical errors.

The mean β_p of the five measurements listed in Table 3 is

$$\langle\beta_p\rangle = 2.94 \pm 0.07.$$

The rms deviation σ of the individual result from the mean is 0.14, i.e., of the same order of magnitude as the error of the individual measurements. This circumstance suggests that the scatter of numerical β_p values is purely statistical in nature, and the individual β_p values can differ significantly from the mean at the statistical errors characteristic of these experiments.

We see that the direct measurements of the proton spectrum yield $\beta = 2.94 \pm 0.07$ for energies 5–20 TeV, which agrees with the above value of $\beta_p = 3.14 \pm 0.09$ for energies 1–5 TeV (the region of the step in the spectrum of all particles).

In [12, 15], the XEC had targets. As was noted in [16, 17], protons with energies close to the detection threshold are detected with a low efficiency in such chambers, which may cause β_p to decrease. Therefore, we determined β_p for $E \geq 20$ TeV, farther from the threshold energy in the spectra of [14, 15]. It turned out that $\beta_p = 3.17 \pm 0.19$ and 3.05 ± 0.19 in these spectra in the above energy range. The mean value is $\beta_p = 3.11 \pm 0.14$ (see [18]).

Within the error limits, all three values of β_p (3.14, 2.94, and 3.11) refer to the same spectral index that characterizes the proton spectrum in an energy range from approximately 1 to 40–50 TeV. The weighted mean β_p of these three values is 3.02 ± 0.05.

There is only one experimental proton spectrum in an energy range of about 0.1–10 TeV. It was obtained more than 30 years ago from the Proton-2 and 3 satellites; this spectrum was published in integral and differential forms in [19] and [11], respectively. We show it in Fig. 2 (from [11]) together with the proton spectrum obtained from the spectrum of all particles: expression (3) at $a = 0.8$ TeV. We see from Fig. 2 that the experimental proton spectrum matches the spectrum obtained from the spectrum of all particles; as was shown above, the latter matches the direct measurements in an energy range of about 5–40 TeV.

Thus, we may assert that all of the direct measurements yield the same result: the proton spectrum has a knee at energy near 1 TeV. The spectral index is $\beta_p = 2.6$ (or possibly 2.7) before the knee and is larger by 0.5–0.6, i.e., 3.0–3.1, after the knee.

It should be added to this conclusion that all of the indirect measurements of high-energy secondary particles in the Earth's atmosphere (hadrons, muons, γ photons) lead us to conclude that $\beta_p = 3.0$ in the TeV energy range [11].

3. THE ORIGIN OF GALACTIC COSMIC-RAY PROTONS

Even a fleeting glimpse of the proton and nuclear spectra reveals a significant difference between them: the nuclei have a purely power-law spectrum over a wide energy range, while the protons have a power-law spectrum, but with a knee near 1 TeV. This difference

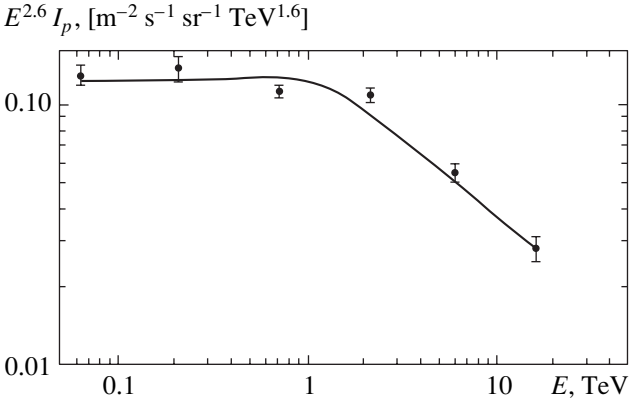


Fig. 2. Measured SEZ-14 proton spectrum [11]. The solid curve represents the fit to the difference between the spectrum of all particles and the spectrum of nuclei with $Z \geq 2$ (formula (3)).

cannot be acquired during the propagation of particles in the Galaxy. In this case, particles of identical rigidity undergo identical disturbances, irrespective of their charge and mass. Therefore, changes in the spectrum, if they were caused by the propagation processes, would have the same effect both on protons and on nuclei. In other words, the observed difference is acquired in the sources. Consequently, the protons and the nuclei have different sources; the sources of nuclei give particles with a purely power-law spectrum, while the sources of protons give particles with a knee in their spectrum.

A characteristic feature of the knee in the proton spectrum is that the energy range in which β_p changes by 0.5–0.6 is narrow. This circumstance is indicative of a universality of the knee formation that depends weakly on the specific properties of the source.

When explaining the formation of the knee in the proton spectrum, we should also point out the cause of the formation of a knee at energy near 1 TeV, and not at some other energy that differs greatly from 1 TeV.

We believe that the particles themselves rather than the sources are mainly responsible for the difference in the spectra of protons and nuclei. What are the differ-

Table 3

Method and reference	E_{min}, TeV	Spectral index	N_0
XEC, [12]	5	$\beta_p - 1 = 1.82 \pm 0.13$	90*
Calorimeter, [9]	4	$\beta_p - 1 = 2.11 \pm 0.15$	90*
Calorimeter, [13]	5	$\beta_p = 2.85 \pm 0.14$	160*
XEC, [14]	10	$\beta_p = 3.14 \pm 0.08$	602
XEC, [15]	6	$\beta_p = 2.80 \pm 0.04$[1]	656

[1] The error of 0.04 given in [15] has no physical meaning, because it is smaller than the statistical error of 0.07.

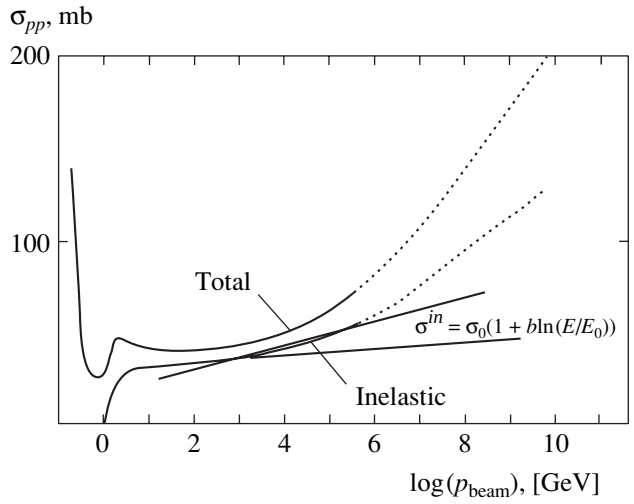


Fig. 3. Effective cross section for p–p interaction versus energy [23].

ences between the protons and the nuclei that can have a crucial effect on their acceleration and escape from the sources? There is one qualitative difference between these types of particles: as they accelerate and escape from the sources, the protons can undergo an infinite number of inelastic collisions while remaining nucleons. In contrast, the nuclei are too fragile: after several inelastic collisions, they break up into their constituent nucleons and cease to exist as nuclei. As a result of this difference, the protons can accelerate in a sufficiently dense medium and traverse a significant thickness of material (~10^2–10^3 g cm^{-2}). In contrast, the nuclei can accelerate only in a low-density medium, i.e., in a greatly expanded supernova envelope.

The possibility of particle acceleration to high energies at the initial phase of a supernova explosion was considered by Colgate and Johnson [20]. They showed that acceleration was possible even in dense stellar layers; in this case, a power-law spectrum of the accelerated particles can be formed.

Without going into the details of the acceleration process, let us consider what particles will escape from such a source and what spectrum they will have.

A supernova explosion is the final phase of stellar evolution. Therefore, old stars, red giants and supergiants, explode. In these stars, hydrogen has long burned out, and the shells are composed of complex nuclei heavier than hydrogen. The explosion energy is released in the stellar core and the adjacent shell regions. Therefore, particle acceleration can begin in sufficiently dense shell layers. The nuclei will accelerate, because the shell consists of nuclei. However, the accelerated nuclei in a dense medium will undergo inelastic collisions and continuously fragment into lighter parts. Therefore, the accelerated nuclei will rapidly turn into a beam of energetic protons. (The neutrons will also turn into protons because of their instability.) Having a speed close to the speed of light, the protons will rapidly escape from the acceleration region and, moving through the stellar envelope, will escape from it. Clearly, during this escape from the exploded supernova, the protons will inevitably have to traverse a significant thickness of material of several hundreds or thousands of g cm^{-2}. What will happen to the original spectrum in this case?

Suppose that the protons in the acceleration region acquire a power-law spectrum of the form

$$I(E) = I_0 E^{-\beta}.$$

The equation that describes the passage of protons through matter is

$$\frac{\partial I(E, x)}{\partial x} = -\frac{I(E, x)}{\lambda} + \int_E^\infty \frac{I(E', x)}{\lambda} P(E', E)dE'. \quad (4)$$

If the effective cross section for inelastic interaction of protons with matter, σ^{in}, does not depend on energy, then the mean free path before inelastic interaction is λ_0 = const. In this case, the solution of Eq. (4) is known to be

$$I(E, x) = I_0 E^{-\beta} e^{-(x/L_0)},$$

where the absorption mean free path L_0 is related to the mean free path before interaction λ_0 by

$$\frac{1}{L_0} = \frac{1 - \langle u^{\beta-1}\rangle}{\lambda_0},$$

$$\langle u^{\beta-1}\rangle = \int_0^1 u^{\beta-1} P(u)du.$$

In other words, in the case under consideration, a proton beam with the same power-law energy distribution as that formed in the acceleration region, but with a lower intensity, would emerge from the stellar envelope.

In reality, however, the effective cross section for inelastic interaction depends on energy, as shown in Fig. 3. In the first approximation, this dependence at $E > E_0$ may be fitted by the function

$$\sigma^{in}(E) = \sigma_0(1 + b\ln(E/E_0)). \quad (5)$$

For a hydrogen medium, as shown in Fig. 3, b = 0.08. For the Earth's atmosphere, b = 0.04–0.05. For this energy dependence of the cross section, the situation must change.

Indeed, the probability of an inelastic collision between a nucleon and nuclei increases with E. Therefore, a higher-energy nucleon undergoes a larger number of collisions in a given layer of material than does a

lower-energy nucleon. Therefore, a higher-energy nucleon loses a larger fraction of its initial energy than does a lower-energy nucleon. Hence, as it emerges from the absorbing layer, a higher-energy nucleon is shifted on the energy scale toward the lower energies by a larger interval than a nucleon with a lower initial energy. Consequently, the initial power-law spectrum emerging from an absorber will be softer than the original spectrum; i.e., the spectral index will be larger than that for the original spectrum. One of us explained the softer spectrum of high-energy hadrons deep in the Earth's atmosphere than the spectrum of primary GCRs by this effect back in 1965 [21]. An approximate solution of Eq. (4) with fit (5) was found in [22]:

$$I(E, x) = I_0 E^{-\beta} e^{-x/L(E)},$$

where

$$L(E) = \frac{L_0}{1 + b\ln(E/E_0)},$$

or, in a different form,

$$I(E, x) = I_0 E^{-(\beta+\delta)} e^{-x/L_0},$$

where

$$\delta = bx/L_0.$$

This approximate solution refers to particles with $E > E_0$ under the boundary conditions $I(E, x = 0) = I_0 E^{-\beta}$. It yields an error in $L(E)$ of only 2% for a layer of absorbing material $x = 700$ g cm^{-2} and $b = 0.04$.

We see from Fig. 3 that $\sigma^{in} = \text{const}$ at $E < E_0$. Therefore, for particles with $E \ll E_0$, the solution of Eq. (4) must correspond to $\sigma^{in} = \text{const}$; i.e., it must be

$$I(E, x) = I_0 E^{-\beta_0} e^{-x/L_0}.$$

Let us apply these solutions to the acceleration of particles at the early phase of a supernova explosion, i.e., in sufficiently deep regions of its envelope.

Suppose that the accelerated particles initially had a power-law spectrum of the form

$$I(E) = I_0 E^{-\beta_0}.$$

From the place of acceleration to the escape of particles from the star, they had to traverse a significant thickness of envelope material. Therefore, the spectrum emerging from the envelope will have different spectral indices in different spectral ranges. The spectral index is $\beta = \beta_0 + \delta$, where $\delta = bx/L_0$ (x is the amount of traversed material) at $E > E_0$ and is constant and equal to $\beta = \beta_0$ at $E \ll E_0$. Thus, the initially power-law spectrum of the accelerated protons as they emerge from the supernova will be a power-law spectrum with a knee.

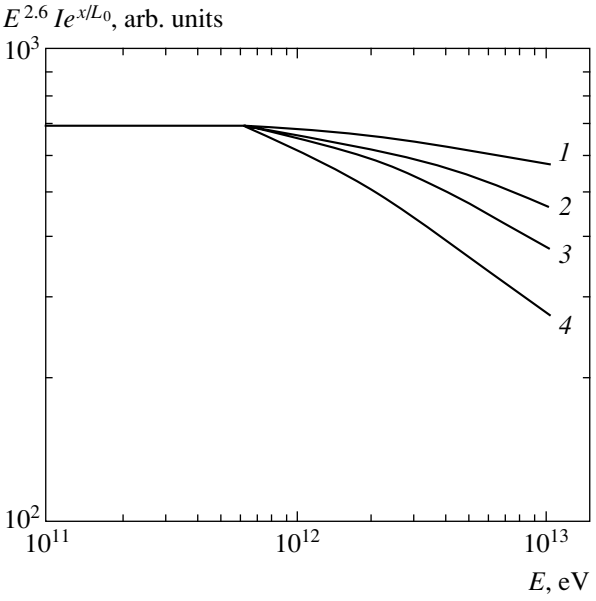


Fig. 4. Proton spectra at various depths x of a hydrogen atmosphere: $x = 100$ (*1*), 200 (*2*), 300 (*3*), and 450 (*4*) g cm^{-2}.

To find out how wide the region in which the spectral index changes from β_0 to $\beta_0 + \delta$ is and how this region changes with the amount of material traversed, we performed Monte Carlo numerical simulations of the passage of a nucleon beam through different thicknesses of material using fit (5). The results of these simulations are shown in Fig. 4.

Using Fig. 4, we may relate the energy E_k at which the knee in the spectrum occurs to the amount of traversed material x by the empirical relation

$$E_k = 3.4(x/L_0)^{-0.8} \text{ TeV}.$$

As we see, the location of the knee in the proton spectrum depends weakly on the amount of material traversed. Therefore, the knee region in the observed spectrum, which is the sum of the spectra from many sources in which the protons traverse different amounts of material, will be smeared only slightly; i.e., the observed location of the knee will be close to E_0 in dependence (5), as observed in the experiment.

A crucial factor in considering various cosmic-ray acceleration mechanisms is usually the spectral shape of the accelerated particles and the spectral index β in this spectrum. However, a different approach to the formation of the observed GCR spectrum is also possible.

According to this approach, particles are generated in the sources with a spectrum far from the power law $I \propto E^{-\beta}$ with $\beta = 2$–2.6. If there is a characteristic parameter ξ in the generated spectrum that determines the spectral shape and if the sources themselves have a power-law distribution of this parameter, i.e., $I_k(\xi) \propto$

$\xi^{-\beta}$, where I_k is the intensity of source k, then the total spectrum can be a power law with an index equal to β.

A glaring example of the formation of a power-law spectrum in this way is the spectrum of energetic γ-ray photons in the Earth's atmosphere from π^0-meson decay. If the π^0 mesons (sources) have a power-law distribution in the Lorentz factor (or, equivalently, in E_{max}—the maximum energy of the generated γ-ray photons), then the total γ-ray spectrum will be a power law with an index that determines the distribution in E_{max}. At the same time, the γ-ray photons in the sources themselves (the rest frame of π^0 mesons) are monoenergetic; i.e., their distribution is very far from a power law.

If we take into account the fact that the power-law distribution of a particular property of the various physical quantities in nature is widespread, then the formation of the observed proton spectrum by the process under consideration also seems likely. In this case, the observed value of β is the mean of many β_i values for the spectra from the individual sources. The value of $\delta = b\langle x/L_0\rangle$ is also the mean of many individual thicknesses of traversed material in individual supernovas.

A peculiar feature of this formation of the observed spectrum is its flattening with increasing energy. This is because the observed spectrum consists of a set of spectra with different β_i. As E increases, the contribution of the components with larger β_i will decrease, and, accordingly, the contribution of the components with smaller β_i will increase. This effect is experimentally observable.

To conclude the discussion of the proton spectrum, we emphasize that the existence of a knee in the proton spectrum at $E_k \sim 1$ TeV is important evidence that the cosmic-ray protons are generated in dense objects in which they traverse hundreds of g cm^{-2} of material. This circumstance can be important evidence for the Galactic origin of the cosmic-ray proton component.

The cosmic-ray particles discussed above constitute the bulk of the beam. Their energies are within the so-called knee in the spectrum of all particles at $E \approx (3–5) \times 10^{15}$ eV. In general, they are investigated by direct methods in experiments on balloons and Earth satellites. According to the popular view, their sources are supernovas of our Galaxy, and their acceleration mechanism involves the shock waves of the expanding supernovas envelope. The range of ultrahigh-energy cosmic-ray particles extends from the knee up to the measured end of the spectrum (~10^{20} eV). This range has been investigated only by indirect methods. In this range, there are interesting problems far from those considered above (see review of current status in [24]).

The formation of a proton spectrum with a knee at energy of ~1 TeV considered here does not affect existing models for the acceleration of nuclei in any way. Moreover, since the acceleration of nuclei is possible only in a low-density medium, the nuclei can be accelerated by shock waves in the envelopes of the same supernovas in which the protons were accelerated, but at a later phase: first, the protons accelerate, and then, after a lapse of time, the nuclei accelerate.

REFERENCES

1. T. Shibata, Nuovo Cimento C **19**, 713 (1996).
2. N. L. Grigorov, V. E. Nesterov, I. D. Rappoport, *et al.*, in *Space Research* (Akademie, Berlin, 1972), Vol. 12, p. 1617.
3. N. L. Grigorov, Kosm. Issled. **33**, 339 (1995).
4. G. Adams, V. I. Zatsepin, M. I. Panasyuk, *et al.*, Izv. Ross. Akad. Nauk, Ser. Fiz. **61**, 1181 (1997).
5. N. Grigorov and E. Tolstaya, in *Proceedings of the 27th ICRC* (Hamburg, Germany, 2001), p. 1647.
6. A. Shomodi, S. Sugar, B. Chadraa, *et al.*, Yad. Fiz. **28**, 445 (1978) [Sov. J. Nucl. Phys. **28**, 225 (1978)].
7. J. Wefel (for the ATIC Collaboration), in *Proceedings of the 27th ICRC* (Hamburg, Germany, 2001), p. 2111.
8. Yu. I. Stozhkov, Kratk. Soobshch. Fiz. (in press).
9. N. L. Grigorov, Yad. Fiz. **51**, 157 (1990) [Sov. J. Nucl. Phys. **51**, 99 (1990)].
10. *GOST* (State Standard) *25645.122-85* (1985); *GOST* (State Standard) *25645.125-85* (1985); *GOST* (State Standard) *25645.144-88* (1988).
11. N. L. Grigorov and E. D. Tolstaya, Pis'ma Zh. Éksp. Teor. Fiz. **74**, 147 (2001) [JETP Lett. **74**, 129 (2001)].
12. Ya. Kawamura, H. Matsutani, and H. Najio, Phys. Rev. D **40**, 729 (1989).
13. I. P. Ivanenko, V. Ya. Shestoperov, I. D. Rappoport, *et al.*, in *Proceedings of the 23rd ICRC* (Calgary, Canada, 1993), Vol. 2, p. 17.
14. V. I. Zatsepin, T. V. Lazareva, G. P. Sazhina, *et al.*, Yad. Fiz. **57**, 684 (1994) [Phys. At. Nucl. **57**, 645 (1994)].
15. M. L. Cherry (for the JACEE Collaboration), in *Proceedings of the 25th ICRC* (Rome, 1997), Vol. 4, p. 1.
16. N. S. Konovalova, Candidate's Dissertation in Physics and Mathematics (Physical Inst., Russian Academy of Sciences, Moscow, 1996).
17. RUNJOB Collaboration, Astropart. Phys. **16**, 13 (2001).
18. N. Grigorov and E. Tolstaya, in *Proceedings of the 26th ICRC* (Salt Lake City, USA, 1999), Vol. 3, p. 183.
19. N. L. Grigorov, V. E. Nesterov, I. D. Rappoport, *et al.*, Yad. Fiz. **11**, 1058 (1970) [Sov. J. Nucl. Phys. **11**, 588 (1970)].
20. S. A. Colgate and M. H. Johnson, Phys. Rev. Lett. **5**, 235 (1960).
21. N. L. Grigorov, I. D. Rappoport, I. A. Savenko, *et al.*, Izv. Akad. Nauk SSSR, Ser. Fiz. **29**, 1656 (1965).
22. N. L. Grigorov, Yad. Fiz. **25**, 788 (1977) [Sov. J. Nucl. Phys. **25**, 419 (1977)].
23. T. Gaisser, *Cosmic Rays and Particle Physics* (Cambridge Univ. Press, Cambridge, 1990).
24. E. Roulet, astro-ph/0310367.

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 651–660.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 744–754.

NUCLEI, PARTICLES, AND THEIR INTERACTION

QED with Vector and Axial Vector Types of Interaction and Dynamical Generation of Particle Masses

D. K. Fedorov* and A. S. Yurkov**
**e-mail: dfedorov@au.ru*
***e-mail: fitec@omskcity.com*
Received August 18, 2003

Abstract—We consider a model of electrodynamics with two types of interaction, the vector $(e\bar{\psi}(\gamma^{\mu}A_{\mu})\psi)$ and axial vector $(e_A\bar{\psi}(\gamma^{\mu}\gamma^5 B_{\mu})\psi)$ interactions, i.e., with two types of vector gauge fields, which corresponds to the local nature of the complete massless-fermion symmetry group $U(1)\otimes U_A(1)$. We present a phenomenological model with spontaneous symmetry breaking through which the fermion and the axial vector field B_{μ} acquire masses. Based on an approximate solution of the Dyson equation for the fermion mass operator, we demonstrate the phenomenon of dynamical chiral symmetry breaking when the field B_{μ} has mass. We show the possibility of eliminating the axial anomalies in the model under consideration when introducing other types of fermions (quarks) within the standard-model fermion generations. We consider the polarization operator for the field B_{μ} and the procedure for removing divergences when calculating it. We demonstrate the emergence of a mass pole in the propagator of the particles that correspond to the field B_{μ} when chiral symmetry is broken and consider the problems of regularizing closed fermion loops with axial vector vertices in connection with chiral symmetry breaking. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

In the universally accepted quantum electrodynamics (QED), the fermion masses are assumed to be given, and the question about the origin of these masses is not considered. At the same time, the idea that the electron mass is generated by electromagnetic interactions dates back to the end of the 19th century–beginning of the 20th century. Despite such a long history, this idea is also embodied in present-day studies [1–4], now on the basis of quantum theory.

It should be noted that the symmetry of the Lagrangian for massless fermions is higher than the symmetry of the Lagrangian for massive fermions. Massless fermions have a chiral symmetry, while massive fermions have no such symmetry. Thus, if the fermion masses are generated dynamically, through interactions, then dynamical chiral symmetry breaking (DCSB) takes place.

One of the first quantum-field models that described DCSB was the so-called Nambu–Jona-Lasinio model [5] with the local interaction Lagrangian

$$L_{NJL}^{\text{int}} = G[(\bar{\psi}\psi)^2 + (\bar{\psi}i\gamma^5\psi)^2],$$

where G is the coupling constant. In a more general form, the interaction Lagrangian for this model appears as the product of two fermion currents,

$$L_{NJL}^{\text{int}} = G(\bar{\psi}\gamma^{\mu}\psi)(\bar{\psi}\gamma_{\mu}\psi).$$

In this form, the interaction Lagrangian is actually a local simplification of the interaction between the two fermion currents through the nonlocal interaction that describes the vector field photon exchange in a phenomenological form. DCSB with the generation of fermion mass m and vacuum condensate $\langle 0|\bar{\psi}\psi|0\rangle$ is observed in the Nambu–Jona-Lasinio model with massless fermions if the coupling constant G exceeds a critical value,

$$G \geq G_{\text{crit}} = 2\pi^2/\Lambda^2,$$

where Λ^2 is the invariant cutoff parameter.

In the quantum electrodynamics of massless fermions, DCSB also exists for strong coupling $\alpha > \alpha_c$, where α_c is the critical coupling constant, and is absent for $\alpha < \alpha_c$ [1–4]. If $\alpha > \alpha_c$ and if there is an invariant cutoff parameter Λ, then a nontrivial solution to the Dyson (Schwinger–Dyson) equation exists for the dynamical fermion mass function $\beta(k^2)$, $1/(\hat{p} - \beta(k^2))$ is the fermion Green function in momentum space.

At the same time, it should be noted that the leading principle of almost any current quantum-field interaction theory is the idea of gauge (local) invariance. In this case, the symmetry of the Lagrangian for free (generally fermion) fields uniquely determines the Lagrangian for the interaction of these fields with the additional vector fields that ensure the satisfaction of the local (gauge) invariance conditions and that are interaction carriers.

1063-7761/04/9804-0651$26.00

The standard quantum electrodynamics is nothing but the gauge $U(1)$ theory. However, in contrast to the case of a nonzero mass, the symmetry group of the Lagrangian for massless fermions is $U(1) \otimes U_A(1)$, where the group $U(1)$ corresponds to the ordinary phase transformations of the field function, and the group $U_A(1)$ corresponds to the chiral transformations. In this case, it seems more natural to consider both $U(1)$ and $U_A(1)$ as equivalent local transformations and to construct the interaction from the group $U(1) \otimes U_A(1)$.

Clearly, new particles (axial photons) will correspond to the additional gauge field, but they acquire a mass through chiral symmetry breaking; their mass may prove to be very large, or, in a sense, the field of these particles may be considered as a kind of a regulator field of the theory.

In this paper, we consider a model of electrodynamics with two types of interaction, the vector $(e\bar{\psi}(\gamma^\mu A_\mu)\psi)$ and axial vector $(e_A\bar{\psi}(\gamma^\mu\gamma^5 B_\mu)\psi)$ interactions, i.e., with two types of interaction-carrying vector gauge fields, which corresponds to the local nature of the complete massless-fermion symmetry group $U(1) \otimes U_A(1)$. A phenomenological (σ-type) model with spontaneous symmetry breaking through which the fermion and the axial vector field B_μ acquire masses is presented in Section 2. In Section 3, based on an approximate solution of the Dyson equation for the fermion mass operator, we demonstrate the phenomenon of DCSB in a model with two gauge fields when the field B_μ has a mass. In Section 4, we show the possibility of eliminating the axial anomalies in the model under consideration when introducing other types of fermions (quarks) within the standard-model fermion generations for an appropriate choice of axial coupling constants e_A (axial charges) $\alpha_A = \alpha$ for each type of fermions. In Section 5, we consider the polarization operator for the field B_μ and the procedure for removing divergences when calculating it. We demonstrate the emergence of a mass pole in the propagator of the particles that correspond to the field B_μ when chiral symmetry is broken. This confirms that the solution of the Dyson equation for the fermion mass operator is adequate in a situation with DCSB. We also consider the problems of regularizing closed fermion loops with axial vector vertices in connection with chiral symmetry breaking.

2. THE PSEUDOVECTOR INTERACTION IN QED AND SPONTANEOUS CHIRAL SYMMETRY BREAKING

The standard quantum electrodynamics with massless fermions described by the Lagrangian

$$L = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \bar{\psi}(i\gamma^\mu\partial_\mu + e\gamma^\mu A_\mu)\psi, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \tag{1}$$

is symmetric relative to the gauge transformations $U(1)$:

$$\psi \longrightarrow (1 + i\Theta(x))\psi, \qquad A_\mu \longrightarrow A_\mu + \frac{1}{e}\partial_\mu\Theta(x), \tag{2}$$

which are represented in (2) in an infinitesimal form. However, there is also a symmetry of Lagrangian (1) relative to the global chiral transformations $U_A(1)$:

$$\psi \longrightarrow (1 + i\gamma^5\Theta_A)\psi. \tag{3}$$

The electromagnetic field described by the potentials A_μ corresponds to the interaction of charge particles.

In this formulation of electrodynamics, the symmetry relative to transformations (2) and (3) is a gauge and nongauge one, respectively, which is connected with an imbalance in the virtually equivalent symmetries of Lagrangian (1).

Let us introduce the second gauge field B_μ that corresponds to the chiral symmetry transformation $U_A(1)$. Lagrangian (1) takes the form

$$L = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} - \frac{1}{4}G^{\mu\nu}G_{\mu\nu} + \bar{\psi}(i\gamma^\mu\partial_\mu + e\gamma^\mu A_\mu + e_A\gamma^\mu\gamma^5 B_\mu)\psi, \qquad G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu; \tag{4}$$

it is invariant relative to the gauge transformations $U(1)$,

$$\psi \longrightarrow (1 + i\Theta(x))\psi, \qquad A_\mu \longrightarrow A_\mu + \frac{1}{e}\partial_\mu\Theta(x), \tag{5a}$$

and $U_A(1)$,

$$\psi \longrightarrow (1 + i\gamma^5\Theta_A(x))\psi, \qquad B_\mu \longrightarrow B_\mu + \frac{1}{e_A}\partial_\mu\Theta_A(x). \tag{5b}$$

The symmetry transformation $U_A(1)$ becomes local with the local parameter $\Theta_A(x)$, while the field B_μ is a pseudovector. The coupling constant e_A, the axial fermion charge, is introduced to couple the field B_μ with the fermion field ψ.

If we naively described the electrodynamics using Lagrangian (4), then we would have two massless gauge fields, A_μ and B_μ; i.e., we would have two types of massless photons corresponding to A_μ and B_μ. However, only the A_μ photons, i.e., ordinary photons, are observable massless photons, so we should take into account the $U_A(1)$ symmetry breaking.

The actual charged fermions have observable finite masses; i.e., the $U_A(1)$ symmetry is broken. As a result

of symmetry breaking, the charged fermions acquire a mass [1–3, 6].

By analogy with the σ-models used to describe strong interactions (see, e.g., [5]), in which the fermions acquire a mass when the symmetry is spontaneously broken, we write the Lagrangian that clearly shows spontaneous chiral symmetry breaking in our case as

$$\begin{aligned} L &= -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} - \frac{1}{4}G^{\mu\nu}G_{\mu\nu} \\ &+ \bar{\psi}(i\gamma^{\mu}\partial_{\mu} + e\gamma^{\mu}A_{\mu} + e_A\gamma^{\mu}\gamma^5 B_{\mu})\psi \\ &+ \frac{1}{2}\begin{pmatrix}\sigma\\ \pi\end{pmatrix}^T(\overleftarrow{\partial}_{\mu} + i\tau_2 2e_A B_{\mu}) \\ &\times(\vec{\partial}^{\mu} - i\tau_2 2e_A B^{\mu})\begin{pmatrix}\sigma\\ \pi\end{pmatrix} \\ &- \frac{\lambda}{4}(\sigma^2 + \pi^2 - \sigma_0^2)^2 - g\bar{\psi}(\sigma + i\gamma^5\pi)\psi. \end{aligned} \tag{6}$$

This Lagrangian is invariant relative to the gauge transformations $U(1)$,

$$\psi \longrightarrow (1 + i\Theta(x))\psi, \quad A_{\mu} \longrightarrow A_{\mu} + \frac{1}{e}\partial_{\mu}\Theta(x), \tag{7a}$$

and $U_A(1)$,

$$\begin{aligned} &\psi \longrightarrow (1 + i\gamma^5\Theta_A(x))\psi, \\ &\beta_{\mu} \longrightarrow B_{\mu} + \frac{1}{e_A}\partial_{\mu}\Theta_A(x), \\ &\sigma \longrightarrow \sigma + 2\Theta_A(x)\pi, \quad \pi \longrightarrow \pi - 2\Theta_A(x)\sigma \end{aligned} \tag{7b}$$

or

$$\begin{pmatrix}\sigma\\ \pi\end{pmatrix} \longrightarrow (1 + i\tau_2 2\Theta_A(x))\begin{pmatrix}\sigma\\ \pi\end{pmatrix}, \tag{7c}$$

where

$$\tau_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}.$$

In (6), the Higgs σ and π bosons (scalar and pseudoscalar, respectively) are the phenomenological fields that are coupled with the fermion field by the coupling constant g; they have no electric charge e, but have an axial charge $2e_A$ equal to twice the fermion axial charge.

The potential

$$V(\sigma, \pi) = (\lambda/4)(\sigma^2 + \pi^2 - \sigma_0^2)^2$$

formed from the Higgs fields have minima at $\sigma^2 + \pi^2 = \sigma_0^2$ that correspond to a vacuum with a spontaneously broken symmetry. By substituting $\sigma' = \sigma - \sigma_0$ for the σ field, we can obtain the fermion mass term in Lagrangian (6),

$$m\bar{\psi}\psi, \quad m = g\sigma_0, \tag{8}$$

and the mass term for the pseudovector boson B_{μ},

$$\frac{1}{2}M^2 B^{\mu}B_{\mu}, \quad M^2 = 4e_A^2\sigma_0^2. \tag{9}$$

Thus, the fermion and the gauge field of the pseudovector boson acquire masses. The σ field remains massive with the mass $M_{\sigma}^2 = 2\lambda\sigma_0^2$, but the π field remains massless, showing the emergence of massless Goldstone modes (excitations) in the case of spontaneous symmetry breaking. As might be expected, the electromagnetic field A_{μ} acquires no mass.

When the σ models for strong interactions are analyzed, π-like fields are associated with low-mass π mesons, but when the electrodynamics is considered, the existence of a physical massless pseudoscalar field is unacceptable. To eliminate the massless pseudoscalar field π, we should take into account the fact that the theory under consideration is a gauge one. In accordance with the standard scheme for demonstrating spontaneous symmetry breaking in theories with gauge fields, we may choose a (unitary) gauge of the field B_{μ} such that the π field vanishes.

Thus, the explicit scheme of spontaneous chiral symmetry breaking with the pseudovector gauge field B_{μ} and a set of Higgs bosons (the phenomenological scalar and the pseudoscalar) show that the fermion and pseudoscalar gauge fields acquire masses and that there are no physical fields corresponding to massless Goldstone modes.

In calculations in the theory described by Lagrangian (6), the propagator of the massive field B_{μ} in the unitary B_{μ} gauge is

$$D_{\mu\nu}^5(k) = \frac{1}{k^2 - M^2 + i0}\left(g_{\mu\nu} - \frac{k_{\mu}k_{\nu}}{M^2}\right);$$

its use causes difficulties with the removal of divergences and renormalization. However, we may use other gauges with a gauge-fixing term of the form

$$L_{GF} = -\frac{1}{2\xi}(\partial^{\mu}B_{\mu})^2.$$

The following propagator of the field B_{μ} corresponds to these gauges:

$$\begin{aligned} D_{\mu\nu}^2(k) &= \frac{1}{k^2 - M^2 + i0} \\ &\times\left(g_{\mu\nu} - (1-\xi)\frac{k_{\mu}k_{\nu}}{k^2 - \xi M^2}\right); \end{aligned} \tag{10}$$

a special case of it is the propagator in the transverse Landau gauge:

$$D^5_{\mu\nu}(k) = \frac{1}{k^2 - M^2 + i0}\left(g_{\mu\nu} - \frac{k_\mu k_\nu}{k^2}\right). \quad (11)$$

Renormalizability is restored in these gauges, but the unphysical π field remains in the theory and interacts with the field B_μ via an interaction term of the form $B^\mu\partial_\mu\pi$. When using propagator (11), this interaction is ineffective, or the field $\partial^\mu B_\mu$ is said to be nonpropagating.

An interaction of the form $B^\mu\partial_\mu\pi$ does not emerge explicitly in the t'Hooft gauges with a gauge-fixing term of the form

$$L_{GF} = -\frac{1}{2\xi}(\partial^\mu B_\mu + \xi M\pi)^2.$$

Propagator (10) also corresponds to the family of t'Hooft gauges.

When $\xi \longrightarrow \infty$ (physical limit), this propagator corresponds to the propagator of a massive vector field. The unphysical π fields are present in these gauges at a finite ξ. However, the physical results should not depend on the gauge, i.e., on the parameter ξ.

3. THE MASS OF THE FIELD B_μ AND DCSB

We will use the Dyson equation [1–3, 6] to demonstrate the DCSB in the model with Lagrangian (4). For the two types of interaction in momentum representation, this equation may be written as

$$\Sigma(k) = G^{-1}(k) - G^{-1}_{\text{int}}(k)$$
$$= -ie^2\int\gamma_\nu G_{\text{int}}(k+p)\Gamma_\mu(k+p, p, k)D^{\nu\mu}_{\text{int}}(p)\frac{d^4p}{(2\pi)^4}$$
$$-ie_A^2\int\gamma_\nu\gamma^5 G_{\text{int}}(k+p)\Gamma^5_\mu(k+p, p, k)D^{\nu\mu}_{5,\text{int}}(p)\frac{d^4p}{(2\pi)^4}, \quad (12)$$

where $\Sigma(k)$ is the fermion mass operator; $G_{\text{int}}(k)$ is the complete fermion Green function (propagator); $G(k)$ is the propagator of a free fermion; $D^{\nu\mu}_{\text{int}}(k)$ and $D^{\nu\mu}_{5,\text{int}}(k)$ are the Green functions for the electromagnetic field (photon) and the field B_μ; and $\Gamma_\mu(k+p, p, k)$ and $\Gamma^5_\mu(k+p, p, k)$ are the vector and axial vector vertex functions, respectively.

The complete fermion propagator $G_{\text{int}}(k)$ may be represented as

$$G_{\text{int}}(k) = 1/[\hat{k} - \Sigma(k)], \quad \Sigma(k) = \alpha(k^2)\hat{k} + \beta(k^2),$$

$$\beta(k^2) = \frac{1}{4}\text{tr}[\Sigma(k)], \quad \hat{k} = \gamma^\mu k_\mu.$$

In the initial approximation, we use the fits for the fermion Green function $\beta(k^2) \approx m$ m is the electromagnetic fermion mass)

$$G_{\text{int}}(k) \approx \frac{1}{\hat{k} - m + i0}, \quad m = \frac{1}{4}\text{tr}[\Sigma(0)], \quad (13)$$

which are similar to those used in demonstrating DCSB in the Nambu–Jona-Lasinio model [5].

The following approximation may be used for the vertex function of the electromagnetic field:

$$\Gamma_\mu(k+p, k, p) \approx \gamma_\mu. \quad (14)$$

Since Lagrangian (4) is symmetric relative to the gauge transformations (5), the Ward axial identity may be written for the vertex function $\Gamma^5_\mu(p+k, k, p)$. In momentum representation, it appears as

$$k^\mu\Gamma^5_\mu(p+k, k, p) = G^{-1}_{\text{int}}(p+k)\gamma^5 + \gamma^5 G^{-1}_{\text{int}}(p). \quad (15)$$

When using fit (13) for the complete fermion propagator, we can find from Eq. (15) that

$$\Gamma^5_\mu(p+k, k, p) = \gamma_\mu\gamma^5 - \gamma^5 k_\mu\frac{\beta((p+k)^2) + \beta(p)}{k^2}$$
$$\approx \gamma_\mu\gamma^5 - \gamma^5\frac{2mk_\mu}{k^2}, \text{ when } k_\mu \longrightarrow 0. \quad (16)$$

If we expect $\beta(p^2)$ and m to be nonzero and to result from DCSB, then the second (pole) term in (16) does not vanish and the emergence of a pole at the vertex $\Gamma^5_\mu(p+k, k, p)$ for $k^2 = 0$ corresponds to the generation of massless Goldstone states in the theory.

However, if we use the propagator of the field B_μ in a transverse gauge similar to (11), then this pole term makes no contribution because of the transverse tensor structure of propagator (11), much as the interaction $B^\mu\partial_\mu\pi$ of the field B_μ with the Goldstone π fields is ineffective in the model with Lagrangian (6).

Based on what we have said above about the propagator of the field B_μ, it would be appropriate to use the fit

$$D^{\nu\mu}_{5,\text{int}}(k) = \frac{1}{(k^2 + i0)(1 - P^5(k^2)/k^2)}$$
$$\times\left(g^{\nu\mu} - \frac{k^\nu k^\mu}{k^2}\right) \longrightarrow \frac{1}{k^2 - M^2 + i0}$$
$$\times\left(g^{\nu\mu} - \frac{k^\nu k^\mu}{k^2}\right), \quad P^5(k^2) \approx M^2, \quad (17)$$

where $P^5(k^2)$ is the polarization operator of the particle (see Section 5) that corresponds to the field B_μ and M is the mass of the field B_μ generated by DCSB. Accord-

ingly, we may use a fit similar to (14) for the vertex $\Gamma^5_\mu(p+k, k, p)$,

$$\Gamma^5_\mu(p+k, k, p) \longrightarrow \gamma_\mu\gamma^5, \tag{18}$$

and the Landau gauge for the Green function of the electromagnetic field,

$$D^{\nu\mu}_{\text{int}}(k) = \frac{1}{(k^2+i0)(1-P(k^2)/k^2)}\left(g^{\nu\mu}-\frac{k^\nu k^\mu}{k^2}\right) \longrightarrow \frac{1}{k^2+i0}\left(g^{\nu\mu}-\frac{k^\nu k^\mu}{k^2}\right), \tag{19}$$

where $P(k^2)$ is the photon polarization operator.

Setting $p_\mu = 0$ in the Dyson equation (12) and substituting fits (13)–(19) for the Green and vertex functions into it, we obtain

$$\frac{1}{4}\text{tr}[\Sigma(p)] \xrightarrow{p=0} m$$

$$= -\frac{ie^2}{4}\text{tr}\left[\int\frac{\gamma_\nu(\hat{k}+m)\gamma_\mu}{(k^2-m^2+i0)(k^2+i0)} \times\left(g^{\nu\mu}-\frac{k^\nu k^\mu}{k^2}\right)\frac{d^4k}{(2\pi)^4}\right] \tag{20}$$

$$-\frac{ie_A^2}{4}\text{tr}\left[\int\frac{\gamma_\nu\gamma^5(\hat{k}+m)\gamma_\mu\gamma^5}{(k^2-m^2+i0)(k^2-M^2+i0)} \times\left(g^{\nu\mu}-\frac{k^\nu k^\mu}{k^2}\right)\frac{d^4k}{(2\pi)^4}\right].$$

The gauge $U(1)$ and $U_A(1)$ symmetries for Lagrangian (4) are equivalent. It would also be natural to assume for the ordinary charge e characterizing the vector interaction and the charge e_A characterizing the axial vector interaction that

$$e_A = \pm e. \tag{21}$$

It should also be noted that relation (21) corresponds to the condition for eliminating the axial anomalies that will be present in the model under consideration as in any theory with the pseudovector interactions $e_A\bar{\psi}(\gamma^\mu\gamma^5 B_\mu)\psi$ (see also formulas (25) and (26) in Section 4).

The sum of the integrals in (20) will converge under condition (21). Integration in (20) yields an equation that relates the masses m and M:

$$m = m\frac{3\alpha}{4\pi}\frac{x^2}{x^2-1}\ln x^2, \quad x^2 = \frac{M^2}{m^2}, \quad \alpha = \frac{e^2}{4\pi}. \tag{22}$$

Apart from the trivial solution $m = 0$, Eq. (22) has a nonzero solution for m at a nonzero value of M.

Thus, a nontrivial solution of Eq. (12) for the mass operator $\Sigma(p = 0) = m$ that corresponds to dynamical chiral symmetry breaking is possible for a nonzero mass of the vector field B_μ. In this case, calculating the integral in (20) does not require the introduction of an ultraviolet cutoff (as would be the case, for example, in the Nambu–Jona-Lasinio model [5]), and a nontrivial solution for $\Sigma(p = 0) = m$ at a nonzero value of M exists for any value of the coupling constant

$$\alpha = e^2/4\pi = e_A^2/4\pi = \alpha_A.$$

Whereas DCSB arises in the Nambu–Jona-Lasinio model at a coupling constant of the model

$$G \geq G_{\text{crit}} = 2\pi^2/\Lambda^2,$$

with the corresponding invariant cutoff parameter Λ^2, such a natural parameter as the mass of the field B_μ plays a role similar to that of Λ^2 in the model considered here.

By m in (13)–(22) we may also mean the sum of the seed mass m_0 inducing symmetry breaking and the variable mass m_{cb} reflecting the degree of symmetry breaking:

$$m = m_0 + m_{cb}.$$

By analogy, we obtain for the mass of the field B_μ

$$M = M_0 + M_{cb};$$

the parameters that induce symmetry breaking, m_0 and M_0, are independent and tend to zero at the end of the calculations (in accordance with the Bogolyubov method when analyzing the DCSB; see, e.g., [6]): $m_0 \longrightarrow 0$ and $M_0 \longrightarrow 0$. If $m_{cb} = m$ and $M_{cb} = M$ are nonzero when passing to a zero limit of the seed parameters, then dynamical chiral symmetry breaking may be said to be present.

4. AXIAL ANOMALY

It follows from the symmetry of Lagrangian (4) (or (6)) relative to transformations (5) (or (7)) that the corresponding axial current is conserved,

$$\partial^\mu J^5_\mu = 0, \tag{23}$$

and that the corresponding Ward identity (15) is valid.

However, when the theory is quantized, equality (23) breaks down due to the triangular loop diagrams with the axial vertices $e_A\bar{\psi}(\gamma^\mu\gamma^5 B_\mu)\psi$, with three axial vertices (loop *BBB* in Fig. 1), and with one axial vertex (loop *AAB* in Fig. 2). The anomaly emerges, because the divergences cannot be removed from these diagrams without explicit symmetry breaking, and exp-

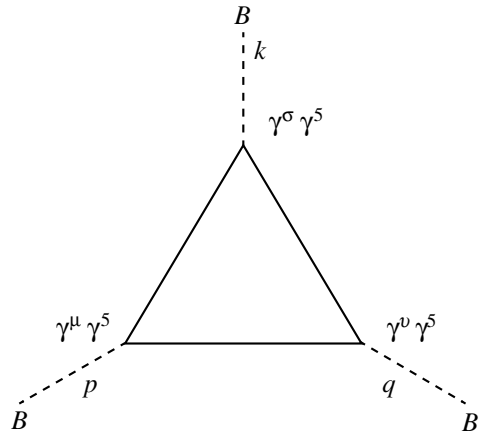


Fig. 1.

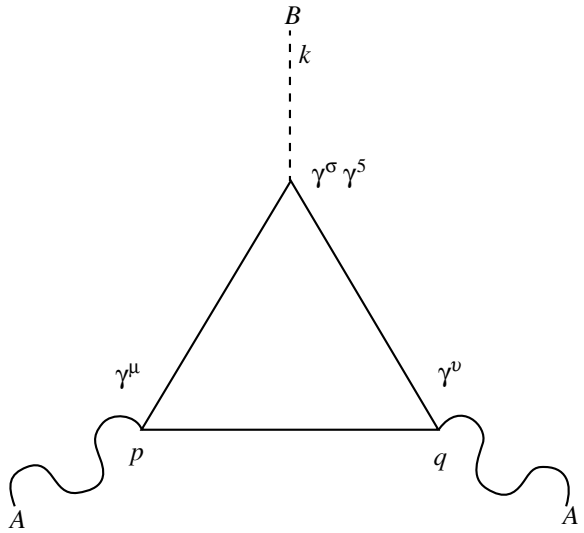


Fig. 2.

ression (23) for the model under consideration will appear as

$$\partial^\mu J^5_\mu = \frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} + \frac{e_A^2}{48\pi^2}\varepsilon^{\mu\nu\rho\sigma}G_{\mu\nu}G_{\rho\sigma}. \quad (24)$$

The emergence of anomalies leads to a number of problems with the B_μ quantization and the renormalizability of the theory [7]. In the model under consideration, the axial interaction is not an external object and the anomaly causes the self-consistency of the theory to be broken.

Including other types of fermions in the theory allows the anomalous components that emerge in the loops in Figs. 1 (loop *BBB*) and 2 (loop *AAB*) to be eliminated. It would be appropriate to treat the fermions included in the model under consideration in accordance with the standard-model fermion generations. Within the first generation of charged fermions of the standard model (*e, u, d*) (i.e., with an electron and two quarks) with charges

$$Q = e\left(-1, \frac{2}{3}, -\frac{1}{3}\right), \quad (25)$$

the anomalies can be eliminated if the following relation is valid for the coupling constants at axial vertices:

$$Q_A = e\left(\pm 1, \mp\frac{2}{3}, \mp\frac{1}{3}\right). \quad (26)$$

Given the three color degrees of freedom of the quarks ($N_c = 3$), the anomalous components of the amplitudes that correspond to both loop *BBB* (Fig. 1) and loop *AAB* (Fig. 2) for all charged fermions vanish (in the massless case, the sum of the diagrams for all fermions with the quark colors becomes equal to zero). This anomaly elimination condition corresponds to assumption (21). Clearly, all of the aforesaid also applies to other generations. The neutrinos as particles that have no charge, but that are members of the standard-model generations, are disregarded; it would be natural to set $e_A = 0$ for all neutrinos. It should also be noted that the anomalies are eliminated precisely within the standard-model generations.

5. THE POLARIZATION OPERATOR OF THE FIELD B_μ AND CHIRAL SYMMETRY BREAKING

The complete Green function $D^{\nu\mu}_{5,\,\mathrm{int}}(k)$ for the field B_μ and the propagator for free particles (in the transverse Landau gauge)

$$D_5^{\nu\mu}(k) = \frac{1}{k^2 + i0}\left(g^{\nu\mu} - \frac{k^\nu k^\mu}{k^2}\right)$$

can be related by

$$D^{\nu\mu}_{5,\,\mathrm{int}}(k) = D_5^{\nu\mu}(k) + D_5^{\nu\theta}(k)P^5_{\theta\rho}(k)D^{\rho\mu}_{5,\,\mathrm{int}}(k),$$
$$D^{\nu\mu}_{5,\,\mathrm{int}}(k) = D_5(k^2)\left(g^{\nu\mu} - \frac{k^\nu k^\mu}{k^2}\right), \quad (27)$$

where $P^5_{\mu\nu}(k)$ is the (axial) polarization operator. The tensor structure of the axial polarization operator (in any case, for the gauge used) may also be assumed to be transverse (i.e., we may consider only the transverse projection of the operator $P^5_{\mu\nu}(k)$; the longitudinal parts, if they exist, make no contribution to the complete Green function in this gauge):

$$P^5_{\mu\nu}(k) = P^5(k)\left(g^{\nu\mu} - \frac{k^\nu k^\mu}{k^2}\right). \quad (28)$$

From (27), we may derive

$$D_5(k^2) = \frac{1}{k^2 + i0} + \frac{1}{k^2 + i0}P^5(k^2)D_5(k^2),$$
$$D_5(k^2) = \frac{1}{(k^2 + i0)(1 - P^5(k^2)/k^2)}. \quad (29)$$

The following Dyson equation similar to (12) is valid for the operator $P^5_{\mu\nu}(k)$:

$$P^5_{\mu\nu}(k)$$
$$= ie_A^2\int \mathrm{tr}[\gamma_\mu\gamma^5 G_{\mathrm{int}}(p+k)\Gamma^5_\nu(p+k, k, p)G_{\mathrm{int}}(p)] \quad (30)$$
$$\times\frac{d^4p}{(2\pi)^4}.$$

Since divergences of the loop integrals arise when calculating $P^5_{\mu\nu}(k)$ in (30), it would be appropriate to determine the normalization conditions. Assuming that the complete propagator for DCSB must describe the propagation of a massive vector particle of the field B_μ, the first normalization condition is

$$P^5(M^2) = M^2, \quad (31)$$

which corresponds to the fit of the complete Green function using in (17) when solving Dyson equation (12).

We redefine the scalar part of the polarization operator as follows:

$$P^5(k^2) = M^2 + (k^2 - M^2)\tilde{P}^5(k^2), \quad (32)$$

where M is the mass acquired by the field B_μ due to chiral symmetry breaking. The complete Green function appears as

$$D^{5,\mathrm{int}}_{\mu\nu} = \frac{1}{(k^2 - M^2 + i0)(1 - \tilde{P}^5(k^2))} \times\left(g_{\nu\mu} - \frac{k_\nu k_\mu}{k^2}\right). \quad (33)$$

The second normalization condition related to the normalization of the axial charge e_A is

$$\tilde{P}^5(0) = 0, \quad (34)$$

or, equivalently,

$$P^5(0) = P^5(M^2) = M^2. \quad (35)$$

Calculating a loop integral similar to (30) requires using regularization. In the presence of gauge $U(1)$ invariance, the corresponding axial current is conserved (for the time being, we disregard the axial anomalies), and the tensor structure of the operator $P^5_{\mu\nu}(k)$ must have only transverse components. However, the commonly used regularization methods (e.g., cutoff regularization) themselves distort the tensor structure of $P^5_{\mu\nu}(k)$ even in the presence of gauge invariance.

For the two normalization conditions, (31) and (35), to be satisfied, the scalar part of the transverse projection of the polarization operator $P^5(k^2)$ after regularization must be

$$P^5(k^2) = c_1M^2 + k^2(c_2 + F(k^2)) + M^2F_1(k^2), \quad (36)$$

where c_1 and c_2 are constants that generally contain divergences when removing regularization, and $F(k^2)$ and $F_1(k^2)$ are finite functions. If there were gauge invariance and if the regularization procedure did not distort the tensor structure of $P^5_{\mu\nu}(k)$, then the constant c_1 (and the function $F_1(k^2)$) would be equal to zero,

$$P^5(k^2) = k^2(c_2 + K(k^2)),$$

and it would be impossible to simultaneously satisfy both conditions, (31) and (35). Physically, this would imply that the B_μ photons remain massless.

Introducing the mass m (or the Bogolyubov seed mass m_0, because m and M are generally independent parameters) in the fermion propagator, we obtain $U_A(1)$ symmetry and gauge invariance breaking. In this case, the operator $P^5_{\mu\nu}(k^2)$ loses its transverse structure, but this loss of transversality will be controllable by the breaking parameter m, while $P^5(k^2)$ after the separation of the transverse part may be represented as (36). Since the regularization procedure can also distort the tensor structure of $P^5_{\mu\nu}(k)$, we impose a condition on the constant c_1,

$$\lim_{m\to 0} c_1 \longrightarrow 0, \quad (37a)$$

and, similarly for the function $F_1(k^2)$,

$$\lim_{m\to 0} F_1(k^2) \longrightarrow 0; \quad (37b)$$

i.e., the components of c_1 ($F_1(k^2)$) that do not satisfy this condition arise in the calculations from the regularization procedure.

In general, using the standard regularization methods that preserve the tensor structure of $P^5_{\mu\nu}(k)$ (the dimensional regularization method or the Pauli–Villars method for fermion loops) is not logically consistent.

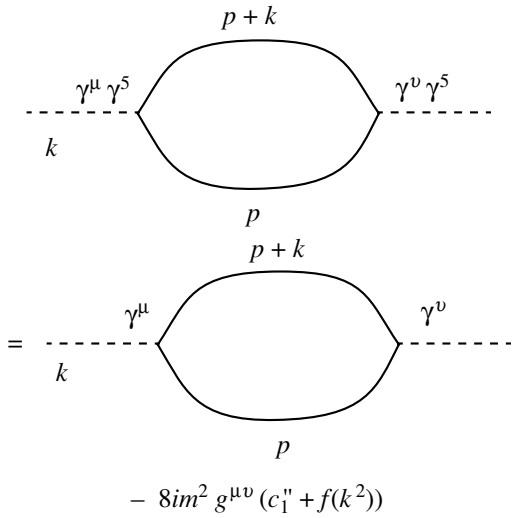


Fig. 3.

The dimensional regularization method runs into problems when determining the matrix γ^5. The Pauli–Villars method for fermion loops requires introducing additional regularizing fermion loops with masses that are regularization parameters that explicitly break the axial gauge $U_A(1)$ invariance, which, in our case, makes it identical to simpler methods like the cutoff regularization method.

Although using the dimensional regularization method is not quite consistent, we will use it to show how the terms corresponding to DCSB emerge in the operator $P^5_{\mu\nu}(k)$. To this end, we note that, in fermion loops with an even number of axial vertices, all of the matrices γ^5 in integral expression (30) may be rearranged in such a way that they will be close and, as a result, will disappear:

$$P^5_{\mu\nu}(k)$$

$$\sim e_A^2\int \mathrm{tr}\left[\gamma_\mu\gamma^5\frac{\hat p+\hat k+m}{(p+k)^2-m^2}\gamma_\nu\gamma^5\frac{\hat p+m}{p^2-m^2}\right]\frac{d^4p}{(2\pi)^4}$$

$$= e_A^2\int \mathrm{tr}\left[\gamma_\mu\frac{\hat p+\hat k+m}{(p+k)^2-m^2}\gamma_\nu\frac{\hat p-m}{p^2-m^2}\right]\frac{d^4p}{(2\pi)^4} \qquad (38)$$

$$= e_A^2\int \mathrm{tr}\left[\gamma_\mu\frac{\hat p+\hat k+m}{(p+k)^2-m^2}\gamma_\nu\frac{\hat p+m}{p^2-m^2}\right]\frac{d^4p}{(2\pi)^4}$$

$$-8m^2e_A^2\int\frac{g_{\mu\nu}}{((p+k)^2-m^2)(p^2-m^2)}\frac{d^4p}{(2\pi)^4}.$$

In expression (38), the integral that formally corresponds to an ordinary vector fermion loop and the addition to it proportional to m^2 appear in the last part of the equality. This equality may be represented in diagram form, as shown in Fig. 3 ($f(k^2)$ is a finite scalar function, and c_1'' is a constant that generally contains divergences). In the massless case, the expressions for loops with vector and axial vector vertices are identical. This identity reflects the facts that there is a symmetry relative to both gauge transformations (5), that the vector and axial vector currents are conserved ($\partial_\mu J_\mu = \partial_\mu J^5_\mu = 0$, $m = 0$), and that both fields A_μ and B_μ are massless. Applying dimensional regularization to the integrals on the right-hand side of the equality in Fig. 3 yields

$$P^5_{\mu\nu}(k)$$

$$= i\mu^{4-d}e_A^2\int \mathrm{tr}\left[\gamma_\mu\frac{\hat p+\hat k+m}{(p+k)^2-m^2}\gamma_\nu\frac{\hat p+m}{p^2-m^2}\right]\frac{d^dp}{(2\pi)^d} \qquad (39)$$

$$-8i\mu^{4-d}m^2e_A^2\int\frac{g_{\mu\nu}}{((p+k)^2-m^2)(p^2-m^2)}\frac{d^dp}{(2\pi)^d},$$

where μ is the mass operator that restores the correct dimensions of the polarization operator; regularization is removed for $d \longrightarrow 4$. The first integral (vector loop) on the right-hand side of equality (39) has a transverse tensor structure owing to the properties of the dimensional regularization, and expression (39) may be written as

$$P^5_{\mu\nu}(k)$$

$$= \frac{e_A^2}{12\pi^2}k^2\left(g_{\mu\nu}-\frac{k_\mu k_\nu}{k^2}\right)\left(c_2'+I\left(\frac{k^2}{4m^2},0\right)\right)$$

$$+ g_{\mu\nu}m^2\frac{e_A^2}{12\pi^2}\left(c_1'+I_1\left(\frac{k^2}{4m^2},0\right)\right), \qquad (40)$$

$$I(t,u) = 6\int_0^1 x(1-x)\ln\left(\frac{1-4x(1-x)t}{1-4x(1-x)u}\right)dx,$$

$$I_1(t,u) = 6\int_0^1 \ln\left(\frac{1-4x(1-x)t}{1-4x(1-x)u}\right)dx,$$

where c_1' and c_2' are constants that contain divergences when removing regularization ($d \longrightarrow 4$), and we have for the $P^5(k^2)$ projection

$$P^5(k^2) = m^2\frac{e_A^2}{12\pi^2}\left(c_1'+I_1\left(\frac{k^2}{4m^2},0\right)\right)$$
$$+\frac{e_A^2}{12\pi^2}k^2\left(c_2'+I\left(\frac{k^2}{4m^2},0\right)\right). \qquad (41)$$

According to (36), it allows both normalization condi-

tions, (31) and (35), to be satisfied:

$$c_1 \sim m^2 \frac{e_A^2}{2\pi^2 M^2}\left(\frac{2}{4-d} - \gamma + \ln\left(\frac{m^2}{\mu^2}\right)\right.$$

$$\left. + \frac{1}{6} I_1\left(\frac{\lambda^2}{4m^2}, 0\right)\right), \qquad (42)$$

$$F_1(k^2) \sim m^2 \frac{e_A^2}{12\pi^2 M^2} I_1\left(\frac{k^2}{4m^2}, \frac{\lambda^2}{4m^2}\right),$$

for $d \longrightarrow 4$, γ is the Euler constant, and λ is the normalization point. Condition (37) is also satisfied.

The expansion of the axial loop presented as a diagram in Fig. 3 is convenient for separating out the components related to the breaking of gauge $U_A(1)$ invariance (5) and to DCSB for any regularization method. The first term in Fig. 3 formally corresponds to the polarization vector of the electromagnetic field A_μ, and the vector current is conserved. Therefore, the final result after the separation of divergences and the removal of regularization for the first term in Fig. 3 must be transverse and have the structure

$$k^2(c_2 + F(k^2))(g_{\mu\nu} - k_\mu k_\nu / k^2),$$

where $F(k^2)$ is a finite function, and the components corresponding to DCSB with a $g_{\mu\nu}$-type tensor structure remain in the second term, which clearly corresponds to condition (37).

Thus, for an arbitrary regularization procedure, the terms without the above structure in the vector loop of the first expansion term in Fig. 3 are assumed to be unphysical and introduced by the regularization procedure. They introduce no problems in the theory and in the corresponding divergences and can be removed by introducing appropriate counterterms. However, in contrast to the addition related to the second term on the right-hand side of the expansion in Fig. 3, these terms have no physical meaning.

When analyzing DCSB, one should use an expression of form (16) with the following pole term of the Goldstone state in (30) for the vertex function:

$$\Gamma_\mu^5(p + k, k, p) \approx \gamma_\mu \gamma^5 - \gamma^5 2mk_\mu / k^2.$$

This nontrivial addition to the vertex $\Gamma_\mu^5(p + k, k, p)$ disrupts the transverse tensor structure of the operator $P_{\mu\nu}^5(k^2)$, and the integral for the Goldstone addition $2mk_\mu/k^2$ diverges when regularization is removed. However, this addition to the polarization operator $P_{\mu\nu}^5(k^2)$ has a purely longitudinal structure; when separating out the transverse part proportional to $P^5(k^2)$, it gives no contribution to the propagator of the field B_μ.

In accordance with conditions (31) and (35), we may obtain for the transverse projection of $P_{\mu\nu}^5(k)$

$$P^5(k^2) = M^2$$

$$+ (k^2 - M^2)\frac{\alpha_A}{3\pi}\left(k^2 \frac{I\left(\frac{k^2}{4m^2}, \frac{M^2}{4m^2}\right)}{(k^2 - M^2)}\right.$$

$$\left. - \frac{m^2}{M^2}\frac{k^2 I_1\left(\frac{M^2}{4m^2}, 0\right) - M^2 I_1\left(\frac{k^2}{4m^2}, 0\right)}{(k^2 - M^2)}\right),$$

$$\tilde{P}^5(k^2) = \frac{\alpha_A}{3\pi}\left(k^2 \frac{I\left(\frac{k^2}{4m^2}, \frac{M^2}{4m^2}\right)}{(k^2 - M^2)}\right. \qquad (43)$$

$$\left. - \frac{m^2}{M^2}\frac{k^2 I_1\left(\frac{M^2}{4m^2}, 0\right) - M^2 I_1\left(\frac{k^2}{4m^2}, 0\right)}{(k^2 - M^2)}\right),$$

$$I(t, u) = 6\int_0^1 x(1-x)\ln\left(\frac{1 - 4x(1-x)t}{1 - 4x(1-x)u}\right)dx,$$

$$I_1(t, u) = 6\int_0^1 \ln\left(\frac{1 - 4x(1-x)t}{1 - 4x(1-x)u}\right)dx.$$

For the proper energy operator of the B_μ photon

$$\Pi_{\mu\nu}^5(x - x') = ie_A^2 \left\langle 0 \middle| T(J_\mu^5(x) J_\nu^5(x')) \middle| 0 \right\rangle$$

in momentum representation, we have

$$D_{5,\,\mathrm{int}}^{\nu\mu}(k) = D_5^{\nu\mu}(k) + D_5^{\nu\theta}(k)\Pi_{\theta\rho}^5(k) D_5^{\rho\mu}(k),$$

$$D_5^{\nu\mu}(k) = \frac{1}{k^2 + i0}\left(g^{\nu\mu} - \frac{k^\nu k^\mu}{k^2}\right). \qquad (44)$$

By analogy with the polarization operator, we may consider only the transverse projection for $\Pi_{\mu\nu}^5(k)$,

$$\Pi_{\mu\nu}^5(k) = \Pi^5(k^2)(g_{\mu\nu} - k_\mu k_\nu / k^2).$$

We obtain for it

$$\Pi^5(k^2) = \frac{k^2 P^5(k^2)}{k^2 + i0 - P^5(k^2)},$$
$$\Pi^5(k^2) = \frac{k^2(M^2 + (k^2 - M^2)\tilde{P}^5(k^2))}{(k^2 - M^2 + i0)(1 - \tilde{P}^5(k^2))}; \quad (45)$$

i.e., $\Pi^5_{\mu\nu}(k)$ also has a pole that corresponds to the massive B_μ photon exchange in momentum representation.

6. CONCLUSIONS

Thus, based on our model for the electrodynamics of massless fermions, which, apart from the gauge field A_μ corresponding to $U(1)$ symmetry, also includes the field B_μ corresponding to chiral gauge $U_A(1)$ symmetry, we have shown that both the fermions and the pseudovector gauge field B_μ can acquire masses. We have demonstrated that the Dyson equation for the fermion mass corresponding to DCSB when the mass of the gauge field B_μ is generated can have a nontrivial solution, and, when the fermion mass is generated through DCSB, the field B_μ can acquire mass.

The massiveness of the field B_μ that results from DCSB is consistent with the existence of only one type of massless particles—the electromagnetic interaction carriers (photons). The removal of anomalies in this QED model, especially within each standard-model fermion generation, also places it in the class of models that can be physically adequate.

Condition (21) for the vector and axial vector coupling constants, which corresponds to the condition for eliminating the axial anomalies in the model under consideration, allows us to dispense with the ultraviolet cutoff when working with the Dyson equation and when analyzing its solutions. The mass of the field B_μ acts as a cutoff factor, while the massiveness of the field B_μ itself is the condition for the existence of a nontrivial solution to the Dyson equation that corresponds to dynamical chiral symmetry breaking.

If the B_μ photons are assumed to be physical particles, then they must be assumed to very massive in order not to come into conflict with the actual experimental situation: according to (22), we have for $M \gg m$

$$\frac{M}{m} \propto \exp\left(\frac{8\pi^2}{3e^2}\right).$$

At the same time, it should be noted that not only electromagnetic, but also other types of interactions can significantly affect the masses of real particles. Thus, actual physical numerical mass estimates for the B_μ photons require further studies.

REFERENCES

1. P. I. Fomin, V. P. Gusynin, V. A. Miransky, and Yu. A. Sitenko, Riv. Nuovo Cimento **6**, 1 (1983).
2. T. Maskawa and H. Nakajima, Prog. Theor. Phys. **52**, 1326 (1974).
3. V. A. Miransky, Phys. Lett. B **91B**, 421 (1980).
4. V. P. Gusynin and V. A. Kushnir, Preprint No. 89-81E, ITF AN SSSR (Inst. for Theoretical Physics, USSR Academy of Sciences, Moscow, 1990).
5. W. Weise, in *Quarks and Nuclei, International Review of Nuclear Physics* (World Sci., Singapore, 1984), Vol. 1, p. 57; Y. Nambu and G. Jona-Lasinio, Phys. Rev. **122**, 345 (1961).
6. V. E. Rochev, Preprint No. 2001-40, IFVÉ (Inst. for High Energy Physics, Protvino, Moscow oblast, 2001).
7. S. Adler, Phys. Rev. **117**, 2426 (1969); S. Bell and R. Jakiw, Nuovo Cimento A **60**, 47 (1969); L. D. Faddeev and A. A. Slavnov, *Gauge Fields: Introduction to Quantum Theory*, 2nd ed. (Nauka, Moscow, 1988; Addison-Wesley, Redwood City, Calif., 1990).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 661–666.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 755–760.

NUCLEI, PARTICLES, AND THEIR INTERACTION

Single Z' Production at Compact Linear Collider Based on e–γ Collisions¶

D. V. Soa[a], H. N. Long[b], D. T. Binh[c], and D. P. Khoi[c,d]

[a]*Department of Physics, Hanoi University of Education, Hanoi, Vietnam*
[b]*Physics Division, NCTS, National Tsing Hua University, Hsinchu, Taiwan;*
e-mail: nlhoang@phys.cts.nthu.edu.tw; on leave from Institute of Physics, NCST, Hanoi, Vietnam
[c]*Institute of Physics, NCST, Hanoi, Vietnam*
[d]*Department of Physics, Vinh University, Vinh, Vietnam*

Received October 10, 2003

Abstract—We analyze the potential of the compact linear collider (CLIC) based on e–γ collisions to search for the new Z' gauge boson. Single Z' production on e–γ colliders in two $SU(3)_C \otimes SU(3)_L \otimes U(1)_N$ models, the minimal model and the model with right-handed neutrinos is studied in detail. The results show that new Z' gauge bosons can be observed on the CLIC and that the cross sections in the model with right-handed neutrinos are bigger than those in the minimal one. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Neutral gauge structures beyond the photon and the Z boson have long been considered one of the best motivated extensions of the standard model of electroweak interactions. They have been predicted in many models going beyond the standard one, including the models based on the $SU(3)_C \otimes SU(3)_L \otimes U(1)_N$ (3–3–1) gauge group [1–5]. These models have some interesting characteristics. First, they predict three families of quarks and leptons if the QCD asymptotic freedom is imposed. Second, the Peccei–Quinn symmetry naturally occurs in these models [6]. Finally, it is characteristic of these models that one generation of quarks is treated differently from the other two. This might lead to a natural explanation for the unbalancing heavy top quark.

The Z' gauge boson is a necessary element of various models that extend the Standard Model. In general, the extra Z' boson may not couple in a universal way. There are, however, strong constraints from flavor-changing neutral current processes that specifically limit the nonuniversality between the first two generations. Lower bounds on the mass of Z' following from analysis of a variety of popular models are found to be in the energy range 500–2000 GeV [7, 8].

It was suggested recently that the 3–3–1 models arise naturally from the gauge theories in spacetime with extra dimensions [9] where the scalar fields are the components in additional dimensions [10]. A few different versions of the 3–3–1 model have been proposed [11].

Recent investigations have indicated that signals of new gauge bosons in models may be observed on the CERN large hadron collider [12] or the next linear collider [13, 14]. In [15], two of us have considered single production of the bilepton and shown that several thousand events are expected at the integrated luminosity $L \approx 9 \times 10^4$ fb^{-1}. In this work, single production of the new Z' gauge boson in the 3–3–1 models is considered. The paper is organized as follows. In Section 2, we give a brief review of two models: the relation among real physical bosons and constraints on their masses. Section 3 is devoted to single production of the Z' boson in the e–γ collisions. Discussion is presented in Section. 4.

2. REVIEW OF 3–3–1 MODELS

To put the context in a proper frame, it is appropriate to briefly recall some relevant features of the two types of 3–3–1 models: the minimal model proposed by Pisano, Pleitez, and Frampton [1, 2], and the model with right-handed neutrinos [4, 5].

2.1. The Minimal 3–3–1 Model

The model treats the leptons as the $SU(3)_L$ antitriplets [1, 2, 16],[1]

$$f_{aL} = \begin{pmatrix} e_{aL} \\ -\nu_{aL} \\ (e^c)_a \end{pmatrix} \approx (1, \bar{3}, 0), \qquad (1)$$

where $a = 1, 2, 3$ is the generation index. Two of the three quark generations transform as triplets, and the

¶This article was submitted by the authors in English.

[1] The leptons may be assigned to a triplet as in [1]; the two models are mathematically identical, however.

third generation is treated differently. It belongs to an antitriplet,

$$Q_{iL} = \begin{pmatrix} u_{iL} \\ d_{iL} \\ D_{iL} \end{pmatrix} \approx \left(3, 3, -\frac{1}{3}\right), \quad (2)$$

$$u_{iR} \approx (3, 1, 2/3), \quad d_{iR} \approx (3, 1, -1/3),$$
$$D_{iR} \approx (3, 1, -1/3), \quad i = 1, 2,$$

$$Q_{3L} = \begin{pmatrix} d_{3L} \\ -u_{3L} \\ T_L \end{pmatrix} \approx (3, \bar{3}, 2/3), \quad (3)$$

$$u_{3R} \approx (3, 1, 2/3), \quad d_{3R} \approx (3, 1, -1/3),$$
$$T_R \approx (3, 1, 2/3).$$

The nine gauge bosons $W^a (a = 1, 2, \ldots, 8)$ and B of $SU(3)_L$ and $U(1)_N$ are split into four light gauge bosons and five heavy gauge bosons after $SU(3)_L \otimes U(1)_N$ has been broken into $U(1)_Q$. The light gauge bosons are those of the Standard Model: the photon (A), Z_1, and $W^\pm$. The remaining five correspond to new heavy gauge bosons Z_2, $Y^\pm$ and doubly charged bileptons $X^{\pm\pm}$. They are expressed in terms of W^a and B as [16]

$$\sqrt{2}W_\mu^+ = W_\mu^1 - iW_\mu^2, \quad \sqrt{2}Y_\mu^+ = W_\mu^6 - iW_\mu^7,$$
$$\sqrt{2}X_\mu^{++} = W_\mu^4 - iW_\mu^5, \quad (4)$$

$$A_\mu = s_W W_\mu^3 + c_W(\sqrt{3}t_W W_\mu^8 + \sqrt{1 - 3t_W^2}B_\mu),$$

$$Z_\mu = c_W W_\mu^3 - s_W(\sqrt{3}t_W W_\mu^8 + \sqrt{1 - 3t_W^2}B_\mu), \quad (5)$$

$$Z'_\mu = -\sqrt{1 - 3t_W^2}W_\mu^8 + \sqrt{3}t_W B_\mu,$$

where we use the notation

$$c_W \equiv \cos\theta_W, \quad s_W \equiv \sin\theta_W, \quad t_W \equiv \tan\theta_W.$$

The physical states are a mixture of Z and Z',

$$Z_1 = Z\cos\phi - Z'\sin\phi,$$

$$Z_2 = Z\sin\phi + Z'\cos\phi,$$

where ϕ is the mixing angle.

Symmetry breaking and fermion mass generation can be achieved by three scalar $SU(3)_L$ triplets Φ, Δ, Δ' and a sextet η,

$$\Phi = \begin{pmatrix} \phi^{++} \\ \phi^{+} \\ \phi^{',0} \end{pmatrix} \approx (1, 3, 1),$$

$$\Delta = \begin{pmatrix} \Delta_1^+ \\ \Delta^0 \\ \Delta_2^- \end{pmatrix} \approx (1, 3, 0),$$

$$\Delta' = \begin{pmatrix} \Delta'^0 \\ \Delta'^- \\ \Delta'^{--} \end{pmatrix} \approx (1, 3, -1),$$

$$\eta = \begin{pmatrix} \eta_1^{++} & \eta_1^+/\sqrt{2} & \eta^0/\sqrt{2} \\ \eta_1^+/\sqrt{2} & \eta'^0 & \eta_2^-/\sqrt{2} \\ \eta^0/\sqrt{2} & \eta_2^-/\sqrt{2} & \eta_2^{--} \end{pmatrix} \approx (1, 6, 0).$$

The sextet η is necessary to give masses to charged leptons [3, 16]. The vacuum expectation value

$$\langle \Phi^T \rangle = (0, 0, u/\sqrt{2})$$

yields masses for exotic quarks, the heavy neutral gauge boson Z', and two new charged gauge bosons X^{++}, Y^+. The masses of the standard gauge bosons and the ordinary fermions are related to the vacuum expectation values of the other scalar fields,

$$\langle \Delta^0 \rangle = v/\sqrt{2}, \quad \langle \Delta'^0 \rangle = v'/\sqrt{2},$$

$$\langle \eta^0 \rangle = \omega/\sqrt{2}, \quad \langle \eta'^0 \rangle = 0.$$

For consistency with low-energy phenomenology, the mass scale of $SU(3)_L \otimes U(1)_N$ breaking must be much larger than that of the electroweak scale, i.e, $u \gg v$, v', ω. The masses of gauge bosons are explicitly given by

$$m_W^2 = \frac{1}{4}g^2(v^2 + v'^2 + \omega^2),$$

$$M_Y^2 = \frac{1}{4}g^2(u^2 + v^2 + \omega^2), \quad (6)$$

$$M_X^2 = \frac{1}{4}g^2(u^2 + v'^2 + 4\omega^2),$$

and

$$m_Z^2 = \frac{g^2}{4c_W^2}(v^2 + v'^2 + \omega^2) = \frac{m_W^2}{c_W^2},$$

$$M_{Z'}^2 = \frac{g^2}{3}\left[\frac{c_W^2}{1-4s_W^2}u^2 + \frac{1-4s_W^2}{4c_W^2}(v^2 + v'^2 + \omega^2) + \frac{3s_W^2}{1-4s_W^2}v'^2\right]. \quad (7)$$

Expressions in (6) yield a splitting between the bilepton masses [17],

$$\left|M_X^2 - M_Y^2\right| \le 3m_W^2. \quad (8)$$

Combining the constraints from direct searches and neutral currents, we obtain the range for the mixing angle [16] as

$$-1.6\times 10^{-2} \le \phi \le 7\times 10^{-4}$$

and a lower bound on M_{Z_2}

$$M_{Z_2} \ge 1.3 \text{ TeV}.$$

Such a small mixing angle can be safely neglected. In that case, Z_1 and Z_2 are the Z boson in the Standard Model and the extra Z' gauge boson, respectively. With the new atomic parity violation in cesium, we obtain a lower bound for the Z_2 mass [18]:

$$M_{Z_2} > 1.2 \text{ TeV}.$$

2.2. The Model with Right-Handed Neutrinos

In this model, the leptons are in triplets and the third member is a right-handed neutrino [4, 5],

$$f_{aL} = \begin{pmatrix} \nu_{aL} \\ e_{aL} \\ (\nu_L^c)_a \end{pmatrix} \approx (1, 3, -1/3), \quad e_{aR} \approx (1, 1, -1). \quad (9)$$

The first two generations of quarks are in antitriplets and the third one is in a triplet,

$$Q_{iL} = \begin{pmatrix} d_{iL} \\ -u_{iL} \\ D_{iL} \end{pmatrix} \approx (3, \bar{3}, 0), \quad (10)$$

$$u_{iR} \approx (3, 1, 2/3), \quad d_{iR} \approx (3, 1, -1/3), \quad D_{iR} \approx (3, 1, -1/3), \quad i = 1, 2,$$

$$Q_{3L} = \begin{pmatrix} u_{3L} \\ d_{3L} \\ T_L \end{pmatrix} \approx (3, 3, 1/3), \quad (11)$$

$$u_{3R} \approx (3, 1, 2/3), \quad d_{3R} \approx (3, 1, -1/3), \quad T_R \approx (3, 1, 2/3).$$

The doubly charged bileptons of the minimal model are here replaced by complex neutral ones as

$$\sqrt{2}W_\mu^+ = W_\mu^1 - iW_\mu^2, \quad \sqrt{2}Y_\mu^- = W_\mu^6 - iW_\mu^7, \quad \sqrt{2}X_\mu^0 = W_\mu^4 - iW_\mu^5. \quad (12)$$

The physical neutral gauge bosons are again related to Z and Z' through the mixing angle ϕ. Together with the photon, they are defined as [5]

$$A_\mu = s_W W_\mu^3 + c_W\left(-\frac{t_W}{\sqrt{3}}W_\mu^8 + \sqrt{1-\frac{t_W^2}{3}}B_\mu\right),$$

$$Z_\mu = c_W W_\mu^3 - s_W\left(-\frac{t_W}{\sqrt{3}}W_\mu^8 + \sqrt{1-\frac{t_W^2}{3}}B_\mu\right), \quad (13)$$

$$Z'_\mu = \sqrt{1-\frac{t_W^2}{3}}W_\mu^8 + \frac{t_W}{\sqrt{3}}B_\mu.$$

Symmetry breaking can be achieved with just three $SU(3)_L$ triplets,

$$\chi = \begin{pmatrix} \chi^0 \\ \chi^- \\ \chi'^0 \end{pmatrix} \approx (1, 3, -1/3), \quad (14)$$

$$\rho = \begin{pmatrix} \rho^+ \\ \rho^0 \\ \rho'^+ \end{pmatrix} \approx (1, 3, 2/3), \quad (15)$$

$$\eta = \begin{pmatrix} \eta^0 \\ \eta^- \\ \eta'^0 \end{pmatrix} \approx (1, 3, -1/3). \quad (16)$$

The necessary vacuum expectation values are

$$\langle\chi\rangle^T = (0, 0, \omega/\sqrt{2}), \quad \langle\rho\rangle^T = (0, u/\sqrt{2}, 0), \quad \langle\eta\rangle^T = (v/\sqrt{2}, 0, 0). \quad (17)$$

The vacuum expectation value $\langle\chi\rangle$ generates masses for the exotic 2/3 and –1/3 quarks, while the values $\langle\rho\rangle$ and $\langle\eta\rangle$ generate masses for all ordinary leptons and

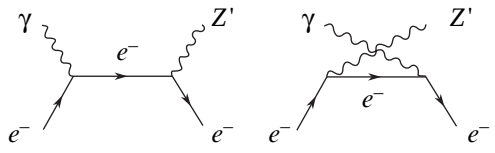


Fig. 1. Feynman diagrams for the reaction $e^-\gamma \longrightarrow Z'e^-$.

quarks. After symmetry breaking, the gauge bosons gain masses as

$$m_W^2 = \frac{1}{4}g^2(u^2+v^2), \quad M_Y^2 = \frac{1}{4}g^2(v^2+\omega^2), \\ M_X^2 = \frac{1}{4}g^2(u^2+\omega^2), \tag{18}$$

and

$$m_Z^2 = \frac{g^2}{4c_W^2}(u^2+v^2) = \frac{m_W^2}{c_W^2}, \tag{19}$$

$$M_{Z'}^2 = \frac{g^2}{4(3-4s_W^2)}\left[4\omega^2 + \frac{u^2}{c_W^2} + \frac{v^2(1-2s_W^2)^2}{c_W^2}\right]. \tag{20}$$

For consistency with low-energy phenomenology, we have to assume that $\langle\chi\rangle \gg \langle\rho\rangle, \langle\eta\rangle$, such that $m_W \ll M_X, M_Y$.

The symmetry-breaking hierarchy gives us a splitting between the bilepton masses [19]

$$\left|M_X^2 - M_Y^2\right| \le m_W^2. \tag{21}$$

It is therefore acceptable to set $M_X \approx M_Y$.

The constraint on the $Z - Z'$ mixing based on the Z decay is given in [5],

$$-2.8\times10^{-3} \le \phi \le 1.8\times10^{-4};$$

in this model, we do not have a limit for $\sin^2\theta_W$. With this small mixing angle, Z_1 and Z_2 are the Z boson in the Standard Model and the extra Z' gauge boson, respectively. From the data on parity violation in the cesium atom, we obtain a lower bound on the Z_2 mass in the range between 1.4 TeV and 2.6 TeV [18]. Data on the kaon mass difference Δm_K gives the bound $M_{Z_2} \le$ 1.02 TeV [8].

3. Z' PRODUCTION IN e–γ COLLISIONS

Now we are interested in the single production of new neutral gauge bosons Z' in e–γ collisions,

$$e^-(p_1,\lambda) + \gamma(p_2,\lambda') \longrightarrow e^-(k_1,\tau) + Z'(k_2,\tau'), \tag{22}$$

where p_i and k_i are the momenta and λ, λ', τ, and τ' are the helicities of the particles. At the tree level, there are two Feynman diagrams contributing to reaction (22), depicted in Fig. 1. The s-channel amplitude is given by

$$M_s^{Z'} = \frac{ieg}{2c_W q_s^2}\epsilon_\mu(p_2)\epsilon_\nu(k_2)\bar u(k_1) \\ \times\gamma^\nu[g_{2V}(e) - g_{2A}(e)\gamma_5]q_s\gamma^\mu u(p_1), \tag{23}$$

where $q_s = p_1 + p_2$. The u-channel amplitude is

$$M_u^{Z'} = \frac{ieg}{2c_W q_u^2}\epsilon_\mu(k_2)\epsilon_\nu(p_2)\bar u(k_1)\gamma^\nu q_u \\ \times\gamma^\mu[g_{2V}(e) - g_{2A}(e)\gamma_5]u(p_1), \tag{24}$$

where $q_u = p_1 - k_2$; $\epsilon_\mu(p_2)$, $\epsilon_\nu(p_2)$ and $\epsilon_\nu(k_2)$, $\epsilon_\mu(k_2)$ are the respective polarization vectors of the photon γ and the Z' boson, and $g_{2V}(e)$, $g_{2A}(e)$ are the coupling constants of Z' to the electron e. In the minimal model, they are given by [16]

$$g_{2V}(e) = \frac{\sqrt3}{2}\sqrt{1-4s_W^2}, \\ g_{2A}(e) = -\frac{1}{2\sqrt3}\sqrt{1-4s_W^2}, \tag{25}$$

and in the model with right-handed neutrinos [5],

$$g_{2V}(e) = \left(-\frac12 + 2s_W^2\right)\frac{1}{\sqrt{3-4s_W^2}}, \\ g_{2A}(e) = \frac{1}{2\sqrt{3-4s_W^2}}. \tag{26}$$

From Eqs. (25) and (26), we see that because of the factor

$$\sqrt{1-4s_W^2} \ll 1,$$

the cross sections in the minimal model are smaller than those in the model with right-handed neutrinos. We work in the center-of-mass frame and let θ denote the scattering angle (the angle between the momenta of the initial electron and the final one). We have evaluated the θ dependence of the differential cross section $d\sigma/d\cos\theta$, the energy, and the Z' boson mass dependence of the total cross section σ.

1) In Fig. 2, we plot $d\sigma/d\cos\theta$ for the minimal model as a function of $\cos\theta$ for the collision energy at CLIC $\sqrt{s}$ = 2733 GeV [20] and the relatively low value of mass $m_{Z'}$ = 800 GeV. From Fig. 2, we see that $d\sigma/d\cos\theta$ is peaked in the backward direction (this is due to the e^- pole term in the u-channel) but is flat in the forward direction. We note that the behavior of $d\sigma/d\cos\theta$ for the model with right-handed neutrinos is similar at other values of $\sqrt{s}$.

2) The energy dependence of the cross section for the minimal model is shown in Fig. 3. The same value

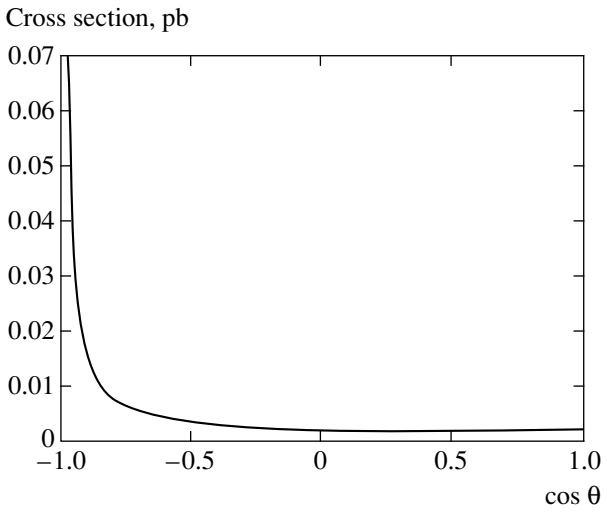


Fig. 2. Differential cross section of the minimal model, $\sqrt{s}$ = 2733 GeV, $m_{Z'}$ = 800 GeV.

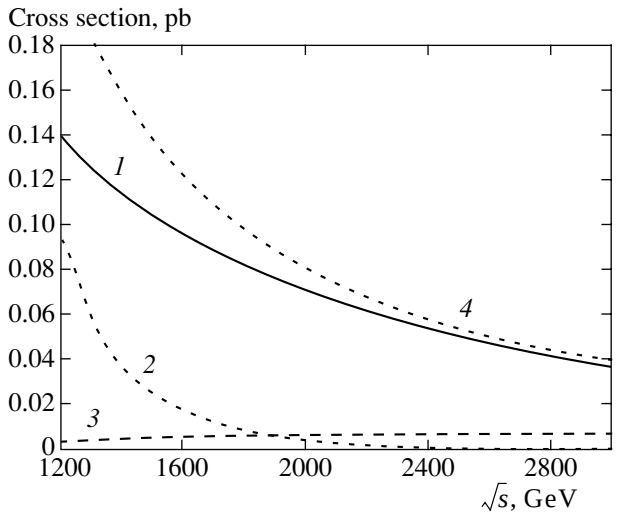


Fig. 3. Cross section $\sigma(e^-\gamma \longrightarrow Z'e^-)$ of the minimal model as a function of $\sqrt{s}$: *1*—total cross section, *2*—cross section in the *u*-channel, *3*—cross section in the *s*-channel, *4*—cross section in Standard Model; $m_{Z'}$ = 800 GeV.

of the mass as in the first case, $m_{Z'}$ = 800 GeV, is chosen. The energy range is

$$1200 \le \sqrt{s} \le 3000 \text{ GeV}.$$

Curve *1* is the total cross section for the minimal model, and curves *2* and *3* represent the respective cross sections of the *u*- and *s*-channels. Curve *4* is the cross section for the Standard Model, reduced three times. The *u*-channel, curve *2*, rapidly decreases with $\sqrt{s}$, while the *s*-channel has a zero point at $\sqrt{s} = m_{Z'}$ and then slowly increases. In the high-energy limit, the *s*-channel makes the main contribution to the total cross section. In Fig. 3, the cross section of the Standard Model reaches 0.18 pb and then slowly decreases to 0.05 pb, while the cross section of the minimal model is only 0.14 pb at $\sqrt{s}$ = 800 GeV and 0.05 pb at $\sqrt{s}$ = 2733 GeV. The same situation occurs in the model with right-handed neutrinos. In this model, we fix $m_{Z'}$ = 800 GeV and illustrate the energy dependence of the cross section in Fig. 4. The energy range is the same as in Fig. 3, $1200 \le \sqrt{s} \le 3000$ GeV. We see from Fig. 4 that the cross section σ decreases with $\sqrt{s}$, from σ = 0.35 pb to σ = 0.08 pb.

3) We have plotted the boson mass dependence of the number of events in the three models in Fig. 5. The energy is fixed as $\sqrt{s}$ = 2733 GeV and the mass range is $800 \le m_{Z'} \le 2000$ GeV. As we mentioned above, due to the coupling constant, the order of the lines of number of events, from bottom to top, is as follows: the minimal model, the Standard Model, and the model with right-handed neutrinos. The smallest number of events is for the minimal model. With the integrated luminosity $L \approx 100$ fb^{-1}, the number of events can be several thousand.

In the final state, Z' decays into leptons and quarks. Its partial decay width is equal to [21]

$$\Gamma(Z' \longrightarrow f\bar{f})$$

$$= \frac{G_F m_{Z'}^2}{6\sqrt{2}\pi} N_c^F [(g_{2A}^f)^2 R_A^f + (g_{2V}^f)^2 R_V^f]$$

$$= \begin{cases} 6.4 \text{ GeV for minimal model,} \\ 11.8 \text{ GeV for right-handed neutrinos model.} \end{cases}$$

Because of the coupling constants, the lifetime of Z' in

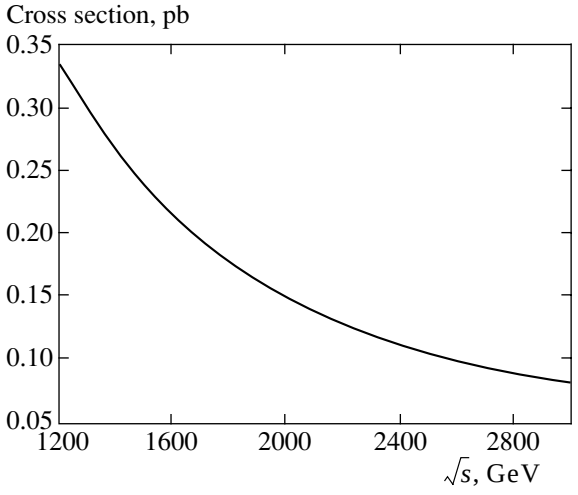


Fig. 4. Cross section $\sigma(e^-\gamma \longrightarrow Z'e^-)$ of the model with right-handed neutrinos as a function of $\sqrt{s}$; $m_{Z'}$ = 800 GeV.

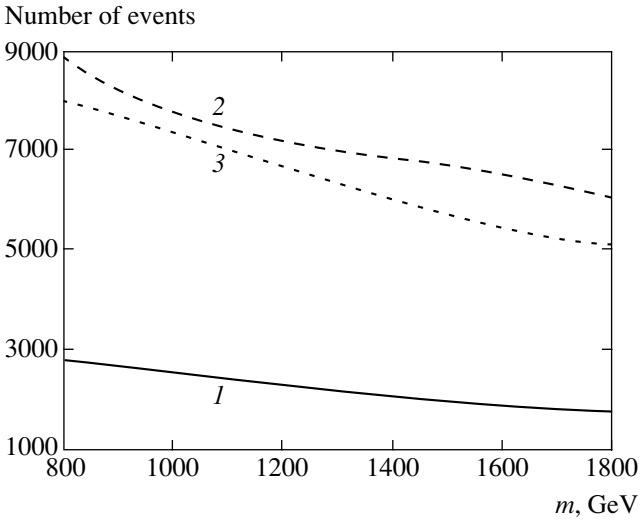


Fig. 5. Number of events of three models: *1*—minimal model, *2*—right-handed neutrinos model, *3*—Standard Model.

the minimal model is longer than that in the model with right-handed neutrinos.

4. CONCLUSIONS

In this paper, we have considered the production of a single Z' boson in the e–γ reaction in the framework of the 3–3–1 models. We see that with this process, the reaction mainly occurs at small scattering angles. The results show that if the mass of the boson is in the range of 800 GeV, then single boson production in e–γ collisions may give observable values at moderately high energies. On a CLIC based on e–γ colliders, with the integrated luminosity $L \approx 100$ fb^{-1}, we expect observable experiments in future colliders. Because of the values of the coupling constants, cross sections in the model with right-handed neutrinos are bigger than in the minimal model.

In conclusion, we have pointed out the usefulness of electron–photon colliders in testing the 3–3–1 models at high energies via the reaction

$$e^-\gamma \longrightarrow e^- Z'.$$

If the Z' boson is not very heavy, this reaction offers a much better discovery reach for Z' than the pair production in e^+–e^- or e^-–e^- collisions.

One of the authors (H. N. L) would like to thank the members of the Physics Department, National Center for Theoretical Sciences (NCTS), Hsinchu, Taiwan, where this work was completed, for warm hospitality during his visit. His work has been supported by the NCTS under a grant from National Science Council of Taiwan. This work was supported in part by the National Council for Natural Sciences of Vietnam.

REFERENCES

1. F. Pisano and V. Pleitez, Phys. Rev. D **46**, 140 (1992).
2. P. H. Frampton, Phys. Rev. Lett. **69**, 2889 (1992).
3. R. Foot, O. F. Hernandez, F. Pisano, and V. Pleitez, Phys. Rev. D **47**, 4158 (1993).
4. R. Foot, H. N. Long, and Tuan A. Tran, Phys. Rev. D **50**, R34 (1994).
5. H. N. Long, Phys. Rev. D **53**, 437 (1996); Phys. Rev. D **54**, 4691 (1996).
6. P. B. Pal, Phys. Rev. D **52**, 1659 (1995).
7. C. Geweriger *et al.* (LEPEWWG ff^- Subgroup), Preprint LEP2FF/00-03.
8. H. N. Long and V. T. Van, J. Phys. G **25**, 2319 (1999).
9. I. Antoniadis, K. Benakli, and M. Quiros, Preprints ETH-TH/01-10, CERN-TH/2001-202; New J. Phys. **3**, 20 (2002); T. Li and L. Wei, Phys. Lett. B **545**, 147 (2002); S. Dimopoulos, D. E. Kaplan, and N. Weiner, Phys. Lett. B **534**, 124 (2002); I. Gogoladze, Y. Mimura, and S. Nandi, Phys. Lett. B **554**, 81 (2003).
10. H. Hatanaka, T. Inami, and C. S. Lim, Mod. Phys. Lett. A **13**, 2601 (1998); H. Hatanaka, Prog. Theor. Phys. **102**, 407 (1999); G. Dvali, S. Randjibar-Daemi, and R. Tabbash, Phys. Rev. D **65**, 064021 (2002); N. Arkani-Hamed, A. G. Cohen, and H. Georgi, Phys. Lett. B **513**, 232 (2001).
11. T. Kitabayshi and M. Yasue, Phys. Rev. D **63**, 095002 (2001); Phys. Lett. B **508**, 85 (2001); Nucl. Phys. B **609**, 61 (2001); Phys. Lett. B **524**, 308 (2002); L. A. Sanchez, W. A. Ponce, and R. Martinez, Phys. Rev. D **64**, 075013 (2001); W. A. Ponce, Y. Giraldo, and L. A. Sanchez, Phys. Rev. D **67**, 075001 (2003).
12. B. Dion, T. Gregoire, D. London, *et al.*, Phys. Rev. D **59**, 075006 (1999).
13. P. Frampton and A. Rasin, Phys. Lett. B **482**, 129 (2000).
14. H. N. Long and D. V. Soa, Nucl. Phys. B **601**, 361 (2001).
15. D. V. Soa, T. Inami, and H. N. Long, hep-ph/0304300.
16. D. Ng, Phys. Rev. D **49**, 4805 (1994).
17. J. T. Liu and D. Ng, Z. Phys. C **62**, 693 (1994); N. A. Ky, H. N. Long, and D. V. Soa, Phys. Lett. B **486**, 140 (2000).
18. H. N. Long and L. P. Trung, Phys. Lett. B **502**, 63 (2001).
19. H. N. Long and T. Inami, Phys. Rev. D **61**, 075002 (2000).
20. Z. Z. Aydin *et al.*, hep-ph/0208041; R. W. Assman *et al.*, CERN 2000-008 (2000), p. 6; J. Ellis, E. Keil, and G. Rolandi, CERN-EP/98-03, CERN-SL/98-004 (AP), CERN-TH/98-33.
21. Particle Data Group, Phys. Rev. D **66**, 010001 (2002).

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 667–677.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 761–773.

ATOMS, SPECTRA, RADIATION

Zeeman Laser Cooling of ^{85}Rb Atoms in Transverse Magnetic Field

P. N. Melentiev, P. A. Borisov, and V. I. Balykin

Institute of Spectroscopy, Russian Academy of Sciences, Troitsk, Moscow oblast, 142190 Russia
e-mail: melentiev@isan.troitsk.ru

Received August 7, 2003

Abstract—The process of Zeeman laser cooling of ^{85}Rb atoms in a new scheme employing a transverse magnetic field has been experimentally studied. Upon cooling, the average velocity of atoms was 12 m/s at a beam intensity of 7.2×10^{12} s^{-1} and an atomic density of 4.7×10^{10} cm^{-3}. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Cold atomic beams characterized by a small average velocity of atoms at a high intensity and high phase space density are widely used in various experiments in atomic beam optics, interferometry, and lithography [1, 2]. Low-energy atomic beams can be obtained by method of laser cooling, which is known in several variants employing the Zeeman effect [3, 4], frequency-chirped laser radiation [5], isotropic light [6], and wideband laser radiation [7]. Unfortunately, the process of cooling by all these techniques is accompanied by unavoidable increase in the transverse temperature of atoms and, hence, by a decrease in the beam brightness and phase space density.

An effective means of solving this problem is offered by schemes employing a two-dimensional magneto-optical trap (2D-MOT) ensuring both transverse compression of the atomic beam and a decrease in the transverse velocity of atoms [8–10]. The degree of compression and cooling in 2D-MOT is usually limited by a finite time of flight of atoms through the apparatus. Effective use of 2D-MOTs requires that atoms in a beam would possess a sufficiently low longitudinal velocity.

An alternative method for obtaining atomic beams of high brightness and high phase space density is based on the extraction of atoms from a three-dimensional MOT (3D-MOT) [11], an advantage of this device being a relatively high phase space density of atoms. However, the use of this technique for obtaining continuous atomic beams of high intensity is limited by the large time required for the accumulation of atoms.

We have developed a new method for obtaining cold atomic beams of high intensity (7.2×10^{12} s^{-1}) and a small average velocity of atoms (12 m/s) and have studied this method in application to a beam of ^{85}Rb atoms. The proposed technique employs the Zeeman laser cooling of thermal ^{85}Rb atoms in *transverse* magnetic field [12]. Application of the *transverse* magnetic field allowed us to obtain the optimum distribution of magnetic field along the beam axis, which is necessary for effective cooling. Using the scheme with transverse magnetic field, we succeeded in creating a compact and effective Zeeman slower ensuring the formation of intense beams of atoms with an average velocity as low as 10 m/s.

2. ZEEMAN LASER COOLING

2.1. Zeeman Slowing in Longitudinal Magnetic Field

According to the method of laser cooling, an atomic beam interacts with the counterpropagating beam of laser radiation with a frequency tuned in resonance with that of a given atomic transition. In the course of deceleration, the absorption frequency exhibits a Doppler shift relative to the laser radiation frequency and, hence, the efficiency of the process tends to decrease. The Doppler shift can be compensated using the linear Zeeman effect. The scheme of atomic beam cooling by laser radiation in a magnetic field, called Zeeman slowing, is now most widely used for obtaining slow atomic beams.

An experimental setup for Zeeman slowing comprises a source of neutral atoms and a Zeeman slower creating the required magnetic field distribution in the zone of interaction between atoms and laser radiation. The cooling laser radiation tuned in resonance with a given atomic transition propagates in the direction opposite to that of the atomic beam. In most Zeeman slower schemes, atoms are decelerated only inside the apparatus and do not interact with the laser radiation outside. The required magnetic field configuration in the Zeeman slower is created using a magnetic solenoid with the distance between turns varied so as to provide

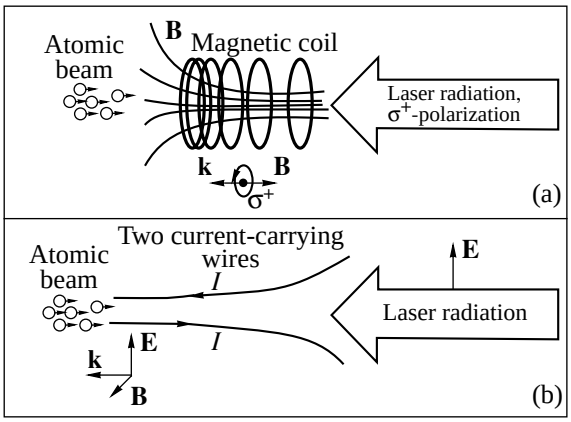


Fig. 1. Schematic diagrams illustrating Zeeman laser cooling of an atomic beam in (a) longitudinal and (b) transverse magnetic fields.

for the optimum distribution of the magnetic field in the axial direction. The solenoid axis coincides with the axes of atomic and laser beams (Fig. 1a). It is possible to use ring permanent magnets instead of the magnetic coil. In both cases, the magnetic field vector in the Zeeman slower is collinear with the wavevector of laser radiation. The laser radiation possesses a σ^+ polarization and excites transitions between the Zeeman sublevels corresponding to a change in the magnetic quantum number $\Delta m = +1$.

The Zeeman shift of the atomic transition frequency is proportional to the magnetic field strength (magnetic induction) B: $\Delta\omega_{\text{Zeeman}} = \alpha B$, where α is a constant determined by the Zeeman effect. The resonance interaction between atoms and laser radiation in the Zeeman slower is determined by the condition

$$\Delta + kV - \alpha B = 0, \tag{1}$$

where $\Delta = \omega_{\text{laser}} - \omega_0$ is the detuning of the laser radiation frequency ω_{laser} from the atomic transition frequency ω_0 in a zero magnetic field, V is the atomic velocity, and $k = 2\pi/\lambda$ is the wavevector. If the magnetic field B varies in space so that condition (1) is valid at all points of the atomic trajectory in the course of deceleration, then atoms occur in resonance with the laser radiation.

The required magnetic field distribution in the Zeeman slower can be readily determined as follows. When condition (1) is fulfilled at all points of the atomic trajectory, the force of light pressure imparts a constant acceleration a to an atom so that its velocity decreases according to the law

$$V(z) = \sqrt{V_0^2 - 2az}.$$

The acceleration is determined by the expression

$$a = \frac{\hbar k\Gamma}{2M}\frac{G}{1 + G + (\Delta + kV - \alpha B)^2/\gamma^2}, \tag{2}$$

where 2γ is the natural width of the given atomic transition, M is the atomic mass, $G = I/I_{\text{sat}}$ is the parameter of saturation of the atomic transition, I is the laser radiation intensity, and I_{sat} is the laser saturation intensity. The latter quantity is given by the formula

$$I_{\text{sat}} = \frac{\hbar\omega_0}{2\tau\sigma}, \tag{3}$$

where $\tau = 1/2\gamma$ is the time of spontaneous decay and ω_0 is the atomic transition frequency, and σ is the absorption cross section.

Let the laser frequency be tuned precisely in resonance to that of the given atomic transition. Using condition (1), we obtain an expression for the required field profile:

$$B(z) = B_0\sqrt{1 - \frac{2az}{V_0^2}}.$$

The existence of a maximum possible acceleration for a given atom in a magnetic field, $a_{\max} = 2\hbar k\Gamma/M$ (for $I \gg I_{\text{sat}}$) poses a limitation on the maximum magnetic field gradient [3],

$$\left(\frac{dB}{dz}\right)_{\max} \le \left(\frac{dB}{d\omega}\right)\frac{a_{\max}}{\lambda V}, \tag{4}$$

where $dB/d\omega = \alpha^{-1}$. When the magnetic field gradient is below maximum, the laser radiation intensity is limited by the condition

$$G \ge \frac{1}{1 + \dfrac{a_{\max}}{V\lambda}\dfrac{dB}{d\omega}\dfrac{dz}{dB}}. \tag{5}$$

The minimum temperature T_{D} of atoms that can be achieved by means of Zeeman slowing, called the Doppler cooling limit, is determined by the formula [13]

$$T_{\text{D}} = \frac{\hbar\gamma}{2k_B}. \tag{6}$$

For ^{85}Rb atoms, $T_{\text{D}} = 141$ μK and the corresponding minimum atomic velocity is $V_{\text{D}} = 0.12$ m/s. However, there are several other limiting factors that hinder obtaining such a low velocity by Zeeman slowing. The main difficulty is encountered in extracting low-energy

atoms from the Zeeman slower [4]: at a low atomic velocity (~10 m/s), the length of interaction between an atom and the laser field on which the velocity is reversed is as small as a fraction of a millimeter. For this reason, intense atomic beams with particle velocities below 50 m/s could not be obtained by means of Zeeman slowing.

2.2. Zeeman Slowing in a Transverse Magnetic Field

The principal difference between the Zeeman slowing in a *transverse* magnetic field and the scheme considered above is that the magnetic field vector is perpendicular to the wavevector of the cooling laser radiation (in the conventional scheme, these vectors are collinear). This mutual orientation of the magnetic field **B** and the wavevector **k** determines, in turn, the required polarization of the laser radiation, and the electric field vector is perpendicular to the magnetic field vector **B** (Fig. 1b). In this configuration, the laser radiation can induce atomic transitions with a change in the magnetic quantum number $\Delta m = +1$ or $\Delta m = -1$.

The method of Zeeman slowing in the transverse magnetic field offers two important advantages over the conventional scheme: (i) simpler realization of the required magnetic field distribution in the Zeeman slower and (ii) higher accuracy of controlling the length of interaction between atoms and laser radiation, facilitating the extraction of low-energy atoms from the Zeeman slower. Let us consider application of the new scheme to cooling ^{85}Rb atoms.

The existence of hyperfine splitting of the ground and excited states in alkali metal atoms leads to transitions between various sublevels of the hyperfine structure. Excitation with single-mode laser radiation leads to optical pumping of atoms to one sublevel of the hyperfine structure of the ground state and drives these atoms out of resonance with the laser radiation. For ^{85}Rb atom (Fig. 2), a transition from the ground state with $F = 3$ to an excited state with $F' = 4$ is a cyclic transition and the $F' = 4 \longrightarrow F = 2$ transition is forbidden in the dipole approximation. Therefore, the former transition can be used for Zeeman slowing of ^{85}Rb atoms.

However, there is a small probability (6×10^{-4} for a laser intensity on a saturation level) of a transition to the sublevel with $F' = 3$ of the excited state. From this state, the atom can equiprobably pass either to the ground state sublevel with $F = 3$ or to an excited sublevel with $F = 2$ spaced at 3 GHz from the sublevel with $F = 3$, which will break cyclic interaction with the laser radiation. A commonly accepted straightforward solution of this problem consists in using two-frequency laser radiation. The dominant (cooling) laser mode (decelerating field) is tuned in resonance to the $F = 3 \longrightarrow F' = 4$ transition. The second (auxiliary) mode is tuned in resonance to the $F = 2 \longrightarrow F' = 3$ transition so as to provide for the optical pumping of the ground state sublevel with $F = 3$.

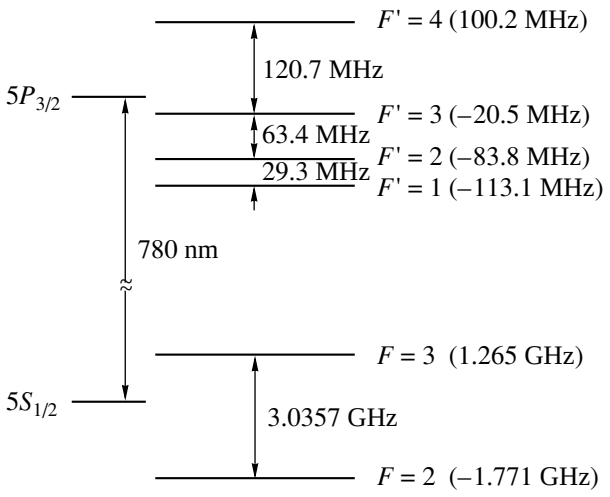


Fig. 2. Schematic diagram of energy levels of the D_2 line of ^{85}Rb atom.

For ^{85}Rb atoms possessing thermal velocities, a Doppler frequency shift is greater than the distance between sublevels of the hyperfine structure of an excited state. For this reason, the positions of energy levels in a magnetic field should be calculated for the cases of weak and strong magnetic fields. In a weak magnetic field, each component of the hyperfine structure of the ground and excited states of ^{85}Rb atom splits into $2F + 1$ Zeeman sublevels characterized by the magnetic quantum number m_F,

$$U_{m_F} = \mu_B g_F m_F B, \tag{7}$$

where $\mu_B = 9.27 \times 10^{-24}$ J/T is the Bohr magneton and g_F is the Lande factor.

In strong magnetic fields such that the energy of an atom in the magnetic field is greater than the energy of electron interaction with the nucleus, the character of splitting changes significantly. A level characterized by the magnetic quantum number J splits into $(2J + 1)(2I + 1)$ sublevels, determined by the quantum numbers m_I and m_J, with the energies

$$U_{m_J m_I} = \mu_B g_J m_J B + A m_I m_J, \tag{8}$$

where A is the hyperfine splitting constant. For the $5P_{3/2}$ level of ^{85}Rb atom, $A = 25$ MHz.

Figure 3 shows the pattern of splitting of the magnetic sublevels of the ground ($5S_{1/2}$) and excited ($5P_{3/2}$) states of ^{85}Rb atom in a magnetic field B. As can be

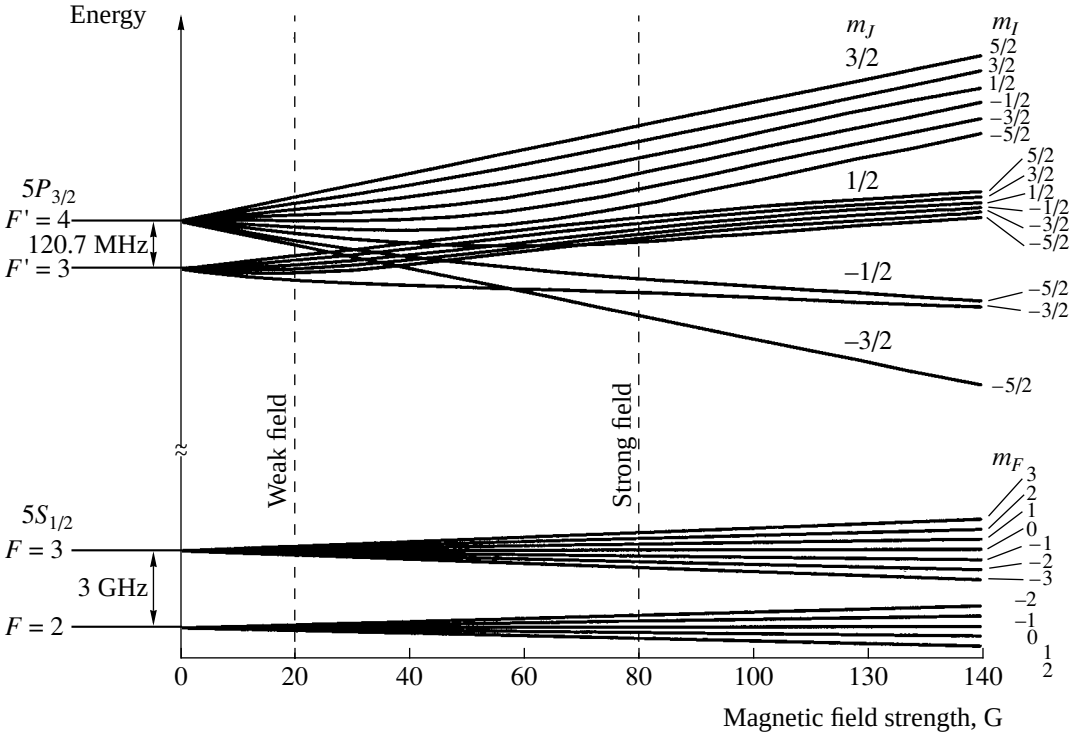


Fig. 3. The hyperfine structure of energy levels of the ground ($5S_{1/2}$) and excited ($5P_{3/2}$) states (the levels with $F' = 3, 4$) of ^{85}Rb atom in a magnetic field.

seen, the weak magnetic field approximation is valid for most of the sublevels under consideration in a field with $B < 20$ G, while the strong magnetic field approximation is applicable when $B > 80$ G.

For the ground state level with $F = 2$, the Lande factor is negative, while for the level with $F = 3$ this factor is positive. As a result, the Zeeman sublevels of the ground state levels with the same magnetic moment projection m_F behave differently in response to increasing magnetic field strength. As can be seen from Fig. 3, the field dependences of the frequencies of transitions between magnetic sublevels with $F = 3$ and $F' = 4$ significantly differ from the analogous dependences for the states with $F = 2$ and $F' = 3$. Since the Doppler shift for these levels is the same, the complicated behavior of magnetic sublevels in the applied magnetic field will lead to a loss in efficiency of excitation for the second mode of the two frequency laser radiation in the course of Zeeman slowing. The efficiency in interaction between atoms and laser radiation can be increased at the expense of the field-induced broadening. According to our calculations, the parameters G_1 and G_2 of the atomic transition saturation for the dominant (cooling) laser mode and the second (auxiliary) mode, respectively, must obey the condition $G_2 \geqslant 0.1G_1$.

When ^{85}Rb atoms interact with a two frequency laser radiation in the presence of a magnetic field, several photon absorption–reemission cycles are sufficient for the optical pumping of the atom to the sublevels with $F = 3$, $m_F = 3$ and $F = 2$, $m_F = 2$. Therefore, an analysis of the Zeeman slowing can be restricted to the $F = 3$, $m_F = 3 \longrightarrow F' = 4$, $m_F = 4$ and $F = 2$, $m_F = 2 \longrightarrow F' = 3$, $m_F = 3$ transitions.

The number of cooled atoms at the output of the Zeeman slower is, together with the average velocity of atoms, among the most important parameters characterizing the cooling efficiency. This number is determined primarily by two factors: (i) the fraction of the initial velocity distribution of atoms cooled by laser radiation in the Zeeman slower and (ii) the fraction of the primary atomic flux injected into the Zeeman slower. The former circumstance dictates the need for increasing the velocity interval of atoms subjected to cooling. However, this usually leads to a considerable increase in the length of the Zeeman slower and, accordingly, to a decrease in the flux of thermal atoms injected into the system. An analysis shows that the shorter the Zeeman slower, the greater the output flux of cold atoms. In selecting the optimum Zeeman slower length, it is also necessary to take into account the fact that the real atomic velocity distribution in a beam is depleted of the low-velocity fraction because of atomic collisions in the beam [14]. With allowance for this fact, we selected a magnetic field configuration in the Zeeman slower such that atoms are decelerated beginning

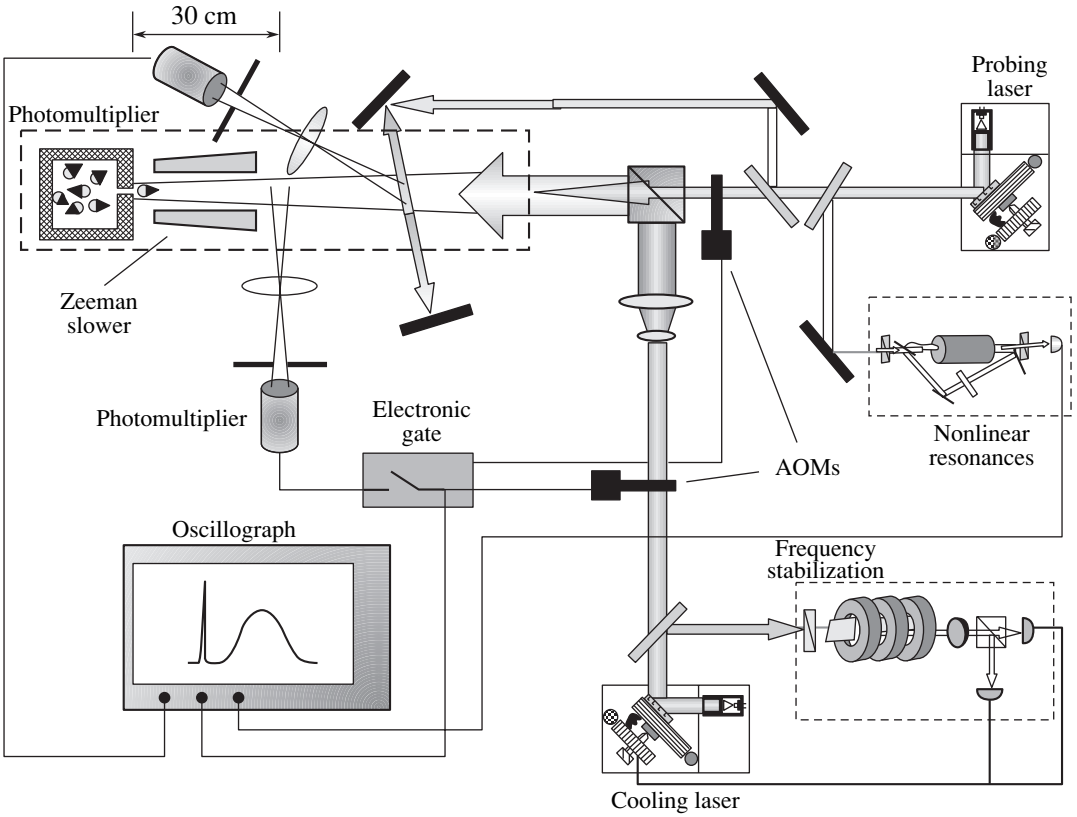


Fig. 4. Schematic diagram of the experimental setup.

with a velocity equal to half of the most probable value for atoms in the beam.

3. EXPERIMENTAL SETUP

The experimental setup for studying the Zeeman slowing of atoms is schematically depicted in Fig. 4. The radiation sources were semiconductor lasers employing the Littrov scheme. Both lasers operated in a two frequency lasing regime ensured by resonance excitation of the relaxation oscillations due to microwave modulation of the injection current [15]. The microwave modulation frequency was equal to the difference in the frequencies of $F = 3 \longrightarrow F' = 4$ and $F = 2 \longrightarrow F' = 3$ transitions (2916 MHz). The microwave generator power was selected such that the intensity at the fundamental frequency would be four times that of the first side band. The dominant laser mode was used to excite the $F = 3 \longrightarrow F' = 4$ transition in ^{85}Rb atoms, while the one at the side band frequency excited the $F = 2 \longrightarrow F' = 3$ transition. The maximum laser output power was 15 mW. The laser beam diameter at the Zeeman slower entrance and exit was 3.5 and 5 mm, respectively. The cooling lasers were operating in the regime of active frequency stabilization with respect to the absorption signal in a cell placed in the magnetic field [16]. The short-term laser frequency stability was 3 MHz, while the long-term frequency drift was within 9 MHz/h. The probing laser frequency was chirped in the vicinity of the frequency of the $F = 3 \longrightarrow F' = 4$ transition in ^{85}Rb. The atomic velocity distribution was determined using the signal of fluorescence detected by a photomultiplier.

In order to eliminate the influence of the cooling laser radiation during the atomic velocity measurements, the cooling laser was switched off by means of an acoustooptical modulator (AOM). The fluorescence signal was measured using a BoxCar electronic gate (Fig. 5). In order to determine a stationary distribution of the atomic velocities, the time for which the cooling laser was switched on was selected sufficiently large, so that slow atoms leaving the slower could reach the detector before the laser was switched off. In our experimental configuration, this time was 6 ms. With this time delay, we measured the stationary distribution of atomic velocities—the same as that established for the constant laser irradiation of the atomic beam. In order to reduce the mechanical action upon atoms from the side probe laser, the probing laser radiation was switched on by an AOM only during the fluorescence measurements.

Since we employed the low-velocity part of the initial atomic velocity distribution for laser cooling, special measures were taken to obtain a beam of thermal atoms with undepleted low-energy fraction of the total

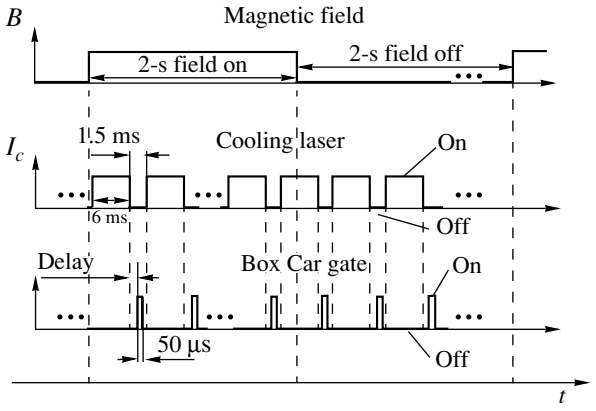


Fig. 5. Time series of the switching of magnetic field $B(t)$, cooling laser field $I_c(t)$, and BoxCar gate for monitoring the atomic velocity distribution during Zeeman slowing.

velocity distribution. This was achieved by using a source of ^{85}Rb atoms analogous to that described in [17]. The atomic source temperature could be varied from 20 to 500°C. The intensity of the atomic beam formed with a 4-mm aperture at a source temperature of $T = 250$°C was 4.5×10^{13} s^{-1} (approximately half of the calculated value). Investigation of the primary atomic beam characteristics showed that this source exhibited no depletion of the low-energy part of the velocity distribution as a result of the beam scattering from vapor in the vicinity of the exit diaphragm of the atomic oven.

The magnetic field profile along the Zeeman slower axis was calculated using the formula

$$B(z) = B_0\sqrt{1 - \frac{a_{\max}z}{V_0^2}},$$

where $a_{\max} = 1.07 \times 10^5$ m/s^2 and the initial velocity is $V_0 = 150$ m/s. The Zeeman slower comprised two 22-cm-long aluminum stripe combs cut to various depths (Fig. 6). The variable cut depth allowed the required field profile along the axis to be obtained because the current passed only via a continuous part of the metal stripe. The comb configuration provided for an increase in the effective mass and the surface area, thus increasing the heat exchange rate under ultrahigh vacuum conditions.

Figure 7 compares the calculated and experimentally measured field profiles along the Zeeman slower axis for a current of $I = 170$ A. As can be seen, the max-

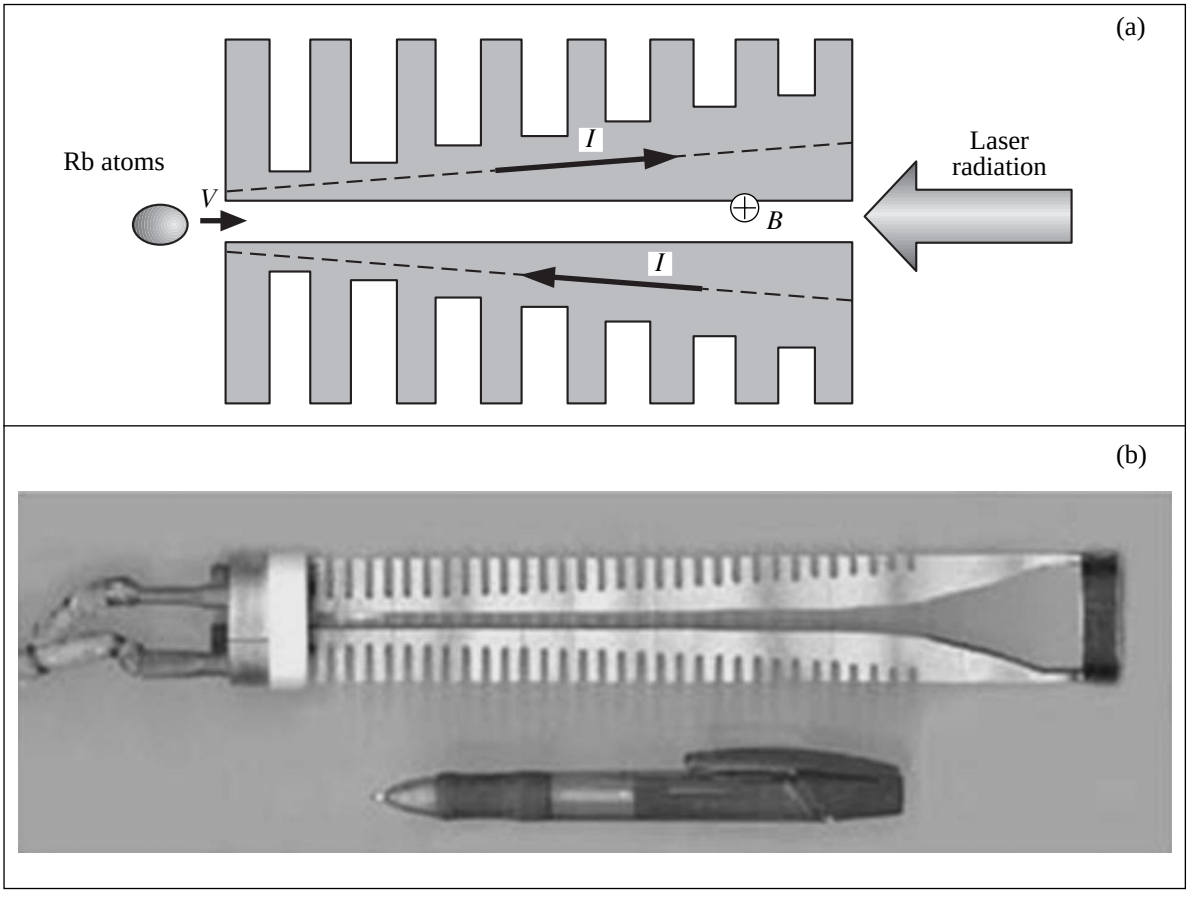


Fig. 6. The Zeeman slower: (a) schematic diagram; (b) general view.

imum deviation of the measured magnetic field strength from the calculated profile amounts to $\Delta B = 15$ G. This deviation may lead to a decrease in the Zeeman slowing efficiency as a result of the laser radiation being out of resonance with the atomic transition frequency. The compensation of the deviation of the magnetic field strength from that required for the effective cooling was achieved by increasing the laser radiation intensity. To this end, the parameter of the atomic transition saturation must be not less than

$$G \approx 14.$$

The electric resistance of the Zeeman slower (together with connecting leads in the vacuum chamber) was $R = 3 \times 10^{-3}$ Ω. For a current of 170 A passing through the device, the electric power converted into heat amounted to 90 W. In order to reduce the influence of Joule heating of the Zeeman slower on the residual gas pressure in the vacuum chamber, the current was supplied in a quasiperiodic regime, with the current switched on for 2 s and off for 8 s. The residual gas pressure in the vacuum chamber was 3×10^{-7} Torr.

For a correct analysis of the measured atomic velocity distributions, it is very important to know the position of zero velocity. For this purpose, a part of the probing laser radiation was introduced perpendicularly to the atomic beam and the corresponding fluorescence component was measured by a separate photomultiplier. Since this scheme eliminates the Doppler broadening, the resonance between the atomic fluorescence signal and the probing radiation indicated the position of the exact resonance corresponding to the $F = 3 \longrightarrow F' = 4$ transition, thus determining the atoms with a zero velocity. It should be noted that deviation of the angle between the probing laser radiation and the atomic beam from 90° leads to an error in the zero velocity determination. In order to minimize this error, the probing laser radiation was adjusted to cross the atomic beam at an angle of about 89° and reflected back by a mirror. This resulted in the appearance of two peaks equally shifted from the zero velocity position. The accuracy of zero velocity calibration in this scheme is determined by the uncertainty of matching of the forward and reflected laser beams. In our experiments, this uncertainty led to an error below 2 m/s in the velocity determination. We have also used an alternative technique for the zero velocity calibration based on the monitoring of nonlinear absorption resonances in a cell with Rb vapor.

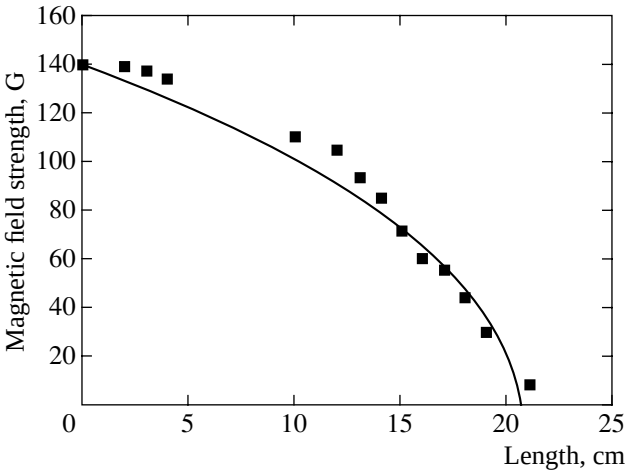


Fig. 7. Magnetic field distribution along the Zeeman slower axis. Solid curve shows the ideal calculated profile; squares show the experimental data.

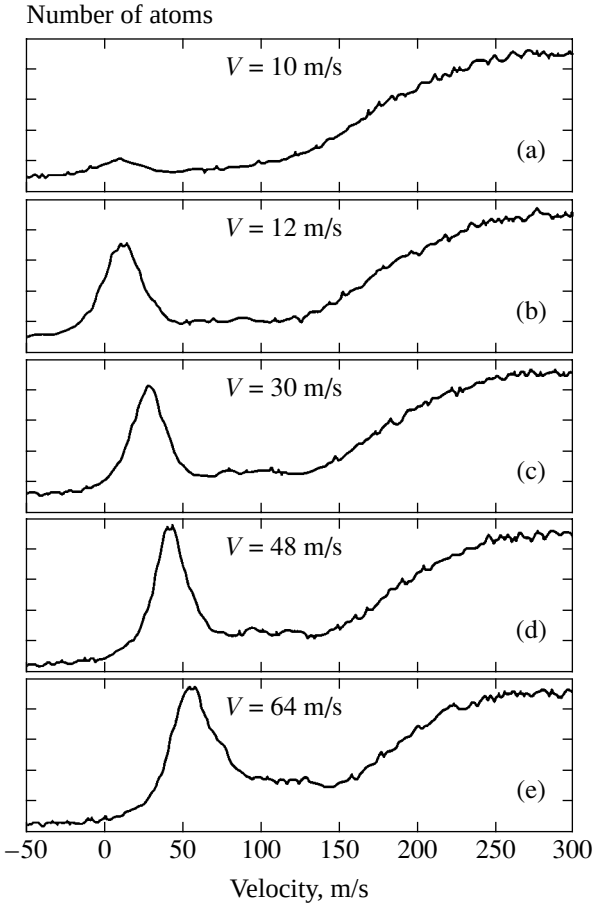


Fig. 8. The velocity distribution of ^{85}Rb atoms in a beam upon Zeeman slowing for various detunings of the cooling laser frequency Δ (MHz): (a) –39, (b) –46, (c) –54, (d) –66, (e) –77.

4. EXPERIMENTAL RESULTS

Figure 8 shows the atomic velocity distributions of a beam of ^{85}Rb atoms upon Zeeman slowing at various detunings Δ of the laser radiation frequency. The atomic source temperature was $T = 250$°C. A peak in the low-velocity part of the distribution corresponds to atoms decelerated as a result of the Zeeman slowing. As can be seen from these data, the average velocity of atoms in this peak, as well as the peak amplitude, depend on the laser frequency detuning. The closer the

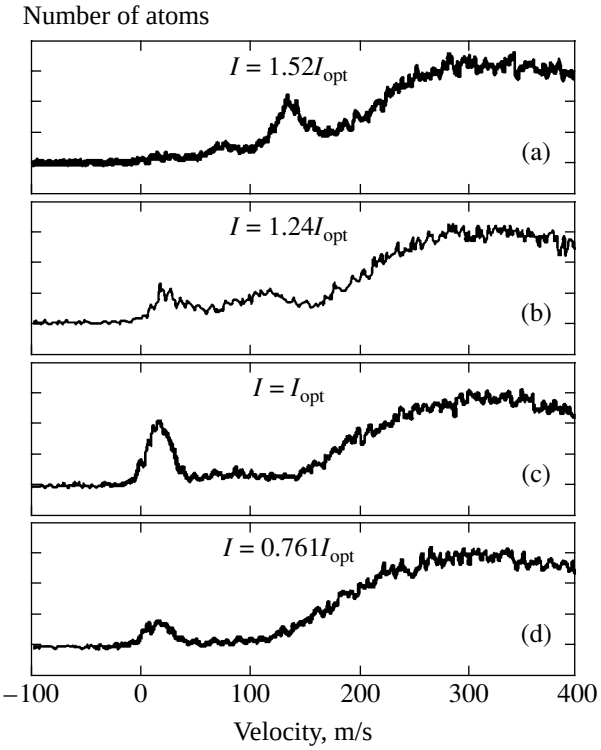


Fig. 9. The effect of magnetic field strength in the Zeeman slower on the atomic velocity distribution. The maximum cooling efficiency is observed for the magnetic field induced by the current $I = I_{opt} = 170$ A.

cooling laser frequency to that of the atomic transition, the lower the average velocity of cooled atoms.

The data in Fig. 8 show that the amplitude of the peak of cold atoms remains virtually unchanged for the laser frequency detunings corresponding to the average velocities of cooled atoms above $V = 12$ m/s. The peak of cold atoms accounts for about 7% of the total number of atoms in the initial velocity distribution, which agrees with theoretical estimates. For the average velocity below 12 m/s, the peak amplitude drops with decreasing detuning Δ. A minimum average velocity for which the peak contains a significant fraction of atoms from the initial distribution is 10 m/s ($\Delta = -39$ MHz). This decrease in the low-velocity peak amplitude is explained by a decrease in the efficiency of detection of low-energy atoms, which is caused by a large divergence of the beam of atoms with longitudinal velocities below 15 m/s. Good agreement of the experimentally observed numbers of cooled atoms with the results of calculations showed that the proposed scheme provides for the effective extraction of cold atoms from the Zeeman slower.

In the spectra of Fig. 8, a full width at half maximum (FWHM) of the peak of cold atoms amounts to $\Delta V = 28 \pm 2$ m/s. This value is significantly greater than the minimum width determined by the Doppler cooling limit. There are several factors responsible for the final (and higher than the Doppler limit) temperature of the atomic beam. According to the results of our calculations, spatial inhomogeneity of the laser radiation intensity leads to a finite width of the atomic velocity distribution on a level of 6 m/s, spatial inhomogeneity of the magnetic field leads to additional broadening of about 3 m/s, and a contribution due to the impulse diffusion amounts to 0.5 m/s. The total calculated width of the low-velocity peak is $\Delta V \approx 10$ m/s, which corresponds to a temperature of $\Delta T \approx 1$ K. This estimate is lower than the value observed in experiment. The discrepancy is related to a finite length of the detection zone, in which atoms continue to interact with the cooling laser radiation. As a result, the average velocity of atoms at various points of the detection zone exhibits various values.

In order to study the dependence of the slowing efficiency on the magnetic field strength in the Zeeman slower, we varied the current passing through the system, all other parameters being fixed. The corresponding velocity spectra are presented in Fig. 9. As can be seen, the curves measured for the current exceeding $I_{opt} = 170$ A exhibit two peaks. This is related to the fact that the process at high currents does not obey the condition (4) for the maximum possible magnetic field gradient during Zeeman slowing, which breaks the interaction between atoms and laser radiation. For currents below the optimum value, the magnetic field gradient drops and, hence, the cooling efficiency decreases.

We have experimentally investigated the dependence of the flux of cold atoms on the laser radiation intensity I_{laser}. In our experiments, the maximum laser radiation intensity corresponded to a saturation parameter of $G = 30$. A twofold decrease in this value reduces the detected flux of cold atoms approximately by half, while the average velocity of cooled atoms remains unchanged. A fourfold decrease in the value of I_{laser} leads to a significant decrease in the number of cold atoms, while the average velocity of these atoms increases by a factor of about 1.4. This behavior is explained by the fact that the magnetic field profile in the Zeeman slower deviated from the ideal shape. By increasing the laser radiation intensity, it was possible to compensate for the nonideal magnetic field distribution by means of field-induced broadening, which led to a decrease in the average velocity and an increase in the flux of cold atoms. According to the estimates presented above, the imperfect magnetic field distribution can be compensated provided that the atomic transition saturation parameter is $G = 14$. When the saturation parameter for the dominant (cooling) laser mode in our experiments was $G_1 = 14$, the corresponding value for the second laser mode was as small as $G_2 = 3.5$ that was insufficient for the effective Zeeman slowing.

We have also studied the dependence of the Zeeman slowing efficiency on the laser radiation polarization and on the mutual orientation of electric vector **E** and

magnetic field **B**. As expected, any deviations of the laser radiation polarization from the optimum (linear) reduced the efficiency of Zeeman slowing. These deviations lead to a decrease in intensity of the laser radiation component exciting the atomic transitions with $\Delta m = +1$ and, hence, to a drop in the cooling efficiency (Fig. 10).

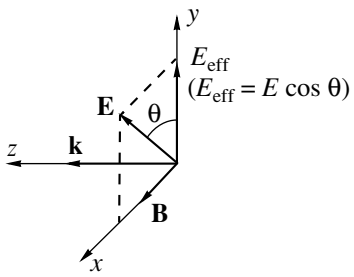


Fig. 10. Schematic diagram illustrating the influence of the mutual orientation of the magnetic induction **B** and the electric vector **E** of the cooling laser radiation on the Zeeman slowing efficiency.

5. PARAMETERS OF A COLD ATOMIC BEAM

We have determined the main parameters of the cold atomic beam, including the intensity, density, divergence, brightness, and phase space density. The intensity of a beam of cold atoms with an average velocity of 12 m/s (at an atomic source temperature of 250°C) was 3×10^{12} s^{-1}. We studied the possibility of improving this parameter by increasing the source temperature. As the source temperature was increased to 400°C, the cold atomic beam intensity exhibited a growth by a factor of 2.4, but further increase in the source temperature led to a decrease in the beam intensity. This is related to the primary beam depletion of the low-velocity atoms as a result of their scattering from fast atoms. Thus, the maximum intensity of the beam of cold atoms in our experiments was $I_{\max} = 7.2 \times 10^{12}$ s^{-1}. The corresponding density of cold atoms was $n_{\max} = I_{\max}/S\bar{V} \approx 4.7 \times 10^{10}$ cm^{-3}.

The brightness of an atomic beam is defined as

$$R = \frac{I}{\pi(\Delta x_\perp)^2 \Delta\Omega},$$

where $\Delta x_\perp$ is the transverse size of the beam and $\Delta\Omega$ is the solid angle in which atoms are confined. The latter solid angle is determined as $\Delta\Omega = \pi(\Delta V_\perp/\bar{V}_\parallel)^2$, where $\Delta V_\perp$ is the width of the transverse velocity component distribution and $\bar{V}_\parallel$ is the average value of the longitudinal velocity component. The maximum transverse velocity of atoms was determined by the beam-forming diaphragms and amounted to $V_\perp \approx 4.5$ m/s. Therefore, for an average longitudinal velocity of $\bar{V}_\parallel = 12$ m/s, cold atoms move within a solid angle of $\Delta\Omega = 0.14\pi$. For the maximum atomic flux of $I_{\max} = 7.2 \times 10^{12}$ s^{-1}, the brightness amounts to $R = 1.3 \times 10^{18}$ (sr m^2 s)$^{-1}$. The spectral brightness of an atomic beam is defined as

$$B_r = R\frac{\bar{p}_\parallel}{\Delta p_\parallel}.$$

In our experiments, the spectral brightness was $B_r = 1.7 \times 10^{18}$ (sr m^2 s)$^{-1}$. The phase space density of an atomic beam is given by the expression

$$\tilde{\Lambda} = B_r \frac{\pi}{m^3 \bar{V}_\parallel^4} h^3.$$

The maximum phase space density observed in our experiments was $\tilde{\Lambda} = 2.4 \times 10^{-11}$.

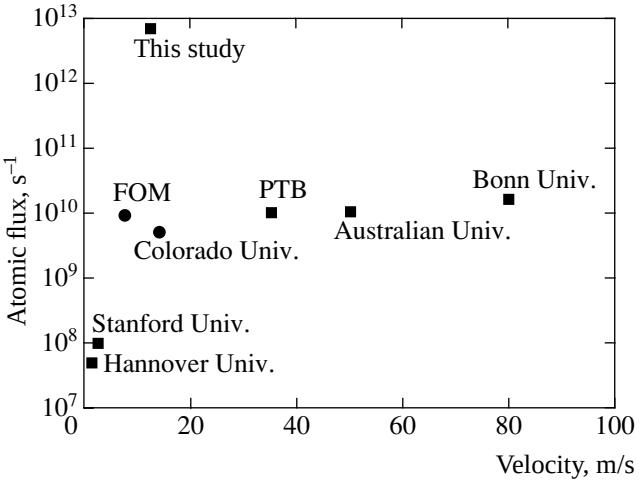


Fig. 11. The flux and average velocity of cooled atoms obtained in this study in comparison to the results obtained by other researchers using cooled atomic beams (■) and atomic beams from 3D-MOT (●): This study; Hannover Univ. [8]; PTB [20]; Australia Univ. [21]; Bonn Univ. [19]; Stanford Univ. [9]; FOM [23];Colorado Univ. [11].

Figure 11 shows a comparison of the parameters (plotted as the beam intensity versus average atomic velocity) of cold beams obtained in various research centers. As can be seen from these data, the flux of cold atoms achieved in this study is more than two orders of magnitude higher than the beam intensities reported by other researchers. This increase in the total flux of cold atoms has become possible for two reasons: first, due to the implemented scheme with transverse magnetic field, which allowed the length of the cooling tract to be significantly reduced; second, due to the use of an

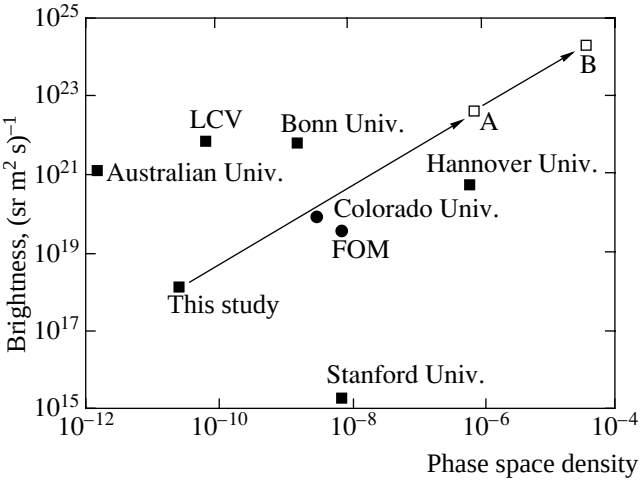


Fig. 12. The brightness and phase space density of cooled atomic beams obtained in this study in comparison to the results obtained by other researchers: This study; Australia Univ. [21]; LCV Univ. [22]; Bonn Univ. [19]; Colorado Univ. [11]; Stanford Univ. [9]; FOM [23]; Hannover Univ. [8]. Arrows show the values of brightness and phase space density of a cold atomic beam predicted for the proposed method in combination with 2D-MOT cooling: (A) for the transverse Doppler cooling; (B) for transverse sub-Doppler cooling.

atomic source providing for an intense primary thermal beam with undepleted low-velocity fraction of the total velocity distribution.

We have estimated the possibility to further increase the level of brightness and phase space density of a cold atomic beam achieved in our experiments. This can be provided by the two-dimensional magnetooptical trap technique (2D-MOT). The density of atoms in a 2D-MOT is limited by the following physical factors: (i) dipole–dipole interaction between atoms; (ii) repulsive potential created by scattered laser radiation; and (iii) attractive potential related to the absorption of laser radiation [18]. The large intensity of a cold beam achieved in our case suggests that the most important factor determining the transverse size and temperature of the beam in the course of transverse laser cooling is reabsorption of photons inside the atomic ensemble. The multiple reabsorption of photons leads to the heating of atoms and to a decrease in the compressive force, so that the maximum possible density of atoms in the beam is restricted to a value on the order of $n_{\max} = 10^{12}$ cm^{-3} [18]. Taking into account this limitation and considering the maximum cold beam intensity and the average atomic velocity obtained in our experiments, a minimum possible transverse size that can be achieved by means of the 2D-MOT technique is $\Delta x_\perp \approx 430$ μm. Upon cooling atoms in a 2D-MOT down to the Doppler limit of laser cooling, the angular divergence of the atomic beam is 2×10^{-2} rad, while reaching sub-Doppler temperatures of about 3 μK allows the divergence to be reduced down to 2.7×10^{-3} rad. Therefore, application of the 2D-MOT technique in our case will allow the brightness and phase space density to be increased by a factor of 3×10^4 and 1.5×10^6, respectively.

Figure 12 presents a summary of data on the brightness and phase space density of cooled atomic beams obtained in various research centers using the 2D-MOT technique. For the comparison, we have also plotted the brightness and phase space density of the atomic beam obtained in this study, as well as expected values of the phase space density that can be achieved using a 2D-MOT technique in the case of Doppler (point A) and sub-Doppler cooling (point B). As can be seen, our cold beam parameters obtained even without using the 2D-MOT technique are comparable to the analogous parameters achieved due to 2D-MOT. The arrows in Fig. 12 show the calculated values of brightness and phase space density of a cold atomic beam obtained by the proposed method in combination with 2D-MOT. According to Figs. 11 and 12, the proposed method of obtaining cooled atomic beams significantly improves the phase space density of a beam.

6. CONCLUSIONS

Using the method of Zeeman laser cooling in a transverse magnetic field, we obtained a source of cold ^{85}Rb atoms with a beam intensity of 7.2×10^{12} s^{-1} at an average atomic velocity of 12 m/s. The density of cold atoms in the source was 4.7×10^{10} cm^{-3}.

ACKNOWLEDGMENTS

The authors are grateful to A.P. Cherkun, I.V. Morozov, and D.V. Serebryakov for their help in preparing to experiments and to M.V. Subbotin for his active participation in the initial stage of experiments.

This study was supported in part by the Russian Foundation for Basic Research (project nos. 01-02-16337 and 02-02-17014), the Presidential Grant for Support of Leading Scientific Schools (project no. NSh-1772.2003.2), and INTAS (grant no. 479).

REFERENCES

1. W. Ketterle, Rev. Mod. Phys. **74**, 1131 (2002).
2. J. T. M. Walraven, in *Quantum Dynamics of Simple Systems*, Ed. by G. L. Oppo and S. M. Barnett (Inst. of Phys., London, 1996), p. 315.
3. W. D. Phillips, J. V. Prodan, and H. J. Metcalf, J. Opt. Soc. Am. B **2**, 1751 (1985).
4. T. E. Barrett, S. W. Dapore-Schwartz, M. D. Ray, *et al.*, Phys. Rev. Lett. **67**, 3483 (1991).
5. W. Ertmer, R. Blatt, J. L. Hall, *et al.*, Phys. Rev. Lett. **54**, 996 (1985).
6. W. Ketterle, A. Martin, M. A. Joffe, *et al.*, Phys. Rev. Lett. **69**, 2483 (1992).
7. M. Zhu, C. W. Oates, and J. S. Hall, Phys. Rev. Lett. **67**, 46 (1991).

8. M. Schiffer, M. Christ, G. Wokurka, *et al.*, Opt. Commun. **134**, 423 (1997).
9. E. Riis, D. S. Weiss, K. A. Moler, *et al.*, Phys. Rev. Lett. **64**, 1658 (1990).
10. J. Nellessen, J. Werner, and W. Ertmer, Opt. Commun. **78**, 300 (1990).
11. Z. T. Lu, K. L. Corwin, M. J. Renn, *et al.*, Phys. Rev. Lett. **77**, 3331 (1996).
12. S. N. Bagayev, V. I. Baraulia, A. E. Bonert, *et al.*, Laser Phys. **11**, 1178 (2001).
13. V. S. Letokhov, V. G. Minogin, and B. D. Pavlik, Zh. Éksp. Teor. Fiz. **72**, 1328 (1977) [Sov. Phys. JETP **45**, 698 (1977)].
14. N. F. Ramsey, *Molecular Beams* (Clarendon Press, Oxford, 1956; Inostrannaya Literatura, Moscow, 1960).
15. P. N. Melentiev, M. V. Subbotin, and V. I. Balykin, Laser Phys. **11**, 891 (2001).
16. K. L. Corwin, Z. T. Lu, C. F. Hand, *et al.*, Appl. Opt. **37**, 3295 (1998).
17. R. D. Swenumson and U. Even, Rev. Sci. Instrum. **52**, 559 (1981).
18. V. I. Balykin and V. G. Minogin, Zh. Éksp. Teor. Fiz. **123**, 13 (2003) [JETP **96**, 8 (2003)].
19. F. Lison, P. Schuh, D. Haubrich, *et al.*, Phys. Rev. A **61**, 13405 (2000).
20. A. Witte, T. Kisters, F. Riehle, *et al.*, J. Opt. Soc. Am. B **9**, 1030 (1992).
21. M. D. Hoogerland, D. Milic, W. Lu, *et al.*, Aust. J. Phys. **49**, 567 (1996).
22. W. Rooijakkers, W. Hogervorst, and W. Vassen, Opt. Commun. **123**, 321 (1996).
23. K. Dieckmann, R. J. C. Spreeuw, M. Weidenmuller, *et al.*, Phys. Rev. A **58**, 3891 (1998).

Translated by P. Pozdeev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 678–686.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 774–784.

ATOMS, SPECTRA, RADIATION

High-Order Harmonic Generation by Atoms in a Two-Color Laser Field: Phase Control of Recombination Radiation Spectrum and Duration

V. D. Taranukhin

Faculty of Physics, Moscow State University, Vorob'evy gory, Moscow, 119992 Russia
e-mail: tvd@ssf.phys.msu.su

Received April 21, 2003

Abstract—The feasibility of phase control over above-threshold tunnel ionization and subsequent recombination emission in two-frequency laser fields is studied. It is shown that, in such fields, we can control the instants of ionization t_0 (within optical cycle T) and recombination t_k. The conditions that minimize the characteristic times $\delta t_0 \ll T$ and $\delta t_k \ll T$, within which effective ionization and recombination occur, were found. Phase control allows recombination radiation to be generated with the selection of a narrow spectral range, while additional high-frequency "background illumination" sets up high harmonic "amplification" conditions. It was shown that special two-frequency pumping with elliptically polarized radiation can generate coherent electromagnetic pulses of attosecond width. The width of the pulses decreases as the intensity of pumping increases and can reach subattosecond values. Experimental generation of such pulses may lead to a breakthrough in the development of new methods for femto- and attosecond diagnostics of fast processes. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The feasibility of generating extremely narrow coherent radiation pulses 10^{-15}–10^{-17} s in width has been extensively studied to develop effective methods of femto- and attosecond metrology [1]. Current suggestions for generating femto- and attosecond pulses are related to the generation of high harmonics in the above-threshold tunnel ionization of atoms in strong laser fields [1, 2]. Generally, the spectrum of high harmonics is a broad slowly sloping plateau that extends from the pumping frequency ω_0 to the "cutoff" frequency $\Omega \approx U_i + 3.17U_p$, where U_i is the ionization potential of the atom and U_p is the ponderomotive potential of pumping radiation. It should be noted from the outset that a certain part of this spectrum is only needed for applications. Subfemtosecond pulses can be generated if the phase matching condition is satisfied for a group of harmonics from the plateau [3] or by applying a polarization gate [4]. The latter technique uses pumping radiation with time-dependent ellipticity. This allows the return of a photoelectron to the parent ion (and, accordingly, the duration of recombination radiation) to be controlled, because such a return is only possible at the instants when the pumping radiation is linearly polarized. In both cases, the duration of recombination radiation is $\tau_g \lesssim T = \lambda/c$ (T and λ are the optical cycle duration and the pumping wavelength, respectively, and c is the speed of light), which corresponds to a value on the order of one femtosecond.

Simultaneously, methods for measuring the width of such short pulses have been under development [1, 5]. These methods use the idea of the phase control of above-threshold tunnel ionization, that is, the control of the ionization instant (the instant of electron transfer into the continuum) within the optical radiation period. Phase control is possible because an electron produced in tunnel ionization begins its motion in the continuum (after the subbarrier evolution stage) at a zero velocity along the instantaneous direction of the electric pumping field. The motion of the electron in the continuum includes both oscillatory and drift components. The drift velocity of the electron contains information about the instant of atom ionization [5] and is recorded by a detector after the photoelectron ceases to interact with radiation. In this work, we study the feasibility of the phase control over the generation of high harmonics in two-frequency fields. It is shown below that the use of two-frequency radiation allows us:

(1) to control the spectrum of high harmonic generation, for instance, generate recombination radiation in a selected narrow spectral range rather than as a broad plateau;

(2) to control the duration of recombination radiation, which opens up a real possibility of generating

coherent pulses of width $\tau_g \sim 1$–10 attoseconds (1 as = 10^{-18} s);

(3) to create conditions for the "amplification" of high harmonics.

The recombination radiation that arises in the tunnel ionization of atoms followed by electron recombination with the parent ion is described by the wave equation with a source, which is the second derivative of the field-induced dipole moment $\mathbf{D}(t)$ with respect to time. The moment $\mathbf{D}$ is calculated by the Schrödinger equation, which, in particular, determines its dependence on time t. This dependence (along with propagation effects) controls the duration of recombination radiation. In this work, propagation effects, which are described by the wave equation and can both broaden and narrow recombination pulses, are not considered. We concentrate on single-atom response $\mathbf{D}(t)$ calculations.

2. A QUANTUM MODEL OF HIGH-ORDER HARMONIC GENERATION IN TWO-FREQUENCY FIELDS

The quantum model of high-order harmonic generation by one atom in a monochromatic field described in [6, 7] can be generalized to more complex pumping, for instance, pumping that includes radiation at two different frequencies. We will use the same approximation as in [6, 7], which corresponds to the tunnel conditions of the ionization of atoms. Under these conditions, high-order harmonic generation admits a quasi-classical description and clear physical interpretation as a process including three stages [8], namely, the ionization of an atom proper (photoelectron transfer into the continuum), kinetic energy gain by the electron to $\varepsilon \sim U_p$ (the above-threshold ionization stage), and the recombination of the electron with the parent ion with the emission of a photon at the frequency $\omega_k = kU_p + U_i$. The k factor varies in the range 0–3.17 depending on the phase (instant) of ionization φ, which results in the generation of a broad spectrum of high harmonics. The maximum generation frequency Ω corresponds to the ionization phase $\varphi \approx \pi/10$ counted from the pumping field maximum. Also note that, because the generation of high harmonics is an essentially nonlinear process, the "fine" structure of the spectrum of high harmonic generation is complex, and, if narrow pumping pulses are used, the spectrum is continuous.

As in [6, 7], we consider pumping radiation of a fairly high intensity I at which atoms experience tunnel ionization. This allows us to ignore the influence on ground state ionization of intermediate resonances, which correspond to the transition to the continuum from all the other bound states [9]. Also note that the probability of recombination into the ground atomic state is much higher than that of recombination into any other discrete state [10]. We can therefore ignore the contributions of all discrete states except the ground state $|0\rangle$ to the wave function $\Psi(\mathbf{r}, t)$,

$$\Psi(\mathbf{r}, t) = a(t)|0\rangle + \Psi_c, \quad \Psi_c = \int d\mathbf{p} b_p(t)|\mathbf{p}\rangle, \quad (1)$$

where $a(t)$ and $b_p(t)$ are the amplitudes of the ground atomic state and continuum states $|\mathbf{p}\rangle$ that correspond to electron states with momentum $\mathbf{p}$. On the other hand, if the intensity of radiation is insufficient for effecting above-the-barrier ionization, and the electron begins its evolution in the continuum fairly far from the parent ion (at the far boundary of a fairly broad potential barrier), then the influence of Coulomb forces on the motion of the electron in the continuum can be ignored. We also ignore bremsstrahlung against the background of recombination radiation. The introduced approximations allow the Schrödinger equation to be solved (the procedure for solving this equation is described in [6]). The amplitude of the $\mathbf{p}$ state of the continuum can be represented in the form

$$b_p(t) = i\int_0^t dt_0 a(t_0)\mathbf{E}(t_0)\mathbf{d}\left(\mathbf{p} - \frac{\mathbf{A}(t_0)}{c}\right) \times \exp[-iS(\mathbf{p}, t, t_0)], \quad (2)$$

where $\mathbf{d}(\mathbf{p}) = \langle \mathbf{p}|\mathbf{r}|0\rangle$ is the matrix element of the dipole moment of the transition from the ground state of the atom to the $\mathbf{p}$ state of the continuum and $\mathbf{E}(t_0)$ and $\mathbf{A}(t_0)$ are the amplitudes of the pumping radiation electric field and vector-potential at the instant of ionization t_0,

$$S(\mathbf{p}, t, t_0) = \int_{t_0}^t dt' \left\{ U_i + \frac{1}{2}\left[\mathbf{p} - \frac{\mathbf{A}(t')}{c}\right]^2 \right\}. \quad (3)$$

Note that the contribution to the b_p amplitude is formed during the whole effective ionization time [the integral over ionization instants t_0 in (2)], which causes the formation of a longitudinal structure of the wave packet in the continuum. Substituting (2) into (1) allows us to write the wave function Ψ_c in the form

$$\Psi_c(r, t) = i\int_0^t dt_0 B(\mathbf{p}_0, t_0) f(\mathbf{r}, t, t_0) \times \exp[-iS(\mathbf{p}_0, t, t_0) + i\mathbf{p}_0 \cdot \mathbf{r}], \quad (4)$$

where

$$B(\mathbf{p}_0, t_0) = a(t_0)\mathbf{E}(t_0)\mathbf{d}\left(\mathbf{p}_0 - \frac{\mathbf{A}(t_0)}{c}\right)$$

is the amplitude of the electron wave packet at the

instant t_0 and the function

$$f(\mathbf{r}, t, t_0) = \int d\mathbf{p} \frac{\mathbf{d}(\mathbf{p} - \mathbf{A}(t_0)/c)}{\mathbf{d}(\mathbf{p}_0 - \mathbf{A}(t_0)/c)} \times \exp[-iS(\mathbf{p}, t, t_0) + iS(\mathbf{p}_0, t, t_0) - i\mathbf{p}_0 \cdot \mathbf{r}]|\mathbf{p}\rangle \quad (5)$$

describes the shape and phase of the packet (packet spreading effects). In these equations, $\mathbf{p}_0(t, t_0)$ is the momentum that makes the major contribution to integral (5) and corresponds to the stationary phase determined by the condition $\partial S/\partial \mathbf{p} = 0$. The stationary phase can be used because the characteristic scale $p^2 \propto 1/(t - t_0)$ of changes in action S at times on the order of the optical period is much smaller than the $p^2 \propto U_i$ scale of changes in the matrix element $\mathbf{d}$. Further, the explicit form of the packet $f(\mathbf{r}, t, t_0)$ will be of no interest to us. It is, however, important that its center, which corresponds to momentum $\mathbf{p}_0$, moves along a classical trajectory (this directly follows from the condition $\partial S/\partial \mathbf{p} = 0$). The packet width will be described by a fairly strict equation [see Eq. (7) below] valid for the tunnel ionization conditions.

Taking (4) into account, the dipole moment

$$\mathbf{D}(t) = \langle \Psi^*(\mathbf{r}, t)|\mathbf{r}|\Psi(\mathbf{r}, t)\rangle$$

is obtained in the form

$$\mathbf{D}(t) = a^*(t)\int_0^t dt_0 B(\mathbf{p}_0, t_0) \times \langle 0|\mathbf{r}|f(\mathbf{r}, t, t_0)\exp[-iS(\mathbf{p}_0, t, t_0) + i\mathbf{p}_0 \cdot \mathbf{r}]\rangle + \text{c.c.} \quad (6)$$

Equation (6) describes the quasi-classical evolution of the photoelectron during above-threshold tunnel ionization and admits a clear physical interpretation. At the first stage of high harmonic generation, the electron wave packet is formed in the continuum. The packet amplitude $B(\mathbf{p}_0, t_0) \propto a(t_0)$ is determined by the probability W_i of tunnel ionization at time t_0 and takes into account ionization saturation (depletion of atomic states). Analytic results for dipole moment (6) were obtained in [6] in the approximation of a comparatively low pumping radiation intensity, at which the probability of ionization was small. This led to the Keldysh formula for the dependence of the probability of tunnel ionization on the instantaneous pumping field value F. The $W_i(F)$ dependence can, however, be substantially different in strong laser fields. The modification of the ionization probability W_i in strong fields is discussed in detail in [11] (also see [12]). Independent ionization probability calculations can be helpful in applying (6).

At the next stage of tunnel above-threshold ionization, the wave packet experiences evolution in the continuum under the action of pumping radiation. The $f(\mathbf{r}, t, t_0)$ function describes the shape of the packet (its spreading) and the evolution of its center. In strong (but not relativistic) fields, the electron path in the continuum L is much larger than the size of its wave packet σ; at the same time, $L \ll \lambda$. We can then use the quasi-classical approach, and the evolution of the center of the wave packet in the continuum can be described by a classical equation of motion; that is, classical trajectories for the wave packet center can be considered. For the Coulomb potential (when the electron is close to the parent ion), the applicability of such an approximation is not obvious. It can, however, be used because the electron largely gains kinetic energy (which is essential to high harmonic generation) far from the parent ion. In addition, the first return of the photoelectron to the parent ion makes the major contribution to high harmonic generation, which allows us to assume that the shape of its wave packet does not change significantly during its evolution. At the same time, the effect of wave packet spreading is of importance. It can be described as the spreading of a free particle packet but at a V_{sp} rate determined by tunnel ionization [13],

$$\sigma(t) = \sqrt{\sigma^2(t_0) + V_{sp}^2(t - t_0)^2}, \quad V_{sp} \approx F_i^{1/2}(2U_i)^{-1/4}, \quad (7)$$

where $F_i = F(t_0)$ is the pumping field at the instant of ionization. We suggest to use the $V_{sp} \approx 1$ nm/fs value (at pumping radiation intensities $I \approx 10^{14}$–10^7 W/cm^2), which closely agrees with the experimental value from [4].

During its evolution in the continuum, the electron gains the energy $\varepsilon = \partial S/\partial t$ from the field. After returning to the parent ion, it can recombine and emit a photon with the energy $\omega_k = U_i + \varepsilon$. The $\langle 0|\mathbf{r}|\ldots\rangle$ matrix element in (6) determines both the probability of recombination and the duration of recombination radiation from a single atom, because this matrix element is only nonzero when the electron closely approaches the parent ion (when the electron wave packet and the wave function of the ground state of the atom noticeably overlap). Lastly, note that the integral in t_0 in (6) describes the coherent contribution to high harmonic generation of electrons released from an atom at various times t_0, that is, the longitudinal structure of the wave packet.

Such a clear physical interpretation of high harmonic generation, which is possible because the process is quasi-classical, allows us to extend the applicability of the high harmonic generation model suggested in [6] to fairly high pumping radiation intensities [11], take into account the three-dimensional character of the evolution of the electron in the continuum, and consider pumping fields of arbitrary configurations [14]. In the next sections, we discuss pumping by combined radiation whose components have different frequencies and are differently polarized.

3. GENERATION OF RECOMBINATION RADIATION IN A TWO-FREQUENCY FIELD WITH DIFFERENT COMPONENT POLARIZATIONS: GENERATION OF ATTOSECOND PULSES

It is known that usual circularly polarized pumping does not cause harmonic generation [4]. This is so because, after ionization, the photoelectron goes far away from the parent ion and never returns to it; that is, at all ionization instants, there are no collisional electron trajectories. Conversely, collisional trajectories exist at any ionization instant if pumping radiation is linearly polarized. Let us show that, under special two-frequency pumping conditions (including circularly polarized fields), high harmonic generation is possible and it becomes feasible to control generation parameters, recombination radiation duration in particular.

Consider two-frequency pumping, which is a combination of a high-frequency elliptically polarized field and a low-frequency linearly polarized field,

$$\mathbf{F} = F_0(\hat{\mathbf{x}}\sqrt{1-\alpha^2}\cos(\omega t) + \hat{\mathbf{y}}\alpha\sin(\omega t)) + \mathbf{F}_{dc}, \quad (8)$$

$$\mathbf{F}_{dc} = -\beta F_0(\hat{\mathbf{x}}\sqrt{1-\gamma^2} + \hat{\mathbf{y}}\gamma), \quad (9)$$

where F_0, ω, and α are the amplitude, frequency, and ellipticity of the high-frequency pumping component; $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are the unit vectors; and β and γ are the parameters that determine the relative amplitude and direction of the linearly polarized low-frequency pumping component $\mathbf{F}_{dc}$. The role of the low-frequency component can be played by CO_2 laser radiation synchronized with the high-frequency field. During one optical cycle of high-frequency radiation at $\lambda \sim 1$ μm, the CO_2 laser field can be considered constant. For simplicity, we only consider collinear propagation of the low-frequency and high-frequency fields. As has been mentioned above, the evolution of the photoelectron in the continuum (in tunnel ionization) can be described by classical trajectories of the center of the electron wave packet. Solving the classical equations of motion in field (8) with zero initial conditions for the coordinates and velocity of the electron [7] leads to the following trajectories:

$$x(t) = -\sqrt{1-\alpha^2} \times [\cos t - \cos t_0 + (t-t_0)\sin t_0 + \beta_1(t-t_0)^2], \quad (10)$$

$$y(t) = -\alpha[\sin t - \sin t_0 - (t-t_0)\cos t_0 + \beta_2(t-t_0)^2], \quad (11)$$

where

$$\beta_1 = \frac{\beta}{2}\sqrt{\frac{1-\gamma^2}{1-\alpha^2}}, \quad \beta_2 = \frac{\beta\gamma}{2\alpha}.$$

Here and throughout this section, we use dimensionless variables, namely, time is normalized with respect to $1/\omega$; coordinates, with respect to $(eF_0/m\omega^2)$; velocity, with respect to $(eF_0/m\omega)$; and energy, with respect to the ponderomotive energy of the high-frequency pumping component $U_p = e^2F_0^2/4m\omega^2$, where m and e are the mass and charge of the electron. For collisional trajectories, both x and y tend to zero simultaneously at time t_k.

It follows from (10) and (11) that, in contrast to one-frequency pumping, collisional electron trajectories (see Fig. 1) only exist for certain sets of the α, β, and γ parameters and certain ionization instants t_0. It is desirable that the following requirements be met when selecting these parameters: The electron energy ε_k at the recombination instant should be higher (for generating harmonics of higher orders), the time spent by the electron in the continuum should be shorter (this decreases wave packet spreading and increases recombination effectiveness), the ratio between the amplitudes of the low- and high-frequency fields should be smaller (because of the difference in the power of 1 and 10 μm lasers), and a single collisional trajectory should exist during the whole period of pumping (for generating a single recombination radiation pulse). An analysis of (10) and (11), however, shows that there is no unambiguous solution to this problem, namely, there exist many (generally, infinitely many) various sets of the α, β, γ, and t_0 parameters that lead to collisional trajectories. Of the greatest importance for generating the shortest recombination radiation pulses is to minimize the time during which the electron occurs close to the parent ion, that is, in the region

$$\Delta r \le \sigma(t_k), \quad (12)$$

where dipole moment (6) is nonzero. This is attained by increasing both the kinetic energy of the electron at the instant of recombination ε_k (this energy is limited by the intensity of radiation used for pumping) and the rate of the "transition" from collisional to noncollisional electron trajectories. The latter factor is in effect when the $dr(t_k)/dt_0$ derivative, where $r = \sqrt{x^2 + y^2}$, is fairly large, because, at given α, β, and γ parameters, the ionization instant t_0 determines whether or not the electron trajectory is collisional. The time of electron interaction with the parent ion can be substantially decreased by selecting the pumping two-component field (8) parameters that determine the spatial orientation of the electron wave packet at the instant of its return to the nucleus. The packet can be oriented in such a way that its over-

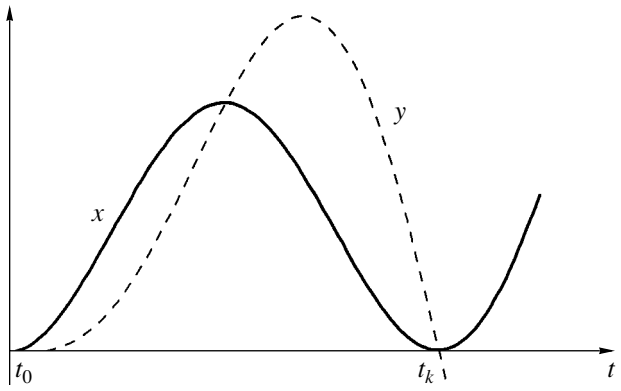


Fig. 1. Time dependence of photoelectron coordinates in the continuum [see (10), (11)] for case *Y* at $\alpha = 1/\sqrt{2}$, $\beta \approx -0.276$, $\gamma \approx -0.96$, and $t_0 = \pi/10$.

lap with the parent ion region (and, therefore, recombination) would only involve a small longitudinal packet part that corresponds to a short interval of ionization instants $\delta t_0 \ll 2\pi$. The time of interaction with the ion (and the duration of recombination) is then determined by the cross size of packet (7).

We found two families of solutions to (10) and (11) that give large $dr(t_k)/dt_0$ derivative values. We denote them by *X* and *Y*. Case *Y* is shown in Fig. 1. It corresponds to electron–ion collision instants (recombination instants) t_k that are found from (10), according to which $x(t_k) = 0$. We then have

$$\frac{dx(t_k)}{dt_0} \equiv 0,$$
$$\frac{dy(t_k)}{dt_0} \equiv \frac{\partial y}{\partial t_0} + \frac{\partial y}{\partial t_k}\frac{\partial t_k}{\partial t_0} = \infty. \quad (13)$$

The first equation in (13) is a corollary to the $x(t_k) = 0$ equation, which determines the $t_k(t_0)$ function, and the second equation ensures a large $dr(t_k)/dt_0$ derivative value. The condition that must be satisfied by the α, β, and γ pumping parameters and the ionization instant t_0 for the collisional trajectory to exist and be characterized by the shortest time during which the electron occurs close to the parent ion is determined by (13) and (10), (11),

$$\sin t_k - \sin t_0 = 2\beta_1\tau, \quad (14)$$

where $\tau = t_k - t_0$.

Case *X* corresponds to the opposite situation,

$$\frac{dy(t_k)}{dt_0} \equiv 0, \quad \frac{dx(t_k)}{dt_0} = \infty. \quad (15)$$

The α, β, γ, and t_0 parameters must then satisfy the condition

$$\cos t_k - \cos t_0 = -2\beta_2\tau. \quad (16)$$

Neither (14) nor (16) determines a unique set of the α, β, γ, and t_0 parameters at which collisional trajectories exist. They, however, considerably facilitate seeking such sets with a computer.

The above reasoning suggests the following procedure. Equation (14) or (16) is used to determine a suitable set of the α, β, and γ pumping parameters and ionization instant t_0. Given these pumping parameters, computer simulation of electronic trajectories (10), (11) is performed for various ionization instants in the vicinity of t_0 to determine the maximum deviation of the ionization instant δt_0 (and the corresponding maximum deviation of the recombination instant δt_k) at which recombination condition (12) is still satisfied; that is, at which the deviation of the electronic trajectory from the parent ion is smaller than the electron wave packet width. Photoelectrons that appear in the continuum outside the δt_0 interval do not collide with the parent ion and do not contribute to recombination radiation. This procedure is used to determine the "elementary" duration of recombination radiation (radiation of a single atom) in the process of high harmonic generation, namely, $\tau_g \approx \delta t_k$.

Below, we present the results of numerical experiments performed for cases *X* and *Y* to determine the duration of recombination radiation τ_g for high harmonic generation under two-frequency pumping conditions (8), (9). Let the high-frequency component be laser radiation with a wavelength of $\lambda = 1$–0.8 μm and intensity of $I_{HF} = 10^{15}$–10^{17} W/cm^2. Such radiation intensities have already been used in experimental and computational studies of above-threshold ionization and high harmonic generation on ions (the use of ions ensures tunnel ionization, which is essential for high harmonic generation). For instance, the emission of high harmonics by He^+ ions under pumping by radiation of intensity $I = 10^{17}$ W/cm^2 was calculated in [15]. In [16], lasers with radiation intensity $I \sim 10^{16}$–10^{17} W/cm^2 were used to effect the tunnel ionization of noble gases (Ar, Kr, Ne, and Xe). High by charged ions (with charges up to $Z = 8$) were observed. High harmonic generation by such ions under pumping with intensity $I \sim 10^{16}$–10^{19} W/cm^2 was calculated in [17, 18].

Case *X*. The set of parameters that satisfied condition (16) for pumping (8), (9) with intensity $I_{HF} = 10^{17}$ W/cm^2 was $\alpha = 1/\sqrt{2}$ (this corresponded to circularly polarized high-frequency radiation), $\beta \approx 0.4$, $\gamma \approx 0.65$. The only ionization instant (during the optical cycle of the high-frequency field) was $t_0 \approx -\pi/5$ (t_0 was counted from the high-frequency field maximum). Numerical examination of electron trajectories (10), (11) with such parameters gave the recombination instant $t_k \approx 1.2\pi$ and the kinetic energy of the electron

at the recombination instant $\varepsilon_k \approx 3.6$. The corresponding deviations of t_0 and t_k [at which recombination condition (12) was still satisfied] were $\delta t_0 \approx \delta t_k \approx 0.02$. Such a δt_k value corresponded to the dimensional duration of recombination radiation by a single atom $\tau_g \approx$ 10 as. Note once more that the obtained set of the α, β, γ, and t_0 parameters, which ensures the generation of such short pulses, is neither unique nor optimal. It follows that the actual τ_g (at the selected high-frequency radiation parameters) can be still shorter. Also note that a consideration of pumping with intensity $I_{HF} \sim 10^{17}$–10^{18} W/cm^2 generally requires taking into account the influence of the magnetic field of radiation on electron trajectories [11], which, however, does not change the τ_g value substantially, although the exact parameter values [and condition (16)] can change.

Case *Y.* Equation (14) gave the following set of parameters for pumping with $I_{HF} = 10^{15}$ W/cm^2: $\alpha = 1/\sqrt{2}$, $\beta \approx -0.22$, $\gamma \approx -1$, and $t_0 = 0$. The simulation of electron trajectories (10), (11) with these parameters gave $t_k \approx 2\pi$, $\varepsilon_k \approx 4$, and $\delta t_0 \approx \delta t_k \approx 0.25$, which corresponded to recombination radiation duration $\tau_g \approx$ 100 as. Increasing the I_{HF} intensity decreases τ_g. For instance, the parameters found for $I_{HF} = 5 \times 10^{17}$ W/cm^2 were $\alpha = 1/\sqrt{2}$, $\beta \approx -0.276$, $\gamma \approx -0.96$, and $t_0 = \pi/10$. These parameters gave $t_k \approx 1.9\pi$, $\varepsilon_k \approx 4.5$, $\delta t_0 \approx 0.01$, and $\delta t_k \approx 0.002$, which corresponded to the recombination radiation duration $\tau_g \approx 0.9$ as. Obtaining such pulses experimentally would mean a breakthrough in the development of new methods for femto- and attosecond diagnostics of fast processes [1]. Note that the τ_g value found above is more than three orders of magnitude smaller than the optical period T of the high-frequency field. Such a pulse, however, contains about 50 recombination radiation optical cycles (whose duration is determined by the ε_k parameter) and is almost monochromatic in this sense.

The above estimates are valid for each high-frequency field optical cycle. In each cycle, we observe a recombination radiation burst of duration τ_g. If the high-frequency radiation pulse is long, this results in the generation of a train of attosecond pulses with the repetition frequency $\sim\omega$. Such trains can, for instance, be used in petahertz spectroscopy with attosecond time resolution. The isolation of one pulse from such a train is, however, a difficult problem.

With extremely short high-frequency radiation pulses (when the ionization of an atom effectively occurs only during one optical cycle), the use of two-frequency pumping (8), (9) opens up the possibility of generating a single attosecond recombination radiation pulse automatically (without additional experimental efforts). For instance, consider an extremely short high-frequency pumping pulse with peak intensity $I_{HF} = 10^{18}$ W/cm^2 and field envelope

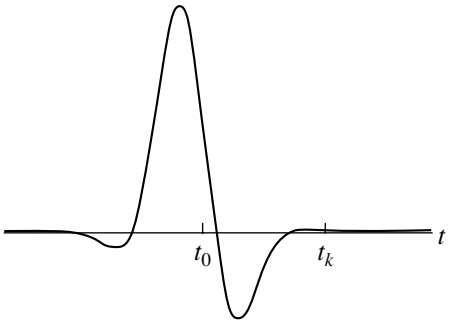


Fig. 2. Example of an extremely narrow high-frequency pumping pulse for generating a single attosecond pulse [see (17)].

$$F_0(t) \sim \exp\left[-\frac{(t-7)^2}{4}\right], \qquad (17)$$

which describes the pulse shown in Fig. 2. The procedure suggested above gives $\alpha = 0.708$, $\beta \approx -0.275$, $\gamma \approx -0.96$, and $t_0 = 7.3575$ for such a pumping pulse. This corresponds to a single collisional trajectory ($t_k \approx 12$) during the whole pulse (17). Varying the t_0 parameter gives $\delta t_0 \approx 0.02$ and $\delta t_k \approx 0.06$; that is, the width of a single recombination radiation pulse is $\tau_g \approx 30$ as.

These simulations lead us to conclude that the suggested approach opens up the possibility in principle of generating 1–10 as coherent electromagnetic radiation pulses and even of overcoming the subattosecond barrier. In essence, this approach is a variety of the method for phase control of tunnel ionization, because it is based on the selection of photoelectrons (within the ionization interval δt_0) that contribute to recombination radiation generation. Let us shortly consider the efficiency of this generation. As with usual pumping, the efficiency is determined by three factors, namely, the probability of ionization of an atom (ion), the rate of electron wave packet spreading in the continuum, and the probability of electron–ion recombination. The use of combined pumping (including circularly polarized waves) does not charge the probability of tunnel ionization to any significant extent, because this probability is determined by instantaneous field strength. Wave packet spreading and recombination probabilities per photoelectron are approximately the same as under usual high harmonic generation conditions. Pumping (8), (9), however, selects only a small part of all photoelectrons that recombine with the parent ion ($\delta t_0/T \sim 0.1$–0.01). This results in both the generation of short recombination radiation pulses and the selection of a narrow region $\delta\omega_g$ in a wide plateau of the spectrum of high harmonic generation. Note that there is no contradiction between the generation of a short recombination radiation pulse (of duration τ_g) and a relatively narrow ($\delta\omega_g$) genera-

tion spectrum; indeed, $\delta\omega_g\tau_g \gg 1$ thanks to the high frequency of recombination radiation. It follows that, when we use pumping (8), (9), the total energy of high harmonic generation decreases (proportionally to $\delta t_0/T$ or $\delta\omega_g/\Omega$). The intensity (recombination radiation harmonics), however, remains approximately the same as under usual pumping conditions within the $\delta\omega_g$ spectral range.

4. ACTIVE PHASE CONTROL OF TUNNEL IONIZATION AND THE GENERATION OF RECOMBINATION RADIATION WITH THE SELECTION OF A NARROW SPECTRAL RANGE

In the preceding section, we considered passive phase control of tunnel ionization, when the ionization of an atom occurred at all time instants during optical cycle T, whereas the selection of photoelectrons occurred at the stage of their recombination with the parent ion. The two-frequency pumping parameters are selected such that only those electrons recombine that "go away" from the atom during a fairly short time interval $\delta t_0/T \ll 1$. The short duration of recombination radiation is then attained because of the high sensitivity of recombination condition (12) to ionization instant t_0 variations. Using two-frequency radiation, we can also exercise active phase control, that is, the selection of photoelectrons already at the atom ionization stage (when ionization only occurs during a small fraction of optical period T). The overall degree of medium ionization is then much lower than under passive control conditions. A substantial concentration of free electrons in a medium is known to cause radiation defocusing, which imposes limitations on increasing the effectiveness of high harmonic generation [2]. It follows that the selection of ionization instants decreases pumping radiation defocusing and allows the phase matching length and, therefore, the effectiveness of high harmonic generation to be substantially increased. While remaining within the framework of analyzing the single-atom response, let us show that two-frequency pumping of a special kind also allows the width of the recombination spectrum to be substantially decreased and thereby the effectiveness of using it to be increased.

Let pumping be performed by the following two-component field: a strong low-frequency field $\mathbf{F}_L$ (which by itself does not cause the tunnel ionization of an atom) and an extremely short high-frequency pulse field $\omega_H > U_i$, capable, for instance, of inducing the one-photon ionization of an atom. We assume that the high-frequency pulse duration is substantially smaller than the optical period T_L of the low-frequency field. We can write

$$F = F_L\cos(\omega_L T) + F_H\cos(\omega_H t + \varphi), \qquad (18)$$

where $F_L \gg F_H$ are the amplitudes and $\omega_L \ll \omega_H$ are the frequencies of the low- and high-frequency fields, respectively. Let both fields be polarized linearly and parallel to each other (along axis x) and synchronized in such a way that the high-frequency pulse corresponds to a certain phase φ of the optical cycle of the low-frequency field. If the F_L amplitude is insufficient for tunneling an electron from an atom, the ionization of the atom is determined by the short high-frequency radiation pulse. Conversely, the above-threshold ionization stage is determined by the low-frequency field. Note that such a scheme of high harmonic generation is fully equivalent to the scheme used in [5] to measure the pulse duration in the subfemtosecond range.

Further calculations and estimates refer to the following experimental situation: The low-frequency component is CO_2 laser radiation of intensity $I_L = 6 \times 10^{13}$ W/cm^2 (the tunnel ionization, for instance, of helium atoms by this radiation can be ignored, but its ponderomotive energy is substantial, $U_p \approx 500$ eV). The high-frequency component is radiation at an $\omega_H \sim 25$ eV frequency of duration $\tau_H \sim 1$–3 fs and intensity $I_H \sim 10^{10}$ W/cm^2. Such parameters are attained in modern high-harmonic generation experiments [2].

If the high-frequency field satisfies the inequality $\omega_H \gtrsim U_i$, the velocity of the photoelectron at the exit to the continuum is $V_0 \approx 0$ or, more exactly, $0 < V_0 < V_{0m}$. The maximum initial velocity of the photoelectron in the absence of phase modulation of the high-frequency pulse, $V_{0m} = (2\hbar/m\tau_H)^{1/2}$, is determined by its width τ_H. The probability of ionization is then proportional to the intensity I_H and the threshold cross section of one-photon ionization. The further evolution of the photoelectron wave packet largely occurs under the action of the low-frequency field (because of the condition $F_L \gg F_H$) and, as has been shown above, can be described by the classical trajectory $x(t)$ of the center of the packet. If Coulomb attraction to the parent ion is ignored (which can be done if $U_p \gg U_i$), the equation for this trajectory has the simple form

$$\frac{d^2x}{dt^2} = \frac{e}{m}F_L\cos(\omega_L t + \varphi). \qquad (19)$$

Solving (19) with the initial conditions $x(t = 0) = 0$ and $V(t = 0) = V_0$ allows calculation of the kinetic energy of the electron ε_k at the instant of its return to the parent ion (that is, at the $x \approx 0$ point where it recombines with the parent ion). Wave packet spreading (7) then only influences the effectiveness of electron recombination and has no significant effect on its kinetic energy ε_k. The ε_k energy depends on the initial photoelectron velocity V_0 and ionization phase φ (Fig. 3). The latter is determined by the high-frequency field (the phases $\varphi_{1,2}$ in Fig. 3 correspond to the high-frequency pulse start and end points), which allows us to control the recombination radiation spectrum. Note that the return of the photoelectron to its parent ion and its recombination

with it only occur when the direction of velocity V_0 is opposite to that of the low-frequency field. As one-photon ionization generates a symmetrical two-lobe wave packet ($\pm V_0$), half the total number of the photoelectrons do not contribute to recombination radiation.

The electron energies ε_k (actually, the recombination spectrum $\omega_k = \varepsilon_k + U_i$) calculated by (19) are plotted in Fig. 3, which shows that the high-frequency pulse gates ionization phases ($\varphi_1 < \varphi < \varphi_2$) and causes generation only in a narrow spectral range $\Delta\Omega$. The spectral components that lie between the $\varepsilon_k(\varphi)$ curves for $V_0 = 0$ and $V_0 = V_{0m}$ and correspond to the ionization phases $\varphi_1 < \varphi < \varphi_2$ only contribute to $\Delta\Omega$. As is shown in Fig. 3, a high-frequency pulse of width $\tau_H = (\varphi_2 - \varphi_1)T_L/2\pi \approx 1.2$ fs close to the relative phase $\varphi \approx \pi/10$ generates recombination radiation in the narrow frequency range $\Delta\Omega \approx 0.12\Omega_1$, where $\Omega_1 = \Omega - U_i$. The central frequency of the recombination spectrum can be retuned by changing the amplitude of the low-frequency field, because $\Omega_1 \propto I_L$. The V_{0m} value used in Fig. 3 corresponds to a high-frequency pulse with certain phase modulation. For a high-frequency pulse of the same width ($\tau_H = 1.2$ fs) but without phase modulation, the width $\Delta\Omega$ of the spectrum is smaller by about 30%.

A decrease in the width of the high-frequency pulse increases the initial electron velocity V_{0m}. At fairly small τ_H, this can broaden the recombination spectrum rather than narrow it. It follows that there exists an optimal high-frequency pulse width at which the recombination spectrum width is minimum. If the $\varepsilon_k(\varphi)$ dependence is close to parabolic at the top (Fig. 3) and there is no phase modulation of the high-frequency pulse, (19) yields

$$\frac{\Delta\Omega}{\Omega_1} = A\left(\tau_H + \frac{B}{\tau_H^{1/2}}\right)^2. \qquad (20)$$

If T_L and τ_H are in femtoseconds and Ω_1 is in electron volts, then $A \approx 55.8/T_L^2$ and $B \approx 0.17T_L/\Omega_1^{1/2}$. It follows from (20) that the smallest recombination spectrum width $\Delta\Omega_{\min}$ is observed at the high-frequency pulse width $\tau_0 = (B/2)^{2/3} \propto T_L^{2/3}\Omega_1^{-1/3}$ and is given by the equation

$$\frac{\Delta\Omega_{\min}}{\Omega_1} = 9A\tau_0^2. \qquad (21)$$

For $T_L \approx 35$ fs (CO_2 laser low-frequency radiation), the following estimates were obtained: $\tau_0 \approx 0.2$ fs and $\Delta\Omega_{\min}/\Omega_1 \approx 0.016$. The absolute recombination spectrum width increases as the intensity of low-frequency radiation grows, $\Delta\Omega_{\min} \propto \Omega_1^{1/3}$, $\Omega_1 \propto I_L$. However simultaneously, the relative spectrum width decreases, $\Delta\Omega_{\min}/\Omega_1 \propto \Omega_1^{-2/3}$, and can be noticeably smaller than 1%.

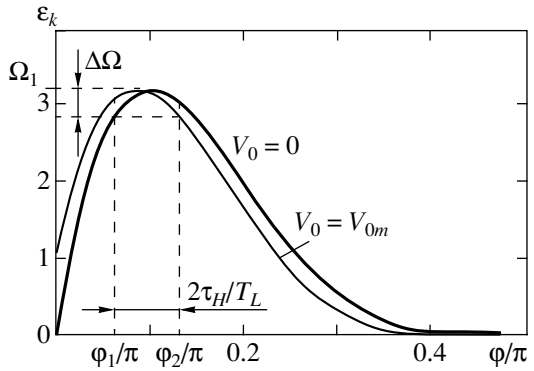


Fig. 3. Dependence of recombination radiation frequency on ionization phase (counted from the low-frequency field maximum) at various initial photoelectron velocities; the gating of the recombination spectrum by a high-frequency pulse of width τ_H is shown.

To summarize, the generation of an extremely narrow recombination spectrum requires the use of high-frequency pulses of optimum width without phase modulation. These pulses should be synchronized with the low-frequency field near the optimum phase $\varphi \lesssim \pi/10$ (the higher the initial electron velocity V_{0m}, the closer the high-frequency pulse must be to the low-frequency field maximum). Also note that a single high-frequency pulse causes a single event of recombination radiation generation. This results in a continuous generation spectrum and decreases generation effectiveness. The effectiveness of generation can be increased if a train of high-frequency pulses is used. The synchronization of two lasers is then, however, a more complex experimental task.

Note in conclusion that the mechanism of ionization of an atom by two-frequency field (18) considered in this work also explains the absorption of recombination radiation (harmonics) in the usual (one-component) scheme for high harmonic generation. Atoms under the action of pumping radiation also experience the influence of weak recombination radiation. Recombination radiation photons change the effective ionization potential of the atom, $U_{i\,\mathrm{eff}} = U_i - \omega_H$ [19], and can therefore participate in the tunnel ionization of an atom by the low-frequency pumping field. Because of the exponential dependence of the probability of tunnel ionization on the ionization potential, this causes a sharp (by several orders of magnitude) increase in the rate of ionization and, therefore, effective absorption of recombination radiation. This effect limits the effectiveness of high harmonic generation when the recombination radiation absorption length is shorter than the active medium length or coherence length. When additional radiation (high-frequency background illumination) is used, this effect can, however, be inverted and used to increase the effectiveness of high harmonic generation.

For instance, if the main pumping is accompanied by comparatively weak radiation at the frequency of one of the harmonics, the rate of the tunnel ionization of atoms considerably increases. This in turn causes an increase in the intensity of all recombination spectrum components. There is no real radiation amplification at the frequency of the harmonic used for additional (background) pumping because of a fairly low efficiency of high harmonic generation. The intensity of the other components of the high harmonic spectrum, however, noticeably increases.

5. CONCLUSIONS

To summarize, two-frequency fields allow us to exercise phase control over the above-threshold tunnel ionization process. A consequence of such control is the unique possibility of controlling recombination radiation parameters. In this work, we showed the feasibility of actively controlling the recombination radiation spectrum (its central frequency and width, which can be decreased by two orders of magnitude) for the example of single atom radiation. We also showed the possibility in principle of generating coherent pulses of width τ_g = 1–10 as. Obtaining such pulses in experiments would signify a fundamentally new step in the development of methods for diagnostics of fast processes. It should especially be noted that conditions (14), (16) for attaining such pulse widths admit further selection of two-frequency pumping parameters. Additional efforts (experimental or in the field of numerical simulations) will allow us to obtain coherent radiation pulses of width actually smaller than one attosecond.

Because of propagation effects and spatial inhomogeneity of laser radiation, the calculation of recombination radiation can change (increase or decrease), which requires a separate study [20]. Of primary importance for any such study are, however, the single-atom response calculations performed in this work.

REFERENCES

1. M. Hentschel, R. Kienberger, Ch. Spielmann, *et al.*, Nature **414**, 509 (2001); M. Drescher, M. Hentschel, R. Kienberger, *et al.*, Science **291**, 1923 (2001); H. Niikura, F. Legare, R. Hasbani, *et al.*, Nature **417**, 917 (2002); A. Baltuska, Th. Udem, M. Uiberacker, *et al.*, Nature **421**, 611 (2003); P. M. Paul, E. S. Toma, P. Breger, *et al.*, Science **292**, 1689 (2001); R. Kienberger, M. Hentschel, M. Uiberacker, *et al.*, Science **297**, 1144 (2002).
2. T. Brabec and F. Krausz, Rev. Mod. Phys. **72**, 545 (2000).
3. Ph. Antoine, A. L'Huillier, and M. Lewenstein, Phys. Rev. Lett. **77**, 1234 (1996).
4. P. B. Corkum, N. H. Burnett, and M. Y. Ivanov, Opt. Lett. **19**, 1870 (1994).
5. E. Constant, V. D. Taranukhin, A. Stolow, and P. B. Corkum, Phys. Rev. A **56**, 3870 (1997).
6. M. Lewenstein, Ph. Balcou, M. Yu. Ivanov, *et al.*, Phys. Rev. A **49**, 2117 (1994).
7. R. V. Kulyagin and V. D. Taranukhin, Kvantovaya Élektron. (Moscow) **23**, 866 (1996) [Quantum Electron. **26**, 866 (1996)].
8. P. B. Corkum, Phys. Rev. Lett. **71**, 1994 (1993).
9. E. Mevel, P. Breger, R. Trainham, *et al.*, Phys. Rev. Lett. **70**, 406 (1993).
10. J. B. Watson, A. Sanpera, and K. Burnett, Phys. Rev. A **51**, 1458 (1995).
11. V. D. Taranukhin, Laser Phys. **10**, 330 (2000).
12. A. Scrinzi, M. Geissler, and T. Brabec, Phys. Rev. Lett. **83**, 706 (1999).
13. A. M. Perelomov, V. S. Popov, and M. V. Terent'ev, Zh. Éksp. Teor. Fiz. **51**, 309 (1966) [Sov. Phys. JETP **24**, 207 (1966)].
14. V. D. Taranukhin and N. Yu. Shubin, J. Opt. Soc. Am. B **17**, 1509 (2000).
15. M. W. Walser, C. H. Keitel, A. Scrinzi, and T. Brabec, Phys. Rev. Lett. **85**, 5082 (2000).
16. S. Augst, D. Strickland, D. D. Meyerhofer, *et al.*, Phys. Rev. Lett. **63**, 2212 (1989).
17. D. B. Milosevic, S. Hu, and W. Becker, Phys. Rev. A **63**, 011403R (2001).
18. V. D. Taranukhin and N. Yu. Shubin, J. Opt. Soc. Am. B **19**, 1132 (2002).
19. N. B. Delone, N. L. Manakov, and A. G. Faĭnshteĭn, Zh. Éksp. Teor. Fiz. **86**, 906 (1984) [Sov. Phys. JETP **59**, 529 (1984)].
20. N. Milosevic, A. Scrinzi, and T. Brabec, Phys. Rev. Lett. **88**, 093905 (2002).

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 687–690.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 785–788.

ATOMS, SPECTRA, RADIATION

Scattering of a Low-Energy Electron in a Strong Coulomb Field

A. I. Milstein and I. S. Terekhov

Budker Institute of Nuclear Physics, Siberian Division, Russian Academy of Sciences, pr. Akademika Lavrent'eva 11, Novosibirsk, 630090 Russia

Novosibirsk State University, ul. Pirogova 2, Novosibirsk, 630090 Russia

e-mail: milstein@inp.nsk.su

Received October 28, 2003

Abstract—The low-energy electron scattering cross section in a strong Coulomb field is analyzed theoretically. It is shown that the exact cross section in a wide energy range significantly differs from the results obtained in the first Born approximation and in the nonrelativistic approximation. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

An explicit expression for the scattering cross section of an electron with an arbitrary energy in a strong Coulomb field was derived long ago by Mott [1] and contained an infinite series in Legendre polynomials. Although the methods of summation of this series were developed in a number of publications, the numerical calculation of the cross section remains a complicated problem. A detailed review of publications devoted to this problem can be found in monograph [2]. The electron scattering cross sections in a Coulomb field were calculated in [3–5] for various scattering angles, nuclear charges Z, and electron energies above 0.023 MeV. In an analysis of backward scattering for $Z = 80$, it was proved that the ratio of the exact relativistic cross section to the nonrelativistic (Rutherford) cross sections increases from 0.15 to 2.35 upon a decrease in the electron energy from 1.675 to 0.023 MeV. In view of such a large difference between the exact result and that obtained in the nonrelativistic approximation, it would be interesting to study the behavior of the exact scattering cross section for slow electrons in a strong Coulomb field. Here, we obtain the answer to this question by calculating the asymptotic form of the cross section for an arbitrary Z and a low electron energy.

2. ELECTRON SCATTERING CROSS SECTION IN A COULOMB FIELD

The electron wave function $\psi_{\lambda\mathbf{p}}(\mathbf{r})$ in an external field can be derived from the Green function $G(\mathbf{r}_2, \mathbf{r}_1|\varepsilon)$ of the Dirac equation in this field. We will use the familiar relation

$$\lim_{r_1\to\infty} G(\mathbf{r}_2, \mathbf{r}_1|\varepsilon) = -\frac{\exp(ipr_1)}{4\pi r_1}\sum_{\lambda=1,2}\psi^{(+)}_{\lambda\mathbf{p}}(\mathbf{r}_2)\bar{u}_{\lambda\mathbf{p}}, \quad (1)$$

$$u_{\lambda\mathbf{p}} = \sqrt{\varepsilon+m}\begin{pmatrix}\varphi_\lambda\\ \dfrac{\boldsymbol{\sigma}\cdot\boldsymbol{\rho}}{\varepsilon+m}\phi_\lambda\end{pmatrix},$$

where $p = \sqrt{\varepsilon^2 - m^2}$; m is the electron mass; $\psi^{(+)}_{\lambda\mathbf{p}}(\mathbf{r})$ denotes the solution to the Dirac equation, containing at infinity a plane wave with momentum $\mathbf{p} = -p\mathbf{n}_1$ ($\mathbf{n}_{1,2} = \mathbf{r}_{1,2}/r_{1,2}$) and a diverging spherical wave; $\lambda = \pm 1$ denotes two independent spinors ϕ_λ; and $\hbar = c = 1$. In the Coulomb field, the right-hand side of expression (1) contains the additional factor $(2pr_1)^{iq}$, where $q = Z\alpha\varepsilon/p$ and $\alpha = 1/137$ is the fine-structure constant. A convenient integral representation of the electron Green function in a Coulomb field was derived in [6]. Using equalities (19)–(22) from [6], we obtain

$$\psi^{(+)}_{\lambda\mathbf{p}}(\mathbf{r}_2) = \sqrt{\varepsilon+m}\sum_{l=1}^{\infty}\begin{pmatrix}f_1\\ \dfrac{p\boldsymbol{\sigma}\cdot\mathbf{n}_2}{\varepsilon+m}f_2\end{pmatrix},$$

$$f_{1,2} = \left[\left(R_1A + i\frac{mZ\alpha}{p}R_2B\right)M_1 \mp iR_2BM_2\right]\phi_\lambda,$$

$$A = l\frac{d}{dy}(P_l(y) + P_{l-1}(y)),$$

$$B = \frac{d}{dy}(P_l(y) - P_{l-1}(y)), \quad (2)$$

$$y = \mathbf{n}_1\cdot\mathbf{n}_2, \quad R_{1,2} = 1 \mp (\boldsymbol{\sigma}\cdot\mathbf{n}_2)(\boldsymbol{\sigma}\cdot\mathbf{n}_1),$$

$$M_{1,2} = i\frac{\exp(ipr_2 - i\pi\nu)}{pr_2}\int_0^\infty t^{(\mp 1 - 2iq)}$$

$$\times \exp(it^2)J_{2\nu}(2t\sqrt{2pr_2})dt.$$

Here, $P_l(x)$ are Legendre polynomials, $J_{2\nu}$ are Bessel functions, and $\nu = \sqrt{l^2 - (Z\alpha)^2}$. The integrals in functions $M_{1,2}$ can be expressed in terms of degenerate hypergeometric functions. Result (2) is in accordance with the well-known solution to the Dirac equation in a Coulomb field.

To find the scattering amplitude, we must calculate coefficient W_λ of the diverging spherical wave $\exp[ipr_2 + iq\ln(2pr_2)]/r_2$ in the asymptotic form of function $\psi^{(+)}_{\lambda\mathbf{p}}(\mathbf{r}_2)$ ($r_2 \longrightarrow \infty$).

A nonzero contribution to W_λ comes from function M_1 (see Eqs. (2)), the required asymptotic form of this function being determined by the integration domain $t \ll 1$. We have

$$W_\lambda = \sqrt{\varepsilon + m}\sum_{l=1}^{\infty}\begin{pmatrix} f \\ \dfrac{p\boldsymbol{\sigma}\cdot\mathbf{n}_2}{\varepsilon + m}f \end{pmatrix}, \tag{3}$$

$$f = \frac{i\exp(-i\pi\nu)\Gamma(\nu - iq)}{2p\Gamma(\nu + 1 + iq)}\left[R_1 A + i\frac{mZ\alpha}{p}R_2 B\right]\phi_\lambda.$$

Calculating the flux of scattered particles, averaging it over spins, and dividing by the incident current density, we obtain the scattering cross section:

$$\frac{d\sigma}{d\Omega} = \frac{1}{4p}\sum_{\lambda=1,2} W^+_\lambda(\boldsymbol{\alpha}\cdot\mathbf{n}_2)W_\lambda$$

$$= \frac{2}{p^2}\left\{(1+x)|F'|^2 + \left(\frac{mZ\alpha}{p}\right)^2\frac{|F|^2}{1-x}\right\},$$

$$F = -\frac{i}{2}\sum_{l=1}^{\infty} l\exp[i\pi(l-\nu)]\frac{\Gamma(\nu - iq)}{\Gamma(\nu + 1 + iq)} \tag{4}$$

$$\times [P_l(x) - P_{l-1}(x)],$$

$$F' = \frac{dF}{dx}, \quad x = \cos\vartheta = -\mathbf{n}_1\cdot\mathbf{n}_2.$$

This result coincides with that obtained earlier in [1].

3. SCATTERING CROSS SECTION IN THE LOW-ENERGY LIMIT

Let us calculate the scattering cross section (4) for $q = Z\alpha\varepsilon/p \gg 1$ and $Z\alpha \sim 1$. This range of parameters corresponds to scattering of a low-energy electron in a strong Coulomb field. We consider the ratio of the exact cross section (in parameter $Z\alpha$) to the cross section

$$\frac{d\sigma_B}{d\Omega} = \frac{q^2}{p^2(1-x)^2}\left[1 - \frac{p^2}{2\varepsilon^2}(1-x)\right] \tag{5}$$

obtained in the first Born approximation. Using the asymptotic form of the Γ function for large values of the argument, the expression for $S = d\sigma/d\sigma_B$ can be reduced to the form

$$S = 1 + \frac{1-x}{q}\mathrm{Im}\left\{\exp\left(-iq\ln\frac{1-x}{2}\right)\right.$$

$$\times\sum_{l=1}^{\infty} l(-1)^l[P_l(x) - P_{l-1}(x)][\exp(-2i\pi\nu) - 1] \tag{6}$$

$$\left.\times\exp\left(\frac{il^2}{q}\right)\right\}.$$

For $1 + x \gg 1/q$, factor $\exp(il^2/q)$ in this formula can be replaced by unity. This gives

$$S = 1 + \frac{1-x}{q}\mathrm{Re}\left\{\exp\left(-iq\ln\frac{1-x}{2}\right)\right.$$

$$\times\left[\pi(Z\alpha)^2\left(\sqrt{\frac{2}{1+x}} - 1\right) - i\pi^2(Z\alpha)^4\ln\sqrt{\frac{1+x}{2}}\right.$$

$$-i\sum_{l=1}^{\infty} l(-1)^l[P_l(x) - P_{l-1}(x)] \tag{7}$$

$$\left.\left.\times\left(\exp(-2i\pi\nu) - 1 - \frac{i\pi(Z\alpha)^2}{l} + \frac{\pi^2(Z\alpha)^4}{2l^2}\right)\right]\right\}.$$

Thus, for $1 + x \gg 1/q$, the correction to function S is proportional to $1/q$. It should be noted that the sum over l in expression (7) converges very rapidly for an arbitrary x. If $1 + x \sim 1/q$ (backward scattering), the main contribution to sum (6) comes from moments $l \sim \sqrt{q} \gg 1$. Using the asymptotic form of the Legendre polynomials for $x \longrightarrow -1$ and replacing summation by integration, we obtain

$$S = 1 + (1-x)\frac{\pi^{3/2}(Z\alpha)^2}{\sqrt{q}}\cos\left(\frac{\pi}{4} + \frac{q(1+x)}{4}\right)$$

$$\times J_0\left(\frac{q(1+x)}{4}\right). \tag{8}$$

It can be seen that the correction to function S for $1 + x \sim 1/q$ is proportional to $1/\sqrt{q} \propto \sqrt{p/\varepsilon}$. Consequently, the value of S tends very slowly to unity as $v \longrightarrow 0$. The exact (in $Z\alpha$) backward scattering cross section signifi-

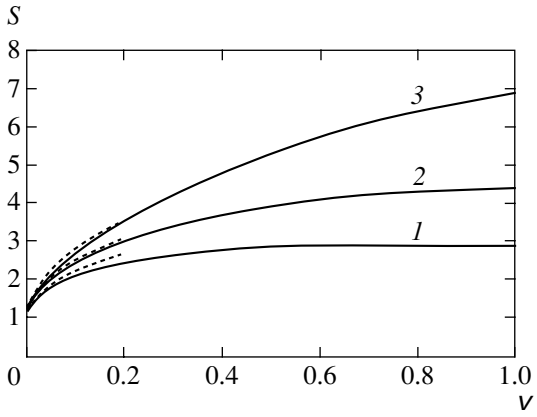


Fig. 1. Dependence of function S on v for $x = -1$ and $Z\alpha = 0.6$ (*1*), 0.7 (*2*), and 0.8 (*3*). Solid curves correspond to exact results and dotted curves, to the asymptotic form.

cantly differs from $d\sigma_B/d\Omega$ even for comparatively low energies.

For $x = -1$ and for arbitrary values of q, the exact expression for function S (see relations (4)) has a simple form:

$$S = 4\left|\sum_{l=1}^{\infty} l\exp(-i\pi\nu)\frac{\Gamma(\nu - iq)}{\Gamma(\nu + 1 + iq)}\right|^2. \qquad (9)$$

Figure 1 shows the dependence of function S on $v = p/\varepsilon$ for $x = -1$ and for several values of Z. The same figure shows the low-energy asymptotic form (8) for $x = -1$:

$$S = 1 + \sqrt{2}\pi^{3/2}(Z\alpha)^{3/2}\sqrt{v}. \qquad (10)$$

It can be seen that, for backward scattering, the exact (in $Z\alpha$) cross section significantly differs from the Born cross section even for small velocities v, the difference decreasing very slowly (in proportion to $\propto\sqrt{v}$). It can be seen from Fig. 1 that the exact result virtually coincides with the asymptotic form for $v < 0.2$.

It is well known that the Born cross section is transformed to the Rutherford cross section for $v \ll 1$:

$$\frac{d\sigma_R}{d\Omega} = \left(\frac{Z\alpha}{mv^2(1-x)}\right)^2. \qquad (11)$$

It is interesting to analyze the ratio $S_1 = d\sigma/d\sigma_R$ of the exact scattering cross section to the Rutherford cross section. Figure 2 shows the dependence of S_1 on $q = Z\alpha/v$ for $x = -1$ and for several values of parameter $Z\alpha$. It can be seen that the ratio S_1 of the cross sections increases with decreasing energy, attains its maximal value, and then slowly tends to unity ($S_1 - 1 \propto 1/\sqrt{q} \propto \sqrt{v}$).

In classical electrodynamics, for scattering with impact parameters ρ satisfying the relation $Z\alpha/mv\rho \gtrsim 1$, a particle attains velocities of $v_{\max} \sim 1$ at minimal distance from a Coulomb center and relativistic effects

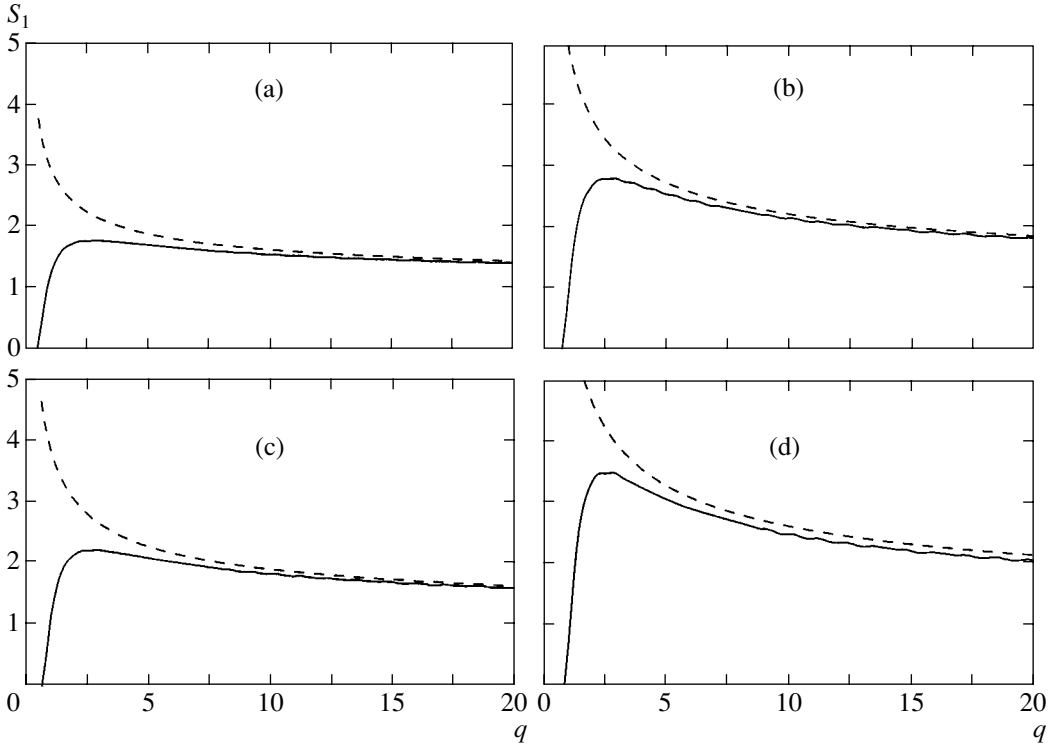


Fig. 2. Dependence of function S_1 on variable $q = Z\alpha/v$ for $Z\alpha = 0.5$ (a), 0.6 (b), 0.7 (c), and 0.8 (d). Solid curves correspond to the exact result and dashed curves describe the asymptotic form.

become significant. In relativistic classical mechanics, for $\rho < (Z\alpha)/mv$ (we consider the case with $v \ll 1$ as before), the phenomenon of falling to the center takes place. In addition, the cross section of scattering through angles close to π is singular. For $1 + x \ll v^{4/3} \ll 1$, the formulas given in [7] readily give

$$\frac{d\sigma_{cl}}{d\Omega} = \left(\frac{Z\alpha}{mv}\right)^2\left(\frac{\pi}{2v^4}\right)^{1/3}\frac{1}{6\sqrt{2(1+x)}} \qquad (12)$$

in this region. However, in the quantum-mechanical case, for $Z\alpha \sim 1$ and $v \longrightarrow 0$, the backward scattering cross section (small values of ρ) tends to the nonrelativistic limit. This is due to the fact that, for a given ρ, indeterminacy $\Delta\phi \sim 1/mv\rho$ in the scattering angle appears, which becomes of the order of unity for $\rho \lesssim Z\alpha/mv$.

ACKNOWLEDGMENTS

The authors are grateful to R.N. Li and V.M. Strakhovenko for fruitful discussions and for their interest in this research.

This study was financed by the Russian Foundation for Basic Research (project no. 03-02-16510).

REFERENCES

1. N. F. Mott, Proc. R. Soc. London, Ser. A **124**, 426 (1929).
2. H. Überall, *Electron Scattering from Complex Nuclei* (Academic, New York, 1971).
3. J. H. Bartlett and R. E. Watson, Proc. Am. Acad. Arts Sci. **74**, 53 (1940).
4. J. A. Doggett and L. V. Spencer, Phys. Rev. **103**, 1597 (1956).
5. N. Sherman, Phys. Rev. **103**, 1601 (1956).
6. A. I. Milstein and V. M. Strakhovenko, Phys. Lett. A **90A**, 447 (1982).
7. L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, 6th ed. (Nauka, Moscow, 1973; Pergamon Press, Oxford, 1975).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 691–704.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 789–804.

ATOMS, SPECTRA, RADIATION

On the Theory of Optical Properties of Fractal Clusters

S. V. Perminov[a], S. G. Rautian[b], and V. P. Safonov[c]

[a]*Institute of Semiconductor Physics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 630090 Russia*

[b]*Lebedev Physical Institute, Russian Academy of Sciences, Moscow, 119991 Russia*

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

e-mail: rautian@direct.ru; safonov@iae.nsk.su

Received September 24, 2003

Abstract—A theory of optical properties of clusters of spherical metal nanoparticles characterized by an arbitrary size distribution is developed in a quasi-static dipole approximation. The equations for coupled dipoles and general relations are formulated in terms of reduced dipole moments. It is shown that the dipole resonant frequencies and amplitudes, the absorbed power, and the acting-field magnitudes strongly depend on the ratios of particle radii in a cluster. Properties of linear, planar, and three-dimensional systems are examined. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Optical properties of disordered fractal clusters of spherical metal nanoparticles have been studied in a vast literature, including a number of reviews and monographs (e.g., see [1–6]). A variety of new linear and nonlinear effects have been discovered: inhomogeneous broadening of the plasmon band, giant field fluctuations with correlation radii much smaller than the wavelengths, surface-enhanced Raman scattering, photomodification of clusters, fast-response highly nonlinear behavior, and nonlinear optical activity. Fine techniques have been developed for producing clusters with controlled values of basic parameters. It was proved that many natural systems with interesting properties have fractal structure [6, 7]. The key role played by the fractal structure of clusters with Hausdorff dimension substantially lower than three ($D \sim 1.5$–1.8) was demonstrated both experimentally and theoretically. In recent years, fractal aggregates of metal nanoparticles were found to be increasingly useful for various applications.

In the pioneering studies [8, 9], it was already revealed that optical properties of fractal clusters are mainly determined by electrodynamic interactions between particles, which is manifested in both linear and nonlinear effects. In particular, the interaction is directly responsible for the strong plasmon-band splitting characteristic of an isolated particle and for a wide, inhomogeneously broadened tails in the absorption spectra of clusters. Numerical simulations showed that the splitting interval is independent of the number N of particles (when N is sufficiently large) and of the optical properties of the particle material (when measured in certain units) [10, 11]. This fundamental fact implies that the dominant role is played by the nearest neighbors of a particle, whereas the contribution of a more distant environment to local field characteristics is not important. The interaction indicated above determines the structure of field fluctuations and explains the essential difference of the optical properties of media consisting of fractal clusters from those of "normal" (weakly inhomogeneous) media.

The observations enumerated above suggest that a collection of fractal clusters should be considered as an unconventional medium that has unusual optical properties and is different from a gas, plasma, liquid, or solid.

The existing theories rely on the assumption of equal particle size (an interaction parameter of primary importance). The actual polydispersity is taken into account by averaging the results over some size distribution (e.g., see [6]). This approach is not self-consistent. Indeed, if particles of difference size exist, then nearest neighbors may have different diameters. However, electrodynamic interaction between particles of equal size is characterized by a certain symmetry, which implies a corresponding degeneracy in the vibrational spectrum and certain selection rules. For example, in a system of two interacting particles (dimer), dipole oscillations of certain types cannot be excited by an external field because of its simple structure. In this respect, a dimer is analogous to a molecule consisting of two identical atoms, for which infrared absorption is prohibited. The analysis presented below shows that particle-size variability has fundamental consequences for spectroscopy of fractal clusters, nonlinear effects, properties of field fluctuations, etc. Thus, the existing theories of optical properties of clusters are essentially

inadequate. Therefore, both foundations of the theory and its applications to real systems must be revised. This problem is addressed in the present study.

2. GENERAL RELATIONS

Consider a system of N spherical particles (monomers) in a medium with a dielectric constant ε_h. The particle material is characterized by a dielectric constant ε. Suppose that each particle radius a_i ($i = 1, 2, \ldots, N$) is sufficiently small as compared to the wavelength, so that the dipole approximation is valid. The dipole–dipole interactions between particles and their interaction with a monochromatic external field $\mathbf{E}_0(\mathbf{r})$ of frequency ω are described by a system of equations for their dipole moments $\mathbf{d}_i$ [10, 11]:

$$\mathbf{d}_i = \alpha_i \varepsilon_h \mathbf{E}_i, \quad \alpha_i = a_i^3 \frac{\varepsilon - \varepsilon_h}{\varepsilon + 2\varepsilon_h}, \tag{2.1}$$

$$\mathbf{E}_i = \mathbf{E}_0(\mathbf{r}_i) + \sum_{j \neq i} \frac{\mathbf{E}_{ij}}{\varepsilon_h},$$
$$\mathbf{E}_{ij} = \frac{3\mathbf{n}_{ij}(\mathbf{n}_{ij} \cdot \mathbf{d}_j)\varphi_{ij} - \mathbf{d}_j \psi_{ij}}{r_{ij}^3}, \tag{2.2}$$

$$\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j, \quad r_{ij} = |\mathbf{r}_{ij}|, \quad \mathbf{n}_{ij} = \mathbf{r}_{ij}/r_{ij}, \tag{2.3}$$

$$\varphi_{ij} = [1 - ikr_{ij} - (kr_{ij})^2/3]\exp(ikr_{ij}),$$
$$\psi_{ij} = [1 - ikr_{ij} - (kr_{ij})^2]\exp(ikr_{ij}). \tag{2.4}$$

The field $\mathbf{E}_i$ acting on the ith dipole is the superposition of $\mathbf{E}_0(\mathbf{r}_i)$ and the fields $\mathbf{E}_{ij}/\varepsilon_h$ generated by all dipoles (indexed by j) at the point $\mathbf{r}_i$ where the ith dipole is located.[1] The retardation effects due to the nonzero distance between the ith and jth particles are represented by φ_{ij} and ψ_{ij}. The general analysis of system (2.1), (2.2) presented in [10, 11] was entirely based on the equality of all particle radii. Since this assumption was inherent in the theory developed in [10, 11], the particle diameter was used as the unit of length. System (2.1), (2.2) is free of this restriction, but it can be transformed into the one that was analyzed in [10, 11] by changing the variables. Define

$$\mathbf{d}_i^r = \frac{\mathbf{d}_i}{a_i^{3/2}}, \quad \xi_{ij} = \frac{(a_i a_j)^{3/2}}{r_{ij}^3},$$
$$\kappa = \frac{a_i^3}{\alpha_i} = \frac{\varepsilon + 2\varepsilon_h}{\varepsilon - \varepsilon_h}. \tag{2.5}$$

[1] The near field of a dipole is inversely proportional to ε_h, and this dependence is factored out in (2.2). The polarizability of a ball is $\alpha_i \varepsilon_h$ (see (2.1)), and the factors ε_h cancel out.

Equations (2.1) and (2.2) are rewritten in terms of $\mathbf{d}_i^r$ as

$$\kappa \mathbf{d}_i^r + \sum_{j \neq i} \xi_{ij}[\mathbf{d}_j^r \psi_{ij} - 3\mathbf{n}_{ij}(\mathbf{n}_{ij} \cdot \mathbf{d}_j^r)\varphi_{ij}] = a_i^{3/2} \varepsilon_h \mathbf{E}_0(\mathbf{r}_i). \tag{2.6}$$

The use of the reduced dipole moment $\mathbf{d}_i^r$ facilitates further analysis. Since its square has the dimension of energy, both energy and amplitude relations are most conveniently formulated in terms of $\mathbf{d}_i^r$. Define column vectors d^r and E^r with the components

$$d_{i\alpha}^r = d_{i\alpha}/a_i^{3/2}, \quad E_{i\alpha}^r = a_i^{3/2} \varepsilon_h E_{0\alpha}(\mathbf{r}_i). \tag{2.7}$$

Following [10], rewrite Eqs. (2.6) in operator form:

$$(\kappa + U)d^r = E^r,$$
$$U_{i\alpha, j\beta} = \xi_{ij}[\delta_{\alpha\beta}\psi_{ij} - 3n_{ij\alpha}n_{ij\beta}\varphi_{ij}], \quad i \neq j. \tag{2.8}$$

The matrix elements $U_{i\alpha, j\beta}$ of the interaction operator U are symmetric under the permutation $i\alpha \longleftrightarrow j\beta$ of the particle and coordinate indices, ij and $\alpha\beta$, and are complex in the general case. As in [10, 11], the operator U can be diagonalized, and the eigenvalues u_m and eigenstates defined by the relation

$$U|m\rangle = u_m|m\rangle$$

can be used to represent the solution to Eqs. (2.6) as

$$d_{i\alpha}^r = \sum \langle i\alpha|m\rangle(\kappa + u_m)^{-1}\langle m|j\beta\rangle E_{j\beta}^r. \tag{2.9}$$

Formally, both Eq. (2.8) and its solution (2.9) are similar to those describing systems of identical particles. Moreover, the proposed model (2.7)–(2.9) has the additional advantage that U, ξ_{ij}, κ, and u_m are dimensionless quantities, whereas their counterparts in [10, 11] are measured in cm^{-3}. Note that Eqs. (2.6) and (2.8) can be separated with respect to parametric dependence (when $kr_{ij} = 0$): the parameter κ depends only on the dielectric properties of particles and medium, while the operator U and its eigenvalues are determined by the system's geometry. Note also that $E_{j\beta}^r \propto a_j^{3/2}$ in (2.9), whereas the column vector on the right-hand side of the equation analogous to (2.8) in the standard theory [10, 11] depends on j only via $\mathbf{r}_j$ in $E_{0\alpha}(\mathbf{r}_j)$. A detailed analysis is easier to perform in terms of $\kappa_m = -u_m$.

It is obvious that the variability of a_i is essential in various respects. Since $E_{j\beta}^r \propto a_j^{3/2}$, the amplitude of the resonance $(\kappa - \kappa_m)^{-1}$ depends on the relations between particle radii. In particular, certain amplitudes that van-

ish when $a_i \equiv a$ [5] are finite in the general case. Furthermore, the coupling parameters

$$\xi_{ij} = (a_i a_j)^{3/2}/r_{ij}^3$$

depend on the ratios of particle radii via both $a_i a_j$ and r_{ij} ($r_{ij} \geq a_i + a_j$). Both eigenvalues u_m and eigenfunctions of U are modified accordingly. In particular, degenerate eigenstates split since the symmetry associated with equality of particle diameters is broken when the ratios a_i/a_j are arbitrary. A detailed analysis of the effects due to particle-size variability is presented in Sections 3 to 5.

In the quasi-static approximation adopted in what follows ($kr_{ij} \longrightarrow 0$), $\varphi_{ij} = \psi_{ij} = 1$. In this case of highest practical importance, both operator U and eigenvalues u_m are real.

It is assumed above that all particles consist of the same material and are characterized by equal permittivities ε. The theory can be naturally extended to particles of different size (a_i) and material (ε_i). This may be important for analyzing the effects of particle size and temperature T on ε_i. It is reasonable to define the following quantities and introduce a relation between them:

$$\mathbf{d}_i' = \frac{\mathbf{d}_i}{\alpha_i^{1/2}}, \quad \alpha_i = a_i^3 \frac{\varepsilon_i - \varepsilon_h}{\varepsilon_i + 2\varepsilon_h},$$

$$(1 + V)d' = \mathscr{E},$$

$$V_{i\alpha, j\beta} = \frac{(\alpha_i \alpha_j)^{1/2}}{r_{ij}^3}[\delta_{\alpha\beta}\psi_{ij} - 3n_{ij\alpha}n_{ij\beta}\varphi_{ij}], \tag{2.10}$$

$$d_{i\alpha}' = \frac{d_{i\alpha}}{\alpha_i^{1/2}}, \quad \mathscr{E}_{i\alpha} = \alpha_i^{1/2}\varepsilon_h E_{0\alpha}(\mathbf{r}_i).$$

The operator V is also invariant under the permutation $i\alpha \longleftrightarrow j\beta$, and all general conclusions made above apply to Eq. (2.10).

In the model adopted here, the work A done on the dipoles $\mathbf{d}_i$ by the external field is consumed to produce scattered radiation and to heat the particles [12, 13]:

$$A = -(\omega/2)\mathrm{Re}\left[i\sum_i \mathbf{d}_i \cdot \mathbf{E}_0^*(\mathbf{r}_i)\right] = Q_s + Q_a. \tag{2.11}$$

An estimate obtained below (see Eq. (2.16)) shows that the scattered power Q_s is much lower than the absorbed power Q_a for systems of interest here, and Q_s is neglected in the analysis that follows.[2]

[2] Following [14], we can take into account the effect of scattering on relaxation characteristics. However, we focus here on the fundamental role played by the variability of particle size and do not expand the model to allow for corrections of this kind.

The total work A done on the dipoles $\mathbf{d}_i$ by the external field $\mathbf{E}_0$ is the sum of contributions due to the acting fields $\mathbf{E}_i$. Indeed, the sum over i and j on the right-hand side in the relation

$$\mathrm{Re}\left[i\sum_i \mathbf{d}_i \cdot \mathbf{E}_0^*(\mathbf{r}_i)\right] = \mathrm{Re}\left[i\sum_i \mathbf{d}_i \cdot \mathbf{E}_i^*\right]$$

$$+ \mathrm{Re}\left\{i\sum_i \mathbf{d}_i \cdot [\mathbf{E}_0^*(\mathbf{r}_i) - \mathbf{E}_i^*]\right\} = \mathrm{Re}\left[i\sum_i \mathbf{d}_i \cdot \mathbf{E}_i^*\right]$$

$$+ \mathrm{Re}\left\{i\varepsilon_h^{-1}\sum_{i,\, j\neq i} [\mathbf{d}_i \cdot \mathbf{d}_j^* - 3(\mathbf{n}_{ij} \cdot \mathbf{d}_i)(\mathbf{n}_{ij} \cdot \mathbf{d}_j^*)]/r_{ij}^3\right\}$$

is obviously real, and the corresponding term vanishes. By virtue of (2.1), $\mathbf{E}_i^*$ can be replaced by $\mathbf{d}_i^*/\varepsilon_h\alpha_i^*$, with ε_h assumed to be real:

$$Q_a = -\frac{\omega}{2}\mathrm{Re}\left[i\sum_i \frac{|\mathbf{d}_i|^2}{\alpha_i^*\varepsilon_h}\right]$$

$$= -\frac{\omega}{2\varepsilon_h}\sum_i |\mathbf{d}_i|^2 \mathrm{Im}\left(\frac{1}{\alpha_i}\right). \tag{2.12}$$

Expression (2.12) can also be derived from a standard expression for the absorbed power per unit volume q [13]:

$$q = \omega\varepsilon''|\mathbf{E}|^2/8\pi, \quad \varepsilon'' = \mathrm{Im}\varepsilon.$$

To calculate the power Q_{ai} absorbed by the ith particle, this expression must be multiplied by its volume, with $\mathbf{E}$ interpreted as the field strength $\mathbf{E}_{i,\,\mathrm{in}}$ inside the particle related to the acting field $\mathbf{E}_i$ [13]:

$$\mathbf{E}_{i,\,\mathrm{in}} = \frac{\mathbf{E}_i 3\varepsilon_h}{\varepsilon + 2\varepsilon_h} = \frac{3\mathbf{d}_i}{\alpha_i(\varepsilon + 2\varepsilon_h)}.$$

As a result, we have

$$Q_{ai} = \frac{\omega}{2}3\varepsilon''\left|\frac{\mathbf{d}_i}{\alpha_i(\varepsilon + 2\varepsilon_h)}\right|^2 a_i^3. \tag{2.13}$$

It is easy to show that expression (2.13) is equivalent to the ith summand in (2.12).

The inverse specific susceptibility κ represented as follows [10]:

$$\kappa = -X - i\delta, \tag{2.14}$$

where the sign of the real part X is consistent with the optical properties of metals ($\varepsilon' < 0$). Now, expression (2.12) has a particularly simple form:

$$Q_a = \sum_i Q_{ai}, \quad Q_{ai} = \frac{\omega\delta}{2\varepsilon_h}\frac{|\mathbf{d}_i|^2}{a_i^3} = \frac{\omega\delta}{2\varepsilon_h}|\mathbf{d}_i^r|^2. \tag{2.15}$$

Expressions (2.12) and (2.15) extend the results of [10, 11] to the case of an arbitrary size distribution.

Formula (2.15) demonstrates that Q_{ai} is determined by the squared absolute value of the reduced dipole moment. Expressions (2.12) and (2.15) determine the power absorbed by each particular particle, i.e., the distribution of power over particles parameterized by their size, field frequency, properties of surrounding particles, and other characteristics. It is obvious that the quantity $\{-(\omega/2)\mathrm{Re}[i\mathbf{d}_i \cdot \mathbf{E}_0^*(\mathbf{r}_i)]\}$ in (2.11) cannot be interpreted as the absorption by the ith particle. In particular, the contributions of some individual summands in (2.11) to the work A done by the field are negative.

It is well known that the power scattered by a particle characterized by a dipole moment $\mathbf{d}_i$ is $Q_{si} = \omega^4|\mathbf{d}_i|^2/3c^3$. For $a_i \approx 10$ nm, $2\pi c/\omega \approx 10^3$ nm, and $\delta \approx 0.1$–0.01 characteristic of metal nanoparticles,

$$\frac{Q_{si}}{Q_{ai}} = \frac{2\varepsilon_h}{3\delta}\left(\frac{\omega a_i}{c}\right)^3 \ll 1. \tag{2.16}$$

Therefore, scattering is negligible as compared to absorption. This conclusion holds even if interference of the fields scattered by different monomers is taken into account.

It is well known that fluctuations and nonlinear phenomena depend on the acting field strength [3]. This quantity of special importance for fractal clusters can be calculated by using solutions $\mathbf{d}_i^r$ to Eqs. (2.6):

$$\mathbf{E}_i = \frac{\mathbf{d}_i}{\alpha_i\varepsilon_h} = \frac{\kappa\mathbf{d}_i^r}{a_i^{3/2}\varepsilon_h}. \tag{2.17}$$

Its absolute value squared is proportional to the absorbed power per unit volume:

$$|\mathbf{E}_i|^2 = \frac{(Q_{ai}/a_i^3)|\kappa|^2}{2\varepsilon_h\omega\delta}. \tag{2.18}$$

According to (2.17) and (2.18), the value of $\mathbf{E}_i$ strongly depends on the monomer radius.

The energy of a dipole placed in an external field is

$$\begin{aligned} U_i &= -\frac{1}{2}\mathrm{Re}(\mathbf{d}_i \cdot \mathbf{E}_0^*) = -\frac{1}{2}\mathrm{Re}\left(\frac{|\mathbf{d}_i|^2}{\alpha^*\varepsilon_h}\right) \\ &= -\frac{1}{2}\frac{|\mathbf{d}_i^r|^2}{\varepsilon_h}\mathrm{Re}\kappa. \end{aligned} \tag{2.19}$$

The ponderomotive force exerted by the external field on the ith dipole and other dipoles is

$$\mathbf{F}_i = -\nabla U_i. \tag{2.20}$$

According to [15], the motion of a particle driven by the force given by (2.20) changes its kinematic properties and leads to certain nonlinear effects. An adequate description of these effects must take into account the variability of particle size, which is suitably characterized in terms of $\mathbf{d}_i^r$.

Analyses of fractal clusters of metal nanoparticles commonly rely on the dielectric constant given by Drude's formula

$$\varepsilon = \varepsilon_0 - \omega_p^2/\omega(\omega + 2i\Gamma), \tag{2.21}$$

where ω_p is the plasma frequency, Γ is the damping constant (generally, a function of particle size), and $\varepsilon_0 - 1$ is the contribution of interband transitions. Setting $\varepsilon_h \approx \varepsilon_0$ for simplicity, we obtain

$$X = \omega^2/\bar{\omega}_p^2 - 1, \quad \delta = 2\omega\Gamma/\bar{\omega}_p^2, \quad \bar{\omega}_p^2 = \omega_p^2/3\varepsilon_h. \tag{2.22}$$

For example, silver is characterized by $\bar{\omega}_p = 2.5 \times 10^4$ cm^{-1} and $\Gamma \approx 500$ cm^{-1}. According to (2.22), the equation $X = -\kappa_m$ yields resonant values of ω^2, which are equivalent to

$$\omega/\bar{\omega}_p = (1 - \kappa_m)^{1/2} \tag{2.23}$$

on the ω scale. When κ_m is positive or negative, the corresponding resonance is shifted toward the low- or high-frequency end of the spectrum, respectively. Substituting the expression for Q_{ai} in (2.15) into (2.22), we obtain

$$Q_{ai} = \left(\frac{\omega}{\bar{\omega}_p}\right)^2 \varepsilon_h^{-1}\Gamma|\mathbf{d}_i^r|^2. \tag{2.24}$$

Since it is obvious that

$$|\mathbf{d}_i^r|^2 \propto \delta^{-2} \propto \omega^{-2}$$

at resonant frequencies, the factor ω^2 in (2.24) is canceled out in resonant values of Q_{ai}.

The model of irregular fractal clusters of identical particles involves the random parameters r_{ij} (spacing between particles), $\mathbf{n}_{ij}$ (relative position of particles), and a (particle radius). In the generalized model constructed here, the ratio a_i/a_j is an additional random parameter. An exact analysis must rely on an N-dimensional size distribution $P(a_1, a_2, \ldots, a_N)$. Spectral properties of clusters depend primarily on the interaction between monomers that are relatively close to each other. Therefore, the minimal required dimension of $P(a_1, a_2, \ldots)$ may be relatively low. This means that a much greater number of random realizations must be computed in numerical experiments of the kind reported in [10, 11].

After this brief discussion of statistical aspects, we focus below on dynamics of clusters. Some dynamical aspects are directly related to experiment, as an analysis of photomodification of clusters based on electron dif-

fraction patterns visualizing individual monomers and aggregates.

3. DIMER

Consider the case of two particles (a dimer) as an illustration of the principal optical properties of clusters due to electrodynamic interaction between particles [8, 16]. We suppose that $a_2 \le a_1$ and drop the subscripts ij in $\mathbf{n}_{ij}$ and $\mathbf{r}_{ij}$ as unnecessary. For the projections of $d_{\mathrm{in}}^{\mathrm{r}}$ and $d_{i\perp}^{\mathrm{r}}$ on $\mathbf{n}$ and the plane perpendicular to it, respectively, Eqs. (2.6) reduce to

$$\begin{aligned}(\kappa - 2\xi)(d_{1n}^{\mathrm{r}} + d_{2n}^{\mathrm{r}}) &= (a_1^{3/2} + a_2^{3/2})\varepsilon_{\mathrm{h}} E_{0n},\\ (\kappa + 2\xi)(d_{1n}^{\mathrm{r}} - d_{2n}^{\mathrm{r}}) &= (a_1^{3/2} - a_2^{3/2})\varepsilon_{\mathrm{h}} E_{0n},\end{aligned} \tag{3.1}$$

$$\begin{aligned}(\kappa + \xi)(d_{1\perp}^{\mathrm{r}} + d_{2\perp}^{\mathrm{r}}) &= (a_1^{3/2} + a_2^{3/2})\varepsilon_{\mathrm{h}} E_{0\perp},\\ (\kappa - \xi)(d_{1\perp}^{\mathrm{r}} - d_{2\perp}^{\mathrm{r}}) &= (a_1^{3/2} - a_2^{3/2})\varepsilon_{\mathrm{h}} E_{0\perp},\end{aligned} \tag{3.2}$$

$$\begin{aligned}\xi &= (a/r)^3, \quad a = (a_1 a_2)^{1/2},\\ E_{0n} &= \mathbf{n}\cdot\mathbf{E}_0 = E_0\cos\theta, \quad E_{0\perp} = E_0\sin\theta.\end{aligned} \tag{3.3}$$

For the n-projections, $\kappa_n = 2\xi$ and $\kappa_n = -2\xi$ correspond to the sum- and difference-frequency oscillations (the latter mode was called antisymmetric in [14]). For the $\perp$-projections, the sum- and difference-frequency modes are characterized by $\kappa_\perp = -\xi$ and $\kappa_\perp = \xi$, respectively.The oscillations of $d_{i\perp}^{\mathrm{r}}$ are twice degenerate by virtue of axial symmetry. It should be noted that the normal modes described by (3.1) and (3.2) are expressed in terms of reduced dipole moments $d_{i\alpha}^{\mathrm{r}}$. This observation applies to systems consisting of a greater number of monomers as well.

The solutions to Eqs. (3.1) and (3.2) can be written as

$$\begin{aligned}d_{in}^{\mathrm{r}} &= -\varepsilon_{\mathrm{h}} E_{0n}\left(\frac{\bar{a}^{3/2}}{X + 2\xi + i\delta} + \frac{a_i^{3/2} - \bar{a}^{3/2}}{X - 2\xi + i\delta}\right),\\ \bar{a}^{3/2} &= \frac{a_1^{3/2} + a_2^{3/2}}{2},\end{aligned} \tag{3.4}$$

$$d_{i\perp}^{\mathrm{r}} = -\varepsilon_{\mathrm{h}} E_{0\perp}\left(\frac{\bar{a}^{3/2}}{X - \xi + i\delta} + \frac{a_i^{3/2} - \bar{a}^{3/2}}{X + \xi + i\delta}\right), \tag{3.5}$$

$$\frac{a_i^{3/2} - \bar{a}^{3/2}}{\bar{a}^{3/2}} = \pm B, \quad B = \frac{1-t}{1+t}, \quad t = \left(\frac{a_2}{a_1}\right)^{3/2}. \tag{3.6}$$

In (3.6), plus and minus correspond to $i = 1$ and $i = 2$, respectively. The components of the total induced dipole moment $\mathbf{d} = \mathbf{d}_1 + \mathbf{d}_2$ of a dimer are

$$\begin{aligned}d_n &= -\varepsilon_{\mathrm{h}} E_{0n}(a_1^3 + a_2^3) f_n(X),\\ f_n(X) &= \frac{1-C}{X + 2\xi + i\delta} + \frac{C}{X - 2\xi + i\delta},\end{aligned} \tag{3.7}$$

$$\begin{aligned}d_\perp &= -\varepsilon_{\mathrm{h}} E_{0\perp}(a_1^3 + a_2^3) f_\perp(X),\\ f_\perp(X) &= \frac{1-C}{X - \xi + i\delta} + \frac{C}{X + \xi + i\delta},\end{aligned} \tag{3.8}$$

$$C = \frac{1}{2}\left[1 - \frac{2(a_1 a_2)^{3/2}}{a_1^3 + a_2^3}\right], \quad \frac{C}{1-C} = B^2. \tag{3.9}$$

The resonance shifts $\pm 2\xi$ and $\pm\xi$ in (3.7) and (3.8) depend on the distance r between the particles, and absorption in different spectral ranges is due to dimers with different values of r. This theoretical conclusion is generally used as a basis for interpreting inhomogeneous broadening of an absorption band. This interpretation holds for $a_1 \ne a_2$, but is more complicated and entails additional properties. In particular, formulas (3.6) and (3.9) yield $B = 0$ and $C = 0$ for $a_1 = a_2$; i.e., two resonant amplitudes vanish. The remaining two resonances correspond to oscillations that are parallel and perpendicular to $\mathbf{n}$ and are characterized by $X_n = -2\xi$ and $X_\perp = \xi$, respectively. If $a_1 \ne a_2$, then $C \ne 0$, oscillation amplitudes change, and the band profile is different. The difference-frequency resonant amplitudes vanish for $a_1 = a_2$ in the approximation of $kr = 0$ adopted here. However, the difference-frequency resonant amplitude does not vanish because of retardation effects [14] even if $a_1 = a_2$. In this case, the amplitude is proportional to $1 - \cos(kr)$. For $\lambda \approx 10^3$ nm and $r \approx 2a_{1,2} \approx 20$ nm, we have $kr \approx 10^{-1}$ and $1 - \cos(kr) \approx 10^{-2}$. An amplitude of similar order of magnitude is obtained when the particle radii differ by about 10%.[3]

The dependence of B^2 on a_2/a_1 is illustrated by curve *1* in Fig. 1. When a_2/a_1 is small, the coefficient B^2 rapidly decreases to $B^2 \approx 0.5$ at $a_2/a_1 = 0.3$. The ratio of resonant amplitudes (3.4) and (3.5) in dipole oscillations is B, and its dependence on the difference in radii is stronger. In (3.7) and (3.8), we expose the factors $(a_1^3 + a_2^3)$ proportional to the total volume of the interacting particles. Thus, the functions $f_n(X)$ and $f_\perp(X)$ are normalized to the unit volume of the particle material. The sums of resonant amplitudes in $f_n(X)$ and $f_\perp(X)$ are constant. Therefore, the appearance of doublet components proportional to B^2 (when $a_1 \ne a_2$) is related to a decrease in absorption in stronger components.

[3] In [5, p. 158], it was claimed that nonvanishing different-frequency resonance amplitudes should be explained solely by retardation effects. This assertion obviously disagrees with reality.

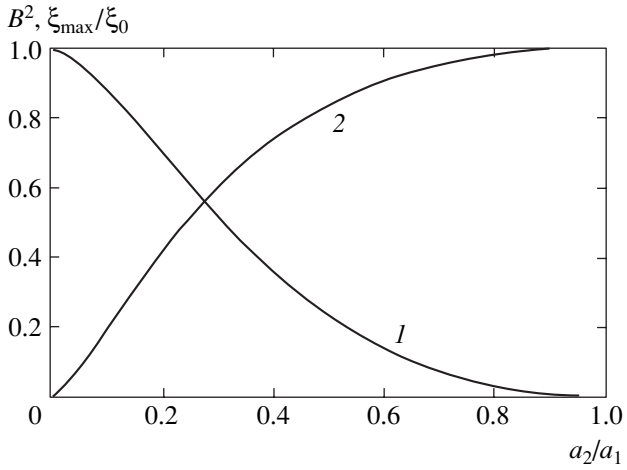


Fig. 1. Resonant-amplitude ratio (curve *1*) and largest shift (curve *2*) versus ratio of particle radii in a dimer.

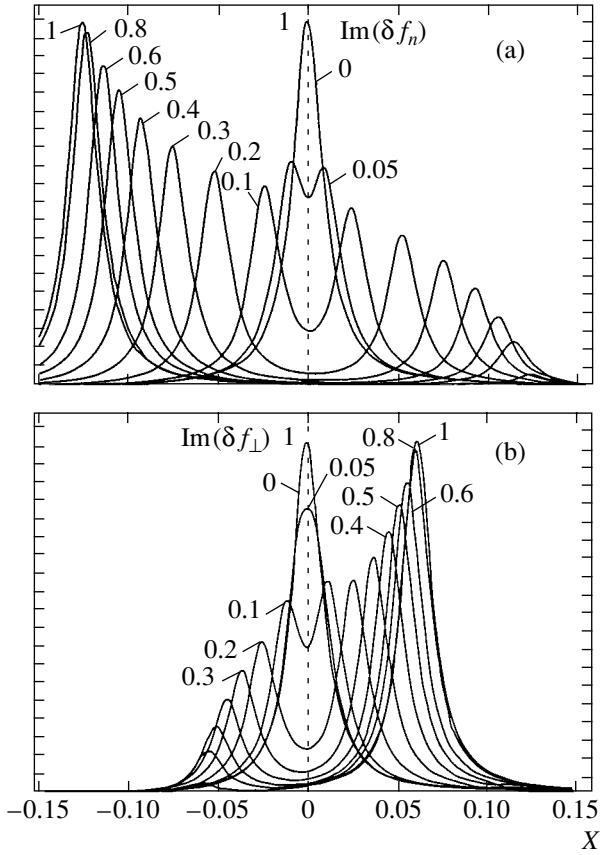


Fig. 2. Normalized absorption coefficient for a dimer oriented (a) parallel and (b) perpendicular to the electric field vector for the values of a_2/a_1 shown at the curves, $\xi = \xi_{max}/2$, and $\delta = 0.01$.

At first glance, it may seem that a decrease in a_2 implies a smaller minimal distance between particles and a wider splitting interval. However, this factor is weaker than the concomitant decrease in volume and polarizability of the smaller particle. Indeed, the value ξ_{max} corresponding to the minimal distance $r_{min} = a_1 + a_2$,

$$\xi_{max} = \left[\frac{(a_1a_2)^{1/2}}{r_{min}}\right]^3 = \left[\frac{2(a_1a_2)^{1/2}}{a_1 + a_2}\right]^3 \xi_0, \qquad \xi_0 = 1/8, \tag{3.10}$$

reaches its maximum when $a_1 = a_2$. Curve *2* in Fig. 1 represents ξ_{max}/ξ_0 as a function of a_2/a_1. It demonstrates that the "new" resonances already have appreciable amplitudes (curve *1*) at $a_2/a_1 = 0.25$–0.50, while the splitting interval remains relatively wide (curve *2*).

Equations (2.1), (2.2), and (2.6) are valid when the field generated by a dipole changes insignificantly over the length of a neighboring particle. When $r \approx a_1 + a_2$, this condition is violated. A detailed analysis shows that the change in the dipole's field is equivalent to an increase in $2a/r$ by a factor of $(6/\pi)^{1/3}$ [3, 17, 18]. Qualitative considerations suggest that a similar renormalization of $(a_1 + a_2)/r$ applies when $a_1 \neq a_2$ [3]. The renormalization changes the value of ξ_0 by a factor of almost 2:

$$\xi_0 = 1/8 \longrightarrow \xi_0 = 3/4\pi.$$

The imaginary parts of $f_n(X)$ and $f_\perp(X)$ determine the spectral profile of absorption on the X scale. Figures 2a and 2b present, respectively, the resonance profiles $\delta\mathrm{Im}f_n(X)$ and $\delta\mathrm{Im}f_\perp(X)$ calculated for several values of a_2/a_1 and a constant difference $(a_1 + a_2) - r$ between the particle surfaces. Figures 2a and 2b show that the n- and $\perp$-polarization resonances "prohibited" for $a_2 = a_1$ can contribute substantially to the high- and low-frequency tails of the band, respectively. Note also that the resonances shift as a_2/a_1 varies from 1 to 0.1 and the doublet components merge as $a_2/a_1 \longrightarrow 0$. The physical explanation of this fact is obvious: as $a_2/a_1 \longrightarrow 0$, a dimer becomes a monomer. In summary, there exist two mechanisms of inhomogeneous broadening of the absorption band: variability of the distance r and polydispersity of particles.

The integral absorption (over X) includes a partial contribution $(a_2/a_1)^3$ due to the smaller particles, whereas their contributions to the resonance shifts are proportional to $(a_2/a_1)^{3/2}$ for $a_2/a_1 \ll 1$, i.e., much greater. For example, when $a_2/a_1 = 1/5$, the contribution of the smaller particles to the integral absorption is less than one per cent, while $\xi_{max} \approx 0.5$, which amounts to 40% of the highest value of ξ_0 (1/8). For sufficiently small δ, shifted resonances give rise to appreciable inhomogeneous broadening. Therefore, relatively small particles can substantially modify the absorption profile without contributing to the integral absorption.

In the conventional model of identical spheres, a change in the central portion of an absorption band is explained by interaction between relatively distant par-

ticles, which implies cooperative effects [10]. This explanation is not complete: when a_2/a_1 is sufficiently small, the resonance shifts are smaller than their half-width,

$$\xi_{\max} \approx (a_2/a_1)^{3/2} < \delta. \tag{3.11}$$

In this situation, the central portion of the absorption band results from the combined contributions of distant particles of nearly equal size and dimers with particles having disparate sizes. For example, condition (3.11) is satisfied for $\delta = 0.1$ when $a_2/a_1 = 0.215$. These trends are demonstrated in graph form by the curves corresponding to $a_2/a_1 = 0.05$ and $\delta = 0.01$ in Figs. 2a and 2b.

Now, let us discuss the individual properties of the dipoles $\mathbf{d}_1$ and $\mathbf{d}_2$. For the projections d_{1n} and $d_{1\perp}$ of the larger particle's dipole, the amplitudes of both resonances have like signs ($a_1^{3/2} > \bar{a}^{3/2}$); for d_{2n} and $d_{2\perp}$, the signs are opposite. Moreover, it is easy to show that both $\mathrm{Im}\, d_{2n}$ and $\mathrm{Im}\, d_{2\perp}$ can be negative.

Consider the power $Q_{ai n}$ absorbed by individual particles. According to expressions (3.4) and (3.5), their respective Lorentzian profiles do not overlap and the resonant amplitudes corresponding to particles 1 and 2 are equal. Therefore,

$$Q_{a1n} = Q_{a2n}, \quad Q_{a1\perp} = Q_{a2\perp} \tag{3.12}$$

near the resonance points. Note also that the ratio of the resonant amplitudes in $Q_{ai n}$ is B^2, as in $\mathrm{Im} f_n(X)$ and $\mathrm{Im} f_\perp(X)$.

The quantity $Q_{ai n}$ is the total power absorbed by the ith particle. Approximate equality of Q_{a1n} and Q_{a2n} implies that the smaller particle absorbs a higher power per unit volume. Therefore, the smaller particle is heated by radiation to a higher degree. This can be manifested in the temperature dependence of dielectric constant and in photomodification of clusters [4, 19], for example, when particles melt or coalesce. Additional factors include the rate of particle cooling, the duration of the irradiating laser pulse, thermal properties of the particle material, etc. The problems related to photomodification of clusters require a special analysis.

Many optical properties of fractal clusters strongly depend on the local acting field $\mathbf{E}_i$. Since Eq. (2.18) entails

$$|\mathbf{E}_i| \propto (Q_{ai}/a_i^3)^{1/2},$$

the field acting on (smaller) particle 2 is stronger than that acting on particle 1 by a factor of $1/t = (a_1/a_2)^{3/2}$ by virtue of (3.12).

4. EIGENFREQUENCIES OF MANY-PARTICLE SYSTEMS

The eigenfrequency range of the interaction operator is independent of the number N of particles in a fractal cluster consisting of many spherical particles [10, 11]. This implies that the nearest neighbors of a particle play a dominant role, whereas distant particles do not contribute substantially to local field characteristics. In the units used here, the width of the range in question is approximately $8\xi_0$, whereas the largest width of the eigenfrequency splitting interval for a dimer is $\pm 2\xi_0$. Therefore, the dimer model cannot be used to explain optical properties of real fractal clusters. It would be interesting to identify the microscopic systems responsible for a splitting width of $\pm 4\xi_0$ and find out if the width varies with the ratio of monomer radii. It should also be expected that the equal of power absorbed by different monomers is specific to dimers, whereas many-particle systems are characterized by nonuniform distributions of Q_{ai} over particles. These questions are addressed in Sections 4 and 5.

Absorption spectra for systems of several monomers in the approximation of dipole–dipole interaction were calculated in numerous studies (see [5] and references therein). Detailed results of eigenvalue calculations for linear, planar, and three-dimensional systems with N varying from 2 to infinity were presented in [20, 21]. It was established in [21] that $|\kappa_m/\xi_0| \approx 4$ are characteristic of various configurations with $N = 6$ and 7. However, since all studies of this kind were conducted for particles of equal size, the problem should be reexamined.

System (2.6) written for N interacting dipoles is on the order of $3N$, and analytical results are difficult to obtain and understand. The present analysis is restricted to several simple configurations, but its results provide a basis for some general conclusions that answer the questions posed in this study.

First, we consider a linear aggregate of N particles. The corresponding system (2.6) breaks up into equations for the projections d_{in}^{r} and $d_{i\perp}^{\mathrm{r}}$ on the chain axis and the plane perpendicular to it:

$$\kappa d_{in}^{\mathrm{r}} - 2\sum_{i \neq j} \xi_{ij} d_{jn}^{\mathrm{r}} = a_i^{3/2} \varepsilon_{\mathrm{h}} E_{0n}, \tag{4.1}$$

$$\xi_{ij} = (a_i a_j)^{3/2}/r_{ij}^3, \quad i, j = 1, 2, \ldots, N,$$

$$\kappa d_{i\perp}^{\mathrm{r}} + \sum_{i \neq j} \xi_{ij} d_{j\perp}^{\mathrm{r}} = a_i^{3/2} \varepsilon_{\mathrm{h}} E_{0\perp}. \tag{4.2}$$

System (4.2) can be obtained from (4.1) by performing the change

$$2\xi_{ij} \longrightarrow -\xi_{ij}, \quad E_{0n} \longrightarrow E_{0\perp}.$$

The analysis presented here is mainly performed for the simpler system (4.2).

Since ξ_{ij} is a cubic function of $1/r_{ij}$, the interactions between neighboring particles play the dominant role. As a first approximation, we set

$$\xi_{ij} = \xi_{ii\pm 1}\delta_{i\pm 1 j}. \tag{4.3}$$

Table

N	1	2	3	4	5	6	7	8	9
κ_m/ξ	0	±1	0	±0.618	0	±0.445	0	±0.351	0
			$\pm 2^{1/2}$	±1.618	±1	±1.245	±0.766	±1	±0.618
					$\pm 3^{1/2}$	±1.800	±1.414	±1.523	±1.176
							±1.845	±1.882	±1.618
									±1.902

Matrices of this form (with κ as diagonal entries) are a special case of Jacobian matrices [22]. We denote the determinant of a matrix of order N by Δ_N and recall a number of properties of Δ_N important for this study [22]. All roots of Δ_N are real and distinct. The sum of the roots of Δ_N is zero. If N is even (odd), then Δ_N contains only even (odd) powers of κ. The interval between two consecutive roots of Δ_N contains exactly one root of Δ_{N-1}. Both the largest positive root and the largest absolute value of a negative root increase with N. The smallest root $\min(\kappa_m)$ of Δ_N is a decreasing function of ξ_{ij}. Its largest root $\max(\kappa_m)$ is an increasing function of ξ_{ij}. The minimal and maxima roots satisfy the inequalities

$$\begin{aligned} -2\max(\xi_{ij})\cos[\pi/(N+1)] &\le \min(\kappa_m) \\ &\le -2\min(\xi_{ij})\cos[\pi/(N+1)], \\ 2\min(\xi_{ij})\cos[\pi/(N+1)] &\le \max(\kappa_m) \\ &\le 2\max(\xi_{ij})\cos[\pi/(N+1)]. \end{aligned} \tag{4.4}$$

The equalities in (4.4) are valid for $\xi_{ij} = \xi$. In this case,

$$\begin{aligned} \kappa_m &= -2\xi\cos[\pi m/(N+1)], \\ m &= 1, 2, \ldots, N, \quad \xi_{ij} = \xi. \end{aligned} \tag{4.5}$$

The table shows the numerical values of κ_m/ξ given by (4.5).[4] It demonstrates that the distribution of κ_m calculated in approximation (4.3) is symmetric about the point $\kappa_m = 0$ and is nearly uniform over the interval bounded by $\pm\kappa_N$ (becoming slightly "condensed" toward $\max|\kappa_N|$. The largest root κ_N almost reaches the limit 2ξ when $N = 6$. With further increase in N, both the largest and smallest roots remain nearly constant. Since the number of roots increases, so does their density in the interval bounded by $\pm 2\xi$. Recall that these results are valid for dipole polarization perpendicular to the chain axis. For axially polarized dipoles with $N \gg 1$, the interval occupied by the roots corresponds to $|\kappa_N|/\xi = 4$, which agrees with the numerical results of [10, 11]; i.e., approximation (4.3) adequately describes linear aggregates in this case.

[4] Approximation (4.3) applies to many physical models, e.g., the linear crystal [23, 24].

The lowest-order corrections (linear in ξ_{ij} with $|i - j| > 1$) amount to about 10% and break the symmetry about $\kappa_m = 0$, in particular, by shifting the zero root. To illustrate the error due to approximation (4.3), we present the exact values of $\max(\kappa_m)$ obtained for $N = 6$ and for an infinite chain of contacting spheres [20, 21]:

$$\max(\kappa_m/\xi) = 2.008, \quad N = 6,$$

$$\max(\kappa_m/\xi) = 2.40, \quad N \longrightarrow \infty.$$

The corresponding approximate values given by (4.5) are 1.800 and 2.000, respectively; i.e., the error increases with N, reaching 10–15% at $N \approx 10$.

Let us use (4.3) and (4.5) to examine the influence of difference in radii on κ_m. For a chain of alternating particles of radius a_1 and a_2,

$$\xi_{ij} = \xi = \left[\frac{(a_1a_2)^{1/2}}{a_1 + a_2}\right]^3. \tag{4.6}$$

In this case, the ratio of κ_m corresponding to $a_1 \neq a_2$ and $a_1 = a_2$ is equal to that for dimers:

$$\frac{\xi}{\xi_0} = \left[\frac{2(a_1a_2)^{1/2}}{a_1 + a_2}\right]^3 < 1$$

(see (3.10) and curve *2* in Fig. 1).

Somewhat similar results are obtained by analyzing two- and three-dimensional configurations. In particular, for the "seven-leafed rosette" consisting of six particles of radius a_1 ($j = 1, 2, \ldots, 6$) located equidistantly on a circumference of radius r and a particle of radius a_2 at the center,

$$\begin{aligned} \xi_{jj+1} &= (a_1/r)^3 = \xi, \\ \xi_{jj+2} &= \xi/(3)^{3/2} = 0.192\xi = \eta, \\ \xi_{jj+3} &= \xi/8 = \zeta, \quad \xi_{j7} = [(a_1a_2)^{1/2}/r]^3 = \xi_1, \\ r &= 2a_1 \quad (a_2/a_1 < 1), \\ r &= a_1 + a_2 \quad (a_2/a_1 > 1). \end{aligned} \tag{4.7}$$

When $a_2/a_1 > 1$, the outer particles are in contact with the central one. When $a_2/a_1 < 1$, the outer particles are in contact with one another. In the case of polarization perpendicular to the rosette's plane, the seventh-order

system of equations breaks up into five first-order equations and a second-order system, which yield

$$\begin{aligned}\kappa_1 &= 2(\xi-\eta)+\xi = 1.741\xi,\\ \kappa_{2,3} &= \xi+\eta-\zeta = 1.067\xi\ (2),\\ \kappa_{4,5} &= -\xi+\eta+\zeta = -0.683\xi\ (2),\\ \kappa_{6,7} &= -\xi-\eta-\zeta/2\\ &\pm[6\xi_1^2+(\xi+\eta+\zeta/2)^2]^{1/2}\\ &= -\{1.255\pm[6(a_2/a_1)^3+1.575]^{1/2}\}\xi.\end{aligned} \tag{4.8}$$

(The numbers in parentheses are the multiplicities of degeneracy.) The roots $\kappa_1, \ldots, \kappa_5$ correspond to difference-frequency dipole oscillations of outer particles and depend only on the parameters of their interaction (for $a_2/a_1 \le 1$). The roots $\kappa_{6,7}$ are associated with the interaction of the central dipole with the resultant peripheral dipole, varying with a_2/a_1. For $a_2/a_1 = 1/2$ and 1 ($r = 2a_1$), we obtain

$$\kappa_6/\xi = -2.780,\quad -4.007,$$
$$\kappa_7/\xi = 0.270,\quad 1.497;$$

i.e., the roots strongly depend on a_2/a_1. When $a_2/a_1 > 1$, the distance r increases with a_2 as $1 + a_2/a_1$ and the frequencies $\kappa_1, \ldots, \kappa_5$ scale with $(1 + a_2/a_1)^{-3}$; i.e., they vary rapidly.

Analogous results are obtained for the five-leafed rosette consisting of four particles of radius a_1 located equidistantly on a circumference and of a particle of radius a_2 at the center. The interaction between the reduced central and resultant peripheral dipoles polarized perpendicular to the rosette's plane is characterized by the roots

$$\kappa_{1,2} = -\eta-\zeta/2 \pm [4\xi^2+(\eta+\xi/2)^2]^{1/2},\quad \kappa_3 = 2\eta-\zeta.$$

The difference-frequency root is doubly degenerate:

$$\kappa_{4,5} = \zeta.$$

In the case of outer particles in contact with the central one,

$$\begin{aligned}\xi &= \left(\frac{2t^{1/2}}{1+t}\right)^3\xi_0,\quad \eta = \left(\frac{2^{1/2}}{1+t}\right)^3\xi_0,\\ \zeta &= (1+t)^{-3}\xi_0,\quad t = a_2/a_1 \ge 2^{1/2}-1.\end{aligned} \tag{4.9}$$

The numerical values for $t = 1/2$, 1, and 2,

$$\kappa_{4,5}/\xi_0 = 0.296,\quad 0.125,\quad 0.037;$$
$$\kappa_3/\xi_0 = 1.380,\quad 0.582,\quad 0.173;$$
$$\kappa_2/\xi_0 = -2.931,\quad -2.459,\quad -1.804;$$
$$\kappa_1/\xi_0 = 0.959,\quad 1.627,\quad 1.557,$$

exhibit a substantial and intricate dependence of κ_m on a_2/a_1.

To illustrate the dependence of roots on particle size for three-dimensional systems, we now consider a cluster with six spherical particles of radius a_1 ($j = 1, 2, \ldots, 6$) lying on the coordinate axes at equal distances from a central spherical particle of radius a_2 (an endohedrally doped octahedron). The system of equations for $d_{i\alpha}^{r}$ breaks up into third-, second-, and first-order systems. The sum-frequency roots are

$$\begin{aligned}\kappa_{1,3} &= -(\eta-\zeta/2)\\ &\pm[12\xi^2+2(\eta+\zeta)^2+(\eta-\zeta/2)^2]^{1/2}\ (3),\\ \kappa_2 &= 0\ (3),\end{aligned} \tag{4.10}$$

and the difference-frequency roots are

$$\begin{aligned}\kappa_4 &= 3\eta+2\zeta\ (2),\\ \kappa_5 &= 3\eta+\zeta\ (3),\\ \kappa_6 &= 2\eta-\zeta\ (3),\\ \kappa_7 &= -3\eta+\zeta\ (3),\\ \kappa_8 &= -6\eta+2\zeta.\end{aligned} \tag{4.11}$$

Since the particles lying in each coordinate plane constitute a five-leafed rosette similar to that considered above, the parameters ξ, η, and ζ are given by (4.9). For $t = \alpha_2/\alpha_1 = 1/2$, 1, and 2, we obtain

$$\kappa_1/\xi_0 = 2.698,\ 3.251,\ 2.825;$$
$$\kappa_3/\xi_0 = -4.078,\ -3.833,\ -2.998;$$
$$\kappa_4/\xi_0 = 3.107,\ 1.311,\ 0.388;$$
$$\kappa_5/\xi_0 = 2.810,\ 1.186,\ 0.351;$$
$$\kappa_6/\xi_0 = 1.380,\ 0.582,\ 0.172;$$
$$\kappa_7/\xi_0 = -2.218,\ -0.936,\ -0.277;$$
$$\kappa_8/\xi_0 = -4.436,\ -1.871,\ -0.554.$$

The sum frequencies $\kappa_{1,3}$ (associated with interaction of the peripheral particles with the central one) vary relatively slowly, because an increase in a_2 affects both the interparticle distance and the central particle's polarizability. Accordingly, $\kappa_{1,3}$ varies approximately as $[2t^{1/2}/(1+t)]^3$. Since the difference modes are determined only by the interactions between peripheral particles, the frequencies $\kappa_4, \ldots, \kappa_8$ vary as $(1+t)^{-3}$ with increasing a_2. These trends are analogous to those characteristic of the seven- and five-particle two-dimensional systems considered above. Note also that

$\max|\kappa_m|$ almost reaches $4\xi_0$ for the endohedrally doped octahedron.

The difference modes strongly depending on the relative particle sizes are not excited in the symmetric systems discussed here, and the corresponding linear phenomena cannot be manifested in optical spectra. It is obvious that these difference modes must be excited when the symmetry is broken (as in dimers), and their frequency will be sensitive to ratios of radii.

The specific cases analyzed here illustrate the general considerations about the influence of relative particle size on u_m and κ_m presented in Section 1. When the eigenfrequency κ_m has an extremum for $a_2 = a_1$, it varies slowly. For other types of oscillations or geometries, κ_m may strongly depend a_2/a_1. For this reason, monomer polydispersity should be taken into account in numerical κ_m calculations analogous to [10, 11]. The results presented in [10, 11] are of fundamental importance for understanding optical properties of fractal clusters, but have a limited scope as applied to real systems. In particular, the range of eigenvalues obtained in [10, 11] would but slightly depend on size distribution, because its boundaries are determined by interaction between particle of equal size. However, allowance for size variability would obviously result in a slower variation of the eigenfrequency spectrum at the boundaries.

5. ABSORBED POWER AND ACTING FIELD FOR MANY-PARTICLE AGGREGATES

In many phenomena, including nonlinear effects and photomodification of clusters, an essential role is played by the local fields $\mathbf{E}_i$ acting on individual particles. According to one hypothesis, photomodification of clusters may be due to melting, evaporation, coalescence, and other changes in particles strongly heated by absorbed radiation. If the pulse duration is too short for heat exchange between particles or heat transfer to the environment to occur, then the initial temperature change is determined by the energy per unit volume absorbed by a particle:

$$\frac{\tau_p Q_{ai}}{a_i^3} \propto \tau_p \left| \frac{\mathbf{d}_i^r}{a_i^{3/2}} \right|^2,$$

where τ_p is the effective pulse duration. According to (2.18), the square root of this quantity determines the magnitude of the acting field $\mathbf{E}_i$.

For transversely and longitudinally polarized dimers with different particle radii, it was shown in Section 3 that $Q_{a1} = Q_{a2}$ near the resonances when $\xi \gg \delta$. In contrast, the distribution of Q_{ai} over particles for linear, planar, and bulk many-particle systems depends on many factors, including ratios of particle sizes.

The analysis that follows is focused on the distributions of $|\mathbf{d}_i^r|$ and $|\mathbf{d}_i^r|/a_i^{3/2}$ over particles near the strongest resonances and (as in Section 4) is restricted to simple cases, but the results provide a basis for some general conclusions. For longitudinally and transversely polarized linear aggregates, the strongest resonances correspond to the largest κ_m and $-\kappa_m$, respectively. Consider a symmetric linear trimer with contacting particles and $a_1 = a_3$. Using approximation (4.3) and assuming that $\xi_{12} = \xi_{23} \gg \delta$, we find that

$$\frac{|d_2^r|}{|d_1^r|} = 2^{1/2}, \quad \frac{Q_{a2}}{Q_{a1}} = 2, \quad |d_3^r| = |d_1^r|. \tag{5.1}$$

These results are independent of polarization and the value of a_2/a_1. The ratios of the acting-field magnitudes and absorbed powers per unit volume are

$$\frac{|E_2|}{|E_1|} = 2^{1/2}\left(\frac{a_1}{a_2}\right)^{3/2},$$
$$\frac{Q_{a2}}{Q_{a1}}\left(\frac{a_1}{a_2}\right)^3 = 2\left(\frac{a_1}{a_2}\right)^3. \tag{5.2}$$

Therefore, the fields acting on the peripheral particles is stronger than that acting on the central one if $a_1/a_2 < 2^{-1/3} = 0.794$. This phenomenon is analogous to the increase in field strength at a metal needlepoint. Under this condition, the peripheral particles are heated stronger than the central one. When the interaction between particles 1 and 3 is taken into account, the resulting correction to (5.1) depends on a_2/a_1 and amounts to about 10%. In particular, $|d_1^r|$ slightly increases, while $|d_2^r|$ slightly decreases; i.e., the "needlepoint effect" becomes more pronounced.

For a symmetric tetramer with contacting particles ($a_1 = a_4$, $a_2 = a_3$), we obtain

$$\frac{|d_2^r|}{|d_1^r|} = \left[1 + \left(\frac{p}{2}\right)^2\right]^{1/2} + \frac{p}{2},$$
$$\frac{Q_{a2}}{Q_{a1}} = 1 + p\left\{\left[1 + \left(\frac{p}{2}\right)^2\right]^{1/2} + \frac{p}{2}\right\}, \tag{5.3}$$
$$|d_4^r| = |d_1^r|, \quad |d_3^r| = |d_2^r|,$$
$$p = \left\{\frac{1}{2}\left[\left(\frac{a_2}{a_1}\right)^{1/2} + \left(\frac{a_1}{a_2}\right)^{1/2}\right]\right\}^3.$$

Since $p \geq 1$ ($p = 1$ when $a_2 = a_1$), it holds that

$$\frac{|d_2^r|}{|d_1^r|} \geq \frac{1 + 5^{1/2}}{2} = 1.618, \quad \frac{Q_{a2}}{Q_{a1}} \geq 2.618. \tag{5.4}$$

Thus, the difference in absorbed power between outer and inner particles is slightly greater than in the case of

a trimer. The ratio of the acting-field magnitudes is derived from (5.3):

$$\frac{|E_2|}{|E_1|} = \left(\frac{a_1}{a_2}\right)^{3/2}\frac{|d_2^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = 2^{-4}\left\{\left[\left(1+\frac{a_1}{a_2}\right)^6 + 2^8\left(\frac{a_1}{a_2}\right)^3\right]^{1/2} + \left(1+\frac{a_1}{a_2}\right)^3\right\}. \quad (5.5)$$

A simple calculation shows that $|E_1|/|E_2| \geq 1$ if

$$a_1/a_2 \leq 0.717. \quad (5.6)$$

Therefore, the needlepoint effect manifests itself at a smaller ratio a_1/a_2 in a tetramer as compared to a trimer.

For the strongest resonance in a symmetric linear system of five contacting monomers ($a_1 = a_5$, $a_2 = a_4$), we obtain

$$\frac{|d_2^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = \left[2\left(\frac{\xi_{23}}{\xi_{12}}\right)^2 + 1\right]^{1/2}, \quad \frac{|d_3^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = \frac{2\xi_{23}}{\xi_{12}},$$
$$|d_5^{\mathrm{r}}| = |d_1^{\mathrm{r}}|, \quad |d_4^{\mathrm{r}}| = |d_2^{\mathrm{r}}|,$$
$$\kappa_1 = -(\xi_{12}^2 + 2\xi_{23}^2)^{1/2}, \quad \xi_{12} = \frac{(a_1a_2)^{3/2}}{(a_1+a_2)^3}, \quad (5.7)$$
$$\xi_{23} = \frac{(a_2a_3)^{3/2}}{(a_2+a_3)^3}.$$

If $a_i = a$, then

$$\frac{|d_2^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = 3^{1/2} = 1.732, \quad \frac{|d_3^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = 2.$$

According to (5.7), $|d_2^{\mathrm{r}}|/|d_1^{\mathrm{r}}| > 1$ for any a_i/a_j, and $|d_3^{\mathrm{r}}|/|d_2^{\mathrm{r}}|$ can be both greater and smaller than unity. Setting $a_2 = a_3$ for simplicity, we obtain

$$\frac{\xi_{12}}{\xi_{23}} = \frac{8(a_1a_2)^{3/2}}{(a_1+a_2)^3} \leq 1, \quad \kappa_1 = -\xi_{23}\left[2+\left(\frac{\xi_{12}}{\xi_{23}}\right)^2\right]^{1/2},$$
$$\xi_{23} = 1/8.$$

Furthermore,

$$\frac{|E_2|}{|E_1|} = \left[2^{-5}\left(1+\frac{a_1}{a_2}\right)^6 + \left(\frac{a_1}{a_2}\right)^3\right]^{1/2},$$
$$\frac{|E_3|}{|E_1|} = \frac{(1+a_1/a_2)^3}{4}, \quad (5.8)$$
$$\frac{|E_3|}{|E_2|} = 2/\left[2+\left(\frac{\xi_{12}}{\xi_{23}}\right)^2\right]^{1/2} \geq \left(\frac{2}{3}\right)^{1/2}.$$

Both $|E_2|/|E_1|$ and $|E_3|/|E_1|$ monotonically increase with a_1/a_2, and $|E_3|/|E_2|$ reaches a minimum when $a_2 = a_1$. The values of a_1/a_2 for which $|E_2|/|E_1| = 1$ and $|E_3|/|E_1| = 1$, respectively, are

$$a_1/a_2 = 0.676, \quad a_1/a_2 = 2^{2/3} - 1 = 0.588. \quad (5.9)$$

Thus, a stronger field acting on an outer particle (as compared to inner ones) is associated with an even smaller size of outer particles in a pentamer as compared to a tetramer or trimer. Otherwise, the fields acting on the inner particles are stronger.

These particular cases suggest the following general rule. The degree of nonuniformity of a $|d_i^{\mathrm{r}}|$ or Q_{ai} distribution over particles increases with the number of monomers in a chain, and $|d_i^{\mathrm{r}}|$ is increases toward the center of a chain. Therefore, nonlinear effects in an aggregate of identical particles are localized at its center. With decreasing outer-particle radius, the localization of nonlinear effects shifts to the corresponding end of a chain. For example, melting must begin either at the center of a chain of identical particles or at the ends of a nonuniform chain with sufficiently small outer particles. This rule applies to two- and three-dimensional systems as well.

Let us discuss this rule as applied to the symmetric five-leafed rosette considered in Section 4. For polarization perpendicular to the rosette's plane, we obtain

$$\frac{|d_2^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = -\frac{4\xi}{\kappa_1} = \frac{2}{(1+q^2)^{1/2}+q},$$
$$q = \frac{1}{2}\left(1+\frac{2^{1/2}}{8}\right)\left(\frac{a_1}{2a_2}\right)^{3/2} = 0.208\left(\frac{a_1}{a_2}\right)^{3/2}, \quad (5.10)$$
$$\frac{|E_2|}{|E_1|} = \frac{2}{[(a_2/a_1)^3 + 0.0433]^{1/2} + 0.208}.$$

Therefore, if $a_2 = a_1$, then

$$\frac{|d_2^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = 1.627,$$

which is close to the values obtained for linear systems. Furthermore, both $|d_2^{\mathrm{r}}|/|d_1^{\mathrm{r}}|$ and $|E_2|/|E_1|$ are monotonically decreasing functions of a_2/a_1, and $|d_2^{\mathrm{r}}|/|d_1^{\mathrm{r}}| \geq 1$ and $|E_2|/|E_1| \leq 1$, respectively, when

$$a_2/a_1 \geq 0.425, \quad a_1/a_2 \leq 0.681. \quad (5.11)$$

Thus, when the radius of the central particle 2 is not sufficiently small, its reduced dipole moment is greater than that of the outer particles. Owing to the needlepoint effect, the acting field and absorbed power per

unit volume are greater for the outer particles if their radius is sufficiently small.

Analogous results are obtained for the seven-leafed rosette considered in Section 4:

$$\frac{|d_2^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = \frac{6}{(6+q^2)^{1/2}+q}, \quad q = 1.255\left(\frac{a_1}{a_2}\right)^{3/2},$$

$$\frac{|d_2^{\mathrm{r}}|}{|d_1^{\mathrm{r}}|} = 1.497 \quad (a_1 = a_2), \tag{5.12}$$

$$\frac{|E_2|}{|E_1|} = \frac{6}{[6(a_2/a_1)^3+1.575]^{1/2}+1.255}.$$

Instead of (5.11), we obtain

$$a_2/a_1 \le 0.632, \quad a_1/a_2 \le 0.659. \tag{5.13}$$

Now, consider the endohedrally doped octahedral cluster with a central particle of radius a_2 and identical peripheral particles of radius a_1. If the external field is parallel to the z axis, then the four particles lying in the xy plane are geometrically equivalent and are called particles 1, the particle at the origin is referred to as particle 2, and those on the z axis are called particles 3. For this symmetric system, we obtain

$$\frac{|d_{2z}^{\mathrm{r}}|}{|d_{1z}^{\mathrm{r}}|} = \frac{12}{[12+0.543(a_1/a_2)^3]^{1/2}+1.248(a_1/a_2)^{3/2}} \quad (2.505),$$

$$\frac{|d_{3z}^{\mathrm{r}}|}{|d_{1z}^{\mathrm{r}}|} = 2 - \frac{2.871}{[12+0.543(a_1/a_2)^3]^{1/2}+1.248(a_1/a_2)^{3/2}} \quad (1.401), \tag{5.14}$$

$$\frac{|E_{2z}|}{|E_{1z}|} = \frac{12}{[12(a_1/a_2)^3+0.543]^{1/2}+1.248} \quad (2.505).$$

The numbers in parentheses are the corresponding ratios for $a_1 = a_2$. Again, we see that the dipole moment of particle 2 is greater than that of particles 1. Therefore, $|E_{1z}|/|E_{2z}| > 1$ only if a_1 is sufficiently small as compared to a_2 (cf. (5.11) and (5.13)):

$$a_1/a_2 \le 0.471. \tag{5.15}$$

Thus, the distributions of absorbed power and acting-field magnitude strongly depend on the number of particles in a cluster, its geometry, and the relative radii of the constituent monomers. A general rule is that the reduced dipole moment increases toward the center of an aggregate. However, nonlinear effects can be more pronounced either for inner or for outer particles, depending on their relative radii. A sufficiently small outer particle can be the "locus" of a nonlinear effect. To perform a more detailed calculation, one must specify the cluster geometry and state a physical problem to be solved.

6. DISCUSSION

In the 1980s and 1990s, when the principal objective was to highlight the properties of fractal clusters as a special optical medium, research was focused on large systems as whole entities, e.g., on the absorption spectra of N-particle aggregates with $N \sim 10^4$ and higher. However, disordered, strongly bonded fractal clusters always contain subsystems of closely spaced, nearly contacting particles, whereas the density of distantly spaced particles is relatively low. This feature distinguishes fractal clusters from uniform media. In particular, closely spaced particles determine the width of the dipole eigenfrequency range and, therefore, the width of the inhomogeneously broadened tail of the plasmon band and the spectral range where resonant nonlinear effects can be observed. It is shown in Section 4 that $\max(|\kappa_m|/\xi_0) \approx 4$ is reached for $N \approx 7$ in linear, planar, and bulk systems; i.e., the number of particles is more important than the aggregate's geometry. This result highlights the dominant role played by the dynamics of dipole–dipole interaction of a relatively small number of monomers, as compared to statistical aspects of the problem. The approach adopted above must be justified with regard to various nonlinear problems. For example, photomodification of clusters is studied by using electron diffraction patterns to analyze changes in aggregate structure, and this analysis is not amenable to statistical averaging of any kind.

In a widely applied model [1–11], the particle radius was assumed to be equal for all monomers in a cluster. In Sections 3 and 4, it is shown that the eigenfrequencies κ_m strongly depend on the ratio of radii and the largest $\max|\kappa_m|$ corresponds to approximately equal radii. Accordingly, the spectrum width (the range of κ_m) is determined by interaction between almost identical contacting particles. The physical explanation of this result lies in the fact that the polarizability of the smaller particle decreases as its radius cubed and this trend has a stronger effect than the possibility of a decrease in interparticle distance. However, the frequency κ_m can vary drastically (as radius cubed) within the interval $(\min\kappa_m, \max\kappa_m)$. This behavior of κ_m spectrum is independent of the aggregate geometry. Thus, inhomogeneous broadening of absorption spectra is explained by variation of both interparticle distance and particle radii.

A difference in radii has an even stronger effect on resonant amplitudes, particularly those corresponding to difference modes. According to (2.7), $d_i^{\mathrm{r}} - d_j^{\mathrm{r}}$ is pro-

portional to $a_i^{3/2} - a_j^{3/2}$. Therefore, difference resonance amplitudes strongly depend on a_i/a_j, and the resonant modes prohibited when $a_i = a_j$ can be excited by an external field. Figure 2 illustrates this effect in the simplest case of a dimer.

The distributions of absorbed power Q_{ai}, absorbed power per unit volume Q_{ai}/a_i^3, and acting-field magnitude

$$|\mathbf{E}_i| \propto [Q_{ai}/a_i^3]^{1/2}$$

are of primary interest for analysis of nonlinear phenomena. According to the general rule stated in Section 5, the value of Q_{ai} increases toward the center of an aggregate of monomers of equal size. Moreover, Q_{ai} may change by several times within an aggregate. If the radii of peripheral particles are sufficiently small, then the field acting on them is stronger than that acting on "inner" particles. This effect is qualitatively consistent with the electrostatic needlepoint effect. Thus, nonlinear phenomena can be more pronounced either for inner or for outer particles in an aggregate, depending on their relative radii.

The distribution of $|\mathbf{E}_i|$ is essential for scattering that is not degenerate with respect to frequency. For example, Raman scattering by molecules adsorbed on a nanoparticle surface depends on the product $|\mathbf{E}_i|^2|\mathbf{E}_{iR}|^2$, where $\mathbf{E}_{iR}$ is the local field at the Stokes frequency [1]. Thus, the strongest effect is attained when both pump and scattered fields are strong at the point where the scattering molecule is located. Since local fields reach maximum values at resonant frequencies, the distribution of modes over the particles that make up a cluster has to be calculated in the theory of giant scattering. A similar problem arises in studies of nonlinear scattering. According to Section 3, a dimer with $a_1 = a_2$ is characterized by a single resonance in the long-wavelength absorption tail, whereas another resonance appears for particles of different size; i.e., high local field magnitudes can be obtained simultaneously at two frequencies. This example demonstrates that the use of aggregates of particles of different size offers new prospects in nonlinear optics. In other words, particle-size variability may be introduced into a model as an element of practical importance rather than a complicating factor that simply cannot be ignored.

Thus, optical properties of a cluster strongly depend on the relative sizes of its constituent particles. This conclusion is supported both by the general analysis presented in Section 2 and by illustrative numerical calculations. Effects due to difference in particle size are so strong that they may change the overall physical picture of optical phenomena in cluster systems, particularly those involving nonlinear optical properties and local field fluctuations.

The diversity of factors that affect the eigenvalue distributions and resonant amplitudes for disordered clusters and the need for statistical averaging over various random parameters suggest that absorption and refraction spectra must be insensitive to the details of models. Wide and featureless inhomogeneously broadened tails of an absorption band merely imply that the particles are strongly bonded in a fractal cluster. To validate a model, one should consider experiments that yield "more local" characteristics. Valuable information of this kind may be extracted from nonlinear effects, such as creation of "holes" in absorption spectra, experimental methods based on near-field optics, or experiments on a relatively small number of particles.

ACKNOWLEDGMENTS

We thank R.V. Markov, A.M. Shalagin, and D.A. Shapiro for helpful discussions of a number of issues addressed in this paper. This work was supported by the Russian Foundation for Basic Research, project nos. 02-02-17885, 03-02-06600, and NSh-439.2003.2).

REFERENCES

1. V. M. Shalaev, *Nonlinear Optics of Random Media: Fractal Composites and Metal–Dielectric Films* (Springer, Berlin, 2000).
2. V. M. Shalaev, Phys. Rep. **272**, 61 (1996).
3. V. M. Shalaev, in *Optical Properties of Nanostructured Random Media*, Ed. by V. M. Shalaev (Springer, Berlin, 2002), p. 93.
4. V. P. Drachev, S. V. Perminov, S. G. Rautian, and V. P. Safonov, in *Optical Properties of Nanostructured Random Media*, Ed. by V. M. Shalaev (Springer, Berlin, 2002), p. 113.
5. U. Kreibig and M. Vollmer, *Optical Properties of Metal Clusters* (Springer, Berlin, 1995).
6. S. V. Karpov and V. V. Slabko, *Optical and Photophysical Properties of Fractal-Structured Sols of Metals* (Sib. Otd. Ross. Akad. Nauk, Novosibirsk, 2003).
7. B. M. Smirnov, *Physics of the Fractal Clusters* (Nauka, Moscow, 1991).
8. V. M. Shalaev and M. I. Shtokman, Zh. Éksp. Teor. Fiz. **92**, 509 (1987) [Sov. Phys. JETP **65**, 287 (1987)].
9. A. V. Butenko, V. M. Shalaev, and M. I. Shtokman, Zh. Éksp. Teor. Fiz. **94**, 107 (1988) [Sov. Phys. JETP **67**, 60 (1988)].
10. V. A. Markel', L. S. Muratov, and M. I. Shtokman, Zh. Éksp. Teor. Fiz. **98**, 819 (1990) [Sov. Phys. JETP **71**, 455 (1990)].
11. V. A. Markel, V. M. Shalaev, E. B. Stechel, *et al.*, Phys. Rev. B **53**, 2425 (1996).
12. J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941; Gostekhizdat, Moscow, 1948).
13. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 8: *Electrodynamics of Continuous Media*, 2nd ed. (Nauka, Moscow, 1982; Pergamon Press, Oxford, 1984).

14. V. A. Markel, J. Opt. Soc. Am. B **12**, 1783 (1995).
15. V. P. Drachev, S. V. Perminov, S. G. Rautian, *et al.*, Zh. Éksp. Teor. Fiz. **121**, 1051 (2002) [JETP **94**, 901 (2002)].
16. S. V. Perminov, S. G. Rautian, and V. P. Safonov, Opt. Spektrosk. **95**, 447 (2003) [Opt. Spectrosc. **95**, 416 (2003)].
17. E. M. Purcell and C. R. Pennipacker, Astrophys. J. **186**, 705 (1973).
18. B. T. Draine, Astrophys. J. **333**, 848 (1988).
19. S. V. Karpov, A. K. Popov, S. G. Rautian, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **48**, 528 (1988) [JETP Lett. **48**, 571 (1988)].
20. P. Clippe, R. Evrard, and A. A. Lucas, Phys. Rev. B **14**, 1715 (1976).
21. M. Ausloos, R. Clippe, and A. A. Lucas, Phys. Rev. B **18**, 7176 (1978).
22. F. R. Gantmakher, *Oscillation Matrices* (Gostekhizdat, Moscow, 1950).
23. E. Fermi, *Molecole e Cristalli* (Zanichelli, Bologna, 1934; Barth, Leipzig, 1938; Inostrannaya Literatura, Moscow, 1947).
24. C. Kittel, *Introduction to Solid State Physics*, 4th ed. (Wiley, New York, 1971; Fizmatgiz, Moscow, 1963).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 705–706.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 805–807.

ATOMS, SPECTRA, RADIATION

Beam Propagation in a Randomly Inhomogeneous Medium

A. A. Stanislavsky

Institute of Radio Astronomy, National Academy of Sciences of Ukraine, Kharkov, 61002 Ukraine
e-mail: alexstan@ira.kharkov.ua

Received October 8, 2003

Abstract—An integrodifferential equation describing the angular distribution of beams is analyzed for a medium with random inhomogeneities. Beams are trapped because inhomogeneities give rise to wave localization at random locations and random times. The expressions obtained for the mean square deviation from the initial direction of beam propagation generalize the "3/2 law." © *2004 MAIK "Nauka/Interperiodica".*

Fluctuations of the beam propagation direction in a randomly inhomogeneous medium are frequently observed in nature. Examples include random refraction of radio waves in the ionosphere and solar corona, stellar scintillation due to atmospheric inhomogeneities, and other phenomena. The propagation of a beam (of light, radio waves, or sound) in such media can be described as a normal diffusion process [1, 2]. One extraordinary property predicted for random media—and later revealed—is the Anderson localization [3], which brings normal diffusion to a complete halt. In the context of wave propagation in a random medium, the Anderson localization is caused by interference of waves resulting from multiple scattering [4]. When two waves propagating in opposite directions along a closed path are in phase, the resultant wave is more likely to return to the starting point than propagate in other directions. The properties of a randomly inhomogeneous medium vary not only from point to point, but also with time. Consequently, localization may take place both at random locations and at random times. Random localization affects diffusive light propagation in a random medium. An approach to describing this effect is developed in this paper.

Suppose that the medium is statistically homogeneous and isotropic. Then, a beam propagating through the medium is deflected at random. Localization implies that the beam is trapped in some region. Since the trapped beam returns to the point where it was trapped, its propagation is "frozen" for some time. After that, a randomly deflected beam leaves the region and propagates further until it is trapped in another region (or at a point), and the localization cycle repeats. The randomly winding beam path due to inhomogeneities is responsible for the random refraction analyzed in this study.

The angle θ of deviation of a beam from its initial direction is characterized by a probability density $W_\alpha(\theta, \sigma)$, where σ is the path traveled by the beam. Let us derive an integrodifferential equation for the probability density. In contrast to rotational Brownian motion, the random walks analyzed here consist of random angle jumps $\Delta\theta_i$ at points separated by segments of random length $\Delta\sigma_i$. Exact knowledge of the distributions of these random variables is not required. It is sufficient to assume that the angle jumps are independent random variables belonging to the domain of attraction of Gaussian probability distributions. The random segment lengths $\Delta\sigma_i$ are also identically distributed independent random variables, with distribution is characterized by an exponent α. Since $\Delta\sigma_i$ is a nonnegative quantity, this distribution is totally asymmetric, and $0 < \alpha \le 1$. Recall that a random variable characterized by a probability distribution $f(x)$ of this kind is described by the Laplace transform

$$\phi(s) = \int_0^\infty \exp(-sx)df(x) = \exp\{-(Ax)^\alpha\},$$

where $x \ge 0$ and $A > 0$ [5]. The total path length is the sum of all $\Delta\sigma_i$. Both $\Delta\sigma_i$ and $\Delta\theta_i$ are Markov processes. However, since the former is the master process with respect to the latter, the resultant process may not preserve the Markov property [6]. Convergence of distributions ensures passing to a continuous limit [7]. This leads to the diffusion equation

$$W_\alpha(\theta, \sigma) - W_\alpha(\theta, 0) = \int_0^\sigma \frac{D}{\Gamma(\alpha)\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial W_\alpha(\theta, \sigma')}{\partial\theta}\right)(\sigma - \sigma')^{\alpha-1}d\sigma',$$

where D is a diffusion coefficient and $\Gamma(x)$ is the gamma function. The solution to this equation can be expressed as an integral transform of the probability

distribution associated with rotational Brownian motion:

$$W_\alpha(\theta, \sigma) = \int_0^\infty F_\alpha(z) W_1(\theta, \sigma^\alpha z) dz,$$

where

$$F_\alpha(z) = \sum_{k=0}^{\infty} \frac{(-z)^k}{k!\Gamma(1-\alpha-k\alpha)}.$$

The diffusion equation yields the mean

$$\overline{\cos\theta} = E_\alpha(-2D\sigma^\alpha),$$

where

$$E_\alpha(-x) = \sum_{n=0}^{\infty} \frac{(-x)^n}{\Gamma(1+n\alpha)}$$

is the Mittag-Leffler function. At large σ, all beam directions are equiprobable. However, in contrast to normal diffusion ($\alpha = 1$), a beam has to travel a longer path σ to reach this state. This process is somewhat analogous to "anomalously slow" relaxation.

Following the method developed in [8], one can find the mean square of the distance r from the starting point to the observation point reached by the beam that has traveled an intricate path of length σ through the medium:

$$\overline{r^2} = \frac{\sigma^\alpha}{D\Gamma(\alpha+1)} - \frac{1}{2D^2}[1 - E_\alpha(-2D\sigma^\alpha)]. \quad (1)$$

If $D\sigma \ll 1$, then

$$\overline{r^2} \approx 2\sigma^{2\alpha}\left[\frac{1}{\Gamma(2\alpha+1)} - \frac{2D\sigma^\alpha}{\Gamma(3\alpha+1)}\right]. \quad (2)$$

If the z axis of a polar coordinate system is aligned with the initial beam direction, then the mean square of the distance passed by the beam along this axis is given by the formula

$$\overline{z^2} = \frac{1}{3D}\left[\frac{\sigma^\alpha}{\Gamma(\alpha+1)} - \frac{1}{6D}(1 - E_\alpha(-6D\sigma^\alpha))\right]. \quad (3)$$

If $D\sigma$ is small, then

$$\overline{z^2} \approx 2\sigma^{2\alpha}\left[\frac{1}{\Gamma(2\alpha+1)} - \frac{6D\sigma^\alpha}{\Gamma(3\alpha+1)}\right]. \quad (4)$$

Now, the mean square deviation of the beam from its initial direction can be calculated by combining (1) with (3):

$$\overline{\rho^2} = \overline{r^2} - \overline{z^2} = \frac{2\sigma^\alpha}{3D\Gamma(\alpha+1)} - \frac{1}{2D^2}[1 - E_\alpha(-2D\alpha^\alpha)] + \frac{1}{18D^2}[1 - E_\alpha(-6D\sigma^\alpha)]. \quad (5)$$

If $D\sigma$ is small, then a generalized 3/2 law [8] is obtained:

$$\sqrt{\overline{\rho^2}} \approx \frac{2\sqrt{2}}{\sqrt{\Gamma(3\alpha+1)}} D^{1/2}\sigma^{3\alpha/2}. \quad (6)$$

The mean squares given by (1), (3), and (5) increase as σ^α at large σ. The case of $\alpha = 1$ corresponds to normal diffusion without wave localization. Thus, the approach developed here subsumes classical results of the theory of beam propagation in a randomly inhomogeneous medium [1, 2].

Finally, it should be recalled that considerable experimental deviations from the 3/2 law (more precisely, from an exponent of 3/2 in the classical power law) were mentioned in [9]. However, they were attributed to systematic measurement errors, probably because of the lack of plausible interpretation. This problem can be revisited in view of the results obtained in this study. Moreover, new accurate experimental studies of beam propagation in suitable randomly inhomogeneous media would be extremely useful for verifying the model proposed here.

REFERENCES

1. S. M. Rytov, *Introduction to Statistical Radiophysics* (Nauka, Moscow, 1966).
2. L. A. Chernov, *Wave Propagation in a Random Medium* (Nauka, Moscow, 1975; McGraw-Hill, New York, 1960).
3. P. W. Anderson, Phys. Rev. **109**, 1492 (1958).
4. D. S. Wiersma, P. Bartolini, A. Lagendijk, and R. Righini, Nature **390**, 671 (1997).
5. V. M. Zolotarev, *One-Dimensional Stable Distributions* (Nauka, Moscow, 1983; Am. Math. Soc., Providence, RI, 1986).
6. W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. (Wiley, New York, 1967; Mir, Moscow, 1984).
7. A. Stanislavsky, Phys. Scr. **67**, 265 (2003).
8. L. A. Chernov, Zh. Éksp. Teor. Fiz. **24**, 210 (1953).
9. I. G. Kolchinskiĭ, Astron. Zh. **29**, 350 (1952).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 707–718.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 808–820.

PLASMA, GASES

Collision Operator for Relativistic Electrons in a Cold Gas of Atomic Particles

L. P. Babich

Russian Federal Nuclear Center, Institute of Experimental Physics, Sarov, Nizhegorodskaya oblast, 607188 Russia
e-mail: kay@sar.ru

Received September 18, 2003

Abstract—The collision operator of relativistic electrons with a cold gas of atomic particles is derived consistently taking into account elastic interactions, excitation of electron shells, and ionization. The creation of secondary electrons is described accurately. In the range of energies exceeding the binding energy of atomic electrons, the operator implicates only the angular scattering by nuclei and the ionization integral that automatically allows for scattering by atomic electrons. The collision operator used earlier for studying the kinetics of avalanches of relativistic runaway electrons is analyzed. A more exact operator derived in the present study is simpler in form and saves time in computer calculations. *© 2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

In collision operators of the kinetic equation (KE), which are reduced to the forms convenient for numerical calculations of transport of relativistic electrons in a substance, infrequent events of creation of high-energy electrons are usually ignored. However, Gurevich *et al.* [1, 2] proved that such events may change the course of an ionization process in the presence of electric field, leading to the development of relativistic runaway electron avalanches (RREAs) and to gas breakdown in weaker fields than those required for the conventional breakdown. The theory of breakdown in air and the mechanisms of ascending atmospheric discharges involving RREAs was developed in terms of the kinetic equation [3, 4]

$$\frac{\partial f}{\partial t} - \left[\frac{1-\mu^2}{p}\frac{\partial}{\partial \mu}f + \mu\frac{\partial}{\partial p}f\right]eE = \mathrm{St}_{\mathrm{fr}} + \mathrm{St}_{\mathrm{sc}} + \mathrm{St}_{\mathrm{ion}} \quad (1.1)$$

with the following components of the electron–molecule collision operator:

$$\mathrm{St}_{\mathrm{fr}} = \frac{1}{p^2}\frac{\partial}{\partial p}[p^2 F(p) f(t, p, \mu)], \quad (1.2)$$

$$\mathrm{St}_{\mathrm{sc}} = \frac{(Z_{\mathrm{mol}}/2 + 1)F(p)}{4\gamma p}\hat{L}_\mu f(t, p, \mu), \quad (1.3)$$

$$\mathrm{St}_{\mathrm{ion}} = N_{\mathrm{mol}}\beta c \int\limits_{2\varepsilon + \varepsilon_{\mathrm{ion}}}^{\infty} d\varepsilon' \left(\frac{\gamma'^2 - 1}{\gamma^2 - 1}\right) \times \sigma_{\mathrm{ion}}(\varepsilon, \varepsilon')\frac{1}{\pi}\int\limits_0^\pi f(t, \varepsilon', \mu')d\alpha. \quad (1.4)$$

Operators (1.2) and (1.3) describe the flux in the momentum space and the angular diffusion due to scattering from atomic nuclei (the component containing factor $Z_{\mathrm{mol}}/2$ in Eq. (1.3)) and electrons. The creation of high-energy electrons is described by ionization integral (1.4). Here, $f(t, p, \mu)$ is the electron distribution function (EDF) over the momentum modulus p and the cosine of the angle between $\mathbf{p}$ and unit vector $\mathbf{e} = -\mathbf{E}/E$ in the direction of the electric force, where $\mathbf{E}$ is the electric field vector; e is the elementary charge; $\varepsilon_{\mathrm{ion}}$ is the ionization threshold; N_{mol} is the molecule concentration; Z_{mol} is the number of electrons per molecule; α is the azimuth angle (see Fig. 1);

$$\hat{L}_\mu = \frac{\partial}{\partial \mu}(1-\mu^2)\frac{\partial}{\partial \mu}$$

is the angular part of the Laplace operator in the spherical system of coordinates; $\gamma = 1/\sqrt{1-\beta^2}$; $\beta = v/c$; $F(p)$ is the drag force describing the average energy losses of an electron [5, 6]; $\sigma_{\mathrm{ion}}(\varepsilon', \varepsilon)$ is the differential ionization cross section; and $\varepsilon = (\gamma - 1)mc^2$ is the kinetic energy. Primes denote the values of the quantities prior to interaction events.

The dependence of the characteristic time t_e of RREA enhancement on $eE/F_{\min}$ was calculated in [3, 4] by numerically solving Eq. (1.1)–(1.4). The discrepancy with the values of t_e obtained by the Monte Carlo method [7, 8] was partly eliminated in subsequent publications [8–12], where the procedure for solving Eq. (1.1)–(1.4) was refined. Satisfactory agreement was reached in [10–12], where the ionization process was described in the KE in 3D geometry, contrary to 2D geometry used in [3, 4], and kinetic equation (1.1)–(1.4) was written in divergent form. The remaining discrep-

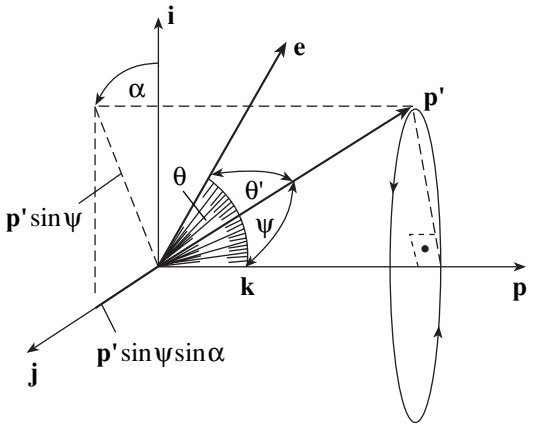


Fig. 1. Scattering geometry: vector **e** defines the direction of the electric force, **p**' and **p** are the electron momenta before and after the interaction, and ψ is the scattering angle.

ancy is probably due to the approximations made in [3, 4] in deriving operators (1.2)–(1.4).

Here, we expound on a consistent derivation of the collision operator in a medium with a uniform external electric field, specifying a preferred direction, and describe all approximations made. We assume that interactions with neutral atoms (molecules) dominate and atoms are immobile, so that their energy distribution is insignificant. We use the procedure developed by Holstein [13], who obtained an exact nonrelativistic operator consisting of rigorous balance components responsible for elastic collisions, excitation, and ionization. In contrast to [13], in reducing the exact operator to the form convenient for calculations, we will not use the Lorentz approximation, which is valid for weakly anisotropic EDFs, but will take advantage of the fact that interactions with small variations of the direction and magnitude of the momentum dominate in the high-energy range. For this reason, the parts of the operator responsible for elastic interactions and excitation of atomic particles can be reduced to differential form. The ionization integral describing the population of an element of the phase volume remains nonreduced with the lower integration limit equal to the accurate value of $\varepsilon + \varepsilon_{ion}$ as in [13] and not to $2\varepsilon + \varepsilon_{ion}$ as in relation (1.4). In contrast to [13], this integral takes into account the relation between the scattering angle and the electron energies as well as the exact relation between the angles formed by the electron momentum vectors participating in ionization event with direction **e**.

To find the differences with relations (1.2)–(1.4), we reduced the collision operator to a form similar to (1.2)–(1.4). For this purpose, we divide the ionization integral into two parts, one of which describes "weak" interactions and the other, "strong" interactions with an energy change of an impinging electron, such that both electrons after ionization event enter the high-energy range (in particular, both electrons can become runaway electrons). The criterion adopted for separating interactions is the same as in [3, 4]. In the latter publications, the explicit dependence of cross section ε_{ion} on ε' and ε as well as the procedure of its factorization were used, although neither of these are necessary.

In the gas of molecules consisting of n identical atoms, molecular quantities N_{mol} and Z_{mol} are connected with atomic quantities N_{at} and Z_{at} via the relations

$$N_{at} = nN_{mol}, \quad Z_{at} = Z_{mol}/n. \tag{1.5}$$

The operator components describing the excitation and ionization contain product NZ, which, in accordance with relations (1.5), is the same for atoms and molecules. The component responsible for scattering from a nucleus contains a factor with number n if calculations are carried out in terms of molecular quantities (see Section 7):

$$N_{at}Z_{at}^2 = N_{mol}Z_{mol}^2/n. \tag{1.6}$$

2. GEOMETRY OF SCATTERING PROCESSES

Figure 1 shows the scattering geometry in the coordinate system defined by unit vectors

$$\mathbf{i} = \mathbf{p} \times [\mathbf{e} \times \mathbf{p}]/p^2 \sin\theta = (\mathbf{e}p^2 - \mathbf{p}(\mathbf{p} \cdot \mathbf{e}))/p^2 \sin\theta,$$

$$\mathbf{j} = [\mathbf{p} \times \mathbf{e}]/p \sin\theta, \quad \mathbf{k} = \mathbf{p}/p,$$

where $\mathbf{p}(p, \theta, \varphi)$ is the electron momentum vector after scattering [13]. In this system, **k** is the polar axis and scattering angle $\psi \in [0, \pi]$ becomes the polar angle, while angle $\alpha \in [0, 2\pi]$ between **j** and the momentum projection $\mathbf{p}'(p', \theta', \varphi')$ onto plane $\mathbf{p} = 0$ before scattering is the azimuth angle. Formula (19b) from [13],

$$\cos\theta' = \cos\theta\cos\psi + \sin\theta\sin\psi\cos\alpha, \tag{2.1}$$

which connects angles α and ψ with angles θ' and θ between vector **e** and the directions of the electron momenta before ($\mathbf{p}'(p', \theta', \varphi')$) and after ($\mathbf{p}(p, \theta, \varphi)$) scattering, was derived under the assumption that $p' = p$, which is violated in inelastic collisions and which is actually superfluous. If we form the scalar product of the regular decomposition of **p**' in unit vectors **i**, **j**, **k**,

$$\mathbf{p}' = \frac{\mathbf{p}}{p}p'\cos\psi + \frac{\mathbf{p} \times \mathbf{e}}{p\sin\theta}p'\sin\psi\sin\alpha + \frac{\mathbf{p} \times [\mathbf{e} \times \mathbf{p}]}{p^2\sin\theta}p'\cos\psi\cos\alpha, \tag{2.2}$$

and vector **e**, result (2.1) is obtained directly without the assumption that $p' = p$.

The change in p' is especially large in ionizing interactions since the energy ε' of a primary electron not only decreases by ε_{ion}, but the remaining energy $\varepsilon' - \varepsilon_{ion}$

is divided between two free electrons. The allowance for the exact relation between angles θ' and θ is very important in the problem of multiplication of runaway electrons since the runaway energy threshold depends on the angle at which an electron moves relative to direction **e**. This relation determines whether both electrons (primary and secondary) are in the runway mode or only one of them will be a runway electron.

3. ELASTIC COLLISIONS

In the approximation of symmetry relative vector **e**, the EDF depends only on the modulus p of the momentum and angle θ between the directions of **p** and **e**. In this section, we denote by ε the total relativistic energy. The electron elastic scattering probability per unit time from the phase volume element $du'dV$ to element $du\,dV$, whose element $du = p^2 dp\,d\omega$ is seen from the coordinate origin in the momentum space at solid angle $d\omega$ (Fig. 2), is defined as

$$\begin{aligned} &T_{\mathrm{el}}(p', p, \psi, dp, d\omega) \\ &= N_{\mathrm{at}} v'(\sigma_{\mathrm{el}}(p', \psi)) d\omega \delta(\varphi_{\mathrm{el}}(\varepsilon, \varepsilon', \psi))\, v dp, \end{aligned} \tag{3.1}$$

where

$$d\sigma_{\mathrm{el}} = \sigma_{\mathrm{el}}(p', \psi) d\omega'$$

is the differential cross section of elastic scattering. Delta function $\delta(\varphi_{\mathrm{el}})$ takes into account the energy and momentum conservation laws:

$$\varphi_{\mathrm{el}}(\varepsilon, \varepsilon', \psi) = \varepsilon - g_{\mathrm{el}}(\varepsilon', \psi) = 0. \tag{3.2}$$

The formulas describing the scattering of a particle of mass m by an immobile particle of mass M in the laboratory reference frame [14] can be used to derive the exact expression for the energy transferred to the second particle. This expression is simplified for $m \ll M$ as follows:

$$\varepsilon' - \varepsilon = p'^2 (1 - \xi) \frac{(1 + \xi)\varepsilon' + M}{(\varepsilon' + M)^2 - \xi^2 p'^2}. \tag{3.3}$$

Here, $\xi = \cos\psi$ and the energy is measured in units of the electron rest energy mc^2. Since $\xi^2 m^2 \ll M^2$, we can replace p'^2 in the denominator by ε'^2 and simplify relation (3.3) so that law (3.2) can be written explicitly in the form

$$\varepsilon = g_{\mathrm{el}}(\varepsilon', \psi) = \varepsilon' - \frac{(\varepsilon'^2 - 1)\frac{m}{M}(1 - \xi)}{1 + \varepsilon' \frac{m}{M}(1 - \xi)}. \tag{3.4}$$

The total number of transitions to $du\,dV$ from all other elements $du'dV$ is given by

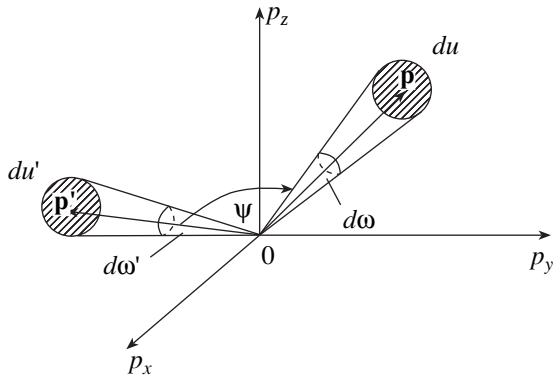


Fig. 2. Scattering geometry: du' and du are the volume elements in the momentum space before and after the interaction, $\mathbf{p}'$ and $\mathbf{p}$ are the electron momenta before and after the interaction, and ψ is the scattering angle.

$$\begin{aligned} &\int f(\mathbf{p}', t) T_{\mathrm{el}}(p', p, \psi, dp, d\omega) du' dV \\ &= dV \int f(p', \mu', t) T_{\mathrm{el}}(p', p, \psi, dp, d\omega) \frac{du'}{du} du \\ &= dV du N_{\mathrm{at}} v \int f(p', \mu', t)(\sigma_{\mathrm{el}}(p', \psi))(p'/p)^2 \\ &\quad \times \delta(\varphi_{\mathrm{el}}(\varepsilon, \varepsilon', \psi))\, v' d\omega' dp'. \end{aligned} \tag{3.5}$$

Here, $\mu' = \cos\theta'$. Function $\delta(\varphi_{\mathrm{el}})$ satisfies the formal relation [14]

$$\begin{aligned} \delta(\varphi_{\mathrm{el}}(\varepsilon, \varepsilon', \psi)) &= \frac{\delta(p' - p_1)}{\left|\frac{\partial g_{\mathrm{el}}(\varepsilon', \psi)}{\partial \varepsilon'} \frac{d\varepsilon'}{dp'}\right|_{p' = p_1}} \\ &= \frac{\delta(p' - p_1)}{\left|\frac{\partial g_{\mathrm{el}}(\varepsilon', \psi)}{\partial \varepsilon'} v'\right|_{p' = p_1}}, \end{aligned} \tag{3.6}$$

where p_1 is the solution to Eq. (3.2). Evaluating the derivative of g_{el}, we obtain the expression

$$\delta(\varphi_{\mathrm{el}}) = \frac{\delta(p' - p_1)}{(p/p')^2 v'}, \tag{3.7}$$

which makes it possible to integrate in Eq. (3.5) with respect to momentum modulus p',

$$\begin{aligned} &dV\, du\, N_{\mathrm{at}} v \int f(p', \mu', t)(\sigma_{\mathrm{el}}(p', \psi))(p'/p)^2 \\ &\quad \times \frac{\delta(p' - p_1)}{(p/p')^2 v'} v' d\omega' dp' \\ &= dV\, du\, N_{\mathrm{at}} v \int_{\omega'} f(p', \mu', t)(\sigma_{\mathrm{el}}(p', \psi))(p'/p)^4 d\omega', \end{aligned} \tag{3.8}$$

where $p' = p_1$. Using relation (3.3), we obtain a cumbersome expression for the derivative of g_{el}. A disadvan-

tage of relation (3.4), which is insignificant as long as $\varepsilon' \ll M$, is the independence of the condition for the smallness of the transferred energy,

$$\frac{\varepsilon' - \varepsilon}{\varepsilon' - m} \ll 1,$$

on the energy itself:

$$(m/M)(1-\xi) \ll 1.$$

Subtracting from relation (3.8) the total number of transitions from $du\,dV$ to other elements, we obtain the following expression for the elastic collision operator:

$$\mathrm{St}_{\mathrm{el}} = N_{\mathrm{at}} v \int [f(p', \mu', t)(p'/p)^4 (\sigma_{\mathrm{el}}(p', \psi)) - f(p, \mu, t)(\sigma_{\mathrm{el}}(p, \psi))]\, d\omega'. \tag{3.9}$$

Taking into account the smallness of $\Delta p = p' - p$ as compared to p', we expand the part of the integrand in expression (3.9) responsible for the population of element $du\,dV$, into a series:

$$f(p', \mu', t)(p'/p)^4 (\sigma_{\mathrm{el}}(p', \psi)) = f(p, \mu', t)(\sigma_{\mathrm{el}}(p, \psi)) + \frac{1}{p^4}\left[\frac{\partial}{\partial p'} f(p', \mu', t)(p'^4 (\sigma_{\mathrm{el}}(p', \psi)))\right]_{p' = p} \Delta p. \tag{3.10}$$

Expressing Δp in terms of $\Delta\varepsilon(\psi) = v\Delta p$ and then, in accordance with relation (3.4), $\Delta\varepsilon(\psi)$ in terms of ε,

$$\Delta\varepsilon(\psi) = \frac{(\varepsilon^2 - m^2c^4)\dfrac{m}{M}(1-\xi)}{mc^2 - \varepsilon\dfrac{m}{M}(1-\xi)}, \tag{3.11}$$

we transform relation (3.9) as follows:

$$\mathrm{St}_{\mathrm{el}} = N_{\mathrm{at}} v \int_{\omega'} (f(p, \mu', t) - f(p, \mu, t))(\sigma_{\mathrm{el}}(p, \psi))\, d\omega' + \frac{N_{\mathrm{at}}}{p^4} \int_{\omega'} \frac{(\varepsilon^2 - m^2c^4)\dfrac{m}{M}(1-\xi)}{mc^2 - \varepsilon\dfrac{m}{M}(1-\xi)} \times \frac{\partial}{\partial p} f(p, \mu', t) p^4 (\sigma_{\mathrm{el}}(p', \psi))\, d\omega'. \tag{3.12}$$

In the coordinates depicted in Fig. 1, an element of the solid angle is given by [13]

$$d\omega' = \sin\psi\, d\psi\, d\alpha = -d\xi\, d\alpha. \tag{3.13}$$

In the nonrelativistic limit, relation (3.12) can be reduced to formula (11) from [13].

Further, we will use the conventional procedure developed for high energies, when the variation of angle θ in a single collision event is small. Expanding the right-hand side of operator (3.12) into a series in $\Delta\mu$, we obtain the following diffusion approximation:

$$\mathrm{St}_{\mathrm{el}(1)} = N_{\mathrm{at}} v \int_{-1}^{1} d\xi \int_{0}^{2\pi} d\alpha \times \left[\frac{\partial f}{d\mu}\Delta\mu + \frac{\partial^2 f}{\partial \mu^2}\frac{(\Delta\mu)^2}{2}\right]\sigma_{\mathrm{el}}(p, \psi). \tag{3.14}$$

Expanding the left-hand side of relation (2.1) into a series, we reduce it to the form

$$\Delta\mu = -\mu(1-\xi) + \sqrt{1-\mu^2}\sqrt{1-\xi^2}\cos\alpha. \tag{3.15}$$

It can be seen that small values of $\Delta\mu$ are realized for $\xi \longrightarrow 1$. The quadratic form of relation (3.15) is as follows:

$$(\Delta\mu)^2 = \mu^2(1-\xi)^2 - 2\mu(1-\xi)\sqrt{1-\mu^2}\sqrt{1-\xi^2}\cos\alpha + (1-\mu^2)(1-\xi^2)\cos^2\alpha. \tag{3.16}$$

Since $f(\mu, p, t)$ and $\sigma_{\mathrm{el}}(p, \psi)$ are independent of $\alpha \in [0, 2\pi]$, we can integrate relations (3.15) and (3.16) with respect to this variable:

$$\int_0^{2\pi} \Delta\mu\, d\alpha = -2\pi\mu(1-\xi), \tag{3.17}$$

$$\int_0^{2\pi} (\Delta\mu)^2 d\alpha = 2\pi\mu^2(1-\xi^2) + \pi(1-\mu^2)(1-\xi^2) \approx 2\pi[\mu^2(1-\xi)^2 + (1-\mu^2)(1-\xi)]. \tag{3.18}$$

Substituting these relations into operator (3.14), we obtain

$$\mathrm{St}_{\mathrm{el}(1)} = N_{\mathrm{at}}\frac{v}{2}\left[\sigma_{\mathrm{tr}}(p)\hat{L}_\mu f + \sigma(p)\mu^2\frac{\partial^2 f}{\partial\mu^2}\right], \tag{3.19}$$

where

$$\sigma_{\mathrm{tr}}(p) = 2\pi\int_{-1}^{1}(1-\xi)\sigma_{\mathrm{el}}(p, \xi)\, d\xi, \tag{3.20}$$

$$\sigma(p) = 2\pi\int_{-1}^{1}(1-\xi)^2\sigma_{\mathrm{el}}(p, \xi)\, d\xi. \tag{3.21}$$

Since transport cross section (3.20) is much larger than cross section (3.21) because $\xi \longrightarrow 1$, we can disregard the second term in Eq. (3.19), which makes it possible

to satisfy the condition of conservation of the number of particles in elastic collisions:

$$\int_{-1}^{1} \mathrm{St}_{\mathrm{el}(1)} d\mu = 0. \tag{3.22}$$

We continue to expand the second term in relation (3.12) in $\Delta\mu$, neglecting the terms quadratic in $(m/M)(1-\xi)\varepsilon$ as compared to mc^2 and taking into account relation (3.17):

$$\mathrm{St}_{\mathrm{el}(2)} \approx \frac{N_{\mathrm{at}}}{p^4}\frac{m}{M}\frac{p^2}{m}\frac{\partial}{\partial p}p^4 \times\left[f(\mu, p, t)\sigma_{\mathrm{tr}}(p) - \mu\frac{\partial f}{\partial \mu}\sigma(p)\right]. \tag{3.23}$$

This operator satisfies the condition of conservation of the number of electrons, so that

$$\int_{0}^{\infty} \mathrm{St}_{\mathrm{el}(2)} p^2 dp = 0 \tag{3.24}$$

due to the fact that $f(p, \mu, t) \longrightarrow 0$ for $p \longrightarrow \infty$. Since values of $\xi \sim 1$ dominate, we disregard the second term in relation (3.23) and obtain the following expression for the total operator of elastic collisions:

$$\mathrm{St}_{\mathrm{el}} \approx N_{\mathrm{at}}\left[\frac{v}{2}\sigma_{\mathrm{tr}}(p)\hat{L}_\mu f(\mu, p, t) + \frac{1}{p^4}\frac{m}{M}\frac{p^2}{m}\frac{\partial}{\partial p}p^4\sigma_{\mathrm{tr}}(p) f(\mu, p, t)\right]. \tag{3.25}$$

4. INELASTIC INTERACTIONS (BOUND-BOUND TRANSITIONS)

Let us derive operator $\mathrm{St}_{\mathrm{ex}}$ responsible for excitation of atomic particles to state (i) with excitation energy $\varepsilon_{\mathrm{ex}}^{(i)}$. In this case, the energy conservation law can be written in the form

$$\varphi_{\mathrm{ex}}^{(i)}(\varepsilon, \varepsilon', \psi) = \varepsilon - g_{\mathrm{ex}}^{(i)}(\varepsilon', \psi) = 0, \tag{4.1}$$

where

$$g_{\mathrm{ex}}^{(i)}(\varepsilon', \psi) = \varepsilon' - \varepsilon_{\mathrm{ex}}^{(i)}. \tag{4.2}$$

Analogously to relation (3.1), the probability of scattering per unit time from $du'dV$ to $dudV$ is given by

$$T_{\mathrm{ex}}^{(i)}(p', p, \psi, dp, d\omega) = N_{\mathrm{at}} v'\sigma_{\mathrm{ex}}^{(i)}(p', \psi)d\omega\delta(\varphi_{\mathrm{ex}}^{(i)}(\varepsilon, \varepsilon', \psi))d\varepsilon, \tag{4.3}$$

where

$$d\sigma_{\mathrm{ex}}^{(i)} = \sigma_{\mathrm{ex}}^{(i)}(p', \psi)d\omega'$$

is the differential cross section of excitation of state (i).

The total number of transitions to $dudV$ from all other elements $du'dV$ is

$$\sum_i \int f(\mathbf{p}', t)T_{\mathrm{ex}}^{(i)} du'dV = dV\sum_i \int f(\mathbf{p}', t)T_{\mathrm{ex}}^{(i)}\frac{du'}{du}du$$

$$= dVdu N_{\mathrm{at}}\sum_i \int f(p', \mu', t)v'\sigma_{\mathrm{ex}}^{(i)}(p', \psi)d\omega \times \delta(\varphi_{\mathrm{ex}}^{(i)})vdp\frac{p'^2 dp' d\omega'}{p^2 dp d\omega}. \tag{4.4}$$

Since $\partial g_{\mathrm{ex}}^{(i)}/\partial\varepsilon' = 1$, the analog of relation (3.7) has the form

$$\delta(\varphi_{\mathrm{ex}}^{(i)}(\varepsilon, \varepsilon', \psi)) = \frac{\delta(p' - p_i)}{v'}, \tag{4.5}$$

where p_i is the solution to Eq. (4.1), and relation (4.4) can be reduced to the integral

$$dVd\gamma N_{\mathrm{at}} v \times \sum_i \int f(p', \mu', t)\sigma_{\mathrm{ex}}^{(i)}(p', \psi)(p'/p)^2 d\omega', \tag{4.6}$$

in which $p' = p_i$. Subtracting the number of transitions from $dudV$ and expanding into a power series in $\Delta p = \Delta\varepsilon/v = \varepsilon_{\mathrm{ex}}^{(i)}/v$, we obtain the following expression for $\mathrm{St}_{\mathrm{ex}}$:

$$\mathrm{St}_{\mathrm{ex}} = N_{\mathrm{at}} v\sum_i \int [f(p, \mu', t)\sigma_{\mathrm{ex}}^{(i)}(p', \psi)(p'/p)^2 - f(p, \mu, t)\sigma_{\mathrm{ex}}^{(i)}(p, \psi)]d\omega'$$

$$= N_{\mathrm{at}} v\sum_i \left\{ \int [f(p, \mu', t) - f(p, \mu, t)]\sigma_{\mathrm{ex}}^{(i)}(p, \psi)d\omega' + \frac{\Delta p}{p^2}\frac{\partial}{\partial p}p^2 f(p, \mu', t)\sigma_{\mathrm{ex}}^{(i)}(p, \psi)d\omega' \right\}. \tag{4.7}$$

Repeating procedure (3.14)–(3.18) for the first part of $\mathrm{St}_{\mathrm{ex}}$, we obtain

$$\mathrm{St}_{\mathrm{ex}(1)} = N_{\mathrm{at}}\frac{v}{2}\sigma_{\mathrm{ex,\,tr}}(p)L_\mu f(p, \mu, t), \tag{4.8}$$

where we have omitted the term quadratic in $(1-\xi)$ and used the following notation:

$$\sigma_{\mathrm{ex,\,tr}}(p) = \sum_i \sigma_{\mathrm{ex,\,tr}}^{(i)}(p) = \sum_i 2\pi\int_{-1}^{1} d\xi(1-\xi)\sigma_{\mathrm{ex}}^{(i)}(p, \xi). \tag{4.9}$$

Analogously to relation (3.23), taking into account the relation

$$v\Delta p = \Delta\varepsilon = \varepsilon_{ex}^{(i)},$$

we obtain for the second part of St_{ex}

$$St_{ex(2)} = \frac{1}{p^2}\frac{\partial}{\partial p}p^2 \times\left[f(p, \mu, t)F_{ex(1)}(p) - F_{ex(2)}(p)\mu\frac{\partial f}{\partial \mu}\right]. \quad (4.10)$$

Here, we introduced the friction forces responsible for inelastic collisions (without ionization):

$$F_{ex(1)} = N_{at}\sum_i \varepsilon_{ex}^{(i)}\sigma_{ex}^{(i)}(p), \quad (4.11)$$

$$\sigma_{ex}^{(i)}(p) = 2\pi\int_{-1}^{1}\sigma_{ex}^{(i)}(p, \xi)d\xi, \quad (4.12)$$

$$F_{ex(2)} = N_{at}\sum_i \varepsilon_{ex}^{(i)}\sigma_{ex,tr}^{(i)}(p). \quad (4.13)$$

Operators (4.8) and (4.10) satisfy the condition of conservation of the number of electrons. Since the scattering through small angles dominates ($\xi \longrightarrow 1$), we have $\sigma_{ex,\,tr}^{(i)}(p) \ll \sigma_{ex}^{(i)}(p)$ and, hence, the second term in relation (4.10) can be neglected. The term proportional to $F_{ex(1)}$ is a part of the term describing small variations in the electron momentum (see Section 7). In the RREA problem, in the atmosphere of the Earth (small values of Z), St_{ex} is superfluous since atomic electrons can be regarded as free in view of the smallness of the binding energy (and, the more so, $\varepsilon_{ex}^{(i)}$) as compared to the energy ε' of impinging electrons. Bound–bound transitions slightly add to the contribution from the collisions transferring atomic electrons to the continuum. Operator St_{ex} can be used in problems in which the binding energy is comparable to ε'.

5. IONIZING COLLISIONS

The probability that, as a result of a collision ionizing shell (i), the electron is moved per unit time from element $du'dV$ in the vicinity of kinetic energy $\varepsilon' = mc^2(\gamma' - 1)$ to element $dudV$ in the vicinity of $\varepsilon = mc^2(\gamma - 1)$ is

$$T_{ion}^{(i)}(\varepsilon', \varepsilon, \psi, d\varepsilon, d\omega) = N_{at}v'(\sigma_{\varepsilon',\,\omega'}(\varepsilon', \varepsilon, \psi))_{ion}^{(i)}d\omega\delta(\varphi(\varepsilon', \varepsilon, \psi))d\varepsilon, \quad (5.1)$$

where

$$d\sigma_{ion}^{(i)}(\varepsilon', \varepsilon, \psi) = (\sigma_{\varepsilon',\,\omega'}(\varepsilon', \varepsilon, \psi))_{ion}^{(i)}d\varepsilon'd\omega' = (\sigma_{\varepsilon'}(\varepsilon', \varepsilon))_{ion}^{(i)}(\delta(\varphi)/2\pi)d\varepsilon'd\omega'$$

is doubly differential (with respect to energy and angle) ionization cross section of shell (i); $\varepsilon_{ion}^{(i)}$ is the ionization threshold of shell (i); and $(\sigma_{\varepsilon'}(\varepsilon', \varepsilon))_{ion}^{(i)}$ is the differential ionization cross section, which is symmetric relative to the secondary electron energy $\varepsilon_s = (\varepsilon' - \varepsilon_{ion}^{(i)})/2$. The energy and momentum conservation law can be written in the form

$$\varphi(\varepsilon', \varepsilon, \psi) = \cos\psi - \mu_0(\varepsilon', \varepsilon) = 0. \quad (5.2)$$

For $\mu_0(\varepsilon', \varepsilon)$, in the approximation $\varepsilon, \varepsilon' \gg \varepsilon_{ion}^{(i)}$, the following expression is valid [5, 14]:

$$\mu_0^2(\varepsilon', \varepsilon) = \frac{\varepsilon(\varepsilon' + 2mc^2)}{\varepsilon'(\varepsilon + 2mc^2)}. \quad (5.3)$$

The total number of electron transitions in ionizing collisions populating per unit time the phase volume element $dudV$ in the vicinity of energy ε is given by

$$N_{at}dV\sum_i\int f(p', \mu', t)v'(\sigma_{\varepsilon'}(\varepsilon', \varepsilon))_{ion}^{(i)}d\omega\delta(\varphi)d\varepsilon du' = dVduN_{at}v\sum_i\int_{\varepsilon + \varepsilon_{ion}^{(i)}}^{\infty}d\varepsilon'\int_{-1}^{1}(\sigma_{\varepsilon'}(\varepsilon', \varepsilon))_{ion}^{(i)}\frac{\gamma'^2 - 1}{\gamma^2 - 1}d\xi \times\int_0^{2\pi} f(p', \mu', t)\frac{\delta(\xi - \mu_0)}{2\pi}d\alpha. \quad (5.4)$$

Summation is carried out over all shells. Integrating with respect to ξ, we obtain the same operator as in relation (1.4), but with a different lower limit of integration with respect to ε':

$$St_{ion(1)} = N_{at}v\sum_i\int_{\varepsilon + \varepsilon_{ion}^{(i)}}^{\infty}d\varepsilon'(\sigma_{\varepsilon'}(\varepsilon', \varepsilon))_{ion}^{(i)} \times\frac{\gamma'^2 - 1}{\gamma^2 - 1}\int_0^{2\pi}d\alpha\frac{f(p', \mu', t)}{2\pi}, \quad (5.5)$$

where, in accordance with relations (2.1), (5.2), and (5.3), we have

$$\mu' = \mu\mu_0 + \sqrt{1 - \mu_0^2}\sqrt{1 - \mu^2}\cos\alpha. \quad (5.6)$$

The operator describing the departure of electrons from *du* has the form

$$\mathrm{St}_{\mathrm{ion}(2)} = N_{\mathrm{at}} v f(p, \mu, t) \sum_i \int_0^{(\varepsilon - \varepsilon_{\mathrm{ion}}^{(i)})/2} (\sigma_\varepsilon(\varepsilon, \varepsilon'))_{\mathrm{ion}}^{(i)} d\varepsilon' = N_{\mathrm{at}} v f(p, \mu, t) \sum_i \sigma_{\mathrm{tot}}^{(i)}(\varepsilon), \quad (5.7)$$

where

$$\sigma_{\mathrm{tot}}^{(i)}(\varepsilon) = \int_0^{(\varepsilon - \varepsilon_{\mathrm{ion}}^{(i)})/2} (\sigma_\varepsilon(\varepsilon, \varepsilon'))_{\mathrm{ion}}^{(i)} d\varepsilon' \quad (5.8)$$

is the total ionization cross section of shell (*i*). In expressions (5.5) and (5.7), ε' and ε are transposed in accordance with the fact that $\mathrm{St}_{\mathrm{ion}(1)}$ is responsible for populating element *du* and $\mathrm{St}_{\mathrm{ion}(2)}$ is responsible for the departure of electrons from *du*.

6. WEAK IONIZING COLLISIONS

In relation (5.5), we single out the "weak" interactions in which the primary electron participating in an ionization event preserves a large part of its energy; i.e.,

$$\varepsilon' \approx \varepsilon. \quad (6.1)$$

We will take advantage of the symmetry of $(\sigma(\varepsilon', \varepsilon_s))_{\mathrm{ion}}^{(i)}$ with respect to the secondary electron energy $\varepsilon_s = (\varepsilon' - \varepsilon_{\mathrm{ion}})/2$ (Fig. 3) [6]. Since electrons are indistinguishable, we assume, for convenience, that the secondary electron is the one possessing the lower kinetic energy,

$$\varepsilon_s \le (\varepsilon' - \varepsilon_{\mathrm{ion}})/2. \quad (6.2)$$

Since the energy range $\varepsilon_s \ll (\varepsilon' - \varepsilon_{\mathrm{ion}})/2$ dominates in $(\sigma(\varepsilon', \varepsilon_s))_{\mathrm{ion}}^{(i)}$, expressions (6.1) and (6.2) can be regarded as compatible. Since

$$\varepsilon' = \varepsilon + \varepsilon_s + \varepsilon_{\mathrm{ion}} = \varepsilon + \Delta\varepsilon, \quad (6.3)$$

the following inequality holds for "weak" interactions:

$$\Delta\varepsilon = \varepsilon_s + \varepsilon_{\mathrm{ion}} \le \frac{\varepsilon' - \varepsilon_{\mathrm{ion}}}{2} + \varepsilon_{\mathrm{ion}} = \frac{\varepsilon' + \varepsilon_{\mathrm{ion}}}{2}. \quad (6.4)$$

Substituting this relation into formula (6.3), we obtain for weak interactions the formula (cf. (1.4))

$$\varepsilon' \le 2\varepsilon + \varepsilon_{\mathrm{ion}}. \quad (6.5)$$

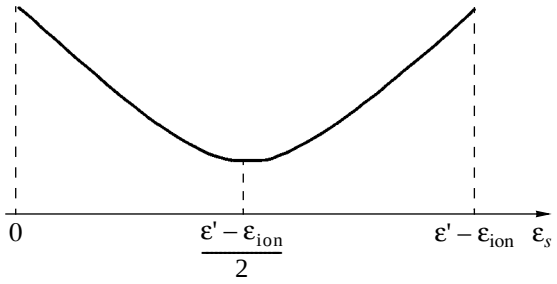


Fig. 3. Dependence of the differential ionization cross section on the secondary electron energy.

Repeating the procedure used in the derivation of elastic collision operator (3.19) and taking into account the fact that, in accordance with relation (5.2), $\xi = \cos\psi = \mu_0$, we single out from expression (5.5) the operator

$$\mathrm{St}_{\mathrm{ion}(1)}^{(\mathrm{weak})} = N_{\mathrm{at}} v \sum_i \int_{\varepsilon + \varepsilon_{\mathrm{ion}}^{(i)}}^{2\varepsilon + \varepsilon_{\mathrm{ion}}^{(i)}} d\varepsilon' (\sigma(\varepsilon', \varepsilon))_{\mathrm{ion}}^{(i)} \left(\frac{p'}{p}\right)^2 \times \left[f(p', \mu, t) + \frac{1}{2}(1 - \mu_0)\hat{L}_\mu f(p, \mu, t) + (1 - \mu_0)^2 \frac{\mu^2}{2} \frac{\partial^2 f}{\partial \mu^2} \right], \quad (6.6)$$

where the second and third terms in the brackets are analogous to the terms of operator (3.19), where integration with respect to ξ is carried out in analogs of relations (3.20) and (3.21).

Let us calculate using relation (5.3) the value of $1 - \mu_0$ in the approximation of small values of $\Delta\varepsilon$. Here, we take into account for the first time the smallness of the binding energy of atomic electrons since relation (5.3) was derived precisely in this approximation:

$$1 - \mu_0 \approx 1 - \sqrt{\mu_0^2(\varepsilon', \varepsilon)} \approx 1 - \left(1 + \frac{\partial \mu_0^2 / \partial \varepsilon'}{2\sqrt{\mu_0^2}} \Delta\varepsilon\right)_{\varepsilon' = \varepsilon} = \frac{2mc^2}{(\varepsilon + 2mc^2)\varepsilon} \frac{\Delta\varepsilon}{2} = \frac{m\Delta\varepsilon}{p^2}. \quad (6.7)$$

Replacing ε by ε_s in $(\sigma(\varepsilon', \varepsilon))_{\mathrm{ion}}^{(i)}$, we expand the integrand in expression (6.6) into a power series in $\Delta\varepsilon$ and replace the integration with respect to ε' by integration with over ε_s, substituting, by virtue of relation (6.1), ε for ε' in the upper limit of integration with respect to ε_s, which is equal to (6.2). As a result, we

arrive at the differential representation of the operator of weak ionizing collisions:

$$\mathrm{St}^{(\mathrm{weak})}_{\mathrm{ion}(1)} = N_{\mathrm{at}} v \sum_i \Bigg\{ f(p, \mu, t) \int_0^{(\varepsilon - \varepsilon^{(i)}_{\mathrm{ion}})/2} (\sigma(\varepsilon, \varepsilon_s))^{(i)}_{\mathrm{ion}} d\varepsilon_s$$

$$+ \frac{1}{p^2}\frac{\partial}{\partial \varepsilon} p^2 f(p, \mu, t) \int_0^{(\varepsilon - \varepsilon^{(i)}_{\mathrm{ion}})/2} \Delta\varepsilon^{(i)} (\sigma(\varepsilon, \varepsilon_s))^{(i)}_{\mathrm{ion}} d\varepsilon_s \quad (6.8)$$

$$+ \frac{m}{2p^2} \hat{L}_\mu f(p, \mu, t) \int_0^{(\varepsilon - \varepsilon^{(i)}_{\mathrm{ion}})/2} \Delta\varepsilon^{(i)} (\sigma(\varepsilon, \varepsilon_s))^{(i)}_{\mathrm{ion}} d\varepsilon_s \Bigg\}.$$

Using effective drag κ for "large" momentum transfer [6], we introduce the friction force acting on electrons as a result of weak ionizing interactions,

$$F_{\mathrm{ion}}(\varepsilon) = N_{\mathrm{at}} \kappa$$
$$= N_{\mathrm{at}} \sum_i \int_0^{(\varepsilon - \varepsilon^{(i)}_{\mathrm{ion}})/2} \Delta\varepsilon^{(i)} (\sigma(\varepsilon, \varepsilon_s))^{(i)}_{\mathrm{ion}} d\varepsilon_s, \quad (6.9)$$

which enables us to rewrite relation (6.8) in a more compact form,

$$\mathrm{St}^{(\mathrm{weak})}_{\mathrm{ion}(1)} = N_{\mathrm{at}} v \sigma_{\mathrm{tot}}(\varepsilon) f(p, \mu, t)$$
$$+ \frac{1}{p^2}\frac{\partial}{\partial p} p^2 F_{\mathrm{ion}}(\varepsilon) f(p, \mu, t) + \frac{F_{\mathrm{ion}}(\varepsilon)}{2\gamma p} \hat{L}_\mu f(p, \mu, t), \quad (6.10)$$

where the total ionization cross section is the sum

$$\sigma_{\mathrm{tot}}(\varepsilon) = \sum_i \sigma^{(i)}_{\mathrm{tot}}(\varepsilon). \quad (6.11)$$

It can be seen that the first term in relation (6.10) is completely compensated by integral (5.7) $\mathrm{St}_{\mathrm{ion}(2)}$ responsible for the departure of electrons from the considered element du.

Result (6.10) can be obtained by factorizing the ionization cross section in relation (6.6),

$$(\sigma(\varepsilon', \varepsilon))^{(i)}_{\mathrm{ion}} = \sigma^{(i)}_{\mathrm{tot}}(\varepsilon') \chi^{(i)}(\varepsilon', \varepsilon_s), \quad (6.12)$$

bearing in mind its symmetry relative to ε and ε_s and taking into account the normalization

$$\int_0^{\varepsilon - \varepsilon^{(i)}_{\mathrm{ion}}} \chi^{(i)}(\varepsilon', \varepsilon_s) d\varepsilon_s = 1. \quad (6.13)$$

A rigorous reduction procedure has made it possible to automatically single out from the ionization integral the differential component responsible for angular scattering by atomic electrons. The absence of this component in [3, 4] is due to the fact that the process was considered on a plane and the EDF $f(p', \mu', t)$ was fixed for two values of μ'; as a result, an opportunity for expanding into a series in μ' was missed.

Operator $\mathrm{St}^{(\mathrm{strong})}_{\mathrm{ion}(1)}$, which is a part of ionization integral (5.5) remaining after the replacement of the lower limit of integration over energies by $2\varepsilon + \varepsilon^{(i)}_{\mathrm{ion}}$ as in [3, 4] (see formula (1.4) in the Introduction), is responsible for strong collisions.

7. DESCRIPTION OF INTERACTIONS OF ELECTRONS WITH ATOMIC PARTICLES

The operators obtained in the previous sections have a large range of application. In this section, we give arguments enabling us to derive simpler and convenient for numerical calculations representations of operators for electron energies considerably exceeding the ionization energies for the electron shells of atoms.

To avoid errors in calculations, it is appropriate to make the following remark. The literature data for cross sections should be used bearing in mind that these data are frequently given in a form integrated over angle α (i.e., contain the factor 2π), while in the formulas derived here (e.g., (3.14), (3.23), (4.7), and (5.4)), the integration with respect to α is carried out as well.

7.1. Elastic Interactions

Using the Rutherford formula with the Mott factor for $\sigma_{\mathrm{el}}(p, \psi)$ [6, 15], we obtain the following expression for transport cross section (3.20):

$$\sigma_{\mathrm{tr}}(\gamma) = 2\pi \frac{Z^2_{\mathrm{at}} e^4 \gamma^2}{(mc^2)^4 (\gamma^2 - 1)^2} \left(2\ln\frac{2}{\psi_{\min}} - \beta^2 \right). \quad (7.1)$$

Here, $\psi_{\min} = 0.0153 Z^{1/3}_{\mathrm{at}}/\beta\gamma$ in accordance with the Thomas–Fermi model [15, 16] and integration with respect to α is performed. In this approximation, we obtain the following dependence of the reciprocal transport length on γ or $\varepsilon = \gamma mc^2$:

$$N_{\mathrm{at}} \sigma_{\mathrm{tr}}(\gamma) = \frac{4\pi N_{\mathrm{at}} Z^2_{\mathrm{at}} e^4 \gamma^2}{(mc^2)^2 (\gamma^2 - 1)^2} \left(\ln\frac{131\gamma\beta}{Z^{1/3}_{\mathrm{at}}} - \frac{\beta^2}{2} \right). \quad (7.2)$$

In accordance with formula (1.6), expression (7.2) for a gas of molecules consisting of two identical atoms ($n = 2$) can be written in terms of molecular quantities:

$$N_{\mathrm{at}} Z^2_{\mathrm{at}} = 2 N_{\mathrm{mol}} (Z_{\mathrm{mol}}/2)^2 = N_{\mathrm{mol}} Z^2_{\mathrm{mol}}/2. \quad (7.3)$$

7.2. Ionization (Bound-Free Transitions)

If electron energies ε' and ε before and after the interaction are considerably higher than the binding energy of atomic electrons, it is expedient, following [3, 4], to use for the ionization cross section $(\sigma_{\varepsilon'}(\varepsilon', \varepsilon))^{(i)}_{\text{ion}}$ the Moeller formula for the cross section of electron scattering by free electron, which was initially at rest [5, 6]. The total differential ionization cross section can be obtained by multiplying this formula by Z_{mol}:

$$\sigma_{\varepsilon'}(\varepsilon', \varepsilon)_{\text{ion}} = \frac{2\pi Z_{\text{mol}} e^4}{mc^2}\frac{\gamma'^2}{\gamma'^2-1}\left[\frac{1}{\varepsilon^2} - \frac{1}{\varepsilon(\varepsilon'-\varepsilon)} \times \frac{(2\varepsilon'+mc^2)mc^2}{(\varepsilon'+mc^2)^2} + \frac{1}{(\varepsilon'-\varepsilon)^2} + \frac{1}{(\varepsilon'+mc^2)^2}\right]. \quad (7.4)$$

In Section 4, we introduced the force $F_{\text{ex}(1)}(p)$ originating from excitation processes (see formula (4.11)). In reduced operator (6.10), we introduced the force $F_{\text{ion}}(p)$ associated with weak ionizing interactions. We will express these forces via the well-known formulas of quantum electrodynamics. Total inelastic ("ionization" in the historically adopted terminology) specific losses of electron energy in the medium of particles with the mean "ionization" energy I can be calculated in terms of effective drag κ, including the range of small momentum transfer (in particular, including transfer as a result of bound–bound transitions) and the range of large momentum transfer. In the range of small momentum transfer, which is defined by the inequality [6]

$$q^2/m = (p'-p)^2/m \ll mc^2,$$

the value of κ is calculated as follows (see, for example, formula (82.20) in monograph [6]):

$$\kappa^{(\text{small})} = \frac{2\pi Z_{\text{mol}} e^4}{mc^2}\frac{\gamma^2}{\gamma^2-1} \times \ln\left(q\frac{m^2c^4(\gamma^2-1)}{2.73I^2} + \frac{1}{\gamma^2}\right). \quad (7.5)$$

This expression includes the processes of excitation and ionization. The product $N_{\text{mol}}\kappa^{(\text{small})}$ gives the sum of the entire force $F_{\text{ex}}(p)$ and a part of force $F_{\text{ion}}(p)$. In the range of large momentum transfer, which is defined by the inequality

$$q^2/m = (p'-p)^2/m \gg I,$$

the value of κ is calculated as the integral of cross section (7.4) multiplied by energy ε_s transferred to the secondary electron (see formula (6.4), where $\Delta\varepsilon \approx \varepsilon_s$) [6],

$$\kappa^{(\text{large})} = \int_{\varepsilon_1}^{\varepsilon/2} \varepsilon_s(\sigma(\varepsilon, \varepsilon_s))_{\text{ion}} d\varepsilon_s, \quad (7.6)$$

where

$$\varepsilon_1 = q_1^2/m, \quad I \ll q_1^2/m \ll mc^2;$$

i.e., I is in the overlap region.

The total drag force appearing in the reduced operator (9.1) (see, for example, the problem on p. 383 in [6]),

$$F(p) = (\kappa^{(\text{small})} + \kappa^{(\text{large})})N_{\text{mol}} = F_{\text{ex}}(p) + F_{\text{ion}}(p) \quad (7.7)$$

is the Bethe drag force $F(\gamma)$ (see formula (1.2)) [5, 6]:

$$F(\gamma) = \frac{2\pi Z_{\text{mol}} e^4 N_{\text{mol}}}{mc^2}\frac{\gamma^2}{\gamma^2-1}\left[\ln\left(\frac{m^2c^4(\gamma^2-1)(\gamma-1)}{2I^2}\right) + \left(\frac{2}{\gamma} - \frac{1}{\gamma^2}\right)\ln 2 + \frac{1}{\gamma^2} + \frac{(\gamma-1)^2}{8\gamma^2}\right]. \quad (7.8)$$

Since formulas (7.4) and (7.8) were derived under the assumption that ε', ε, $\varepsilon_s \gg \varepsilon_{\text{ion}}$, quantity $\varepsilon^{(i)}_{\text{ion}}$ should be omitted in the lower limit of the ionization integral as well as the summation over index i. The inclusion of $\varepsilon^{(i)}_{\text{ion}}$ leads to an excessive accuracy, while the summation was in fact carried out when the Moeller cross section was multiplied by Z_{mol}.

7.3. Excitation Processes (Bound–Bound Transitions)

These processes are taken into account via force $F_{\text{ex}(1)}(p)$ and cross section $\sigma_{\text{ex, tr}}(p)$. Since $\sigma_{\text{el}}(p, \psi)$ and $\sigma_{\text{ex}}(p, \psi)$ are proportional to Z^2 and Z, respectively (in the range of small angles, both cross sections are described by the Rutherford formula [15]), quantity $\sigma_{\text{ex, tr}}(p)$ can be neglected for large values of Z. The quantity $F_{\text{ex}(1)}(p)$ appears in the first component of the operator with reduced ionization integral (9.1) and (9.2) via the total Bethe force $F(p)$ (see Section 9). It follows from expressions (9.3)–(9.6) that the component responsible for elastic energy losses is much smaller than $F(p)$:

$$\frac{m}{M}N_{\text{at}}\frac{p^2}{m}\sigma_{\text{tr}}(p)/F(\gamma) = \frac{m}{M}\frac{p^2 Z}{m\gamma v p}\Gamma(\gamma)F(\gamma)/F(\gamma) = \frac{Z_{\text{mol}}}{A_{\text{mol}}}\frac{m}{m_{\text{nucl}}}\Gamma(\gamma) \approx 10^{-4}. \quad (7.9)$$

Here, $Z_{mol}/A_{mol} \approx 1/2$ is the ratio of the number of electrons to the number of nucleons in a molecule, $m/m_{nucl} \approx 1/1840$ is the ratio of the electron and nucleon masses, and the following notation is used [16]:

$$\Gamma(\gamma) = \frac{\frac{(mc^2)^2(\gamma-1)(\gamma^2-1)}{2I^2} - \left(\frac{2}{\gamma} - \frac{1}{\gamma^2}\right)\ln 2 + \frac{1}{\gamma^2} + \frac{1}{8}\left(1-\frac{1}{\gamma}\right)^2}{\ln(131 Z_{at}^{-1/3}\sqrt{\gamma^2-1})}. \quad (7.10)$$

In accordance with calculations made in [16], Γ weakly depends on γ, increasing for air from 0.262 to 0.271 for electron energies in the range from 51 keV to 1.53 MeV. In the factor preceding operator $\hat{L}_\mu$ in the second component of Eq. (9.2), force $F_{ion}(p)$ can be replaced by the total Bethe force since the main contribution to this force for large values of Z comes from ionization processes (see the arguments on p. 732 in monograph [15]). In accordance with formula (9.5), in the second component of operator (9.2), we have

$$N_{at}\frac{v}{2}\sigma_{tr}(p) \gg \frac{F}{2\gamma p}, \quad (7.11)$$

i.e., angular scattering by nuclei dominates over the scattering from atomic electrons. By virtue of relation (7.11), the first component in the total operator with nonreduced ionization integral (8.2) cannot be disregarded (see below). In the RREA kinetics, this component is most important, elevating the runaway threshold for electrons and reducing the rate of avalanche enhancement [7, 8].

The operator in form (8.2) with the nonreduced ionization integral can be obtained from relation (8.1) due to the fact that the first term, the only one containing $F_{ex(1)}$, can be ignored in the high-energy range since

$$F_{ex(1)}(p) + \frac{m}{M}N_{at}\frac{p^2}{m}\sigma_{tr}(p) \ll F(p). \quad (7.12)$$

Indeed, it was mentioned above that excitation processes make a small contribution to the first component of reduced operator (9.2), which includes the total Bethe force. Since weak ionizing collisions for which this component is responsible are contained in the nonreduced ionization integral of total operator (8.1), this disregard, subject to relation (7.12), is justified with a high accuracy.

8. TOTAL COLLISION OPERATOR

Let us combine the parts of operator (3.25) for $St_{el}\{f\}$, (4.8) and (4.10) for $St_{ex}\{f\}$, and (5.5) and (5.7) for $St_{ion}\{f\}$, disregarding the terms quadratic in $1-\xi$. As a result, we obtain the following representation for the total collision operator:

$$\begin{aligned} &St\{f(p,\mu,t)\} \\ &= \frac{1}{p^2}\frac{\partial}{\partial p}p^2\left(F_{ex(1)}(p) + \frac{m}{M}N_{at}\frac{p^2}{m}\sigma_{tr}(p)\right)f(p,\mu,t) \\ &\quad + \left[N_{at}\frac{v}{2}(\sigma_{tr}(p) + \sigma_{ex,\,tr}(p))\right]\hat{L}_\mu f(p,\mu,t) \\ &\quad + N_{at}v\sum_i \int_{\varepsilon+\varepsilon_{ion}^{(i)}}^{\infty} d\varepsilon'(\sigma_{\varepsilon'}(\varepsilon',\varepsilon))_{ion}^{(i)}\frac{\gamma'^2-1}{\gamma^2-1} \\ &\quad \times \int_0^{2\pi}\frac{d\alpha}{2\pi}f(p',\mu',t) - N_{at}v\sigma_{tot}(\varepsilon)f(p,\mu,t). \end{aligned} \quad (8.1)$$

Here and below, $\sigma_{tot}(\varepsilon)$ is the total ionization cross section (6.11); $\cos\theta' = \mu'(\varepsilon', \varepsilon, \mu, \alpha)$ as a function of ε', ε, $\mu = \cos\theta$, and α is defined by formulas (5.3) and (5.6); and cross sections $\sigma_{tr}(p)$, $\sigma_{ex,\,tr}(p)$ and forces $F_{ex(1)}(p)$, $F_{ion}(p)$ are described by formulas (3.20), (4.9), (4.11), and (6.9), respectively.

The ionization integral in relation (8.1) was derived without any limitations except those imposed by the conservation laws. The accuracy of a description is determined by the differential cross section and dependence $\mu_0(\varepsilon', \varepsilon)$. If we use formula (5.3) for $\mu_0(\varepsilon', \varepsilon)$ and Moeller formula (7.4) for the cross section, the ionization integral is limited by the condition $\varepsilon', \varepsilon, \varepsilon_s \gg \varepsilon_{ion}^{(i)}$. To preserve the strictness in problems where the kinetics should be taken into account in the range of energies on the order of $\varepsilon_{ion}^{(i)}$, we must use the appropriate set of $((\sigma_{\varepsilon'}(\varepsilon', \varepsilon))_{ion}^{(i)}$ and the exact formula for $\mu_0(\varepsilon', \varepsilon)$ taking into account the coupling of atomic electrons.

The first component of relation (8.1) is not significant while describing the kinetics of high-energy electrons. We can also disregard $\sigma_{ex,\,tr}(p)$ in the second com-

ponent according to the arguments given in Section 7. As a result, the total operator

$$\begin{aligned}\mathrm{St}\{f(p,\mu,t)\} &= N_{\mathrm{at}}\frac{v}{2}\sigma_{\mathrm{tr}}(p)\hat{L}_\mu f(p,\mu,t)\\ &+ N_{\mathrm{at}}v\sum_i\int\limits_{\varepsilon+\varepsilon_{\mathrm{ion}}^{(i)}}^{\infty}d\varepsilon'(\sigma_{\varepsilon'}(\varepsilon',\varepsilon))_{\mathrm{ion}}^{(i)}\frac{\gamma'^2-1}{\gamma^2-1}\\ &\times\int_0^{2\pi}\frac{d\alpha}{2\pi}f(p',\mu',t)-N_{\mathrm{at}}v\sigma_{\mathrm{tot}}(\varepsilon)f(p,\mu,t)\end{aligned}\qquad(8.2)$$

turns out to be simpler than operator (1.2)–(1.4). In contrast to relations (1.2)–(1.4), where processes of angular scattering by nuclei and atomic electrons are artificially combined in factor $(Z_{\mathrm{mol}}/2+1)F(\varepsilon)/4\gamma p$, the scattering by nuclei is described in relation (8.2) by the transport cross section, while the scattering by atomic electrons remains in the ionization integral.

9. ANALYSIS OF THE TOTAL REDUCED COLLISION OPERATOR

If we use the reduced form of the ionization operator with extracted weak interactions, the total operator assumes the form

$$\begin{aligned}&\mathrm{St}[f(p,\mu,t)]\\ &= \frac{1}{p^2}\frac{\partial}{\partial p}p^2\left(F(p)+\frac{m}{M}N_{\mathrm{at}}\frac{p^2}{m}\sigma_{\mathrm{tr}}(p)\right)f(p,\mu,t)\\ &+\left[N_{\mathrm{at}}\frac{v}{2}(\sigma_{\mathrm{tr}}(p)+\sigma_{\mathrm{ex,\,tr}}(p))+\frac{F_{\mathrm{ion}}(p)}{2\gamma p}\right]\hat{L}_\mu f(p,\mu,t)\\ &+N_{\mathrm{at}}v\sum_i\int\limits_{2\varepsilon+\varepsilon_{\mathrm{ion}}^{(i)}}^{\infty}d\varepsilon'(\sigma_{\varepsilon'}(\varepsilon',\varepsilon))_{\mathrm{ion}}^{(i)}\frac{\gamma'^2-1}{\gamma^2-1}\int_0^{2\pi}\frac{d\alpha}{2\pi}f(p',\mu',t),\end{aligned}\qquad(9.1)$$

in which after the separation of weak interactions (6.6) from relation (5.5); the lower limit of integration in the remaining part of relation (5.5) describing strong interactions is equal to $2\varepsilon+\varepsilon_{\mathrm{ion}}^{(i)}$ (cf. relation (1.4)). In relation (9.1),

$$F(p) = F_{\mathrm{ion}}(p)+F_{\mathrm{ex}(1)}(p)$$

is the Bethe force.

Omitting elastic energy losses in the first component of operator (9.1) and $\sigma_{\mathrm{ex,\,tr}}(p)$ in the second component in accordance with the arguments given in Section 7, we obtain the operator

$$\begin{aligned}\mathrm{St}\{f(p,\mu,t)\} &= \frac{1}{p^2}\frac{\partial}{\partial p}p^2F(p)f(p,\mu,t)\\ &+\left(N_{\mathrm{at}}\frac{v}{2}\sigma_{\mathrm{tr}}(p)+\frac{F_{\mathrm{ion}}(p)}{2\gamma p}\right)\hat{L}_\mu f(p,\mu,t)\\ &+N_{\mathrm{at}}v\sum_i\int\limits_{2\varepsilon+\varepsilon_{\mathrm{ion}}^{(i)}}^{\infty}d\varepsilon'(\sigma_{\varepsilon'}(\varepsilon',\varepsilon))_{\mathrm{ion}}^{(i)}\frac{\gamma'^2-1}{\gamma^2-1}\int_0^{2\pi}\frac{d\alpha}{2\pi}f(p',\mu',t),\end{aligned}\qquad(9.2)$$

which differs from relations (1.2)–(1.4) only in the second component responsible for angular scattering. The same role is played by operator (1.3) combining the description of angular scattering by the nucleus and by atomic electrons in factor $(Z_{\mathrm{mol}}/2+1)F(p)/4\gamma p$ derived in [3, 4] using the results of analysis of Longmire and Longley [16]). It should be recalled that angular scattering is also contained in ionization integral (1.4) appearing in (9.2) as well.

For molecules consisting of two identical atoms (see relation (7.3)), quantity $(v/2)N_{\mathrm{at}}\sigma_{\mathrm{tr}}(\gamma)$ in relation (9.2) can be written, in accordance with the results obtained in [16], in terms of Bethe force $F(\gamma)$:

$$\begin{aligned}\frac{v}{2}N_{\mathrm{at}}\sigma_{\mathrm{tr}}(\gamma) &= \frac{v}{2}\frac{2(Z_{\mathrm{mol}}/2)}{mc^2(\gamma^2-1)}F(\gamma)\Gamma(\gamma)\\ &= \frac{Z_{\mathrm{mol}}}{2\gamma p}F(\gamma)\Gamma(\gamma).\end{aligned}\qquad(9.3)$$

In report [16], the contribution of the Mott factor is absent. Apparently, it was assumed in [3, 4] that $\Gamma = 0.25$ since relation (9.3) in these publications has the form (see formula (1.3) in the Introduction)

$$\frac{v}{2}N_{\mathrm{at}}\sigma_{\mathrm{tr}}(\gamma) = \frac{Z_{\mathrm{mol}}}{8\gamma p}F(\gamma).\qquad(9.4)$$

If we use this formula and assume that $F(p) = F_{\mathrm{ion}}(p)$, which is valid for small values of Z_{mol} and $(\gamma-1)mc^2 \gg \varepsilon_{\mathrm{ion,\,max}}$, the quantity

$$N_{\mathrm{at}}\frac{v}{2}\sigma_{\mathrm{tr}}(p)+\frac{F_{\mathrm{ion}}(p)}{2\gamma p}$$

in relation (9.2) will be reduced to the expression

$$\frac{(Z_{\mathrm{mol}}/2+2)F(\varepsilon)}{4\gamma p},\qquad(9.5)$$

differing from relation (1.3) in the addend "2" in the numerator; however, this should not strongly affect the results of the solution of the kinetic equation since $Z_{\mathrm{mol}}/2$ is usually much larger than unity.

A consistent procedure of reducing the ionization integral permitted not only comparison with operator (1.2)–(1.4), but also made it possible to estimate the contribution from the components of the total collision operator

(see Section 7) and, which is important, to considerably simplify the total operator by omitting some of these components (cf. expressions (8.1) and (8.2)). The separation of weak ionizing interactions into an individual differential operator is inexpedient since this lowers the accuracy of the description without providing any computational advantages.

10. CONCLUSIONS

We have derived the collision operator (8.1) for electrons in a dense weakly ionized plasma with predominant interactions with atomic particles, which takes into account the elastic scattering by nuclei, the excitation of atomic particles, and their ionization.

The operator in form (8.2) is intended for describing the kinetics of high-energy electrons, for which the cross sections and arguments given in Section 7 are valid. Predominant interactions in relation (8.2) are the elastic scattering by nuclei and the ionization of atoms. Operator (8.2) is simpler than operator (1.2)–(1.4) used earlier [3, 4, 8–12] and requires less computer time in numerical calculations since it does not contain an analog of component (1.2) separately describing small variations of the momentum modulus in ionizing collisions.

A consistent procedure of separating "weak" interactions from the ionization integral made it possible to obtain the operator component responsible for angular scattering, which differs from relation (1.3) in the factor in front of $\hat{L}_\mu$ and in a larger contribution of ionizing collisions to the angular scattering of electrons.

Operator (8.1) can be used for describing the kinetics of electrons in a wide range of energies considerably higher than the excitation energies. Limitations are imposed by the requirement of smallness of the scattering angle.

Operator (8.2) can be used not only in studies pertaining to the problem of breakdown in planetary atmospheres in thunderstorm fields, which is only a particular physical problem in spite of its fundamental importance for atmospheric electricity. A KE with operator (8.2) is applicable in problems of high-energy electron transport through dense gaseous media both in the presence and absence of an electric field (the description of the electron–positron component of cosmic-ray showers; analysis of high-voltage discharges in dense gases with runaway electrons and gas discharges sustained by an electron beam, including those intended for pumping high-power gas lasers; analysis of propagation of relativistic electron beams in the atmosphere; and description of the kinetics of Compton electrons from a nuclear explosion in the atmosphere).

ACKNOWLEDGMENTS

The author is deeply indebted to Acad. A.V. Gurevich and Dr. R.A. Roussel-Dupré for valuable discussions and helpful recommendations, and to Acad. R.I. Il'kaev, director of the VNIIEF, and Dr. S.J. Gitomer for their support in studies in the physics of atmospheric electricity.

REFERENCES

1. A. V. Gurevich, G. M. Milikh, and R. Roussel-Dupré, Phys. Lett. A **165**, 463 (1992).
2. A. V. Gurevich, G. M. Milikh, and R. Roussel-Dupré, Phys. Lett. A **187**, 197 (1994).
3. R. A. Roussel-Dupré, A. V. Gurevich, T. Tunnell, and G. M. Milikh, *Kinetic Theory of Runaway Air Breakdown and the Implications for Lightning Initiation* (Los Alamos National Laboratory, 1993), Report LA-12601-MS, p. 51.
4. R. A. Roussel-Dupré, A. V. Gurevich, T. Tunnell, and G. M. Milikh, Phys. Rev. E **49**, 2257 (1994).
5. H. Bethe and U. Ashkin, in *Experimental Nuclear Physics*, Ed. by E. Segré (Wiley, New York, 1953; Inostrannaya Literatura, Moscow, 1955), Vol. 1, Part 2.
6. V. B. Berestetskiĭ, E. M. Lifshitz, and L. P. Pitaevskiĭ, *Quantum Electrodynamics*, 3rd ed. (Nauka, Moscow, 1989; Pergamon Press, Oxford, 1982), Vol. 4.
7. L. P. Babich, I. M. Kutsyk, E. N. Donskoy, and A. Yu. Kudryavtsev, Phys. Lett. A **245**, 460 (1998).
8. E. M. D. Symbalisty, R. A. Roussel-Dupré, L. P. Babich, *et al.*, EOS Trans. Am. Geophys. Union **78**, 4760 (1997).
9. E. M. D. Symbalisty, R. A. Roussel-Dupré, and V. Yukhimuk, IEEE Trans. Plasma Sci. **26**, 1575 (1998).
10. L. P. Babich, E. N. Donskoy, A. Yu. Kudryavtsev, *et al.*, Tr. Ross. Fed. Yad. Tsentra VNIIÉF, No. 1, 432 (2001).
11. L. P. Babich, E. N. Donskoy, R. I. Il'kaev, *et al.*, Dokl. Akad. Nauk **379**, 606 (2001) [Dokl. Phys. **46**, 536 (2001)].
12. L. P. Babich, E. N. Donskoy, I. M. Kutsyk, *et al.*, IEEE Trans. Plasma Sci. **29**, 430 (2001).
13. T. Holstein, Phys. Rev. **70**, 367 (1946).
14. L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, 7th ed. (Nauka, Moscow, 1988; Pergamon Press, Oxford, 1975), Vol. 2.
15. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 3: *Quantum Mechanics: Non-Relativistic Theory*, 4th ed. (Nauka, Moscow, 1989; Pergamon Press, Oxford, 1977).
16. C. L. Longmire and H. J. Longley, *Improvements in the Treatment of Compton Current and Air Conductivity in EMP Problems* (Defense Nuclear Agency, 1973), Report No. 3192T.

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 719–727.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 821–830.

PLASMA, GASES

Thermodynamics and Correlation Functions of an Ultracold Nonideal Rydberg Plasma

M. Bonitz[a], B. B. Zelener[b,*], B. V. Zelener[b], É. A. Manykin[c], V. S. Filinov[d], and V. E. Fortov[d]

[a]*Lehrstuhl Statistische Physik, Institut für Teoretische Physik und Astrophysik, Christian-Albrechts-Universität Kiel, Kiel, D-24098 Germany*

[b]*Joint Institute for High Temperatures, Russian Academy of Sciences, Moscow, 125412 Russia*

[c] *Russian Research Centre Kurchatov Institute, Moscow, 123182 Russia*

[d]*Institute of Thermal Physics of Extreme States, Joint Institute for High Temperatures, Russian Academy of Sciences, Moscow, 125412 Russia*

**e-mail: bobozel@mail.ru*

Received October 2, 2003

Abstract—A pseudopotential model is suggested to describe the thermodynamics and correlation functions of an ultracold, strongly nonideal Rydberg plasma. The Monte Carlo method is used to determine the energy, pressure, and correlation functions in the ranges of temperatures $T = 0.1$–10 K and densities $n = 10^{-2}$–10^{16} cm^{-3}. For a weakly nonideal plasma, the results closely agree with the Debye asymptotic behavior. For a strongly nonideal plasma, many-particle clusters and a spatial order in the arrangement of plasma electrons and ions have been found to be formed. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The unique experimental works [1–3] have given us an insight into some of the physical properties of a dense ionized gas at cryogenic temperatures. Until now, a dense ionized gas has been traditionally produced at high temperatures and high pressures.

In [1–3], in which a Xe plasma was studied, about 5×10^6 metastable Xe atoms (the $6S[3/2]_2$ level, a lifetime of 43 s) were produced, decelerated by using the Zeeman technique, collected in a magneto-optical trap, and radiatively cooled on the $6S[3/2]_2$ –$6P[5/2]_3$ ($\lambda_1 \approx$ 882 nm) transition down to a temperature of 100 μK. The maximum atomic density reached 5×10^{10} cm^{-3}; the density distribution was Gaussian with a root-mean-square radius of $\sigma \approx 180$ μm.

To produce a plasma, more than 20% of the atoms were photoionized over 10 ns. First, the $6P[5/2]_3$ ($\lambda_1 \approx$ 882 nm) level was populated, and then the atom was ionized by photons with a wavelength $\lambda_2 \approx 514$ nm. The difference between the photon energy and the ionization potential, ΔE, was distributed between electrons and ions. Because of the small electron-to-ion mass ratio, only an energy of $4 \times 10^{-6}\Delta E$ was acquired by ions, while the remaining energy was acquired by electrons. In [1–3], $\Delta E/k$ was varied in a controllable way between 0.1 and 1000 K. The maximum charged particle density was

$$n = n_e + n_i = 2 \times 10^9 \text{ cm}^{-3}.$$

An anomalous slowdown of the recombination in the produced plasma was found in [1–3]. The recombination time was on the order of 100 μs. The recombination time estimated by using a formula valid for a rarefied plasma is several nanoseconds, which is many orders of magnitude shorter than the observed value.

Note that the produced plasma is strongly nonideal. In this plasma, the ratio of the mean potential energy of the particles to their kinetic energy (nonideality parameter), $\gamma = \beta e^2 n^{1/3}$ (where $\beta = 1/kT$ is the inverse temperature, and e is the electron charge), is much larger than unity. Thus, at $T = 0.1$ K and $n = 2 \times 10^9$ cm^{-3}, $\gamma = 21$. At the same time, the electrons in this plasma are nondegenerate. The ratio of the thermal de Broglie wavelength $\lambda\!\!\!^{-}_e$ of an electron to the mean particle separation (degeneracy parameter) at $T = 0.1$ K and $n = 2 \times 10^9$ cm^{-3} is much smaller than unity:

$$n^{1/3}\lambda\!\!\!^{-}_e = \frac{n^{1/3}\hbar}{\sqrt{2m_e kT}} = 0.2 \times 10^{-2}. \quad (1)$$

Here, m_e is the electron mass, and $\hbar$ is the Planck constant.

In a strongly nonideal plasma, the estimates of any processes based on the formulas obtained in the approximation of $\gamma \ll 1$ are inapplicable for such nonideality parameters γ.

2. THERMAL RELAXATION

Based on the properties of a nonideal plasma [4], the authors of [1–3] concluded that thermal equilibrium sets in several tens of microseconds. It seems to us that this conclusion is unjustified, because a weakly nonideal plasma with $\gamma \ll 1$ was considered in [4]. Since there is no quantitative kinetic theory for $\gamma \geq 1$, only qualitative estimates can be obtained. In this case, it follows from general physical considerations that the thermal relaxation time in a strongly coupled system (e.g., at fixed temperature T with variation in particle density n) will be shorter than that in a weakly coupled system.

To estimate the thermal relaxation time in a gas of electrons and ions, we use a standard expression from [5] for $\gamma \ll 1$ and $n\lambda^3 \ll 1$:

$$\tau_{ei}^{\varepsilon} = \frac{T_e^{3/2} M}{8 n_i z^2 e^4 (2\pi m_e)^{1/2}} \frac{1}{L_e}. \quad (2)$$

Here, T_e is the electron temperature, M is the ion mass, z is the ion charge, n_i is the ion density, and L_e is the Coulomb logarithm:

$$L_e = \ln\left(\frac{aT_e}{ze^2}\right) = \ln\frac{1}{\sqrt{\pi\gamma^{3/2}}} \quad \left(\frac{ze^2}{\hbar v_{Te}} \gg 1\right), \quad (3)$$

where v_{Te} is the electron velocity that corresponds to T_e, and

$$a = \frac{1}{\sqrt{4\pi n_e \beta e^2}} \quad (4)$$

is the Debye screening length.

If the plasma is electrically neutral, then $n_e = n_i$. Formula (2) contains a Coulomb logarithm L_e that has no physical meaning for $\gamma \geq 1$. The Coulomb logarithm L_e arises in this formula when the transport cross section is calculated. The latter is generally determined in the rarefied case for $a \geq \bar{n}$ (where $\bar{n}$ is the mean density). However, it is clear that the transport cross section always has a physical meaning and is finite.

By analogy, the estimate of the thermal relaxation time τ_{ei}^{ε} for $\gamma \geq 1$ can be represented by using the characteristic physical quantities as (2); in this case, however, L_e^* (the effective Coulomb logarithm that includes the collective effects in a nonideal plasma) should be substituted for L_e:

$$\tau_{ei}^{\varepsilon} = \frac{(\tau_{ei}^{\varepsilon})_0}{L_e^*}, \quad (5)$$

where

$$(\tau_{ei}^{\varepsilon})_0 = \frac{T_e^{3/2} M}{8 n_i z^2 e^4 (2\pi m_e)^{1/2}},$$

$(\tau_{ei}^{\varepsilon})_0 = 3.4 \times 10^{-2}$ s at $T_e = 0.1$ K and $n_i = 10^4$, $z = 1$ ($\gamma = 0.67$, $M = 131.3$ amu). At $T_e = 0.1$ K and $n_i = 10^{10}$, $(\tau_{ei}^{\varepsilon})_0 = 34$ ns.

We also assume that the differential scattering cross section remains Coulomb, but either corrections to it arise or the integration limits change. However, the dependence of the relaxation time on n at $T = \text{const}$ remains logarithmic.

The value of τ_{ei}^{ε} can change via the substitution of L_e^* for L_e by no more than a factor of 10 to 20. Of course, all of this reasoning must be confirmed by rigorous estimations or numerical calculations. Thus, we may assume that thermal equilibrium at $T = 0.1$ K and $n = 10^9$–10^{10} sets in less than 1 μs.

3. THEORETICAL APPROACHES TO STUDYING AN ULTRACOLD NONIDEAL RYDBERG PLASMA

The Xe plasma produced in the experiments [1–3] consists of singly charged Xe ions, electrons, and highly excited ($n > 100$) hydrogen-like Xe states. These states are called Rydberg atoms. The possibility of the existence of condensed excited states of matter was first considered in [6]. At present, these condensed states of substance for Rydberg atoms (the so-called Rydberg substance) have been extensively studied theoretically [7, 8] and experimentally [9–11]. Subsequently, this idea has been developed by several authors (see the review [12]). According to this approach, the interaction between Rydberg atoms as their density increases ultimately leads to a change in the phase state of the system and to a qualitative change in all parameters. Moreover, in contrast to free Rydberg atoms whose lifetime in an excited state is about 10 ns, the lifetime of a Rydberg substance is macroscopically long. Despite the low density that is the gas density by its parameters, the condensed excited state is a metastable ordered state of the substance. At present, there are experimental data [13, 14] for a cesium plasma at $T = 500$–1000 K that suggest the existence of a Rydberg substance.

In [15, 16], the experimental data [13, 14] were found to correlate with the ideas of an isolated region of metastable states of an ultracold nonideal plasma. In [6, 12], the methods of solid-state physics were used to produce a Rydberg substance by assuming the presence of an electron Fermi liquid. In [17, 18], the recombination time of a dense plasma was calculated numerically and was shown to be much longer than that for a

weakly nonideal plasma. The authors of [19, 20] suggested a different approach to studying the thermodynamics of a Rydberg substance in the case where the electron gas is nondegenerate. This approach uses previously developed methods for a strongly nonideal nondegenerate plasma [21].

We consider the thermodynamic equilibrium of a nonideal gas of electrons and ions ($\gamma \geq 1$). This gas is peculiar in that there is absolutely no atomic discrete spectrum or there are discrete states with $n \geq n^*$, where $n^* = 100$ or more. Since $n^* \geq 100$, the states may be said to be Rydberg ones. Strictly speaking, there is no full thermodynamic equilibrium in our case. Therefore, when we talk about thermodynamic equilibrium degrees of freedom, we primarily have in mind the translational degrees of freedom. The thermodynamic equilibrium of all the remaining degrees of freedom (rotational, vibrational, dissociation and chemical reactions, ionization and electron excitation) arises much later. It may be assumed that in easily excited degrees of freedom, equilibrium exists at each instant of time, while the slow relaxation processes do not proceed at all over the period under consideration. We will use results from [21–29] to describe this gas.

4. THE PSEUDOPOTENTIAL MODEL AND THE RANGE OF ITS APPLICABILITY

The thermodynamics of an equilibrium quantum-mechanical system is completely determined if the partition function is known:

$$Z_N \approx \mathrm{Tr}(\exp(-\beta\hat{H})) = \int_{V^N} \sum_{n=1}^{\infty} |\psi_n|^2 \exp(-\beta E_n) dq_N, \quad (6)$$

where $\mathrm{Tr}(\exp(-\beta\hat{H}))$ is the trace of the density matrix $\exp(-\beta\hat{H})$; $\psi_n(q_N)$ and E_n are the wave functions and energy levels of N particles, respectively; $\hat{H}$ is the Hamiltonian of the N-particle system; q_N are the coordinates of the N particles; and V is the volume of the system.

The authors of [25–29] developed an approach that allows the thermodynamic properties of a dense plasma to be described over a wide range of nonideality and degeneracy parameters, including the region of strong nonideality and degeneracy. However, a simpler approach developed in [21–24] can be used for a nondegenerate plasma. The authors of these papers suggested a pseudopotential model to calculate the partition function of a nonideal nondegenerate plasma. It involves the Slater partition function

$$S_N = N!\lambda^{3N} \sum_{n=1}^{\infty} |\psi_n|^2 \exp(-\beta E_n), \quad (7)$$

where λ is the particle thermal wavelength.

The essence of the pseudopotential model is that the partition function (7) is represented as a product of the pair electron–electron, ion–ion, and electron–ion Slater partition functions:

$$S_{N_e+N_i} = \prod_{i<j=1}^{2N} S_{ij} = \prod_{i<j=1}^{N_e} S_{ee} \times \prod_{i<j=1}^{N_i} S_{ii} \prod_{i<j=1}^{N=N_i=N_e} S_{ei}. \quad (8)$$

In the experimental conditions under consideration, this approximation is valid not only for $\gamma < 1$, but also for $\gamma \geq 1$, because there are neither pair nor many-particle bound states in the Xe gas. By analogy with the classical case, product (8) may be substituted with

$$S_{N_i+N_e} = \exp\left(-\beta \sum_{ij} U_{ij}\right), \quad (9)$$

where

$$U_{ij} = -\frac{\ln S_{ij}}{\beta} \quad (10)$$

is the pseudopotential.

The pair Slater partition functions for electron–ion, electron–electron, and ion–ion interactions can be calculated accurately. The expression for the long-range pseudopotential is identical to the Coulomb law, while the short-range pseudopotential is finite and temperature-dependent.

4.1. The Electron–Ion Pseudopotential

For the interaction of an electron with an ion, the pseudopotential Φ_{ei} is defined by the relation

$$\exp(-\beta\Phi_{ei}(r, T)) = S_{ei} = 8\pi^{3/2}\lambda_{ei}^3 \sum_{E_\alpha = E_0}^{\infty} |\Psi_\alpha(r)|^2 \exp(-\beta E_\alpha), \quad (11)$$

where S_{ei} is the two-particle Slater partition function, $E_\alpha(r)$ is the energy of the electron in the field of the ion

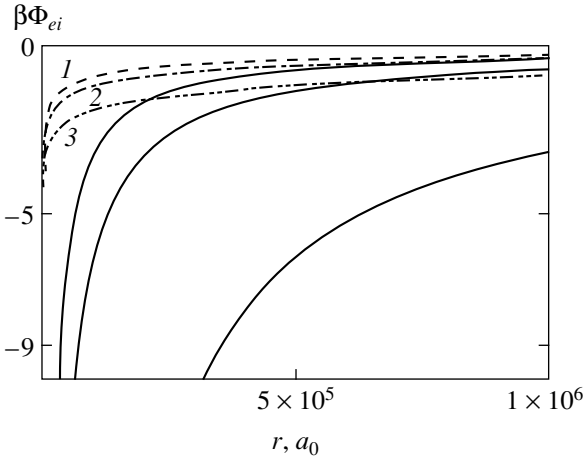


Fig. 1. The ion–electron pseudopotentials at various temperatures, $T = 1$ (*1*), 0.5 (*2*), and 0.1 K (*3*), compared to the Coulomb potentials (solid lines)

in state α, $\Psi_\alpha(r)$ is the corresponding wave function, and $\lambda_{ei} = \lambda_e/\sqrt{2}$. The summation is over all possible states E_α. This pseudopotential is finite at $r = 0$, while the expression for the long-range pseudopotential is identical to the Coulomb law.

For a plasma without bound states up to $n = n_0$, the expression for $S_{ei}(r, T)$ can be written as [21]

$$S_{ei}(r, T) = S_d + S_c. \tag{12}$$

Here,

$$S_d = 8\pi^{3/2}\lambda_{ei}^3 \sum_{E_\alpha = E_0}^{E'} |\psi_\alpha(r)|^2 \exp(-\beta E_\alpha) \tag{13}$$

is the contribution of the part of the discrete spectrum from E_0 to E', which can be calculated from the known wave functions $\Psi_\alpha(r)$ for the hydrogen atom, and the contribution of the remaining part of the spectrum is

$$S_c(z, T) = \exp\left(-\frac{\beta e^2}{r}\right), \tag{14a}$$

if

$$\frac{\beta E' r}{\lambda_{ei}} \geq \frac{|\beta e^2|}{\lambda_{ei}},$$

and

$$S_c(r, T) = \left[1 - y\left(\sqrt{\beta E' + \frac{\beta e^2}{r}}\right)\right]\exp\left(-\frac{\beta e^2}{r}\right) + 2\pi^{-1/2}\left(\beta E' + \frac{\beta e^2}{r}\right)^{1/2}, \tag{14b}$$

if

$$\frac{\beta E' r}{\lambda_{ei}} < \frac{|\beta e^2|}{\lambda_{ei}},$$

where

$$y(z) = 2\pi^{-1/2}\int_0^z \exp(-t^2)dt.$$

The partition function $S_c(r, T)$ was derived in the quasi-classical approximation for bound states at $E_\alpha > E'$ and continuum states.

Expressions (13) and (14) for $|\beta e^2/\lambda_{ei}| > 1$ at $r = 0$ take the form

$$S_d\left(0, \frac{\beta e^2}{\lambda_{ei}}\right) \approx \pi^{1/2}\left(\frac{\beta e^2}{\lambda_{ei}}\right)^3 n_0^{-3}\exp\left(\frac{\beta e^2}{n_0^3}\right),$$
$$S_c\left(0, \frac{\beta e^2}{\lambda_{ei}}\right) \approx 2\pi^{1/2}\frac{\beta e^2}{\lambda_{ei}}. \tag{15}$$

At $n_0 \geq 100$, it will suffice to use (14) and (15) and the hydrogen wave functions and energy levels to determine the potential $\Phi_{ei}(r, T)$. When determining $\Phi_{ei}(r, T)$ for the Xe ion, we must also take into account the fact that its size is finite (its crystallographic radius is about 2 Å). Figure 1 shows the electron–ion pseudopotentials at $T = 0.1$, 0.5, and 1 K and, for comparison, the Coulomb potentials.

4.2. The Electron–Electron and Ion–Ion Pseudopotentials

The Slater partition function for the interaction between two electrons is [21]

$$S_{ee}(r, T) = 16\pi^{3/2}\lambda_{ee}^3 \sum_\alpha |\Psi_\alpha(r, \sigma_1, \sigma_2)|^2 \exp(-\beta E_\alpha) \equiv \exp(-\beta\Phi_{ee}(r, T)). \tag{16}$$

The wave functions in expression (16) depend on the electron spins σ_1 and σ_2 and must be antisymmetric. Expression (16) can be written as a sum of the contributions from the wave functions with symmetric and nonsymmetric parts:

$$S_{ee}(r, T) = \frac{1}{4}S_{ee}^c(r, T) + \frac{3}{4}S_{ee}^a(r, t) \equiv \exp(-\beta\Phi_{ee}(r, T)). \tag{17}$$

The potential $\Phi_{ee}(r, T)$ was numerically calculated by Barker [22] over a wide temperature range, $T = 10^2$–10^5 K. He suggested the following fitting formula for $\Phi_{ee}(r, T)$:

$$\Phi_{ee}(r, T) = \frac{2}{r}(1 - \exp(-8.35 \times 10^{-4} r T^{0.625})),$$
$$\Phi_{ee}(0, T) = 16.7 \times 10^{-4} T^{0.625}, \quad (18)$$

where r is measured in a_0, T is in Kelvins, and Φ_{ee} is in Rydbergs (Re = $0.5me^4/\hbar^2$).

At long range, formula (18) is identical to the Coulomb law. For $\beta e^2/\lambda_{ee} > 1$, expression (17) at $r = 0$ can be written [23] as

$$S_{ee}\left(0, \frac{\beta e^2}{\lambda_{ee}}\right) \approx \left(\frac{4\pi}{3}\right)^{1/2} \left(\frac{\beta e^2}{\lambda_{ee}}\right)^{4/3} \times \left(\frac{\pi}{2}\right)^{1/3} \exp\left(-3\left(\pi \frac{\beta e^2}{2\lambda_{ee}}\right)^{2/3}\right). \quad (19)$$

Thus, for $\beta e^2/\lambda_{ee} > 1$, fit (18) can be used down to $T = 0.1$ K.

Figure 2 shows the electron–electron pseudopotentials at $T = 0.1$, 0.5, and 1 K and, for comparison, the Coulomb potentials.

According to [21], the expressions for the ion–ion pseudopotentials are identical to the Coulomb law. We must only take into account the fact that the ion size (e.g., the crystallographic radius) is finite.

5. CALCULATING THE THERMODYNAMIC QUANTITIES AND CORRELATION FUNCTIONS

5.1. The Method of Calculation

In the pseudopotential approach, the quantum partition function reduces to an expression that is classical in form [21]. Therefore, all of the methods developed in the statistical thermodynamics of classical systems (both analytical and numerical) can be used to determine the relevant thermodynamic quantities.

We used the Monte Carlo method for a multicomponent plasma in a canonical ensemble developed in [24, 25] to calculate the thermodynamic quantities and correlation functions of an ultracold Rydberg plasma.

In this case, determining the various thermodynamic quantities reduces to calculating the mean values of the known functions of coordinates q. For example, for the energy, we obtain

$$\bar{E} = Q^{-1}(N, V, T) \int_V \ldots \int E_N(q) S_N(a, T) d^N q, \quad (20)$$

where $Q(N, V, T)$ is the path integral.

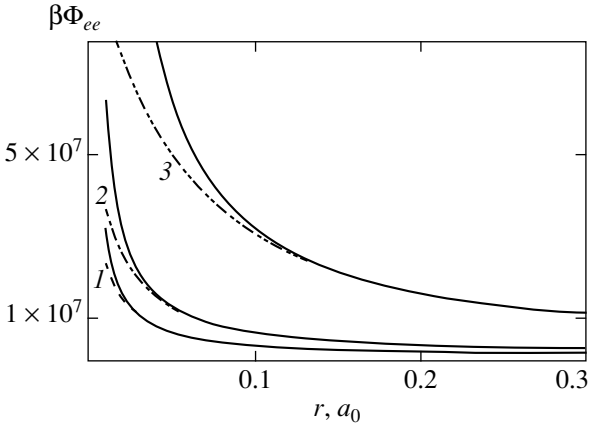


Fig. 2. The electron–electron pseudopotentials at various temperatures, $T = 1$ (*1*), 0.5 (*2*), and 0.1 K (*3*), compared to the Coulomb potentials (solid lines).

The Monte Carlo method is a numerical method that uses Markov chains [24]. It allows us to select only the principal, most typical terms that define the integral sum. Therefore, it is also called a method of significant selection. Another peculiarity of the method is the use of periodic boundary conditions. The entire three-dimensional space is broken down into equal cells of volume V with N particles in each cell. If one of the particles exits from a cell due to a change in its coordinates, then its image from a neighboring cell simultaneously enters through the opposite cell face, and the number of particles in the cell is conserved.

The errors in the Monte Carlo results [24] are attributable to the choice of the number of particles in the cell and to the finite length of the Markov chain. To estimate the error in choosing the number of particles, we performed calculations for various $N = 16$, 32, 64, and 128 and showed convergence ($(\propto N^{-1})$). Our estimate of the statistical error due to the finite length of the Markov chain [24] allowed us to choose Markov chains of the required length. In addition, we discarded the nonequilibrium part. We also calculated the electron–electron, $g_{ee}(r)$, ion–ion, $g_{ii}(r)$, and electron–ion, $g_{ei}(r)$, radial correlation functions.

5.2. Results of the Calculations

Thus, we consider the pseudopotential model of an ultracold Rydberg plasma. Experimental data suggest that this plasma consists of electrons, singly charged ions, and atoms in highly excited ($n > 100$) states. There are no low-excitation ($n < 100$) states in this plasma, because it was produced through the laser excitation of atoms at a certain wavelength, and because an anomalous increase in the recombination time was observed in the experiment.

We performed calculations for the temperature range $T = 0.1$–10 K and the density range $n = 10^{-2}$–10^{16} cm^{-3}. The calculations in the range of low densities

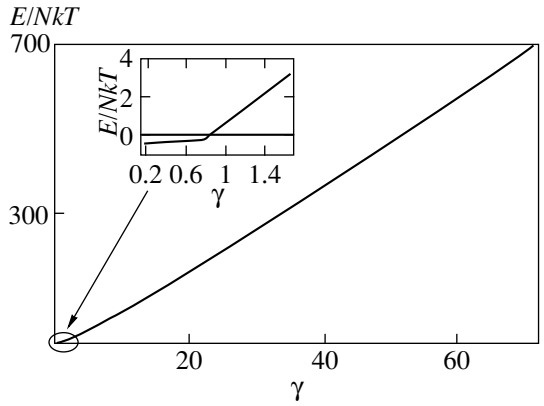


Fig. 3. Internal energy per particle, *E*/*NkT*, versus nonideality parameter γ.

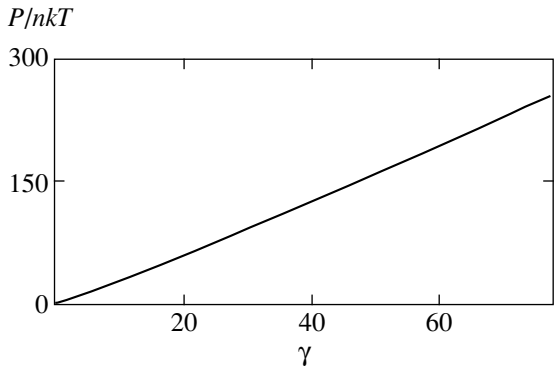


Fig. 4. Pressure *P*/*nkT* versus nonideality parameter γ.

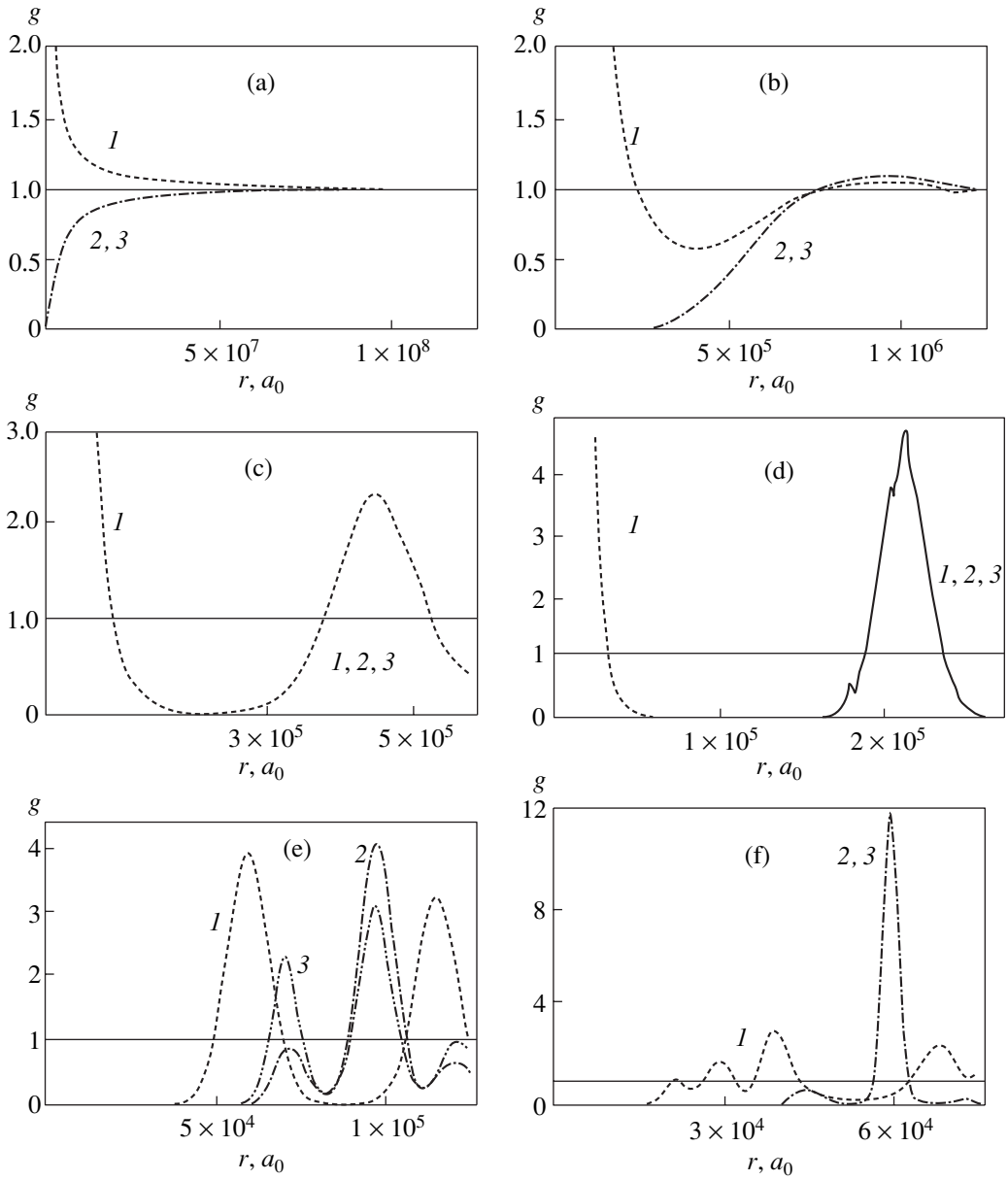


Fig. 5. The correlation functions at various densities and temperature $T = 0.1$ K: $g_{ei}(r)$ (*1*), $g_{ee}(r)$ (*2*), and $g_{ii}(r)$ (*3*); (a) $n = 10$, $\gamma = 0.036$; (b) $n = 10^7$, $\gamma = 3.6$; (c) $n = 10^8$, $\gamma = 7.7$; (d) $n = 10^9$, $\gamma = 16.7$; (e) $n = 10^{10}$, $\gamma = 36$; and (f) $n = 10^{11}$, $\gamma = 77$.

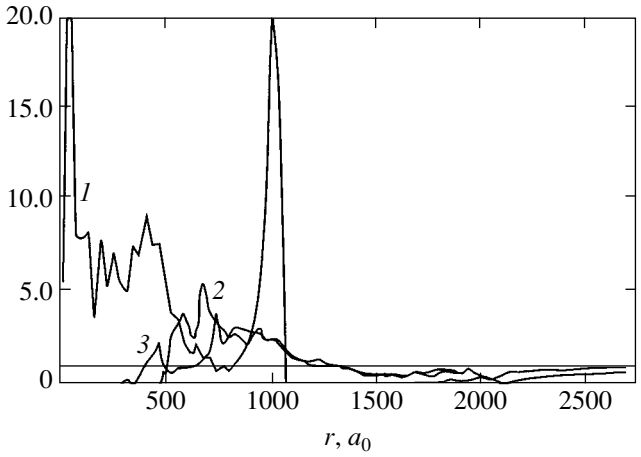


Fig. 6. The correlation functions at density $n = 10^{15}$ cm^{-3} and temperature $T = 10$ K: $g_{ei}(r)$ (*1*), $g_{ee}(r)$ (*2*), and $g_{ii}(r)$ (*3*).

were needed to pass to the limit of the values that are consistent with the Debye–Hückel approximation for $\gamma \ll 1$ (see, e.g., [21]).

In Figs. 3 and 4, internal energy E/NkT per particle and pressure P/nkT are plotted against nonideality parameter γ. In the limit of small γ, there is agreement with the Debye–Hückel approximation (see the inset in Fig. 3). At $\gamma < 0.5$, the dimensionless energy E/NkT reaches the Debye value

$$\frac{E}{NkT} = -\sqrt{\pi}\gamma^{3/2}. \quad (21)$$

Figures 5 and 6 show the correlation functions $g_{ee}(r)$, $g_{ii}(r)$, and $g_{ei}(r)$ for various densities and temperatures $T = 0.1$ and 10 K, respectively. For $\gamma \ll 1$, there is good agreement with the Debye–Hückel approxima-

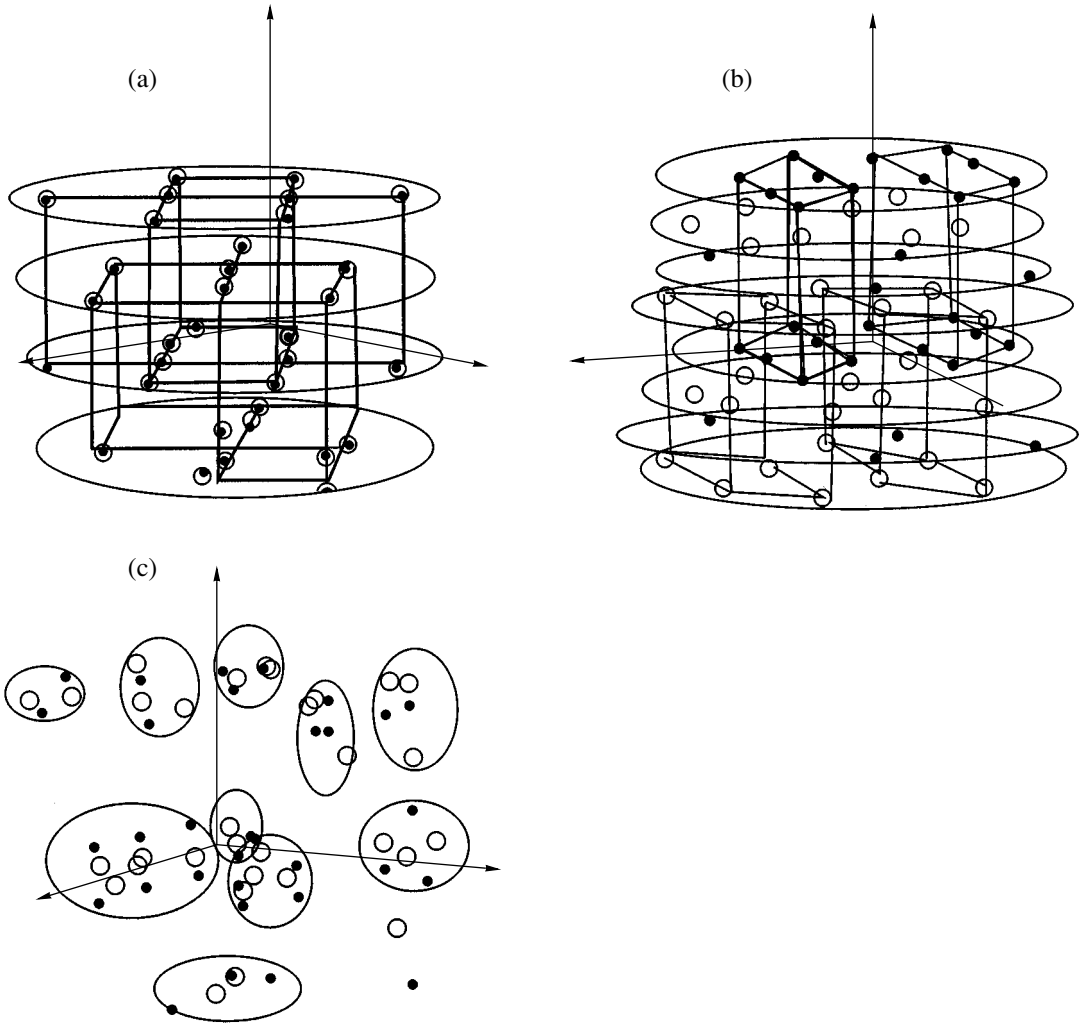


Fig. 7. The graphical images of the particle coordinates that correspond to the correlation functions $g_{ee}(r)$, $g_{ii}(r)$, and $g_{ei}(r)$. The open and filled circles represent the ions and electrons, respectively: (a) $T = 0.1$ K, $n = 10^9$ cm^{-3}, the ions and electrons are at the lattice site; (b) $T = 0.1$ K, $n = 10^{10}$ cm^{-3}, the ions are at the sites of one lattice, while the electrons are at the sites of the other lattice; (c) $T = 10$ K, $n = 10^{15}$ cm^{-3}, the ions and electrons form droplets with lattice site nuclei, the transition state between the short-range order and the lattice.

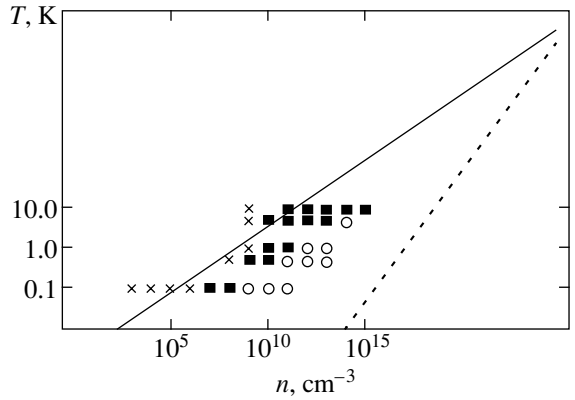


Fig. 8. The *n–T* diagram. The crosses, squares, and circles represent the gaseous, liquid, and solid (lattice) states of the plasma, respectively. The solid and dashed lines correspond to $\gamma = 1$ and $n\lambda^3 = 1$, respectively.

tion (linearized or nonlinearized). For $\gamma \geq 1$, the shape of the correlation functions suggests that a short-range order is formed among particles of both the same and opposite signs. This order is enhanced with increasing γ. The maxima of the correlation functions increase, while their minima become zero, which is attributable to the formation of a strict order in the spatial arrangement of particles. There are almost no particles in the region of zero correlation functions.

We used a visualization program to better understand the situation related to ordering with increasing γ. This program visualizes the arrangement of particles in various equilibrium configurations. Figures 7a–7c show some of the equilibrium configurations for various T and n.

Let us discuss the results for the $T = 0.1$ K isotherm. The order that corresponds to a lattice of size $L = 2.2 \times 10^5 a_0$ at $n = 10^9$ cm^{-3} arises as the density increases (Figs. 5d and 7a). The pairs of electrons and ions are located at the lattice sites at distance $r = 2.2 \times 10^5 a_0$. As the density increases to $n = 10^{10}$ cm^{-3} (Figs. 5e and 7b), the electrons and ions that form the pair move apart, and two (electron and ion) lattices are formed. This probably suggests that two nested lattices constituted the initial lattice at the sites of which the pairs were located.

As the temperature increases, the formation of an ordered structure shifts toward higher particle densities. Thus, for $T = 10$ K and at a much higher density, $n = 10^{15}$ cm^{-3}, only a short-range order in the form of electron–ion clusters (as we see from the equilibrium configuration) shown in Figs. 6 and 7c is established. These clusters are droplets of oppositely charged particles, with the electrons and ions in these droplets lining up in minilattices.

As was noted above, the energy E/NkT at $\gamma \geq 0.1$ in the range $T = 0.1$–10 K is a linear function of γ (see Fig. 3). This implies that, eliminating the temperature from this function, we will obtain an expression similar to the standard Madelung law for an ionic crystal [30]:

$$\frac{E}{N} = Ae^2 n^{1/3}, \tag{22}$$

where $A = 8$–9 is a constant (an analogue of the Madelung constant).

As follows from our calculations, the lattice constant is proportional to $n^{-1/3}$. The form of Eq. (22) suggests that an order similar to a crystal lattice is established in the ionized gas produced at $\gamma \geq 1$. In this case, the energy is a function of the mean particle separation, which is approximately equal to the lattice constant.

Figure 8 shows an *n–T* diagram. The region of parameters that corresponds to a Debye plasma is indicated by crosses; the regions where droplets and lattices appear are indicated by squares and circles, respectively. The $\gamma = 1$ and $n\lambda^3 = 1$ lines are also shown in the figure. We see from this diagram that the formation of a short-range order begins only at $\gamma \approx 1$; as was described above, the formation of an ordered structures shifts toward higher particle densities as the temperature increases. In addition, under the given conditions, a long-range order is formed long before the onset of degeneracy.

Our results also give us an insight into what the authors of [1–3] call the anomalously slowed down recombination. There are no Rydberg atoms that must recombine in an ionized gas at $\gamma \geq 1$. However, there is a short-range order (and a long-range order at $\gamma \gg 1$) for charged particles of both the same and opposite signs, which reduces the probability of the approach and recombination of oppositely charged particles.

6. CONCLUSIONS

We have considered a Rydberg ionized gas formed from continuum electrons and ions. We investigated the temperature range $T = 0.1$–10 K and the density range $n = 10^{-2}$–10^{16} cm^{-3}. As a result, we found the formation of a structure at $\gamma \geq 1$, which probably leads to the experimentally observed slowdown of the recombination. The structure is formed in the region where the electron gas is far from being degenerate ($n\lambda^2 \ll 1$) and where the structure itself changes from a short-range order (similar to the structure in a liquid) to a long-range order (similar to the lattice in solid bodies). Adding states of the discrete spectrum to the gas under consideration will change the properties of this gas. When these states are taken into account for specific densities and temperatures, the energy decreases, which may cause the phase diagram to change.

The suggested model contains no specific parameters of the elements. Therefore, it may be used for a gas of any element.

ACKNOWLEDGMENTS

This work was supported in part by the Russian Foundation for Basic Research (project nos. 02-02-16320 and 04-02-17474).

REFERENCES

1. T. C. Killian, S. Kulin, S. D. Bergeson, *et al.*, Phys. Rev. Lett. **83**, 4776 (1999).
2. S. Kulin, T. C. Killian, S. D. Bergeson, and S. L. Rolston, Phys. Rev. Lett. **85**, 318 (2000).
3. T. C. Killian, M. J. Lim, S. Kulin, *et al.*, Phys. Rev. Lett. **86**, 3759 (2001).
4. L. Spitzer, Jr., *Physics of Fully Ionized Gases*, 2nd ed. (Wiley, New York, 1962; Mir, Moscow, 1957), p. 35.
5. E. M. Lifshitz and L. P. Pitaevskiĭ, *Physical Kinetics* (Nauka, Moscow, 1979; Pergamon Press, Oxford, 1981).
6. É. A. Manykin, M. I. Ozhovan, and P. P. Poluéktov, Dokl. Akad. Nauk SSSR **260**, 1096 (1981) [Sov. Phys. Dokl. **26**, 974 (1981)].
7. É. A. Manykin, M. I. Ozhovan, and P. P. Poluéktov, Zh. Éksp. Teor. Fiz. **84**, 442 (1983) [Sov. Phys. JETP **57**, 256 (1983)].
8. É. A. Manykin, M. I. Ozhovan, and P. P. Poluéktov, Zh. Éksp. Teor. Fiz. **102**, 804 (1992) [Sov. Phys. JETP **75**, 440 (1992)].
9. C. Aman, J. B. C. Pettersson, and L. Holmlid, Chem. Phys. **147**, 189 (1990).
10. R. S. Svensson, L. Holmlid, and L. Lundgren, J. Appl. Phys. **70**, 1489 (1991).
11. C. Aman, J. B. C. Pettersson, H. Lindroth, and L. Holmlid, J. Mater. Res. **7**, 100 (1992).
12. É. A. Manykin, M. I. Ozhovan, and P. P. Poluéktov, Khim. Fiz. **18**, 88 (1999).
13. R. Svensson and L. Holmlid, Phys. Rev. Lett. **83**, 1739 (1999).
14. 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)].
15. G. É. Norman, in *Proceedings of XVI International Conference on Impact of Intensive Energy Fluxes on the Matter*, Ed. by V. E. Fortov (Inst. Probl. Khim. Fiz. Ross. Akad. Nauk, Chernogolovka, 2001), p. 110.
16. G. É. Norman, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 13 (2001) [JETP Lett. **73**, 10 (2001)].
17. A. N. Tkachev and S. I. Yakovlenko, Kvantovaya Élektron. (Moscow) **30**, 1077 (2000).
18. A. N. Tkachev and S. I. Yakovlenko, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 71 (2001) [JETP Lett. **73**, 66 (2001)].
19. B. B. Zelener, B. V. Zelener, and É. A. Manykin, in *Proceedings of XVII International Conference on Equations of State of Substance*, Ed. by V. E. Fortov (Inst. Probl. Khim. Fiz. Ross. Akad. Nauk, Chernogolovka, 2002).
20. V. S. Filinov, E. A. Manykin, B. B. Zelener, and B. V. Zelener, in *Proceedings of 12th International Laser Physics Workshop* (Hamburg, 2003).
21. B. V. Zelener, G. É. Norman, and V. S. Filinov, *Perturbation Theory and Pseudopotential in Statistical Thermodynamics* (Nauka, Moscow, 1981), p. 101.
22. A. A. Barker, J. Chem. Phys. **55**, 1751 (1971).
23. V. S. Vorob'ev, G. É. Norman, and V. S. Filinov, Zh. Éksp. Teor. Fiz. **57**, 838 (1969) [Sov. Phys. JETP **30**, 459 (1969)].
24. V. M. Zamalin, G. É. Norman, and V. S. Filinov, *The Monte Carlo Method in Statistical Thermodynamics* (Nauka, Moscow, 1977), p. 129.
25. V. S. Filinov, M. Bonitz, W. Ebeling, and V. E. Fortov, Plasma Phys. Controlled Fusion **43**, 743 (2001).
26. V. S. Filinov, M. Bonitz, P. Levashov, *et al.*, J. Phys. A: Math. Gen. **36**, 6069 (2003).
27. V. S. Filinov, V. E. Fortov, M. Bonitz, and P. R. Levashov, Pis'ma Zh. Éksp. Teor. Fiz. **74**, 422 (2001) [JETP Lett. **74**, 384 (2001)].
28. V. S. Filinov, V. E. Fortov, and M. Bonitz, Pis'ma Zh. Éksp. Teor. Fiz. **72**, 361 (2000) [JETP Lett. **72**, 245 (2000)].
29. V. S. Filinov, V. E. Fortov, M. Bonitz, and D. Kremp, Phys. Lett. **274**, 228 (2000).
30. C. Kittel, *Introduction to Solid State Physics*, 5th ed. (Wiley, New York, 1976; Nauka, Moscow, 1978).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 728–744.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 831–849.

FLUIDS

A Nonlinear Theory of Turbulent Diffusion

N. A. Silant'ev

Instituto Nacional de Astrofísica, Óptica y Electrónica,
Apartado Postal 51 y 216, CP 72000, Puebla, México
Pulkovo Astronomical Observatory, Russian Academy of Sciences,
Pulkovskoe shosse 65, St. Petersburg, 196140 Russia
e-mail: silant@inaoep.mx

Received May 15, 2003

Abstract—It is shown that the well-known conservation laws for magnetic helicity and passive-scalar fluctuation intensity in the case of negligible molecular diffusion require that the hierarchy of nonlinear equations for the averaged Green function and the hierarchy of Bethe–Salpeter-type equations for fluctuation intensity be treated in a mutually consistent manner. These hierarchies are obtained to the sixth order in turbulent velocity correlators. Asymptotic formulas describing the evolution of scalar fluctuations and magnetic field are presented. A number of new effects are revealed that strongly affect diffusion, but are beyond the scope of the frequently used model of a delta-correlated turbulent field. © *2004 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

Diffusion of passive fields, such as particle density, temperature, or magnetic field, is a problem of practical importance in turbulence theory. Passivity means that the inverse effect of a passive field on turbulent flow can be neglected. This condition is most restrictive as applied in the theory of magnetic field diffusion in turbulent plasmas, because turbulent motion may strengthen both large- and small-scale components of a magnetic field. When the magnetic energy, $B^2/8\pi$, is comparable to the plasma kinetic energy, $\rho u^2/2$, the effect of magnetic field on turbulence must be taken into account.

It was argued in [1, 2] that magnetic field fluctuation intensity rapidly grows to a level comparable to turbulent energy, and therefore magnetic field can almost never be treated as a passive variable. The results of these studies are analogous to the well-known estimate for the energy of magnetic field fluctuations in a two-dimensional fluid [3]. Computations of time-dependent magnetic fields in turbulent flows [4–6] do not support these estimates. In [7, 8], the predictions made in [2, 3] was thoroughly analyzed and the estimates for magnetic energy fluctuations presented therein were shown to be highly overestimated. Thus, the evolution of magnetic field in a turbulent flow involves a period when magnetic field can be treated as passive (particularly in the case of zero turbulent helicity). In numerous studies, evolution of magnetic field is analyzed in this particular approximation.

The starting stochastic equations for passive-scalar density $n(\mathbf{r}, t)$ and magnetic induction $\mathbf{B}(\mathbf{r}, t)$ are

$$\frac{\partial n(\mathbf{r}, t)}{\partial t} - D_\mathrm{m}\nabla^2 n(\mathbf{r}, t) = -\mathrm{div}[\mathbf{u}(\mathbf{r}, t)n(\mathbf{r}, t)] \quad (1)$$

and

$$\frac{\partial \mathbf{B}(\mathbf{r}, t)}{\partial t} - D_\mathrm{m}\nabla^2 \mathbf{B}(\mathbf{r}, t) = \nabla \times [\mathbf{u}(\mathbf{r}, t) \times \mathbf{B}(\mathbf{r}, t)]. \quad (2)$$

Here, $\mathbf{u}(\mathbf{r}, t)$ is the turbulent velocity field with prescribed statistical properties and D_m is molecular diffusivity ($D_\mathrm{m} = c^4/4\pi\sigma$ for magnetic field, where σ is plasma conductivity and c is the speed of light).

The problem of turbulent diffusion essentially reduces to finding evolution equations for the mean fields $\langle n\rangle$ and $\langle \mathbf{B}\rangle$ and for the second-order quantities

$$\begin{aligned} V(\mathbf{r}, t; \mathbf{r}', t') &= \langle n(\mathbf{r}, t)n(\mathbf{r}', t')\rangle \\ &\equiv \langle n(1)\rangle\langle n(2)\rangle + \langle n'(1)n'(2)\rangle, \end{aligned} \quad (3)$$

$$\begin{aligned} H_{ij}(\mathbf{r}, t; \mathbf{r}', t') &= \langle B_i(\mathbf{r}, t)B_j(\mathbf{r}', t')\rangle \\ &\equiv \langle B_i(1)\rangle\langle B_j(2)\rangle + \langle B_i'(1)B_j'(2)\rangle. \end{aligned} \quad (4)$$

All variables are routinely represented as sums of mean and fluctuating components: $n = \langle n\rangle + n'$, $\mathbf{B} = \langle \mathbf{B}\rangle + \mathbf{B}'$, where $\langle n'\rangle = 0$ and $\langle \mathbf{B}'\rangle = 0$. Suppose that turbulent flow can be characterized by an rms velocity fluctuation u_0 ($u_0^2 = \langle u^2(\mathbf{r}, t)\rangle$, $\langle \mathbf{u}\rangle = 0$) and the length and time scales R_0 and τ_0 associated with two-point correlations. The assumption that the mean fields are smooth over the characteristic scales R_0 and τ_0 (or $t_0 = R_0/u_0$) leads to diffusion equations for these fields, with diffusivity $D_\mathrm{m} + D_\mathrm{t}$, where D_t is a turbulent diffusivity and $D_\mathrm{t} \gg D_\mathrm{m}$ (e.g., see [9]). Thus, the problem reduces to calculation of D_t.

In the model of turbulence in an unbounded flow domain considered here, exact formulas can be

obtained for D_t in both Lagrangian [10, 11] and Eulerian [9, 12] representations. Since Eulerian calculations are more suitable from a practical perspective, the present analysis relies on the Eulerian representation. In this representation, the exact value of D_t is determined by finding the stochastic Green function $G(\mathbf{r}, t; \mathbf{r}', t') \equiv G(1, 2)$ of Eq. (1) or (2) and averaging the result over the statistical ensemble of the $\mathbf{u}(\mathbf{r}, t)$ components.

Hereinafter, we use the following convenient abbreviated notation:

$$f(1) = f(\mathbf{r}_1, t_1), \quad f(1-2) = f(\mathbf{r}_1 - \mathbf{r}_2, t_1 - t_2),$$

$$dm = d\mathbf{r}_m dt_m, \quad d\mathbf{m} = d\mathbf{r}_m, \quad \mathbf{R} = \mathbf{r}_1 - \mathbf{r}_2,$$

$$\tau = t_1 - t_2, \ldots.$$

The derivation and solution of evolution equations for correlators (3) and (4) is a much more complicated task than the determination of the mean fields $\langle n\rangle$ and $\langle \mathbf{B}\rangle$. Indeed, it is shown below that correlators (3) and (4) cannot be fully described in the diffusion approximation, and complex Bethe–Salpeter-type integral (or integrodifferential) equations must be solved (e.g., see [13, 14]). Various approximate forms of this equation have been proposed and analyzed in [15–17]. In this equation, fluctuation intensity is determined by the difference of two terms of similar order of magnitude. On the other hand, Eqs. (1) and (2) written for homogeneous turbulence have the following important corollaries:

$$\frac{d}{dt}\langle n^2(\mathbf{r}, t)\rangle = -2D_m\langle(\nabla n(\mathbf{r}, t))^2\rangle - \langle n^2(\mathbf{r}, t)\mathrm{div}\mathbf{u}(\mathbf{r}, t)\rangle, \tag{5}$$

$$\frac{d}{dt}\langle \mathbf{A}\cdot\mathbf{B}\rangle = -2D_m\langle \mathbf{B}\cdot(\nabla\times\mathbf{B})\rangle. \tag{6}$$

According Eq. (5), decrease in the mean square of passive-scalar density at a fixed point in an incompressible turbulent flow ($\mathrm{div}\mathbf{u} = 0$) can be caused only by molecular diffusion. When molecular diffusion is neglected, $\langle n^2(\mathbf{r}, t)\rangle$ is a conserved quantity. Accordingly, Eqs. (5) and (6) are referred to as conservation laws here. Equation (6) reduces to the well-known conservation law for magnetic helicity $\mathbf{A}\cdot\mathbf{B}$ in a perfectly conducting, homogeneous plasma as $\sigma \longrightarrow \infty$ and $D_m \longrightarrow 0$ (e.g., see [18, 19]). Here, $\mathbf{A}$ is the magnetic vector potential: $\mathbf{B} = \nabla\times\mathbf{A}$.

Evolution equations for $\langle n(1)n(2)\rangle$ and $\langle B_i(1)B_j(2)\rangle$ should be derived under the condition that conservation laws (5) and (6) hold; i.e., the aforementioned difference of two terms of similar order of magnitude vanishes in an integral sense. Note that the conservation laws are satisfied automatically in the frequently employed approximation of $\delta(\tau)$-correlated velocity field ($\langle u_i(1)u_j(2)\rangle \propto \delta(t_1 - t_2)$), since $G(\mathbf{R}, 0) = \delta(\mathbf{R})$ for any Green function by definition. In [20] and other studies, a model was proposed for the evolution of magnetic field fluctuations in a non-$\delta(\tau)$-correlated velocity field. However, the Bethe–Salpeter-type equation derived therein is not consistent with conservation law (6). Therefore, the results obtained in these studies must be incorrect, at least, quantitatively.

Linear stochastic equations (1) and (2) imply that the averaged quantities $\langle n\rangle$ and $\langle \mathbf{B}\rangle$ are related to the fluctuations n' and $\mathbf{B}'$, and vice versa. Therefore, an attempt to write a separate equation for the averaged Green function $\langle G(1, 2)\rangle$ of Eqs. (1) and (2) leads to a hierarchy of nonlinear equations for $\langle G\rangle$. This situation is completely analogous to the classical closure problem in turbulence theory.

It is frequently assumed that the turbulent velocity ensemble is Gaussian; i.e., the averaged product of an odd number of velocity components vanishes while the averaged product of an even number of components is expressed as the sum of terms containing all possible two-point correlators. Under this assumption, a Dyson equation can be written for $\langle G(1, 2)\rangle$ [13]. Formally, it is a linear integral equation with a kernel expressed as the sum of an infinite number of terms describing elementary interactions between the turbulent flow and the passive field (so-called connected, or irreducible, interactions). Similarly, one can easily derive a Bethe–Salpeter-type linear integral equation for the averaged quantity $\langle G(1, 3)G(2, 4)\rangle$, with a kernel also expressed as the sum of an infinite number of terms corresponding to irreducible interactions. Retaining a progressively increasing number of terms in the kernels of these equations, one obtains a hierarchy of linear equations for $\langle G(1, 2)\rangle$ and $\langle G(1, 3)G(2, 4)\rangle$. However, the linear equations for $\langle G(1, 2)\rangle$ are not very useful. In particular, retaining the first term in the kernel (in the Bourret approximation [15]), one can calculate D_t only for $\xi_0 = u_0\tau_0/R_0 \ll 1$. This follows already from the fact that the Green function of the Bourret equation tends to that of the standard diffusion equation in the diffusion (long time, long distance) limit, with the diffusivity $D_t = u_0^2\tau_0/3$ obtained as the first term of the expansion in terms of ξ_0 in the general theory [9, 12].

A more effective calculation of D_t relies on the hierarchy of nonlinear equations for $\langle G(1, 2)\rangle$. In particular, even the first equation of the hierarchy (direct interaction approximation equation [16]), which involves a quadratic nonlinearity, can be used to calculate $D_t \equiv D_t^{(0)}$ for any value of ξ_0. When the next term (containing irreducible interaction of the fourth order in velocity) is retained in the hierarchy, the resulting correction to D_t (denoted here by $D_t^{(1)}$) is a negative quantity increasing from zero at $\xi_0 = 0$ to approximately $0.1D_t^{(0)}$ as $\xi_0 \longrightarrow \infty$ [21]. Thus, solving even the first two equations of the nonlinear hierarchy for $\langle G(1, 2)\rangle$ makes it

possible to calculate turbulent diffusivity almost exactly as $D_t = D_t^{(0)} + D_t^{(1)}$.

The nonlinear equations take into account a much greater number of elementary interactions between the turbulent flow and the passive field, as compared to the corresponding ones in the hierarchy of linear Dyson equations. Mathematically, this implies that the asymptotic behavior of the averaged Green function at $\xi_0 \gg 1$ is different from that of the Green function of the Bourret equation (see details below). This makes it possible to calculate D_t for large values of ξ_0. However, since D_t is primarily determined by large-scale turbulent motion, the accuracy of description of small-scale diffusion by the first equations of the nonlinear hierarchy remains an open question. The expansion $D_t = D_t^{(0)} + D_t^{(1)} + \ldots$ is an asymptotic series even if it is based on the nonlinear hierarchy for the Green function, because the number of different interactions allowed for in successive approximations rapidly increases. (Recall that the nth approximation for a Gaussian velocity field contains $(2n-1)!!$ different terms!) Even though this number is somewhat reduced in a nonlinear analysis (see below), a mere improvement of the asymptotic convergence of the series is achieved as a result.

In contrast to the hierarchy of linear Dyson equations, a hierarchy of nonlinear equations can be derived for a non-Gaussian velocity ensemble as well. This possibility can be used to analyze the influence of non-Gaussian velocity statistics on the behavior of the averaged Green function, in particular, on the value of turbulent diffusivity. The nonlinear hierarchy is derived in this paper. It is well known that real turbulence is non-Gaussian [22]. Furthermore, the nonlinear hierarchy is characterized by a reduced number of irreducible interactions in the kernels of the equations. For example, the nonlinear hierarchy contains only one irreducible fourth-order correlator instead of two in the corresponding Dyson equation, four sixth-order correlators instead of nine, and so on.

Evolution of $\langle n'(1)n'(2)\rangle$ and $\langle B_i'(1)B_j'(2)\rangle$ is determined by the averaged quantity $\langle G'(1, 3)G'(2, 4)\rangle$ (i.e., by the fluctuating parts of the Green function). Since conservation laws (5) and (6) involve $\langle n^2(1)\rangle = \langle n(1)\rangle\langle n(1)\rangle + \langle n'(1)n'(1)\rangle$ and $\langle B_i(1)B_j(1)\rangle$, it is reasonable to consider the correlation functions $V(1, 2)$ and $H_{ij}(1, 2)$ defined by (3) and (4). The mean quantities $\langle n\rangle$ and $\langle \mathbf{B}\rangle$ satisfy diffusion equations and decay over a short characteristic time $t_* \approx R_0^2/12D_t$. At $t \geq t_*$, the correlators $V(1, 2)$ and $H_{ij}(1, 2)$ represent only the fluctuations $\langle n'(1)n'(2)\rangle$ and $\langle B_i'(1)B_j'(2)\rangle$.

The present derivation of integrodifferential equations for $V(1, 2)$ and $H_{ij}(1, 2)$ (resulting in a hierarchy of Bethe–Salpeter-type equations in the case of a Gaussian ensemble) shows that conservation laws (5) and (6) hold only if the averaged Green function $\langle G(1, 2)\rangle \equiv G_0(1, 2)$ satisfies the hierarchy of nonlinear equations derived in the next section. In particular, this means that the widely used Bourret equation for the averaged Green function, being a linear one, cannot be applied to the evolution of fluctuation intensity, because this leads to violation of conservation laws (5) and (6).

Note that the simplest equations for $V(1, 2)$ and $H_{ij}(1, 2)$ (where only the ladder diagrams describing interactions between passive field and turbulence are retained) require that the function $\langle G(1, 2)\rangle$ satisfy the first equation in the nonlinear hierarchy (the DIA equation). Allowance for interactions of the next order, which are described by a fourth-order irreducible correlator of $\mathbf{u}(\mathbf{r}, t)$, requires that $\langle G(1, 2)\rangle$ satisfy the next equation in the hierarchy, which contains fourth-order correlators; and so on. Thus, conservation laws (5) and (6) will hold if the hierarchical equations for $V(1, 2)$ and $H_{ij}(1, 2)$ are consistent with the corresponding equations in the nonlinear hierarchy for $\langle G(1, 2)\rangle$. It has been pointed out that the Bethe–Salpeter equations may not be consistent with conservation laws (see [23] and references cited therein). However, integrodifferential equations for $V(1, 2)$ and $H_{ij}(1, 2)$ consistent with (5) and (6) are derived for the first time in this paper.

The paper is organized as follows. First, a hierarchy of nonlinear equations for the averaged Green function $\langle G(1, 2)\rangle$ is derived from Eqs. (1) and (2) simultaneously with increasingly complex expressions for the fluctuating part $G'(1, 2)$ of the Green function. These results are somewhat similar to those obtained in [12], but are obtained by a different method. It is important that the resulting hierarchy of equations is not restricted to Gaussian turbulence. Next, correct equations are derived for the correlators $V(1, 2)$ and $H_{ij}(1, 2)$ up to sixth-order irreducible velocity correlators. After that, equations for fluctuation spectra are written out and some asymptotic expressions are presented for $\langle n^2(\mathbf{r}, t)\rangle$ and $\langle B^2(\mathbf{r}, t)\rangle$ predicted by models of delta-correlated turbulent field and turbulence with a finite correlation time. (Owing to its relative mathematical simplicity, the latter model is still widely used to describe turbulent advection of passive fields, as in [24].) By comparing these expressions, it is shown that the delta-correlated approximation applies only to turbulence with $\xi_0 \ll 1$ and substantially overestimates the fluctuation intensity for both passive scalar and magnetic field.

2. NONLINEAR EQUATIONS FOR THE GREEN FUNCTION

The equations for the Green functions $G(1, 2)$ of Eqs. (1) and (2) have the general form

$$\left(\frac{\partial}{\partial t_1} - D_m\nabla_1^2\right)G(1, 2) \equiv L_0(1)G(1, 2) = L(1)G(1, 2) + I\delta(\mathbf{R})\delta(\tau), \quad (7)$$

where $L(1)G(1, 2) = -\mathrm{div}[\mathbf{u}(1)G(1, 2)]$ for Eq. (1), while both Green function and operator $L(1)$ of Eq. (2) are tensors:

$$L_{ik}(1)G_{kj}(1, 2) = [\nabla_k^{(1)}u_i(1) - \delta_{ik}\nabla_t^{(1)}u_t(1)]G_{kj}(1, 2).$$

Recall that $\mathbf{R} = \mathbf{r}_1 - \mathbf{r}_2$, $\tau = t_1 - t_2$, summation over repeated indices is assumed, and $I_{ij} = \delta_{ij}$. The Green functions vanish at negative τ; i.e., they can be represented as $G(1, 2) \equiv \theta(\tau)g(1, 2)$, where $\theta(\tau)$ is the Heaviside step function ($\theta(\tau) = 0$ at $\tau < 0$ and $\theta(\tau) = 1$ at $\tau > 0$).

The Green function of the left-hand side of Eq. (7),

$$G_\mathrm{m}(1-2) = \theta(\tau)(4\pi D_\mathrm{m}\tau)^{-3/2}\exp(-R^2/4D_\mathrm{m}\tau),$$

can be used to write an integral equation for $G(1, 2)$:

$$G(1, 2) = G_\mathrm{m}(1-2) + \int d3\,G_\mathrm{m}(1-3)L(3)G(3, 2). \tag{8}$$

In what follows, diagrammatic notation is used to abbreviate cumbersome equations and demonstrate their symmetry:

$$G(1, 2) = \bigcirc\!\!\!\!-\!\!\!-, \quad G_\mathrm{m}(1-2) = —,$$

$$L(1) = \circ, \quad \langle G(1, 2)\rangle \equiv G_0(1, 2) = \oplus,$$

$$\bigcirc\!\!\!\!-\!\!\!- = \oplus + \bigcirc\!\!\!\!-\!\!\!-'.$$

In this notation, Eq. (8) has the form

$$\bigcirc\!\!\!\!-\!\!\!- = — + —\circ\bigcirc\!\!\!\!-\!\!\!-. \tag{9}$$

A comparison of (8) with (9) shows that the integral is taken over the coordinates of the interaction operator (circle) while the outer coordinates 1 and 2 are held fixed. Iterating Eq. (9), one obtains the series

$$\bigcirc\!\!\!\!-\!\!\!- = — + —\circ— + —\circ—\circ— + \ldots. \tag{10}$$

Averaging the series over the ensemble of $\mathbf{u}(\mathbf{r}, t)$ leads to the expressions

$$\oplus = — + —\langle\circ\bigcirc\!\!\!\!-\!\!\!-\rangle \equiv — + —\langle\circ\bigcirc\!\!\!\!-\!\!\!-\circ\rangle—. \tag{11}$$

It is assumed here that the medium as a whole is at rest, i.e., $\langle\mathbf{u}\rangle = 0$. Applying the operator L to series (10), one obtains the relation

$$—\circ\bigcirc\!\!\!\!-\!\!\!- = \bigcirc\!\!\!\!-\!\!\!-\circ— = \bigcirc\!\!\!\!-\!\!\!- - —. \tag{12}$$

Averaging this relation and using (11), one has

$$—\langle\circ\bigcirc\!\!\!\!-\!\!\!-\rangle = \langle\bigcirc\!\!\!\!-\!\!\!-\circ\rangle— = —\langle\circ\bigcirc\!\!\!\!-\!\!\!-\circ\rangle—. \tag{13}$$

Combining (11) with (12) yields two expressions for the fluctuating part of the Green function $G'(1, 2) = \bigcirc\!\!\!\!-\!\!\!-' = \bigcirc\!\!\!\!-\!\!\!- - \oplus$:

$$\begin{aligned}\bigcirc\!\!\!\!-\!\!\!-' &= —\circ\bigcirc\!\!\!\!-\!\!\!- - —\langle\circ\bigcirc\!\!\!\!-\!\!\!-\circ\rangle—,\\ \bigcirc\!\!\!\!-\!\!\!-' &= \bigcirc\!\!\!\!-\!\!\!-\circ— - —\langle\circ\bigcirc\!\!\!\!-\!\!\!-\circ\rangle—.\end{aligned} \tag{14}$$

To eliminate the molecular-diffusion Green function $G_\mathrm{m}(1–2)$ from these formulas, replace the outer $G_\mathrm{m} =$ — in the second term of the first equation with

$$— = \bigcirc\!\!\!\!-\!\!\!- - —\circ\bigcirc\!\!\!\!-\!\!\!-,$$

which follows from (12); replace — in the inner part of the second term in the second equation with

$$— = \bigcirc\!\!\!\!-\!\!\!- - \bigcirc\!\!\!\!-\!\!\!-\circ—,$$

and use (11) and (13) in the resulting expressions to obtain the equivalent formulas

$$\begin{aligned}\bigcirc\!\!\!\!-\!\!\!-' &= \oplus\circ\bigcirc\!\!\!\!-\!\!\!- - \langle\bigcirc\!\!\!\!-\!\!\!-\circ\rangle\bigcirc\!\!\!\!-\!\!\!-,\\ \bigcirc\!\!\!\!-\!\!\!-' &= \bigcirc\!\!\!\!-\!\!\!-\circ\oplus - \bigcirc\!\!\!\!-\!\!\!-\langle\circ\bigcirc\!\!\!\!-\!\!\!-\rangle.\end{aligned} \tag{15}$$

Adding the averaged Green function $G_0(1, 2)$ to these formulas, one finally obtains equivalent equations for the total Green function that do not contain G_m:

$$\begin{aligned}\bigcirc\!\!\!\!-\!\!\!- &= \oplus + \oplus\circ\bigcirc\!\!\!\!-\!\!\!- - \langle\bigcirc\!\!\!\!-\!\!\!-\circ\rangle\bigcirc\!\!\!\!-\!\!\!-,\\ \bigcirc\!\!\!\!-\!\!\!- &= \oplus + \bigcirc\!\!\!\!-\!\!\!-\circ\oplus - \bigcirc\!\!\!\!-\!\!\!-\langle\circ\bigcirc\!\!\!\!-\!\!\!-\rangle.\end{aligned} \tag{16}$$

Equations (15) and (16) play a key role in the derivation of a hierarchy of nonlinear equations for $\langle G(1, 2)\rangle \equiv G_0(1, 2)$. Equations (16) written out in literal form are

$$\begin{aligned}G(1, 2) &= G_0(1, 2)\\ &+ \int d3[G_0(1, 3)L(3) - \langle G(1, 3)L(3)\rangle]G(3, 2),\\ G(1, 2) &= G_0(1, 2)\\ &+ \int d3\,G(1, 3)[L(3)G_0(3, 2) - \langle L(3)G(3, 2)\rangle].\end{aligned} \tag{17}$$

To verify that $\langle G'(1, 2)\rangle = 0$, by virtue of (15), one must show that

$$\oplus\langle\circ\bigcirc\!\!\!\!-\!\!\!-\rangle = \langle\bigcirc\!\!\!\!-\!\!\!-\circ\rangle\oplus. \tag{18}$$

This is done by inserting (11) into both sides of Eq. (18) and using (13).

Note that expressions (15)–(18) are valid when $\langle\mathbf{u}(\mathbf{r}, t)\rangle \neq 0$ as well. Their derivation follows that presented above for $\langle\mathbf{u}\rangle = 0$, where one can set $G_0(1, 2) \equiv G_0(1 - 2)$ for stationary homogeneous turbulence.

Represent $G(1, 2)$ as a series in powers of L:

$$G(1, 2) = G_0(1-2) + G^{(1)}(1, 2) + G^2(1, 2) + \ldots. \tag{19}$$

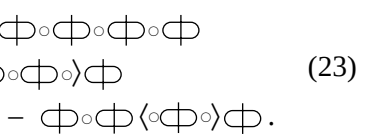
Equations (17) entail the recursion relation

$$G^{(n)}(1, 2) = \int d3\left[G^{(n-1)}(1, 3)L(3)G_0(3-2) - \sum_{k=2}^{n} G^{(n-k)}(1, 3)\langle L(3)G^{(k-1)}(3, 2)\rangle\right]. \tag{20}$$

The expression for $G^{(n)}(1, 2)$ approximates the fluctuating part of the Green function and contains n interaction operators L. Using (18), one can readily show that $\langle G^{(n)}(1, 2)\rangle = 0$. The analysis that follows relies on diagrammatic representations for the first three $G^{(n)}$ in the case of $\langle \mathbf{u}\rangle = 0$:


(21)

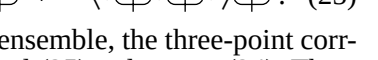

(22)


(23)

Substituting (21) into the first expression in (11), one obtains a nonlinear equation with a quadratic nonlinearity for the averaged Green function,

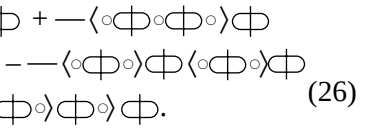

(24)

known as the DIA equation [16, 25]. Similarly, the substitution of the sum of (21) and (22) into (11) leads to an equation with a cubic nonlinearity:


(25)

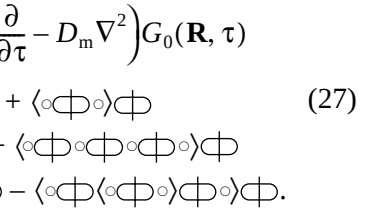
For a Gaussian velocity ensemble, the three-point correlator in (25) vanishes and (25) reduces to (24). Thus, Eq. (25) is the lowest order equation that reflects the non-Gaussian nature of real turbulence in the nonlinear hierarchy.

The next equation of the hierarchy is obtained by substituting $G'(1, 2) = G^{(1)} + G^{(2)} + G^{(3)}$ into (11):


(26)

The next-to-last term in this equation is disconnected. The 17 sixth-order terms contained in the next equation of the hierarchy include seven disconnected ones. Thus, the resulting hierarchy of nonlinear equations for a non-Gaussian ensemble of $\mathbf{u}(\mathbf{r}, t)$ is not a Dyson equation, because it contains disconnected terms. In the Gaussian limit, the disconnected (reducible) terms cancel out, and a nonlinear analogue of the hierarchy of Dyson equations is obtained.

Applying the operator $L_0(1)$ defined by (7) to Eq. (26), one obtains a nonlinear integrodifferential equation for $G_0(1 - 2)$:


(27)

In the case of a Gaussian velocity ensemble, this equation is much simpler and the sixth-order correlators can be included:

(28)

Here, the use of nested angle brackets $\langle\langle\ldots\rangle\rangle$ and superior braces is dictated by the complexity of combinations of averaged pairs of operators L.

If series (10) were averaged directly without proceeding to the derivation of the nonlinear hierarchy, and the connected (irreducible) terms were singled out in the averaged series, then the standard Dyson equation would be obtained [13]. This equation can be derived from (28) by replacing every inner $G_0(1, 2) =$ with the molecular-diffusion Green function $G_m(1 - 2) =$ and adding to the right-hand side the fourth-order irreducible term

(29)

and five additional sixth-order connected terms, which are not written out here. Terms analogous to (29) are already contained in (24), which can be demonstrated by developing an iterative series for this nonlinear equation. Similarly, additional sixth-order terms arise as the next equation of the hierarchy, Eq. (26), is iterated for a Gaussian velocity ensemble. Thus, the kernels in nonlinear equations contain fewer connected terms.

The connectedness of terms in (28) ensures that the kernels are narrowly supported. Therefore, the outer Green function can be expanded into a Taylor series in coordinates and time, and the first terms of the resulting expansion can be used to obtain a diffusion approximation:

$$\left(\frac{\partial}{\partial\tau} - D_m\nabla^2\right)G_0(R, \tau) = I\delta(\mathbf{R})\delta(\tau) + D_t(\tau)\nabla^2 G_0(R, \tau). \quad (30)$$

Retaining only the first equation in the hierarchy (with

quadratic nonlinearity), one obtains

$$D_t^{(0)}(\tau) = \frac{1}{3}\int d\mathbf{R}\int_0^\tau d\tau[\langle u_i(1)u_i(2)\rangle - \mathbf{R}\cdot\langle \mathbf{u}(1)\mathrm{div}\mathbf{u}(2)\rangle]G_0(R,\tau). \quad (31)$$

In the diffusion approximation, only the first term should be retained in the Taylor series expansion with respect to time, i.e., the value of the outer Green function at the point of the kernel's maximum on the time axis should be factored out (see details in [26]). An expression for $D_t^{(1)}$ (contribution of the fourth-order nonlinear terms in the hierarchy) was given in [27]. The accuracy of the expressions for $D_t^{(0)}$ and $G_t^{(1)}$ is discussed in Introduction. Expressions for $D_t^{(0)}$ and the α-effect factor $\alpha_t^{(0)}$ for magnetic field diffusion can be found in [26].

3. EQUATIONS FOR THE CORRELATORS $V(1, 2)$ AND $H_{ij}(1, 2)$

Suppose that the distribution of passive-scalar density in a turbulent medium at the initial moment $t = 0$ is $n_0(\mathbf{r})$. When the Green function $G(\mathbf{r}, t; \mathbf{r}', t')$ of Eq. (1) is known, the statistical properties of $n(\mathbf{r}, t)$ are determined by the expression

$$n(\mathbf{r}, t) \equiv \langle n(\mathbf{r}, t)\rangle + n'(\mathbf{r}, t) = \int d\mathbf{r}' G(\mathbf{r}, t; \mathbf{r}', 0)n_0(\mathbf{r}'). \quad (32)$$

An analogous expression relates the magnetic field $\mathbf{B}(\mathbf{r}, t)$ to its initial distribution $\mathbf{B}_0(\mathbf{r})$.

Using (32), one can easily write expressions for $V(1, 2) = \langle n(1)n(2)\rangle$ and $H_{ij}(1, 2)$. The stochastic Green function $G(1, 2)$ is given by two expressions: series (10) in molecular-diffusion Green function $G_m(1 - 2)$ and series (19) in the averaged Green function $G_0(1, 2)$. For the commonly used Gaussian ensemble of $\mathbf{u}(\mathbf{r}, t)$, the averaging of the expressions for $n(1)n(2)$ and $B_i(1)B_j(2)$ leads to a hierarchy of Bethe–Salpeter-type linear integral equations (see [13]) differing only by the number of connected (irreducible) terms retained in their kernels. These equations have the following general form:

$$V(1, 2) = \int d\mathbf{3}\int d\mathbf{4}G_0(1; \mathbf{3}, 0)G_0(2; \mathbf{4}, 0)V_0(\mathbf{3}, \mathbf{4}) + \int d3\int d4\int d5\int d6G_0(1, 3)G_0(2, 4) \times K(3, 4; 5, 6)V(5, 6), \quad (33)$$

where $V_0(\mathbf{3}, \mathbf{4}) = \langle\langle n_0(\mathbf{r}_3)n_0(\mathbf{r}_4)\rangle\rangle$. Here, double brackets denote averaging over the ensemble of initial particle-density distributions. The ensemble is generally assumed to be homogeneous and isotropic, in which case $V_0(\mathbf{3}, \mathbf{4}) = V_0(|\mathbf{3} - \mathbf{4}|)$. The absolute and variable terms in (33) are $\langle n(1)\rangle\langle n(2)\rangle$ and the fluctuation correlator $\langle n'(1)n'(2)\rangle$, respectively. Equation (33) can readily be used to write an integral equation for $\langle n'(1)n'(2)\rangle$, which is more complex than (33). This equation is not employed here.

The analysis is simplified by applying the operator $L_0(1)$ or $L_0(2)$ to Eq. (33) and using an equation for the averaged Green function (Eq. (28) in the present context) to obtain an integrodifferential equation for $V(1, 2)$ independent of $V_0(\mathbf{3}, \mathbf{4})$.

The simplest form (ladder approximation) of Eq. (33) is obtained when only the first connected term is retained in the kernel. In the case of a $\delta(\tau)$-correlated field, the ladder approximation yields an exact solution, because the contribution of remaining connected terms to (33) vanishes (e.g., see [17]). In the non-$\delta(\tau)$-correlated case, this equation is written as

$$V(1, 2) = \int d\mathbf{3}\int d\mathbf{4}G_0(1; \mathbf{3}, 0)G_0(2; \mathbf{4}, 0)V_0(\mathbf{3}, \mathbf{4}) + \int d3\int d4G_0(1, 3)G_0(2, 4) \times \nabla_i^{(3)}\nabla_j^{(4)}\langle u_i(3)u_j(4)\rangle V(3, 4). \quad (34)$$

To abbreviate diagrammatic representation of more complicated equations, an additional symbol is defined by writing Eq. (34) as

(35)

Henceforth, it should be borne in mind that coordinates 1 and 2 correspond to the extreme left and extreme right elements of a diagram, respectively. The operators on the right of a box (which represents $V(3, 4)$) act on the terms on its left, and time decreases toward the box. The box crossed by a vertical line represents $V_0(\mathbf{3}, \mathbf{4})$.

When operator $L_0(1)$ or $L_0(2)$ is applied to (35), one has to determine how many terms should be retained in Eq. (28). It was noted in Introduction that conservation law (5) will hold only if the orders of the nonlinear equation for the averaged Green function $\oplus \equiv G_0$ and in the obtained integrodifferential equation for $V(1, 2)$ are matched. This implies that only the first integral term must be retained in (28), i.e., the analysis must be restricted to the DIA equation for $G_0(1, 2)$. The resulting equations are

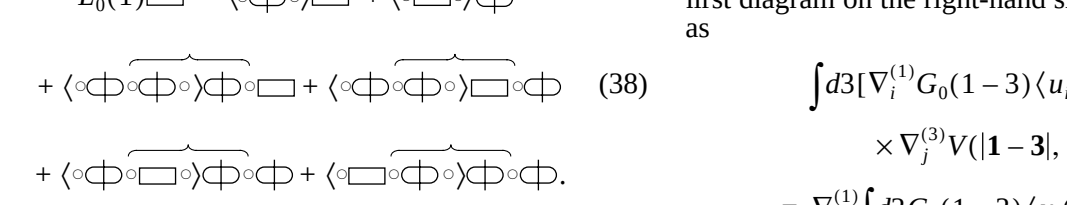
(36)

(37)

By virtue of the symmetry $V(1, 2) = V(2, 1)$, Eq. (37) is obtained from Eq. (36) by interchanging the points

indexed by 1 and 2, which means inversion of diagrams. This is true in any approximation.

The terms on the right-hand side of (36) can be obtained from the first diagram on the right-hand side of (28) by successively replacing one of the averaged Green functions $G_0(1, 2)$ with $V(1, 2)$. A direct verification shows that this rule applies to higher order equations in the hierarchy of Bethe–Salpeter equations as well. In particular, allowing for fourth-order irreducible interactions (see (28)), one obtains



(38)

The equation containing the sixth-order correlators has an analogous form. Since Eq. (28) has four sixth-order terms with six Green functions G_0 in each, the total number of sixth-order terms in (38) is 24. This equation is not written out here, because even its diagrammatic representation is too cumbersome. Note also that the nonlinear equations for $G_0(1, 2)$ again contain fewer connected terms, as compared to the Bethe–Salpeter equations derived directly from series (10) (see [13]). However, this reduction starts only from the sixth-order correlators: the 26 connected terms contained in the linear theory [13] reduce to 20. Each eliminated term has the form

i.e., it contains a disconnected element inside a pair of averaging brackets. Note that the Bethe–Salpeter equations are commonly represented as "two-level" diagrams (see [13, 17]). The diagrammatic expressions presented here are more compact, and their symmetry properties are demonstrated more explicitly. In particular, this representation is consistent with the rule of successive replacement of terms in the Bethe–Salpeter equation with $V(1, 2)$ (see Eq. (38)). Of course, these equations can readily be recast in the standard two-level form.

The evolution of $\langle n^2(\mathbf{r}, t)\rangle \equiv V(1, 1)$ is of primary interest. The derivative $d\langle n^2(\mathbf{r}, t)\rangle/dt$ is obtained by adding Eqs. (36) and (37) and setting $t_1 = t_2 = t$. This makes the right-hand sides of (36) and (37) equal, and twice the right-hand side of (36) (or (38) in the next equation of the hierarchy) can be taken. By virtue of conservation law (5), the right-hand side of the equation for $d\langle n^2(\mathbf{r}, t)\rangle/dt$ must vanish if $D_m = 0$ for incompressible turbulence. It is important that the particle distribution must also be statistically homogeneous, i.e., $V(1, 2) = V(R, t_1, t_2)$. Violation of (5) would lead to spurious effects: $\langle n^2(\mathbf{r}, t)\rangle$ would rapidly increase if the first term on the right-hand side of (36) is less than the second one, and vice versa. Equations (36), (38), and others obtained here ensure that conservation law (5) is consistent with each equation in the Bethe–Salpeter hierarchy and the spurious effects are ruled out. However, it remains unclear if these equations can adequately describe the spectral distribution of scalar field fluctuations.

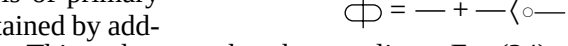
To show that Eq. (36) (the simplest one in the hierarchy) is consistent with conservation law (5), set $1 = 2$ (i.e., $\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}$ and $t_1 = t_2 = t$) and $\operatorname{div}\boldsymbol{u} = 0$. Then, the first diagram on the right-hand side can be represented as

$$\begin{aligned}
&\int d3\,[\nabla_i^{(1)} G_0(1-3)\langle u_i(1)u_j(3)\rangle] \\
&\qquad \times \nabla_j^{(3)} V(|\mathbf{1}-\mathbf{3}|, t_3, t_1) \\
&= \nabla_i^{(1)} \int d3\, G_0(1-3)\langle u_i(1)u_j(3)\rangle \\
&\qquad \times \nabla_j^{(3)} V(|\mathbf{1}-\mathbf{3}|, t_3, t_1) \\
&\qquad - \int d3\, G_0(1-3)\langle u_i(1)u_j(3)\rangle \\
&\qquad \times \nabla_i^{(1)}\nabla_j^{(3)} V(|\mathbf{1}-\mathbf{3}|, t_3, t_1).
\end{aligned} \tag{39}$$

Since the correlator $\langle u_i(1)u_j(3)\rangle \equiv B_{ij}(1-3)$ depends on the difference of the arguments, the integral in the first term on the right-hand side of (39) is independent of $\mathbf{r}_1$ and therefore vanishes. The second term is obviously equal to minus the second diagram in (36) (since the latter is invariant under inversion if $1 = 2$). Thus, the right-hand side of Eq. (36) vanishes when $1 = 2$. In other words, the extreme left operator L (circle) can be placed at the extreme right position in the diagram (with opposite sign).

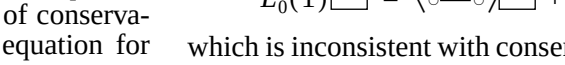
Similarly, the first fourth-order term in (38) cancels out with the last one, and the second and third terms cancel out as well. When $1 = 2$, the terms of the equation containing 24 sixth-order correlators also cancel out pairwise; i.e., conservation law (5) holds for these equations as well.

In [20] and other studies, Eq. (35) was considered for the magnetic field correlator $H_{ij}(1, 2) = \langle B_i(1)B_j(2)\rangle$, but the averaged Green function was assumed to satisfy the Bourret equation



(40)

rather than nonlinear Eq. (24), which would be correct in the ladder approximation. This led to the equation



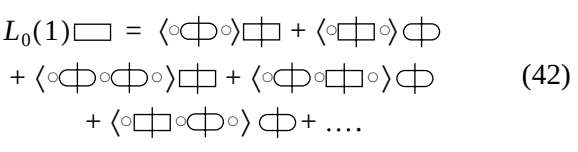
(41)

which is inconsistent with conservation law (6), instead of (36) (here, boxes represent correlators H_{ij}). Therefore, the quantitative results obtained in [20] are not valid.

The foregoing analysis shows that conservation laws (5) and (6) ensure that the hierarchy of Bethe–Salpeter equations is correct if the orders of the nonlinear

equation for the averaged Green function $G_0(1-2)$ and the corresponding Bethe–Salpeter equation are equal. One may use any plausible expression for the averaged Green function (e.g., a diffusion Green function) in properly structured Eqs. (36) and (38). The conservation laws will hold. It is important that the expressions for the averaged Green functions in the kernels of the equations be identical, whereas different expressions are contained in (41). Equation (41) is equivalent to (36) only in the model of delta-correlated turbulent field ($\langle u_i(1)u_j(2)\rangle \propto \delta(t_1 - t_2)$), when the Green functions in the first terms on the right-hand sides of these equations reduce to $\delta(\mathbf{R})$.

The analysis presented above made use of the common assumption of Gaussian turbulent-velocity ensemble. Real turbulent fields are not Gaussian [22]. The nonlinear equations for the averaged Green function $G_0(1, 2)$ (see Eqs. (26) and (27)) and the expressions for the fluctuating part $G'(1, 2)$ of the Green function (see (21)–(23)) are obtained here for an arbitrary velocity ensemble. Using (21)–(23), one can easily verify that $\langle G'(1, 3)G'(2, 4)\rangle$ cannot be decomposed into connected and disconnected parts. For example, the terms of fourth order in L do not involve the corresponding fourth-order ladder diagram. Thus, no equation of Bethe–Salpeter type can be written. One can only directly use the approximate expressions for $G'(1, 2)$ given by (21)–(23) to write down an approximate series expansion for $\langle n(1)n(2)\rangle = V(1, 2)$. Applying the operator $L_0(1)$ or $L_0(2)$ to the series and making use of (27), one obtains



(42)

Here, the ten fourth-order terms are not written out. In contrast to the Gaussian case, these terms do not satisfy the rule of successive replacement of one of the averaged Green functions in (27) with the initial correlator $V_0(1-2) =$  . The method applied to Eq. (39) can readily be used here to show that expression (42) is consistent with conservation law (5). Note that the third term on the right-hand side of (42) vanishes when 1 = 2. As in the Gaussian case, the order of expression (42) equals that of the equation for $G_0(1-2)$.

Once again, recall that the nonlinearity of (27) ensures good asymptotic convergence of (42) even when the parameter $\xi_0 = u_0\tau_0/R_0$ is large. Equation (42) can be used as a basis for analyzing non-Gaussian behavior of the correlation functions $V(1, 2)$ and H_{ij} if an expression for the three-point velocity correlator is known or prescribed.

All results obtained in this section can be applied to describe the correlator $H_{ij}(1, 2) = \langle B_i(1)B_j(2)\rangle$, in which case both operator $L(1)$ and Green functions are tensors. Magnetic-helicity conservation (6) was verified only for Eq. (36) (with a quadratic nonlinearity), because of the complexity of further calculations.

4. EQUATIONS FOR PASSIVE-SCALAR SPECTRA

Restricting analysis to the case of a Gaussian velocity ensemble, consider first Eqs. (24) and (36) (with quadratic nonlinearity) and then Eqs. (28) and (38) (of fourth order in $L(1) = -\nabla_i^{(1)} u_i(1)$). The spectrum of the correlator $V(1, 2) = V(R; t_1, t_2)$ is related to the Fourier integral

$$V(R; t_1, t_2) = \frac{1}{(2\pi)^3}\int d\mathbf{p} e^{i\mathbf{p}\cdot\mathbf{R}}\tilde{V}(p; t_1, t_2) \qquad (43)$$

as follows:

$$V(0; t_1, t_2) = \int_0^\infty dp E_V(p; t_1, t_2),$$
$$E_V(p; t_1, t_2) = \frac{p^2}{2\pi^2}\tilde{V}(p; t_1, t_2). \qquad (44)$$

Therefore, it is sufficient to analyze the equations for $\tilde{V}(p; t_1, t_2)$.

A stationary, homogeneous, and isotropic turbulent velocity ensemble is characterized by the correlator $B_{ij}(\mathbf{R}, t) = \langle u_i(1)u_j(2)\rangle$ ($\mathbf{R} = \mathbf{r}_1 - \mathbf{r}_2$, $\tau = t_1 - t_2$), and its Fourier transform has the form [28]

$$\tilde{B}_{kj}(\mathbf{p}, \tau) = (\delta_{kj}p^2 - p_k p_j) f(p, \tau) + p_k p_j W(p, \tau) + ie_{kjt}p_t D(p, \tau), \qquad (45)$$

where e_{kjt} is the Levi-Civita permutation symbol ($e_{xyz} = -e_{yxz} = 1$ etc.). The functions $f(p, \tau)$, $W(p, \tau)$, and $D(p, \tau)$ characterize turbulence spectra:

$$\langle \mathbf{u}(\mathbf{r}, t)\cdot\mathbf{u}(\mathbf{r}, t+\tau)\rangle = \int_0^\infty dp E(p, \tau),$$
$$E(p, \tau) = E_{\text{inc}}(p, \tau) + E_{\text{compr}}(p, \tau),$$
$$E_{\text{inc}}(p, \tau) = p^4 f(p, \tau)/\pi^2,$$
$$E_{\text{compr}}(p, \tau) = p^4 W(p, \tau)/2\pi^2. \qquad (46)$$

For incompressible turbulent flow, $W(p, \tau) = 0$. The helicity spectrum is characterized by the function $D(p, \tau)$:

$$H(\tau) \equiv \langle \mathbf{u}(\mathbf{r}, t)\cdot(\nabla\times\mathbf{u}(\mathbf{r}, t+\tau))\rangle = \int_0^\infty dp E_{\text{h}}(p, \tau),$$
$$E_{\text{h}}(p, \tau) = -p^4 D(p, \tau)/\pi^2. \qquad (47)$$

The equation for $\tilde{V}(p, t_1, t_2)$ in the ladder approxi-

mation is the Fourier transform of Eq. (36):

$$\left(\frac{\partial}{\partial t_1} + D_m p^2\right)\tilde{V}(p, t_1, t_2) = \frac{1}{(2\pi)^3}$$
$$\times \int d\mathbf{q}\left[-\int_0^{t_1} dt' p_i \tilde{B}_{ij}(\mathbf{q}, |t_1 - t'|)(\mathbf{p} - \mathbf{q})_j \right. \quad (48)$$
$$\times \tilde{g}(|\mathbf{p} - \mathbf{q}|, t_1 - t')\tilde{V}(p, t', t_2)$$
$$\left. + \int_0^{t_2} dt' p_i \tilde{B}_{ij}(\mathbf{q}, |t_1 - t'|) p_j \tilde{g}(p, t_2 - t')\tilde{V}(|\mathbf{p} - \mathbf{q}|, t_1, t')\right],$$

where the representation $\tilde{G}_0(p, \tau) \equiv \theta(\tau)\tilde{g}(p, \tau)$ is used ($\theta(\tau)$ is the Heaviside step function, see the beginning of Section 2). It can readily be shown that (48) does not contain turbulent helicity; i.e., $\tilde{B}_{ij}$ can be treated as a symmetric tensor. Equation (48) holds for $\tilde{L}_0(p, t_2)\tilde{V}(p, t_1, t_2) \equiv \tilde{L}_0(p, t_2)\tilde{V}(p, t_2, t_1)$ (see (37)), under the change $t_1 \longrightarrow t_2$ and $t_2 \longrightarrow t_1$.

By virtue of conservation law (5), the integral of the right-hand side of (48) over $\mathbf{p}$ must vanish if $\mathrm{div}\mathbf{u} = 0$. This can readily be verified. It may seem that Eq. (48) can be analyzed in a diffusion approximation under the assumption that $p \ll p_0 \approx 1/R_0$, $t_1 \gg \tau_0$, and $t_2 \gg \tau_0$ by analogy with Eq. (28) for the averaged Green function. However, the resulting equation is integrodifferential because the second term contains $\tilde{V}(q, t_1, t')$. More importantly, this equation is not consistent with conservation law (5).

The use of the Bourret equation (see (41)) is equivalent to replacing $\tilde{g}(|\mathbf{p} - \mathbf{q}|, t_1 - t')$ in (48) with the molecular-diffusion Green function $\tilde{g}_m(|\mathbf{p} - \mathbf{q}|, t_1 - t') = \exp[-D_m(\mathbf{p} - \mathbf{q})^2(t_1 - t')]$, which can be approximated by unity if the diffusivity D_m is small. The first term in (48) describes the decrease in particle density due to turbulent diffusion. In the Bourret approximation defined by (41), its contribution is overestimated and conservation law (5) is violated again.

Setting $t_1 = t_2 = t$ in the sum of (48) with its counterpart for $\tilde{L}_0(2)\tilde{V}(p, t_1, t_2)$ and integrating the result over $\mathbf{p}$, one obtains

$$\frac{d}{dt}\langle n^2(\mathbf{r}, t)\rangle = -2D_m\langle(\nabla n(\mathbf{r}, t))^2\rangle$$
$$+ \int_0^t d\tau \int_0^\infty dq E_V(q, t, t - \tau) \quad (49)$$
$$\times \int_0^\infty dp \int_{-1}^1 d\mu E_{compr}(p, \tau)\tilde{g}(|\mathbf{p} + \mathbf{q}|, \tau)(p^2 + pq\mu),$$

where μ is the cosine of the angle between the vectors $\mathbf{p}$ and $\mathbf{q}$. For an incompressible flow ($\mathrm{div}\mathbf{u} = 0$), $E_{compr}(p, \tau) \equiv 0$ and Eq. (49) reduces to conservation law (5). The second term on the right-hand side of (49) is an approximate expression for the term with $\mathrm{div}\mathbf{u}$ in (5) that corresponds to the ladder approximation of the Bethe–Salpeter equation. The term containing $\mathrm{div}\mathbf{u}$ in (5) is responsible for particle clustering [14], i.e., formation of regions of excessively high and low particle density in a compressible turbulent flow.

In the frequently used model of delta-correlated turbulent field ($\tilde{B}_{ij}(\mathbf{p}, \tau) = \tau_0\delta(\tau)\tilde{B}_{ij}(\mathbf{p})$), Eq. (48) is exact since the remaining equations of the hierarchy vanish. In this model, a closed time-dependent equation can be written for $\tilde{V}(p, t, t)$:

$$\frac{\partial\tilde{V}(p, t, t)}{\partial t} = -2(D_m + D_t^{(0)})p^2\tilde{V}(p, t, t)$$
$$+ \frac{1}{2}p^2\tau_0\int_0^\infty dq \int_{-1}^1 d\mu[(1 - \mu^2)E_{inc}(q) + 2\mu^2 E_{compr}(q)] \quad (50)$$
$$\times \tilde{V}(|\mathbf{p} - \mathbf{q}|, t, t).$$

The diffusivity $D_t^{(0)}$ given by (31) can also be written as

$$D_t^{(0)}(t) = \frac{1}{3}\int_0^t d\tau \int_0^\infty dp\left\{[E_{inc}(p, \tau)\right.$$
$$\left. + E_{compr}(p, \tau)]\tilde{g}(p, \tau) + E_{compr}(p, \tau)p\frac{\partial\tilde{g}(p, \tau)}{\partial p}\right\}. \quad (51)$$

It should be noted that the last term in (51) vanishes ($\tilde{g}(p, 0) \equiv 1$) in the case of a $\delta(\tau)$-correlated field and compressibility ($\mathrm{div}\mathbf{u} \neq 0$) does not affect the diffusivity since $D_t^{(0)} = u_0^2\tau_0/3$ (diffusivity is determined by the total turbulent kinetic energy). However, this is not true for correlated scalar fluctuations, because the contributions of $E_{inc}(q)$ and $E_{compr}(q)$ to the exact Eq. (50) have different angle-dependent weights.

In the $\delta(\tau)$-correlated model under consideration, (49) is also an exact equation:

$$\frac{d}{dt}\langle n^2(\mathbf{r}, t)\rangle = -2D_m\langle(\nabla n(\mathbf{r}, t))^2\rangle$$
$$+ 2\tau_0\langle\mathrm{div}^2\mathbf{u}\rangle\langle n^2(\mathbf{r}, t)\rangle. \quad (52)$$

According to (52), the fluctuating part of $\langle n^2\rangle$ (i.e., clustering) grows at the initial stage, when the density is nearly uniform. However, $\langle(\nabla n)^2\rangle$ increases with the contribution of small-scale scalar fluctuations, and diffusive dissipation of fluctuations tends to play a dominant role.

At $t_1 \approx t_2 \gg \tau_0$ (or $t_1 \approx t_2 \gg t_0 = R_0/u_0$), both (48) and (49) can be approximated in terms of $\tilde{V}(p, t, t)$. Let us consider two cases. At moderate t ($t \geq \tau_0$), when the scalar turbulence has not yet concentrated in small-scale motions, it can be assumed that $\tilde{V}(q, t, t-\tau) \approx \tilde{V}(q, t, t)$ has a maximum at $q \ll q_0 \approx 1/R_0$. Using the series expansion of $\tilde{g}(|\mathbf{p}-\mathbf{q}|, \tau)$ in powers of the small parameter q, one obtains

$$\frac{d}{dt}\langle n^2(\mathbf{r}, t)\rangle = -2(D_m - D_t')\langle(\nabla n(\mathbf{r}, t))^2\rangle + 2C\langle n^2(\mathbf{r}, t)\rangle, \quad (53)$$

where

$$C = \int_0^\infty dp \int_0^\infty d\tau p^2 E_{compr}(p, \tau)\tilde{g}(p, \tau), \quad (54)$$

$$D_t' = \frac{1}{3}\int_0^\infty dp \int_0^\infty d\tau E_{compr}(p, \tau) p \frac{\partial \tilde{g}(p, \tau)}{\partial p}. \quad (55)$$

Here, D_t' is the direct contribution of compressibility ($\mathrm{div}\mathbf{u} \neq 0$) to the turbulent diffusivity $D_t^{(0)}$. It vanishes for a $\delta(\tau)$-correlated field. When acoustic effects in turbulence are negligible, $\partial\tilde{g}(p, \tau)/\partial p < 0$ and, therefore, $D_t' < 0$. This implies that chaotic shock waves strongly damp fluctuations in a compressible gas flow. Note that the coefficient C in (53) may be much smaller than the corresponding coefficient in (52), which leads to additional damping of fluctuations. In the limit of $\delta(\tau)$-correlated turbulence, Eq. (53) reduces to (52). In other words, Eqs. (53)–(55) can be interpreted as an extension of Eq. (52) to the initial stage of scalar-field evolution in turbulent flows with finite correlation times.

For "acoustic" turbulence,

$$E_{compr}(p, \tau) = E_{compr}(p)\cos(cp\tau)\exp[-k(p)p^2\tau], \quad (56)$$

and the method for calculating $D_t^{(0)}$ developed in [29] can be used to obtain

$$D_t' = \frac{\pi M^3}{9}\frac{u_0}{p_0}\int_0^\infty dx E_{compr}^2(x), \quad (57)$$

$$C = u_0 p_0 M \times \int_0^\infty dx x^2 E_{compr}(x)\left(\eta(x) + \frac{\pi}{6}M^2 E_{compr}(x)\right), \quad (58)$$

where $M = u_0/c$ is the Mach number, $x = p/p_0$, $E_{compr}(p) = E_{compr}(x)u_0^2/p_0$, and $\eta(x) = k(x)p_0/c$ (c is the sound speed). According to the model proposed in [30], $\eta(x) = M^2 E_{compr}(x)$.

It is obvious that the correlator (56) cannot be approximated by $\delta(\tau)$, and the growth of fluctuations can be evaluated only by using Eq. (53). For "acoustic" turbulence, $D_t' > 0$, i.e., the fluctuation intensity at $t \geq \tau_0$ (or $t = t_0$) increases owing to both $C\langle n^2\rangle$ and $D_t'\langle(\nabla n)^2\rangle$.

Thus, a more realistic model of turbulence with a finite correlation time reveals new qualitative tends in the initial evolution: fluctuations are damped by turbulent diffusion when acoustic effects are negligible, whereas additional growth of $\langle n^2(\mathbf{r}, t)\rangle$ is predicted for turbulence with a substantial acoustic component.

When t is much greater than the smaller of the times τ_0 and $t_0 = R_0/u_0$, only small-scale scalar fluctuations may be taken into account, i.e., one may assume that $q \gg p_0$ in (49). A series expansion in terms of $p/q \ll 1$ yields

$$\frac{d}{dt}\langle n^2(\mathbf{r}, t)\rangle = -2D_m\langle(\nabla n)^2\rangle + \frac{4}{\sqrt{3}u_0}\langle \mathrm{div}^2\mathbf{u}\rangle\int_0^\infty dp\frac{E_V(p, t, t)}{p}. \quad (59)$$

This equation is derived by assuming that the function $\tilde{g}(q, \tau)$ has a narrower support in τ as compared to $E_{compr}(p, \tau)$ and using the approximation

$$\int_0^\infty d\tau\tilde{g}(q, \tau) \approx \frac{\sqrt{3}}{u_0 q}, \quad q \gg q_0, \quad (60)$$

which follows from DIA equation (24) written for the Laplace transform of $\tilde{g}(p, \tau)$:

$$\tilde{\tilde{g}}(p, s) = \left[s + D_m p^2 + \frac{p}{4}\int_0^\infty dq\int_{-1}^1 d\mu\int_0^\infty d\tau[(1-\mu^2)pE_{inc}(q, \tau) + 2\mu(p\mu - q)E_{compr}(q, \tau)]e^{-s\tau}\tilde{g}(|\mathbf{p}-\mathbf{q}|, \tau)\right]^{-1}. \quad (61)$$

Thus, Eq. (59) also substantially differs from its counterpart, Eq. (52), written for a $\delta(\tau)$-correlated process.

In the $\delta(\tau)$-correlated case, the exact solution to Eq. (61) is (see [17])

$$\tilde{\tilde{g}}(p, s) = [s + (D_m + D_t^{(0)})p^2]^{-1}, \quad (62)$$

which entails

$$\tilde{g}(p, \tau) = \exp[-(D_m + D_t^{(0)})p^2\tau],$$

i.e., the solution to the diffusion equation with the diffusivity $D_m + D_t^{(0)}$, where $D_t^{(0)} = u_0^2\tau_0/3$. This diffusive solution is identical to that predicted by the model of turbulence with a finite correlation time for $p \ll p_0$. However, the largest contributions to expressions (54) and (55) are due to $p \geq p_0$ (rather than $p \ll p_0$). Since diffusive solution (62) is not valid in this case, the $\delta(\tau)$-correlated approximation cannot be used either. Indeed, when the diffusion Green function is substituted into (54), the resulting expression equals the diffusivity in (52) only for $(p/p_0)^2\xi_0 \ll 1$. In most turbulence models, it is assumed that $\xi_0 = u_0\tau_0/R_0 \sim 1$, i.e., the $\delta(\tau)$-correlated approximation is not valid a priori (e.g. see [31]). In summary, the $\delta(\tau)$-correlated approximation is applicable only to turbulence with $\xi_0 \ll 1$, in which case it is sufficient to use expansion (10) in terms of molecular-diffusion Green functions.

The Green function of the Bourret equation (40) is given by expression (61), where $\tilde{g}(|\mathbf{p} - \mathbf{q}|, \tau)$ on the right-hand side should be replaced with $\tilde{g}_m(|\mathbf{p} - \mathbf{q}|, \tau)$ or with unity (if the small molecular diffusivity D_m is neglected). In the diffusion approximation ($p \ll p_0$, $s \ll 1/\tau_0$), the Green function of the Bourret equation is given by (62).

When $\xi_0 \gg 1$, the Green functions of the DIA equation (61) and the Bourret equation (40) exhibit different asymptotic behavior. While the exact asymptotic expressions are not written out here (see [21]), it should be noted that, roughly, $\tilde{\tilde{g}}(p, s) \propto (\xi_0 p/p_0)^{-1}$ in the former case and $\tilde{\tilde{g}}(p, s) \propto (\xi_0 p/p_0)^{-2}$ in the latter. This difference explains why general expression (51) yields the physically reasonable $D_t^{(0)} = \text{const}\, u_0/p_0$ as $\xi_0 \longrightarrow \infty$ in the former case (see also [12]), whereas the result obtained in the latter case tends to zero as $1/\xi_0$. The fact that the DIA equation yields qualitatively correct values of turbulent diffusivity both for $\xi_0 \ll 1$ and for $\xi_0 \gg 1$ was first noted in [16], where the DIA equation for scalar diffusion was derived on the basis of an earlier paper by Kraichnan focused on a nonlinear theory of turbulence per se [25]. Thus, the nonlinear theory of turbulent diffusion does not rely on any expansion in terms of ξ_0. The nonlinear analysis takes into account a much greater number of elementary interactions between the scalar field and the flow, as compared to a Bourret-like linear theory, leading to an expression for D_t that is valid for any ξ_0.

The equation for $\tilde{V}(p, t_1, t_2)$ up to fourth-order irreducible correlators is obtained by adding the four terms corresponding to the last four terms in (38) to the right-hand side of Eq. (48). These additional terms are denoted by A, B, C, and D in accordance with their order in (38):

$$\begin{aligned} A(p, t_1, t_2) = {} & \int_0^{t_1} d\tau \int_0^{t_1-\tau} d\tau' \int_0^{t_1-\tau-\tau'} d\tau'' \int \frac{d\mathbf{q}}{(2\pi)^3} \int \frac{d\mathbf{s}}{(2\pi)^3} \\ & \times \tilde{g}(|\mathbf{p} - \mathbf{s}|, \tau)\tilde{g}(|\mathbf{p} - \mathbf{q} - \mathbf{s}|, \tau')\tilde{g}(|\mathbf{p} - \mathbf{q}|, \tau'') \\ & \times p_i \tilde{B}_{ij}(\mathbf{s}, \tau + \tau')(\mathbf{p} - \mathbf{q} - \mathbf{s})_j \\ & \times (\mathbf{p} - \mathbf{s})_n \tilde{B}_{nm}(\mathbf{q}, \tau' + \tau'')(\mathbf{p} - \mathbf{q})_m \\ & \times \tilde{V}(p, t_1 - \tau - \tau' - \tau'', t_2), \end{aligned} \quad (63)$$

$$\begin{aligned} D(p, t_1, t_2) = {} & -\int_0^{t_2} d\tau \int_0^{t_2-\tau} d\tau' \int_0^{t_2-\tau-\tau'} d\tau'' \int \frac{d\mathbf{q}}{(2\pi)^3} \int \frac{d\mathbf{s}}{(2\pi)^3} \\ & \times \tilde{g}(p, \tau)\tilde{g}(|\mathbf{p} + \mathbf{q} - \mathbf{s}|, \tau'')\tilde{g}(|\mathbf{p} + \mathbf{q}|, \tau')p_i \\ & \times \tilde{B}_{ij}(\mathbf{s}, |t_1 - t_2 + \tau + \tau'|)(\mathbf{p} + \mathbf{q})_j \\ & \times \mathbf{p}_n \tilde{B}_{nm}(\mathbf{q}, \tau' + \tau'')(\mathbf{p} + \mathbf{q} - \mathbf{s})_m \\ & \times \tilde{V}(|\mathbf{p} - \mathbf{s}|, t_2 - \tau - \tau' - \tau'', t_1), \end{aligned} \quad (64)$$

$$\begin{aligned} B(p, t_1, t_2) = {} & -\int_0^{t_1} d\tau \int_0^{t_1-\tau} d\tau' \int_0^{t_2} d\tau'' \int \frac{d\mathbf{q}}{(2\pi)^3} \int \frac{d\mathbf{s}}{(2\pi)^3} \\ & \times \tilde{g}(|\mathbf{p} - \mathbf{s}|, \tau)\tilde{g}(|\mathbf{p} - \mathbf{q} - \mathbf{s}|, \tau')\tilde{g}(p, \tau'')p_i \\ & \times \tilde{B}_{ij}(\mathbf{s}, \tau + \tau')(\mathbf{p} - \mathbf{q} - \mathbf{s})_j \\ & \times (\mathbf{p} - \mathbf{s})_n \tilde{B}_{nm}(\mathbf{q}, |t_1 - t_2 - \tau + \tau''|)p_m \\ & \times \tilde{V}(|\mathbf{p} - \mathbf{q}|, t_1 - \tau - \tau', t_2 - \tau''), \end{aligned} \quad (65)$$

$$\begin{aligned} C(p, t_1, t_2) = {} & \int_0^{t_2} d\tau \int_0^{t_2-\tau} d\tau' \int_0^{t_1} d\tau'' \int \frac{d\mathbf{q}}{(2\pi)^3} \int \frac{d\mathbf{s}}{(2\pi)^3} \\ & \times \tilde{g}(p, \tau)\tilde{g}(|\mathbf{p} - \mathbf{q}|, \tau')\tilde{g}(|\mathbf{p} + \mathbf{s}|, \tau'')(\mathbf{p} - \mathbf{q})_i \\ & \times \tilde{B}_{ij}(\mathbf{s}, |t_1 - t_2 + \tau + \tau'|)p_j \\ & \times (\mathbf{p} + \mathbf{s})_n \tilde{B}_{nm}(\mathbf{q}, |t_1 - t_2 + \tau - \tau''|)p_m \\ & \times \tilde{V}(|\mathbf{p} - \mathbf{q} + \mathbf{s}|, t_1 - \tau'', t_2 - \tau - \tau'). \end{aligned} \quad (66)$$

Consistency with conservation law (5) is verified by showing that the integrals of $A + D$ and $B + C$ over $\mathbf{p}$ vanish if $t_1 = t_2 = t$ and $\text{div}\mathbf{u} = 0$. In contrast to Eq. (48), additional terms (63)–(66) depend on turbulent helicity. When (63)–(66) are taken into account,

the following expressions must be added to the right-hand side of (49):

$$E(t) = \int_0^t d\tau \int_0^{t-\tau} d\tau' \int_0^{t-\tau-\tau'} d\tau'' \int \frac{d\mathbf{p}}{(2\pi)^3} \int \frac{d\mathbf{q}}{(2\pi)^3} \int \frac{d\mathbf{s}}{(2\pi)^3} \times \tilde{g}(|\mathbf{p}-\mathbf{s}|, \tau)\tilde{g}(|\mathbf{p}-\mathbf{q}-\mathbf{s}|, \tau')\tilde{g}(|\mathbf{p}-\mathbf{q}|, \tau'')s_i \times \tilde{B}_{ij}(\mathbf{s}, \tau+\tau')(\mathbf{p}-\mathbf{q}-\mathbf{s})_j \times (\mathbf{p}-\mathbf{s})_n \tilde{B}_{nm}(\mathbf{q}, \tau'+\tau'')(\mathbf{p}-\mathbf{q})_m \times \tilde{V}(p, t-\tau-\tau'-\tau'', t), \quad (67)$$

$$F(t) = -\int_0^t d\tau \int_0^{t-\tau} d\tau' \int_0^t d\tau'' \int \frac{d\mathbf{p}}{(2\pi)^3} \int \frac{d\mathbf{q}}{(2\pi)^3} \int \frac{d\mathbf{s}}{(2\pi)^3} \times \tilde{g}(|\mathbf{p}-\mathbf{s}|, \tau)\tilde{g}(|\mathbf{p}-\mathbf{q}-\mathbf{s}|, \tau')\tilde{g}(p, \tau'')s_i \times \tilde{B}_{ij}(\mathbf{s}, \tau+\tau')(\mathbf{p}-\mathbf{q}-\mathbf{s})_j \times (\mathbf{p}-\mathbf{s})_n \tilde{B}_{nm}(\mathbf{q}, |\tau'-\tau''|)p_m \times \tilde{V}(|\mathbf{p}-\mathbf{q}|, t-\tau'', t-\tau-\tau'). \quad (68)$$

The functions $E(t)$ and $F(t)$ are obtained by integrating the sums $A + D$ and $B + C$, respectively. In contrast to $\tilde{V}(p, t_1, t_2)$, the quantity $\langle n^2(\mathbf{r}, t)\rangle$ is independent of turbulent helicity even if the fourth-order correlators are taken into account. It should be noted that both $E(t)$ and $F(t)$ depend on the product of the energy spectrum $E_{\text{compr}}(p, \tau)$ (associated with potential motions in compressible turbulence) with a correlator containing $E_{\text{inc}}(p, \tau)$. Expressions (67) and (68) can be used to calculate corrections to asymptotic equations (53) and (59). The additional contribution to (59) has a relatively simple analytical form:

$$-\frac{2}{\sqrt{3}u_0^3}\langle \text{div}^2\mathbf{u}\rangle\langle \text{curl}^2\mathbf{u}\rangle \int_0^\infty dq \frac{\tilde{E}_V(q, t, t)}{q^3}. \quad (69)$$

Expression (69) demonstrates that turbulent fluctuations are additionally damped by vortex motion. When the molecular diffusivity is small, this effect plays a dominant role in damping the fluctuations (dispersing clusters of particles) at the initial stage. At a later stage, molecular diffusion takes over as a factor that flattens out the particle density. This effect is beyond the scope of the $\delta(\tau)$-correlated model, because vortex motion cannot be described by this model.

5. EQUATIONS FOR MAGNETIC ENERGY AND MAGNETIC HELICITY SPECTRA

Magnetic field diffusion is described by a tensor Green function $G_{ij}(1, 2)$ and a tensor interaction operator $L_{ij}(1) = \nabla_j^{(1)} u_i(1) - \delta_{ij}\nabla_t^{(1)} u_t(1)$. For stationary, homogeneous, and isotropic turbulence, $\langle G_{ij}(1, 2)\rangle = \theta(\tau)g_{ij}(\mathbf{R}, \tau)$ and $H_{ij}(1, 2) = \langle B_i(1)B_j(2)\rangle = H_{ij}(\mathbf{R}, t_1, t_2)$, where $\mathbf{R} = \mathbf{r}_1 - \mathbf{r}_2$ and $\tau = t_1 - t_2$. The Fourier transforms of these quantities with respect to $\mathbf{R}$ are expressed as follows:

$$\tilde{g}_{jk}(\mathbf{p}, \tau) = \delta_{jk}\tilde{g}_m(p, \tau) + (\delta_{jk}p^2 - p_j p_k)\tilde{g}_2(p, \tau) + ie_{jkt}p_t\tilde{g}_1(p, \tau), \quad (70)$$

$$\tilde{H}_{jk}(\mathbf{p}, t_1, t_2) = (\delta_{jk}p^2 - p_j p_k)\tilde{H}_0(p, t_1, t_2) + ie_{jkt}p_t\tilde{H}_1(p, t_1, t_2). \quad (71)$$

The term containing $\tilde{g}_m(p, \tau)$ is due to the absolute term in Eq. (9) for the Green function. The condition $\text{div}\mathbf{B} = 0$ ($\mathbf{p}\cdot\tilde{\mathbf{B}} = 0$) is satisfied, because the initial magnetic field $\mathbf{B}_0(\mathbf{r})$ is solenoidal. Further analysis is facilitated by introducing the function

$$\tilde{g}_0(p, \tau) = \tilde{g}_m(p, \tau) + p^2\tilde{g}_2(p, \tau) \quad (72)$$

and deriving a system of equations for $\tilde{g}_0(p, \tau)$ and $\tilde{g}_1(p, \tau)$ from DIA equation (24). Defining $\tilde{g}_\pm(p, \tau) = \tilde{g}_0(p, \tau) \pm p\tilde{g}_1(p, \tau)$, one obtains

$$\tilde{g}_\pm(p, \tau) = \tilde{g}_m(p, \tau) - \int_0^\tau d\tau' \int_0^{\tau-\tau'} d\tau''\tilde{g}_m(p, \tau') \times [p^2 S_0(p, \tau'') \pm pS_1(p, \tau'')]\tilde{g}_\pm(p, \tau-\tau'-\tau''), \quad (73)$$

where

$$p^2 S_0(p, \tau) = \frac{1}{4}\int_0^\infty dq \int_{-1}^1 d\mu\{p^2(1-\mu^2) \times [E_{\text{inc}}(q, \tau)\tilde{g}_0(|\mathbf{p}-\mathbf{q}|, \tau) - E_h(q, \tau)\tilde{g}_1(|\mathbf{p}-\mathbf{q}|, \tau)] + 2p\mu(p\mu - q) \times E_{\text{compr}}(q, \tau)g_0(|\mathbf{p}-\mathbf{q}|, \tau)\}, \quad (74)$$

$$S_1(p, \tau) = \frac{1}{4}\int_0^\infty dq \int_{-1}^1 d\mu\{(1-\mu^2)[(p^2 + q^2 - pq\mu) \times [E_{\text{inc}}(q, \tau)\tilde{g}_1(|\mathbf{p}-\mathbf{q}|, \tau) - E_h(q, \tau)\tilde{g}_0(|\mathbf{p}-\mathbf{q}|, \tau)] + [2p^2\mu^2 + (1+\mu^2)(q^2 - 2pq\mu)] \times E_{\text{compr}}(q, \tau)\tilde{g}_1(|\mathbf{p}-\mathbf{q}|, \tau)\}. \quad (75)$$

Expressions (73)–(75) have much simpler form for incompressible turbulent flow. The Laplace transform of Eq. (73) is

$$\tilde{\tilde{g}}_\pm(p, s) = [s + D_m p^2 + p^2\tilde{S}_0(p, s) \pm p\tilde{S}_1(p, s)]^{-1}. \quad (76)$$

Under the solenoidality condition $\mathbf{p}\cdot\mathbf{B}=0$, the term $p_i p_j \tilde{g}_2(p,\tau)$ vanishes and Green function (70) reduces to

$$\tilde{g}_{jk}(p,\tau) = \delta_{jk}\tilde{g}_0(p,\tau) + ie_{jkt}p_t\tilde{g}_1(p,\tau). \quad (77)$$

In the case of zero turbulent helicity ($D(p,\tau)=0$), both $\tilde{g}_1(p,\tau)=0$ and $S_1(p,\tau)=0$, and the equations for $\tilde{g}_0(p,\tau)$ are identical to Eqs. (61) and (62) for scalar diffusion.

In the diffusion approximation, expression (76) has the form

$$\tilde{\tilde{g}}_\pm(p,s) = [s + (D_\mathrm{m} + D_\mathrm{t}^{(0)})p^2 \pm \alpha_\mathrm{t}^{(0)}p]^{-1}, \quad (78)$$

where

$$D_\mathrm{t}^{(0)} = \frac{1}{3}\int_0^\infty dp\int_0^\infty d\tau$$

$$\times\Big\{[E_\mathrm{inc}(p,\tau)+E_\mathrm{compr}(p,\tau)]\tilde{g}_0(p,\tau) \quad (79)$$

$$+ E_\mathrm{compr}(p,\tau)p\frac{\partial\tilde{g}_0(p,\tau)}{\partial p} - E_\mathrm{h}(p,\tau)\tilde{g}_1(p,\tau)\Big\}.$$

In the case of zero turbulent helicity ($E_\mathrm{h}(p,\tau)=0$), this expression is identical to turbulent diffusivity (51) for a passive scalar (different coefficients D_t are obtained for magnetic and scalar fields only if irreducible interactions of fourth order in velocity are taken into account [26]). The factor describing amplification of the mean magnetic field by turbulent helicity is

$$\alpha_\mathrm{t}^{(0)} = \frac{1}{3}\int_0^\infty dp\int_0^\infty d\tau\{-E_\mathrm{h}(p,\tau)\tilde{g}_0(p,\tau) \quad (80)$$
$$+ p^2[E_\mathrm{inc}(p,\tau)+2E_\mathrm{compr}(p,\tau)]\tilde{g}_1(p,\tau)\}.$$

For a $\delta(\tau)$-correlated process,

$$D_\mathrm{t}^{(0)} = u_0^2\tau_0/3, \quad \alpha_\mathrm{t}^{(0)} = -\langle\mathbf{u}\cdot(\nabla\times\mathbf{u})\rangle\tau_0/3.$$

The diffusion Green function corresponding to (78) has the form of (77) with

$$\tilde{g}_0(p,\tau) = \cosh(\alpha_\mathrm{t}^{(0)}p\tau)\exp(-D_\mathrm{t}^{(0)}p^2\tau),$$
$$\tilde{g}_1(p,\tau) = -\frac{1}{p}\sinh(\alpha_\mathrm{t}^{(0)}p\tau)\exp(-D_\mathrm{t}^{(0)}p^2\tau). \quad (81)$$

Now, consider Eq. (38) for the tensor $\tilde{H}_{ij}(p,t_1,t_2)$, retaining only second-order correlators (ladder approximation for the Bethe–Salpeter equation). First of all, note that the function $\tilde{H}_0(p,t_1,t_2)$ determines the spectrum of magnetic energy fluctuations:

$$\langle\mathbf{B}(\mathbf{r},t_1,t_2)\cdot\mathbf{B}(\mathbf{r},t_1,t_2)\rangle = \int_0^\infty dpE_B(p,t_1,t_2), \quad (82)$$
$$E_B = p^4\tilde{H}_0(p,t_1,t_2)/\pi^2.$$

Use the relation $\mathbf{B}=\nabla\times\mathbf{A}$ to express the spectrum of magnetic helicity fluctuations in terms of $\tilde{H}_1(p,t,t)$:

$$H_\mathrm{Mh}(t)\equiv\langle\mathbf{A}(\mathbf{r},t)\cdot\mathbf{B}(\mathbf{r},t)\rangle = \int_0^\infty dpE_\mathrm{Mh}(p,t), \quad (83)$$
$$E_\mathrm{Mh}(p,t) = -p^2\tilde{H}_1(p,t,t)/\pi^2.$$

When $D_\mathrm{m}=0$ in a homogeneous turbulent flow, magnetic helicity conservation law (6) implies that

$$\frac{d}{dt}\int d\mathbf{p}\tilde{H}_1(p,t,t) = 0. \quad (84)$$

The validity of this relation is demonstrated below.

Equation (38) entails the following system of equations for $\tilde{H}_0(p,t_1,t_2)$ and $\tilde{H}_1(p,t_1,t_2)$:

$$\left(\frac{\partial}{\partial t_1}+D_\mathrm{m}p^2\right)\tilde{H}_0(p,t_1,t_2)$$
$$= -\int_0^{t_1}d\tau[p^2S_0(p,\tau)\tilde{H}_0(p,t_1-\tau,t_2)$$
$$+ S_1(p,\tau)\tilde{H}_1(p,t_1-\tau,t_2)]$$
$$+\int_0^{t_2}d\tau[T_0(p,|t_1-t_2+\tau|,t_1,t_2-\tau)\tilde{g}_0(p,\tau)$$
$$+ T_1(p,|t_1-t_2+\tau|,t_1,t_2-\tau)\tilde{g}_1(p,\tau)],$$
$$\left(\frac{\partial}{\partial t_1}+D_\mathrm{m}p^2\right)\tilde{H}_1(p,t_1,t_2) \quad (85)$$
$$= -\int_0^{t_1}d\tau p^2[S_1(p,\tau)\tilde{H}_0(p,t_1-\tau,t_2)$$
$$+ S_0(p,\tau)\tilde{H}_1(p,t_1-\tau,t_2)]$$
$$+\int_0^{t_2}d\tau[T_1(p,|t_1-t_2+\tau|,t_1,t_2-\tau)\tilde{g}_0(p,\tau)$$
$$+ T_0(p,|t_1-t_2+\tau|,t_1,t_2-\tau)\tilde{g}_1(p,\tau)],$$

where

$$T_0(p,|t|,t_1,t_2-\tau) = \frac{1}{4}\int_0^\infty dq \times \int_{-1}^{1} d\mu\{(1-\mu^2)[(p^2+q^2-pq\mu) \times E_{\rm inc}(q,|t|)\tilde{H}_0(|\mathbf{p}-\mathbf{q}|,t_1,t_2-\tau) - E_{\rm h}(q,|t|)\tilde{H}_1(|\mathbf{p}-\mathbf{q}|,t_1,t_2-\tau)] + [2p^2\mu^2+(1+\mu^2)(q^2-2pq\mu)] \times E_{\rm compr}(q,|t|)\tilde{H}_0(|\mathbf{p}-\mathbf{q}|,t_1,t_2-\tau)\}, \quad (86)$$

$$T_1(p,|t|,t_1,t_2-\tau) = \frac{1}{4}\int_0^\infty dq \times \int_{-1}^{1} d\mu\{p^2(1-\mu^2)[E_{\rm inc}(q,|t|)\tilde{H}_1(|\mathbf{p}-\mathbf{q}|,t_1,t_2-\tau) - E_{\rm h}(q,|t|)\tilde{H}_0(|\mathbf{p}-\mathbf{q}|,t_1,t_2-\tau) + 2p\mu(p\mu-q)E_{\rm compr}(q,|t|)\tilde{H}_1(|\mathbf{p}-\mathbf{q}|,t_1,t_2-\tau)\}. \quad (87)$$

Note that S_0 and S_1 are associated with the diagram $\langle\circ\supset\!\!\!\!\!\!\bigcirc\circ\rangle$; T_0 and T_1, with $\langle\circ\square\circ\rangle$ (see (38)). Therefore, the corresponding expressions have similar structure. In the case of zero turbulent helicity ($D(p,\tau)=0$, $\tilde{g}_1(p,\tau)=0$, $S_1(p,\tau)=0$), system (85) splits into separate equations for $\tilde{H}_0(p,t_1,t_2)$ and $\tilde{H}_1(p,t_1,t_2)$.

In the frequently used case of a $\delta(\tau)$-correlated velocity ensemble ($\tilde{B}_{ij}(\mathbf{p},\tau)=\tau_0\delta(\tau)\tilde{B}_{ij}(\mathbf{p})$), system (85) simplifies to

$$\left(\frac{\partial}{\partial t}+2D_{\rm m}p^2\right)\tilde{H}_0(p,t,t) = -2[D_{\rm t}^{(0)}p^2\tilde{H}_0(p,t,t)+\alpha_{\rm t}^{(0)}\tilde{H}_1(p,t,t)] + \frac{\tau_0}{2}\int_0^\infty dq\int_{-1}^{1} d\mu\{(1-\mu^2)[(p^2+q^2-pq\mu) \times E_{\rm inc}(q)\tilde{H}_0(|\mathbf{p}-\mathbf{q}|,t,t) - E_{\rm h}(q)\tilde{H}_1(|\mathbf{p}-\mathbf{q}|,t,t)] + [2p^2\mu^2+(1+\mu^2)(q^2-2pq\mu)] \times E_{\rm compr}(q)\tilde{H}_0(|\mathbf{p}-\mathbf{q}|,t,t)\}, \quad (88)$$

$$\left(\frac{\partial}{\partial t}+2D_{\rm m}p^2\right)\tilde{H}_1(p,t,t) = -2p^2[D_{\rm t}^{(0)}\tilde{H}_1(p,t,t)+\alpha_{\rm t}^{(0)}\tilde{H}_0(p,t,t)] + \frac{\tau_0}{2}\int_0^\infty dq\int_{-1}^{1} d\mu\{p^2(1-\mu^2)[E_{\rm inc}(q)\tilde{H}_1(|\mathbf{p}-\mathbf{q}|,t,t) \times E_{\rm h}(q)\tilde{H}_0(|\mathbf{p}-\mathbf{q}|,t,t)+2(p^2\mu^2-pq\mu) \times E_{\rm compr}(q)\tilde{H}_1(|\mathbf{p}-\mathbf{q}|,t,t)\},$$

where

$$D_{\rm t}^{(0)} = u_0^2\tau_0/3, \quad \alpha_{\rm t}^{(0)} = -\langle\mathbf{u}\cdot(\nabla\times\mathbf{u})\rangle\tau_0/3.$$

For incompressible turbulent flow with zero helicity, the first equation in (88) written for $p^2H_0(p,t,t)$ is identical to Eq. (14) in [17].

Now, magnetic helicity conservation law (84) for flows with $D_{\rm m}=0$ can be obtained by performing some simple changes of variables in the integral of the second equation in (85). It should be noted that even calculation of magnetic energy (82) must be consistent with conservation law (84), because the functions $\tilde{H}_0(p,t_1,t_2)$ and $\tilde{H}_1(p,t_1,t_2)$ are interrelated.

Equations (85)–(87) can be used to obtain an equation for $\langle B^2(\mathbf{r},t)\rangle$:

$$\frac{d}{dt}\langle B^2(\mathbf{r},t)\rangle = -2D_{\rm m}\langle(\nabla\times\mathbf{B}(\mathbf{r},t))^2\rangle + \frac{1}{2}\int_0^\infty dp\int_0^\infty dq\int_{-1}^{1} d\mu\int_0^t d\tau\{E_B(p,t,t-\tau) \times \tilde{g}_0(|\mathbf{p}-\mathbf{q}|,\tau)[(1-\mu^2)(q^2-pq\mu) \times E_{\rm inc}(q,\tau)+(q^2(1+\mu^2)-2pq\mu^3)E_{\rm compr}(q,\tau)] + \tilde{g}_1(|\mathbf{p}-\mathbf{q}|,\tau)E_{\rm Mh}(p,t,t-\tau)[p^3q(1-\mu^2) \times E_{\rm inc}(q,\tau)+(p^2q^2(1-\mu^2)+2p\mu(q^3+2p^3\mu+qp^2\mu^2)) \times E_{\rm compr}(q,\tau)]-(1-\mu^2)(q^2-2pq\mu) \times E_{\rm h}(q,\tau)E_B(p,t,t-\tau)\tilde{g}_1(|\mathbf{p}-\mathbf{q}|,\tau)\}. \quad (89)$$

In the model of delta-correlated turbulent field, it simplifies to

$$\frac{d}{dt}\langle B^2\rangle = -2D_{\rm m}\langle(\nabla\times\mathbf{B})^2\rangle + \frac{2\tau_0}{3}[\langle(\nabla\times\mathbf{u})^2\rangle+2\langle{\rm div}^2\mathbf{u}\rangle]\langle B^2\rangle. \quad (90)$$

Note that Eq. (90) is independent of turbulent helicity, which determines the amplification factor for the

mean magnetic field $\langle \mathbf{B} \rangle$ (see (80)). Since $\langle B^2 \rangle = \langle \mathbf{B} \rangle \cdot \langle \mathbf{B} \rangle + \langle B'^2 \rangle$, this implies that the increase in the mean (large-scale) magnetic energy due to the α-effect is compensated for by the damping of magnetic fluctuations due to the same mechanism. This observation was originally made in [7], where the diffusion approximation was used without assuming short turbulent correlation time. Note that Eq. (90) was recently obtained in [24], where extension of Kazantsev's equation [17] to two- and three-dimensional compressible turbulent flows with zero helicity was analyzed in detail.

It should be noted that magnetic fields are generally associated with rotating media (rotating stellar atmospheres, rotation of the Galaxy, etc.). Since rotation of a turbulent medium as a whole gives rise to helicity, system of equations (88) for $\tilde{H}_0(p, t, t)$ and $\tilde{H}_1(p, t, t)$ must be solved even in $\delta(\tau)$-correlated models. However, effects due to helicity are totally ignored in [20, 24] and other studies.

As in the analysis of scalar transport [see (53)], Eq. (90) can be extended to describe an initial stage of field evolution in the model of delta-correlated turbulent field:

$$\frac{d}{dt}\langle B^2 \rangle = -2(D_\mathrm{m} - D_\mathrm{t}'')\langle (\nabla \times \mathbf{B})^2 \rangle + \gamma \langle \mathbf{B} \cdot (\nabla \times \mathbf{B}) \rangle + (C_1 + C_2)\langle B^2 \rangle, \tag{91}$$

where

$$D_\mathrm{t}'' = \frac{1}{15}\int_0^\infty dp \int_0^\infty d\tau \left\{ [E_\mathrm{inc}(p, \tau) + 3E_\mathrm{compr}(p, \tau)] \times p\frac{\partial \tilde{g}_0(p, \tau)}{\partial p} - 2E_\mathrm{h}(p, \tau) p \frac{\partial \tilde{g}_1(p, \tau)}{\partial p} \right\}, \tag{92}$$

$$\gamma = \frac{2}{3}\int_0^\infty dp \int_0^\infty d\tau p^2 E_\mathrm{compr}(p, \tau) \times \left[\tilde{g}_1(p, \tau) - p\frac{\partial \tilde{g}_1(p, \tau)}{\partial p} \right], \tag{93}$$

$$C_1 = \frac{2}{3}\int_0^\infty dp \int_0^\infty d\tau p^2 \times [E_\mathrm{inc}(p, \tau) + 2E_\mathrm{compr}(p, \tau)]\tilde{g}_0(p, \tau), \tag{94}$$

$$C_2 = -\frac{2}{3}\int_0^\infty dp \int_0^\infty d\tau p^2 E_\mathrm{h}(p, \tau)\tilde{g}_1(p, \tau). \tag{95}$$

Note that two small terms on the order of $\langle \nabla^2 \mathbf{B} \cdot (\nabla \times \mathbf{B}) \rangle$ are ignored in Eq. (91). Furthermore, the term with γ is entirely determined by compressible turbulent motion with nonzero helicity. In contrast to (90), the term proportional to $\langle B^2 \rangle$ in (91) contains a contribution due to turbulent helicity, because the function g_1 is proportional to the helicity spectrum. This implies that $C_2 \le 0$; i.e., turbulent helicity inhibits the increase in magnetic energy at initial stages of evolution. Since $\partial \tilde{g}_0(p, \tau)/\partial p \le 0$, it may be expected that $D_\mathrm{t}'' \le 0$ (the term with E_h must be smaller than the first term in (92)). Since D_t'' is much greater than D_m, the former coefficient controls the initial decay of magnetic energy. Note that Eq. (91) reduces to (90) in the limit of short correlation time.

In summary, the increase in magnetic energy in turbulence with a finite correlation time is substantially slower than that predicted by the model of delta-correlated turbulent field.

When acoustic effects are important, one obtains $D_\mathrm{t}'' = 3D_\mathrm{t}'/5$, $C_2 \equiv 0$, and $C_1 = 4C/3$, where the positive coefficients D_t' and C are given by (57) and (58), respectively. Note that the magnetic field diffusivity in "acoustic" turbulence is equal to the scalar diffusivity obtained in [27]. Thus, additional growth of fluctuations due to acoustic effects is predicted for both scalar and magnetic field diffusion.

In the case of zero helicity, magnetic energy must concentrate in small-scale fluctuations in the long-time limit. Equation (89) simplifies accordingly:

$$\frac{d}{dt}\langle B^2 \rangle = -2D_\mathrm{m}\langle (\nabla \times \mathbf{B})^2 \rangle + \frac{8}{5\sqrt{3}u_0}[\langle (\nabla \times \mathbf{u})^2 \rangle + 2\langle \mathrm{div}^2 \mathbf{u} \rangle]\int_0^\infty dp \frac{E_B(p, t, t)}{p}. \tag{96}$$

In contrast to scalar fluctuations (see (55)), magnetic field fluctuations can be amplified not only by compressible turbulent motions (e.g., shock waves), but also by rotational motions of a turbulent plasma. Since $E_B(p, t, t)$ has a maximum at $p_\mathrm{max} \ll p_0$, the value of $1/p$ at some intermediate point p_1 ($p_0 \ll p_1 \le p_\mathrm{max}$) can be factored out of the integral. As a result, the second term in (96) will become similar in structure to its counterpart in (90), but its value will differ by a factor of $(p_0/p_1)/\xi_0$. When $\xi_0 \approx 1$ (as usually assumed for well-developed turbulence [31]), the resulting amplification factor will be much smaller than that predicted by Eq. (90). A similar result is obtained for scalar fluctuations.

Equations (91) and (96) differ substantially from (90), which corresponds to a $\delta(\tau)$-correlated process. All critical remarks on the model of delta-correlated turbulent

field made in the preceding section apply to the model of magnetic field diffusion.

6. CONCLUSIONS

The principal results of this study are summarized as follows. Most importantly, it is shown for the first time that the Bethe–Salpeter-type time-dependent integrodifferential equations for magnetic field and scalar fluctuation intensities must be consistent with conservation laws (5) and (6). The kernels of these equations are differences of two terms of similar order describing the balance of growth and decay of fluctuations at a fixed point in a turbulent medium. The conservation laws impose integral constraints on these processes. Consistency with these laws rules out spurious growth or damping of fluctuations.

It is found that a hierarchy of Bethe–Salpeter equations are consistent with the conservation laws only if the averaged Green function $\langle G(1, 2)\rangle$ satisfies a hierarchy of Dyson-type nonlinear equations. Moreover, the highest orders of velocity correlators retained in both hierarchies must be equal. In particular, it is shown that the widely used approximate Bourret equation for $\langle G(1, 2)\rangle$ does not ensure consistency of Bethe–Salpeter equations with the conservation laws even in the case of $\xi_0 = u_0\tau_0/R_0 \ll 1$, when this equation yields a correct value of turbulent diffusivity.

The simple derivation of a hierarchy of nonlinear equations for the averaged Green function is not restricted to the case of a Gaussian velocity ensemble. Correct time-dependent equations are obtained for scalar and magnetic fluctuations in compressible turbulent media with nonzero helicity. These equations are consistent with the conservation laws. They can also be used to analyze the influence of non-Gaussian velocity statistics on turbulent diffusivities and time evolution of scalar fluctuations.

These equations are used to derive asymptotic formulas describing the evolution of scalar fluctuation intensity at a fixed point of a turbulent medium.

It is shown that the model of delta-correlated turbulent field ($\langle u_i(1)u_j(2)\rangle \propto \delta(t_1 - t_2)$), which is frequently applied because of its mathematical simplicity, fails to predict a number of important effects of turbulent diffusion on the time of energy transfer to small-scale fluctuations. The overall effect of a finite velocity correlation time is to inhibit the growth of fluctuation intensity, because fluctuations are damped by turbulent diffusion (in the $\delta(\tau)$-correlated model, damping is due to molecular diffusion only) and the amplification factor is reduced by taking into account nonlocal mechanisms of turbulent transport. In contrast, turbulent diffusion in "acoustic" turbulence counteracts molecular dissipation and simultaneously increases the amplification factor. However, when small-scale fluctuations play a dominant role at the final stage of evolution, their decay is controlled by molecular diffusion. Note also that the amplification factor for well-developed turbulence with $\xi_0 \approx 1$ is much smaller than that predicted by the $\delta(\tau)$-correlated model.

The correct hierarchies of equations for $\langle G(1, 2)\rangle$ and fluctuation intensities provide a solid basis for more detailed analysis of time-dependent fluctuation intensities. The method used to derive and match the hierarchies can also be used to analyze the background turbulence. In particular, it would be interesting to generalize Kraichnan's equation (which predicts $p^{-3/2}$ instead of Kolmogorov's $p^{-5/3}$ for the energy spectrum in the inertial subrange [25]) by retaining the fourth-order correlators. Such a generalization should be expected to yield a power exponent closer to Kolmogorov's, which would mean a more accurate description of real turbulence.

REFERENCES

1. S. I. Vainshtein and F. Cattaneo, Astrophys. J. **393**, 165 (1992).
2. A. V. Gruzinov and P. H. Diamond, Phys. Rev. Lett. **72**, 1651 (1994).
3. Ya. B. Zel'dovich, Zh. Éksp. Teor. Fiz. **31**, 154 (1956) [Sov. Phys. JETP **4**, 160 (1956)].
4. R. S. Peckover and N. O. Weiss, Mon. Not. R. Astron. Soc. **182**, 189 (1978).
5. L. L. Kichatinov, V. V. Pipin, and G. Rüdiger, Astron. Nachr. **315**, 157 (1994).
6. J. Cho and A. Lazarian, Astrophys. J. **589**, L77 (2003).
7. N. A. Silant'ev, Pis'ma Zh. Éksp. Teor. Fiz. **72**, 60 (2000) [JETP Lett. **72**, 42 (2000)].
8. N. A. Silant'ev, Astron. Astrophys. **370**, 533 (2001).
9. A. Z. Dolginov and N. A. Silant'ev, Geophys. Astrophys. Fluid Dyn. **63**, 139 (1992).
10. G. I. Taylor, Proc. London Math. Soc. A **20**, 196 (1921).
11. H. K. Moffatt, J. Fluid Mech. **65**, 1 (1974).
12. N. A. Silant'ev, Zh. Éksp. Teor. Fiz. **101**, 1216 (1992) [Sov. Phys. JETP **74**, 650 (1992)].
13. V. I. Tatarskiĭ, *Wave Propagation in a Turbulent Medium* (Nauka, Moscow, 1967; McGraw-Hill, New York, 1961).
14. V. I. Klyatskin, *Stochastic Equations by the Eyes of a Physicist* (Fizmatlit, Moscow, 2001).
15. R. Bourret, Can. J. Phys. **38**, 665 (1960).
16. P. H. Roberts, J. Fluid Mech. **11**, 257 (1961).
17. A. P. Kazantsev, Zh. Éksp. Teor. Fiz. **53**, 1806 (1967) [Sov. Phys. JETP **26**, 1031 (1967)].
18. H. K. Moffatt, *Magnetic Field Generation in Electrically Conducting Fluids* (Cambridge Univ. Press, Cambridge, 1978).
19. N. Seehafer, Phys. Rev. E **53**, 1283 (1996).

20. A. P. Kazantsev, A. A. Ruzmaĭkin, and D. V. Sokolov, Zh. Éksp. Teor. Fiz. **88**, 487 (1985) [Sov. Phys. JETP **61**, 285 (1985)].
21. N. A. Silant'ev, Zh. Éksp. Teor. Fiz. **111**, 871 (1997) [JETP **84**, 479 (1997)].
22. C. C. Lin, *Turbulent Flows and Heat Transfer* (Princeton Univ. Press, Princeton, 1959).
23. G. Rickayzen, *Green Functions and Condensed Matter* (Academic, New York, 1980).
24. A. A. Schekochihin, S. A. Boldyrev, and R. M. Kulsrud, Astrophys. J. **567**, 828 (2002).
25. R. H. Kraichnan, J. Fluid Mech. **5**, 497 (1959).
26. N. A. Silant'ev, Zh. Éksp. Teor. Fiz. **112**, 1312 (1997) [JETP **85**, 712 (1997)].
27. N. A. Silant'ev, Zh. Éksp. Teor. Fiz. **114**, 930 (1998) [JETP **87**, 505 (1998)].
28. G. K. Batchelor, *The Theory of Homogeneous Turbulence* (Cambridge Univ. Press, Cambridge, 1953; Inostrannaya Literatura, Moscow, 1955).
29. N. A. Silant'ev, Zh. Éksp. Teor. Fiz. **122**, 1107 (2002) [JETP **95**, 957 (2002)].
30. V. E. Zakharov and R. Z. Sagdeev, Dokl. Akad. Nauk SSSR **192**, 296 (1970) [Sov. Phys. Dokl. **15**, 439 (1970)].
31. S. A. Molchanov, A. A. Ruzmaĭkin, and D. D. Sokolov, Usp. Fiz. Nauk **145**, 593 (1985) [Sov. Phys. Usp. **28**, 307 (1985)].

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 745–747.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 850–853.

SOLIDS
Structure

Critical Dynamics of Three-Dimensional Spin Systems with Long-Range Interactions

S. V. Belim
Omsk State University, pr. Mira 55, Omsk, 644077 Russia
e-mail: belim@univer.omsk.su
Received October 21, 2003

Abstract—A field-theoretic description of critical behavior of Ising systems with long-range interactions is obtained by using the Padé–Borel summation technique in the two-loop approximation directly in the three-dimensional space. It is shown that long-range interactions affect the relaxation time of the system. © *2004 MAIK "Nauka/Interperiodica".*

It was shown in [1] that effects due to long-range interaction are essential for the critical behavior of Ising systems. The renormalization-group approach to spin systems with long-range interactions developed in [2] directly in the three-dimensional space made it possible to calculate the static critical exponents in the two-loop approximation. However, analogous calculations of critical dynamics have never been performed for these systems.

In this paper, a field-theoretic description of critical behavior of homogeneous systems with long-range interactions is developed directly for $D = 3$ in the two-loop approximation. The model under analysis is the classical spin system with the exchange integral depending on the distance between spins. The corresponding Hamiltonian is

$$H = \frac{1}{2}\sum_{i,j} J(|r_i - r_j|)S_iS_j, \tag{1}$$

where S_i is a spin variable and $J(|r_i - r_j|)$ is the exchange integral. This model is thermodynamically equivalent to the $O(n)$-symmetric Ginzburg–Landau–Wilson model defined by the effective Hamiltonian

$$H = \int d^Dq\left\{\frac{1}{2}(\tau_0 + q^a)\varphi^a + u_0\varphi^4\right\}, \tag{2}$$

where φ is a fluctuating order parameter, D is the space dimension, $\tau_0 \sim |T - T_c|$ (T_c is the critical temperature), and u_0 is a positive constant. Critical behavior strongly depends on the parameter a characterizing the interaction as a function of distance. It was shown in [3] that long-range interaction is essential when $0 < a < 2$, whereas systems with $a \geq 2$ exhibit critical behavior characteristic of short-range interactions. For this reason, the analysis that follows is restricted to the case of $0 < a < 2$.

Relaxational dynamics of spin systems near the critical temperature can be described by a Langevin-type equation for the order parameter:

$$\frac{\partial\varphi}{\partial t} = -\lambda_0\frac{\delta H}{\delta\varphi} + \eta + \lambda_0\mathbf{h}, \tag{3}$$

where λ_0 is a kinetic coefficient, $\eta(x, t)$ is a gaussian random force (representing the effect of a heat reservoir) defined by the probability distribution

$$P_\eta = A_\eta\exp[-(4\lambda_0)^{-1}\int d^dxdt\eta^2(x, t)] \tag{4}$$

with a normalization factor A_η, and $\mathbf{h}(t)$ is an external field thermodynamically conjugate to the order parameter. The temporal correlation function $G(x, t)$ of the order-parameter field can be found by solving Eq. (3) with $H[\varphi]$ given by (2) for $\varphi[\eta, \mathbf{h}]$, averaging the result over P_η, and retaining the component linear in $\mathbf{h}(0)$:

$$G(x, t) = \frac{\delta}{\delta\mathbf{h}(0)}[\langle\varphi(x, t)\rangle]|_{h=0}, \tag{5}$$

where

$$[\langle\varphi(x, t)\rangle] = B^{-1}\int D\{\eta\}\varphi(x, t)P_\eta, \tag{6}$$

$$B = \int D\{\eta\}P_\eta. \tag{7}$$

When applying the standard renormalization-group procedure to this dynamical model, one must deal with substantial difficulties. However, it was shown in [4] that the model of critical dynamics in homogeneous

systems without long-range interaction based on a Langevin-type equation is equivalent to that described by the standard Lagrangian [5]

$$L = \int d^d x dt \left\{ \lambda_0^{-1} \varphi^2 + i\varphi^* \left(\lambda_0^{-1} \frac{\partial \varphi}{\partial t} + \frac{\delta H}{\delta \varphi} \right) \right\}, \quad (8)$$

where φ^* denotes an auxiliary field. The corresponding correlation function $G(x, t)$ of the order parameter is defined for a homogeneous system as

$$G(x, t) = \langle \varphi(0, 0)\varphi(x, t) \rangle$$
$$= \Omega^{-1} \int D\{\varphi\} D\{\varphi^*\} \varphi(0, 0)\varphi(x, t) \exp(-L[\varphi, \varphi^*]),$$

where

$$\Omega = \int D\{\varphi\} D\{\varphi^*\} \exp(-L[\varphi, \varphi^*]). \quad (9)$$

Instead of dealing with the correlation function, it is reasonable to invoke the Feynman diagram technique and represent the corresponding vertex in the two-loop approximation as

$$\Gamma^{(2)}(k, \omega; \tau_0, u_0, \lambda_0) = \tau_0 + k^a - \frac{i\omega}{\lambda_0} - 96 u_0^2 D_0, \quad (10)$$

$$D_0 = \frac{3}{4} \int \frac{d^D q d^D p}{(1 + |\mathbf{q}|^a)(1 + |\mathbf{p}|^a)(3 + |\mathbf{q}|^a + |\mathbf{p}|^a + |\mathbf{p} + \mathbf{q}|^a - i\omega/\lambda)}.$$

The next step in the field-theoretic approach is the calculation of the scaling functions β, γ_τ, γ_φ, and γ_λ in the renormalization-group differential equation for vertexes:

$$\left[\mu \frac{\partial}{\partial \mu} + \beta \frac{\partial}{\partial u} - \gamma_\tau \tau \frac{\partial}{\partial \tau} + \gamma_\lambda \lambda \frac{\partial}{\partial \lambda} - \frac{m}{2} \gamma_\varphi \right] \times \Gamma^{(m)}(k, \omega; \tau, u, \lambda, \mu) = 0, \quad (11)$$

where the scaling parameter μ is introduced to change to dimensionless variables.

Further analysis requires the use of the function β and the dynamic scaling function γ_λ.

An expression for β in the two-loop approximation was obtained in [2]:

$$\beta = -(4 - D)\left[1 - 36 u J_0 + 1728 \left(2J_1 - J_0^2 - \frac{2}{9} G \right) u^2 \right],$$

$$J_1 = \int \frac{d^D q d^D p}{(1 + |\mathbf{q}|^a)^2 (1 + |\mathbf{p}|^a)(1 + |q^2 + p^2 + 2\mathbf{p} \cdot \mathbf{q}|^{a/2})},$$

$$J_0 = \int \frac{d^D q}{(1 + |\mathbf{q}|^a)^2},$$

$$G = -\frac{\partial}{\partial |\mathbf{k}|^a} \int \frac{d^D q d^D p}{(1 + |q^2 + k^2 + 2\mathbf{k} \cdot \mathbf{q}|^{a/2})(1 + |\mathbf{p}|^a)(1 + |q^2 + p^2 + 2\mathbf{p} \cdot \mathbf{q}|^{a/2})}.$$

The function γ_λ calculated in the two-loop approximation is

$$\gamma_\lambda = (4 - D) 2 (D' - G) u^2,$$
$$D' = \left. \frac{\partial D_0}{\partial (-i\omega/\lambda)} \right|_{k = 0, \omega = 0}. \quad (12)$$

Defining the effective interaction vertex

$$v = \frac{u}{J_0}, \quad (13)$$

one obtains the following expressions for β and γ_λ:

$$\beta = -(4 - D)\left[1 - 36 v + 1728 \left(2\tilde{J}_1 - 1 - \frac{2}{9} \tilde{G} \right) v^2 \right],$$

$$\gamma_\lambda = (4 - D) 96 (\tilde{D} - \tilde{G}) v^2, \quad (14)$$

$$\tilde{J}_1 = \frac{J_1}{J_0^2}, \quad \tilde{G} = \frac{G}{J_0^2}, \quad \tilde{D} = \frac{D'}{J_0^2}.$$

This redefinition is meaningful for $a \le D/2$. In this case, J_0, J_1, G and D' are divergent functions. Introducing the cutoff parameter Λ, we obtain finite expressions for the ratios

$$\frac{J_1}{J_0^2} = \frac{\int_0^\Lambda\int_0^\Lambda d^Dq d^Dp/((1+|\mathbf{q}|^a)^2(1+|\mathbf{p}|^a)(1+|q^2+p^2+2\mathbf{p}\cdot\mathbf{q}|^a))}{\left[\int_0^\Lambda d^Dq/(1+|\mathbf{q}|^a)^2\right]^2},$$

$$\frac{G}{J_0^2} = \frac{-\partial/(\partial|\mathbf{k}|^a)\int_0^\Lambda\int_0^\Lambda d^Dq d^Dp/((1+|q^2+k^2+2\mathbf{k}\cdot\mathbf{q}|^a)^2(1+|\mathbf{p}|^a)(1+|q^2+p^2+2\mathbf{p}\cdot\mathbf{q}|^a))}{\left[\int_0^\Lambda d^Dq/(1+|\mathbf{q}|^a)^2\right]^2}, \quad (15)$$

$$\frac{D'}{J_0^2} = \frac{3/4\int d^Dq d^Dp/((1+|\mathbf{q}|^a)(1+|\mathbf{p}|^a)(3+|\mathbf{q}|^a+|\mathbf{p}|^a+|\mathbf{p}+\mathbf{q}|^a)^2)}{\left[\int_0^\Lambda d^Dq/(1+|\mathbf{q}|^a)^2\right]^2},$$

as $\Lambda \longrightarrow \infty$.

The integrals are performed numerically. For $a \le D/2$, a sequence of J_1/J_0^2 and G/J_0^2 corresponding to various values of Λ is calculated and extrapolated to infinity.

Critical behavior is completely determined by the stable fixed points of the renormalization group (RG) transformation. These points can be found from the condition

$$\beta(v^*) = 0. \quad (16)$$

The effective interaction vertexes were evaluated at the stable fixed points of the RG transformation in [2].

The dynamic critical exponent z characterizing critical slowing-down of relaxation is determined by substituting the effective charges at a fixed point into the scaling function γ_λ:

$$z = 2 + \gamma_\lambda. \quad (17)$$

The table shows the stable fixed points of the RG transformation and the values of the dynamic critical exponent for $1.5 \le a \le 1.9$. When $0 < a < 1.5$, the only fixed point is the unstable gaussian one, $v^* = 0$.

Fixed points and values of the dynamic critical exponent for three-dimensional systems

a	v^*	z
1.5	0.015151	2.000072
1.6	0.015974	2.000180
1.7	0.020485	2.000777
1.8	0.023230	2.001529
1.9	0.042067	2.006628

A comparison of the present results with the value $z = 2.017$ of the dynamic critical exponent for three-dimensional systems with short-range interactions obtained in [6] demonstrates the essential role played by long-range interactions in critical dynamics of spin systems. In particular, the system's relaxation time increases according to the scaling $t \sim |T - T_c|^{\nu z}$, where ν is the critical exponent characterizing the increase in the correlation radius near a critical point. In three-dimensional systems with long-range interaction [2], both the critical dynamics and static behavior become increasingly gaussian as the long-range interaction parameter a decreases. When $a \le 1.8$, the critical behavior is virtually gaussian.

REFERENCES

1. E. Luijten and H. Mebingfeld, Phys. Rev. Lett. **86**, 5305 (2001).
2. S. V. Belim, Pis'ma Zh. Éksp. Teor. Fiz. **77**, 118 (2003) [JETP Lett. **77**, 112 (2003)].
3. M. E. Fisher, S.-K. Ma, and B. G. Nickel, Phys. Rev. Lett. **29**, 917 (1972).
4. C. De Dominicis, Lett. Nuovo Cimento **12**, 567 (1975).
5. E. Brezin, J. C. LeGuillon, and J. Zinn-Justin, Phys. Rev. D **8**, 434, 2418 (1973).
6. V. V. Prudnikov, S. V. Belim, A. V. Ivanov, *et al.*, Zh. Éksp. Teor. Fiz. **114**, 972 (1998) [JETP **87**, 527 (1998)].

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 748–759.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 854–867.

SOLIDS
Electronic Properties

Superconductivity in the Pseudogap State in the Hot-Spot Model: Ginzburg–Landau Expansion

É. Z. Kuchinskiĭ, M. V. Sadovskiĭ, and N. A. Strigina

Institute of Electrophysics, Ural Division, Russian Academy of Sciences, Yekaterinburg, 620016 Russia

e-mail: kuchinsk@iep.uran.ru; sadovski@iep.uran.ru; strigina@iep.uran.ru

Received May 15, 2003

Abstract—Peculiarities of the superconducting state (*s* and *d* pairing) are considered in the model of the pseudogap state induced by short-range order fluctuations of the dielectric (AFM (SDW) or CDW) type, which is based on the model of the Fermi surface with "hot spots." A microscopic derivation of the Ginzburg–Landau expansion is given with allowance for all Feynman diagrams in perturbation theory in the electron interaction with short-range order fluctuations responsible for strong scattering in the vicinity of hot spots. The superconducting transition temperature is determined as a function of the effective pseudogap width and the correlation length of short-range order fluctuations. Similar dependences are derived for the main parameters of a superconductor in the vicinity of the superconducting transition temperature. It is shown, in particular, that the specific heat jump at the transition point is considerably suppressed upon a transition to the pseudogap region on the phase diagram. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The pseudogap state observed in a wide region on the phase diagram of HTSC cuprates leads to numerous anomalies in the properties of these compounds both in the normal and in the superconducting state [1, 2]. In our opinion, the most feasible scenario for the formation of the pseudogap state in HTSC oxides is that [2] based on the existence of strong scattering of charge carriers under short-range order fluctuation of the "dielectric" type (antiferromagnetic AFM (SDW) or of the type of charge density waves (CDW)) in this region of the phase diagram. In the momentum space, this scattering occurs in the vicinity of the characteristic vector $\mathbf{Q} = (\pi/a, \pi/a)$ (a is the parameter of the 2D lattice), corresponding to period doubling (the antiferromagnetism vector) and is a predecessor of the spectrum rearrangement occurring during the establishment of the long-range AFM (SDW) order. Accordingly, an essentially non-Fermi-liquid rearrangement of the electron spectrum takes place in definite regions of the momentum space in the vicinity of the so-called hot spots at the Fermi surface [2]. In a number of recent experiments [3–5], precisely this scenario of pseudogap formation was convincingly confirmed. In the framework of the picture described above, it is possible to construct a simplified model of the pseudogap state, which describes the main features of this state [2] and takes into account the contribution from all Feynman diagrams in perturbation theory relative to scattering from (Gaussian) short-range order fluctuations with a characteristic scattering momentum from a neighborhood of vector $\mathbf{Q}$, which is determined by the corresponding correlation length ξ [6, 7].

Most of the previous theoretical publications were devoted to analysis of the models of the pseudogap state in the normal phase at $T > T_c$. In our earlier publications [8–11], we considered superconductivity using a simplified model of the pseudogap state, which is based on the assumption of the existence of hot (plane) regions at the Fermi surface. In the framework of this model, we constructed the Ginzburg–Landau expansion for various types of Cooper pairing [8, 10] and studied peculiarities of the superconducting state in the region of $T < T_c$ on the basis of analysis of the solutions to the Gor'kov equations [9–11]. It should be noted above all that we considered an extremely simplified model of Gaussian short-range order fluctuations with an infinitely large correlation length, for which an exact solution can be obtained for the pseudogap state [8, 9]. A more realistic case of finite correlation lengths was analyzed both for model [10] (under the assumption of self-averaging of the superconducting order parameter in short-range order fluctuations) and for an extremely simplified, exactly solvable model [11], in which the role of non-self-averaging effects could be analyzed [9, 11].

The present study aims at analyzing the basic properties of the superconducting state (for various types of pairing) arising against the background of a "dielectric" pseudogap in a more realistic model of hot spots at the Fermi surface. We will confine our analysis to a very close neighborhood of the superconducting transition temperature T_c based on the microscopic derivation of

the Ginzburg–Landau expansion, assuming that the superconducting order parameter is self-averaging, thus generalizing the approach proposed for the model of a hot region developed in [10].

2. HOT-SPOT MODEL AND PAIRING INTERACTION

In the model of an "almost antiferromagnetic" Fermi liquid, which is actively used for explaining the microscopic mechanism of HTSC [12, 13], the effective interaction of electrons with spin fluctuations is introduced. This interaction is described by the dynamic susceptibility characterized by the correlation length ξ of spin fluctuations (which must be determined from experiment), the vector $\mathbf{Q} = (\pi/a, \pi/a)$ of antiferromagnetic ordering in the dielectric phase, and the characteristic frequency ω_{sf} of spin fluctuations. This dynamic susceptibility and, hence, the effective interaction have peaks in the region of $\mathbf{q} \sim \mathbf{Q}$. Accordingly, two types of quasiparticles appear in the system: hot quasiparticles whose momenta lie in the vicinity of hot spots at the Fermi surface (Fig. 1) and cold quasiparticles whose momenta are in the vicinity of regions at the Fermi surface, surrounding the diagonals of the Brillouin zone [6]. As a matter of fact, quasiparticles from the neighborhoods of hot spots are strongly scattered over a vector on the order of $\mathbf{Q}$ due to their interaction with spin fluctuations, while this interaction for particles with momenta far away from hot spots is quite weak.

Considering the range of high temperatures $2\pi T \gg \omega_{sf}$, we can disregard the spin dynamics [6], confining our analysis to the static approximation. Computations can be considerably simplified and the contributions from higher orders of perturbation theory can be analyzed if we pass to the model interaction of electrons with spin (or charge) fluctuations of the form [7]

$$V_{\text{eff}}(\mathbf{q}) = W^2 \frac{2\xi^{-1}}{\xi^{-2} + (q_x - Q_x)^2} \frac{2\xi^{-1}}{\xi^{-2} + (q_y - Q_y)^2}, \quad (1)$$

where W is an effective parameter having the dimension of energy. Here, as in [6, 7], W and ξ are treated as phenomenological parameters (which are determined from experiment). Expression (1) is qualitatively similar to the static limit of the interaction considered in [12, 13] and quantitatively differs insignificantly from this limit in the most interesting region $|\mathbf{q} - \mathbf{Q}| < \xi^{-1}$, which determines scattering in the vicinity of hot spots, if the parameters appearing in this expression are appropriately defined. In fact, we are talking about the replacement of the actual interaction with dynamic short-range order fluctuations by the electron scattering from the static random (Gaussian) field of such fluctuations. The least justified assumption from the standpoint of physics is the one that concerns the static (and Gaussian) nature of fluctuations, which can be used only for quite high temperatures [6, 7]. At low temperatures (including those corresponding to the superconducting phase), the spin dynamics and the non-Gaussian nature of fluctuations may also become significant for the microscopy of Cooper pairing in the model of a nearly antiferromagnetic Fermi liquid [12, 13]. However, in our opinion, the static Gaussian approximation considered here might be sufficient for analyzing the qualitative effect of the pseudogap formation on superconductivity (in particular, in the vicinity of the superconducting transition temperature), which will be henceforth described by using the simple approach of the BCS theory and the Ginzburg–Landau phenomenology.

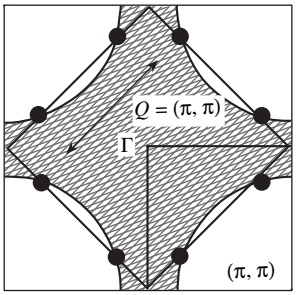


Fig. 1. Fermi surface with hot spots connected by a scattering momentum on the order of $\mathbf{Q} = (\pi/a, \pi/a)$.

The spectrum of the initial (free) quasiparticles will be taken in the form [6]

$$\xi_{\mathbf{p}} = -2t(\cos p_x a + \cos p_y a) - 4t' \cos p_x a \cos p_y a - \mu, \quad (2)$$

where t is the integral of transfer between the nearest neighbors, t' is the same for the next to nearest neighbors in a square lattice, a is the lattice parameter, and μ is the chemical potential. This expression provides a good approximation to the results of band calculations for real HTSC systems. For example, for $YBa_2Cu_3O_{6+\delta}$, we have $t = 0.25$ eV and $t' = -0.45t$ [6]. Chemical potential μ is determined by the carrier concentration.

In the limit of an infinitely large correlation length ($\xi \longrightarrow \infty$), the model of scattering from short-range order fluctuations of the type considered here has an exact solution [14]. For finite values of ξ, we can construct an approximate solution [7] generalizing the 1D approach proposed in [15]. In this case, it is possible to sum (approximately) the entire diagrammatic series for the one-particle electron Green function. As a result, the following recurrent procedure arises for the one-

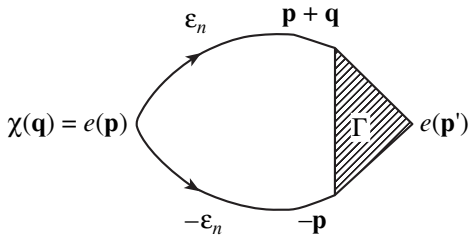


Fig. 2. Diagrammatic representation for generalized susceptibility $\chi(\mathbf{q})$ in a Cooper channel.

particle Green function $G(\varepsilon_n\mathbf{p})$ (representation in the form of a chain fraction) [6, 7, 15]:

$$G_k(\varepsilon_n\mathbf{p}) = \frac{1}{i\varepsilon_n - \xi_k(\mathbf{p}) + ikv_k\kappa - W^2 s(k+1)G_{k+1}(\varepsilon_n\mathbf{p})}; \quad (3)$$

here, $\kappa = \xi^{-1}$, $\varepsilon_n = 2\pi T(n + 1/2)$ (for definiteness, we assume that $\varepsilon_n > 0$),

$$\xi_k(\mathbf{p}) = \begin{cases} \xi_{\mathbf{p}+\mathbf{Q}} \text{ for odd } k, \\ \xi_{\mathbf{p}} \text{ for even } k, \end{cases} \quad (4)$$

$$v_k = \begin{cases} |v_x(\mathbf{p}+\mathbf{Q})| + |v_y(\mathbf{p}+\mathbf{Q})| \text{ for odd } k, \\ |v_x(\mathbf{p})| + |v_y(\mathbf{p})| \text{ for even } k, \end{cases} \quad (5)$$

where $\mathbf{v}(\mathbf{p}) = \partial\xi_{\mathbf{p}}/\partial\mathbf{p}$ is the velocity of a free quasiparticle.

The "physical" Green function is defined from relation (3) as $G(\varepsilon_n\mathbf{p}) \equiv G_0(\varepsilon_n\mathbf{p})$.

The combinatorial factor is given by

$$s(k) = k \quad (6)$$

in the case of commensurate functions with $\mathbf{Q} = (\pi/a, \pi/a)$ [15], which will be considered below, if we disregard their spin structure [6] (i.e., if we confine our analysis to fluctuations of the CDW type). If we take into account the spin structure of the interaction in the model of a nearly antiferromagnetic Fermi liquid (the spin-fermion model [6]), the combinatorics of diagrams becomes more intricate. In particular, the spin and charge two-particle vertices differ considerably in this model. The spin interaction was described in [6] by using the isotropic Heisenberg model. If we adopt the Ising model for this interaction, we will be left only with scattering processes with electron spin conservation, for which the commensurate combinatorics of diagrams (6) is valid both for the one-particle Green function and for spin and charge vertices. For this reason, we will confine our analysis only to the case of commensurate (6) "Ising" spin fluctuations (AFM, SDW) and commensurate charge fluctuations (CDW). The details corresponding to incommensurate fluctuations of the CDW type can be found in [7, 14, 15].

The conditions for applicability of this approximation were discussed in detail in [6, 7]. In the limit $\xi \longrightarrow \infty$, relation (3) is reduced to the exact solution [14], while in the limit $\xi \longrightarrow 0$ for a fixed value of W, relation (3) gives a physically correct limit of free electrons.

Passing to superconductivity in the system with developed short-range order fluctuations considered here, we assume that the superconducting pairing is due to the attractive potential acting between electrons with opposite spins of the simplest (BCS) form,

$$V_{sc}(\mathbf{p}, \mathbf{p}') = -Ve(\mathbf{p})e(\mathbf{p}'), \quad (7)$$

where for $e(\mathbf{p})$ we assume that

$$e(\mathbf{p}) = \begin{cases} 1 \quad (s \text{ pairing}), \\ \cos(p_x a) - \cos(p_y a) \quad (d_{x^2-y^2} \text{ pairing}), \\ \sin(p_x a)\sin(p_y a) \quad (d_{xy} \text{ pairing}), \\ \cos(p_x a) + \cos(p_y a) \\ (\text{anisotropic } s \text{ pairing}). \end{cases} \quad (8)$$

As usual, the attraction constant V is assumed to be nonzero in a certain layer of width $2\omega_c$ in the vicinity of the Fermi level (ω_c is the characteristic frequency of quanta, which ensures attraction between electrons). In the general case, the superconducting gap is anisotropic and has the form $\Delta(\mathbf{p}) = \Delta e(\mathbf{p})$.

The subsequent analysis will be carried out under the assumption of self-averaging of the energy gap of the superconductor over short-range order fluctuations, which allows us to use the standard approach of the theory of disordered superconductors [16, 17]. Under the conditions when the short-range order correlation length is $\xi \ll \xi_0$, where $\xi_0 \sim v_F/\Delta_0$ is the coherence length of the BCS theory (i.e., when fluctuations correlate over distances smaller than the characteristic size of Cooper pairs), the assumption concerning self-averaging of Δ must be preserved, being violated only in the region[1] $\xi > \xi_0$ [9–11].

3. COOPER INSTABILITY. RECURRENCE PROCEDURE FOR THE VERTEX PART

It is well known that the superconducting transition temperature can be determined from the equation for Cooper instability of the normal phase,

$$1 - V\chi(0; T) = 0, \quad (9)$$

[1] The absence of self-averaging of the superconducting gap even in the region $\xi < \xi_0$, which was obtained in our previous publication [11], is apparently due to the specific nature of the short-range order model used in that work.

where the generalized Cooper susceptibility is defined in Fig. 2 and is given by

$$\chi(\mathbf{q}; T) = -T\sum_{\varepsilon_n}\sum_{\mathbf{p},\mathbf{p}'} e(\mathbf{p})e(\mathbf{p}')\Phi_{\mathbf{p},\mathbf{p}'}(\varepsilon_n, -\varepsilon_n, \mathbf{q}); \quad (10)$$

here, $\Phi_{\mathbf{p},\mathbf{p}'}(\varepsilon_n, -\varepsilon_n, \mathbf{q})$ is a two-particle Green function in the Cooper channel, which takes into account the scattering from short-range order fluctuations.

We will first consider the case of charge fluctuations (CDW), where the interaction is independent of spin variables. For the *s* and d_{xy} pairing, the superconducting gap remains unchanged upon a transfer over **Q** (i.e., $e(\mathbf{p} + \mathbf{Q}) = e(\mathbf{p})$) and $e(\mathbf{p}') \approx e(\mathbf{p})$. In the case of anisotropic *s* and $d_{x^2-y^2}$ pairing, the superconducting gap reverses its sign upon a transfer over **Q** ($e(\mathbf{p} + \mathbf{Q}) = -e(\mathbf{p})$); consequently, $e(\mathbf{p}') \approx e(\mathbf{p})$ for $\mathbf{p}' \approx \mathbf{p}$ and $e(\mathbf{p}') \approx -e(\mathbf{p})$ for $\mathbf{p}' \approx \mathbf{p} + \mathbf{Q}$. Thus, for diagrams containing an even number of interaction lines connecting the upper (ε_n) and lower ($-\varepsilon_n$) electron lines, we have $\mathbf{p}' \approx \mathbf{p}$; Thus, we arrive at the same expression for the contribution to susceptibility as in the case of the *s* and d_{xy} pairing. On the other hand, for diagrams with an odd number of such interaction lines, we obtain an expression with the opposite sign for the contribution to susceptibility. This sign reversal can be attributed simply to the sign reversal for the interaction connecting the upper and lower electron lines of the loop in Fig. 2. In this case, we obtain for the generalized susceptibility the expression

$$\chi(\mathbf{q}; T) = -T\sum_{\varepsilon_n}\sum_{\mathbf{p}} G(\varepsilon_n\mathbf{p} + \mathbf{q})G(-\varepsilon_n, -\mathbf{p})e^2(\mathbf{p}) \times \Gamma^{\pm}(\varepsilon_n, -\varepsilon_n, \mathbf{q}), \quad (11)$$

where $\Gamma^{\pm}(\varepsilon_n, -\varepsilon_n, \mathbf{q})$ is the triangular vertex part taking into account the interaction with short-range order fluctuations, the superscript "±" allowing for the abovementioned difference in the signs of interactions connecting the upper and lower electron lines.

Let us now consider the scattering from spin fluctuations (AFM (SDW)). In this case, the line of interaction with the longitudinal spin component S^z, which embraces the vertex and changes the direction of the spin, should be supplemented with an additional factor of (–1) [6]. From this point of view, in the case of interaction with spin fluctuations, the types of pairing considered above "change places"[2] and the generalized Cooper susceptibility is determined by triangular vertex Γ^- for *s* and d_{xy} pairing and by triangular vertex Γ^+ for anisotropic *s* and $d_{x^2-y^2}$ pairing.

[2] This is due to the fact that the sign of the spin projection is reversed at the vertex of the interaction with the superconducting gap (we consider only the singlet pairing).

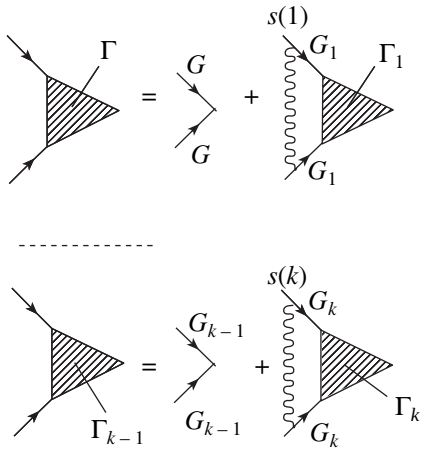


Fig. 3. Recurrence equations for the vertex part.

Thus, we must calculate the triangular vertices taking into account all diagrams (including cross diagrams) describing the interaction with dielectric fluctuations. The corresponding recurrence procedure for a 1D analog of our problem (and for real-valued frequencies, $T = 0$) was formulated for the first time in [18]. For the 2D model of the pseudogap with hot spots at the Fermi surface considered here, a generalization of this recurrence procedure is given in [19] in connection with optical conductivity calculations. The details of the corresponding derivation can also be found in [19]. A generalization to the case of Matsubara frequencies required for our problem can be carried out directly. For definiteness, we will henceforth assume, as before, that $\varepsilon_n > 0$. Ultimately, for a triangular vertex, we obtain the recurrence relation represented by the graphs in Fig. 3 (where the wavy line indicates the interaction with pseudogap fluctuations) and having the following analytic form:

$$\Gamma^{\pm}_{k-1}(\varepsilon_n, -\varepsilon_n, \mathbf{q}) = 1 \pm W^2 s(k) G_k \bar{G}_k \times \left\{1 + \frac{2ik v_k \kappa}{2i\varepsilon_n - \mathbf{v}_k \cdot \mathbf{q} - W^2 s(k+1)(G_{k+1} - \bar{G}_{k+1})}\right\} \times \Gamma^{\pm}_k(\varepsilon_n, -\varepsilon_n, \mathbf{q}); \quad (12)$$

here, $G_k = G_k(\varepsilon_n\mathbf{p} + \mathbf{q})$ and $\bar{G}_k = G_k(-\varepsilon_n, -\mathbf{p})$ are calculated in accordance with expression (3), v_k is defined by formula (5), and $\mathbf{v}_k$ have the form

$$\mathbf{v}_k = \begin{cases} \mathbf{v}(\mathbf{p} + \mathbf{Q}) \text{ for odd } k, \\ \mathbf{v}(\mathbf{p}) \text{ for even } k. \end{cases} \quad (13)$$

A "physical" vertex is defined as $\Gamma^{\pm}(\varepsilon_n, -\varepsilon_n, \mathbf{q}) \equiv \Gamma^{\pm}_0(\varepsilon_n, -\varepsilon_n, \mathbf{q})$.

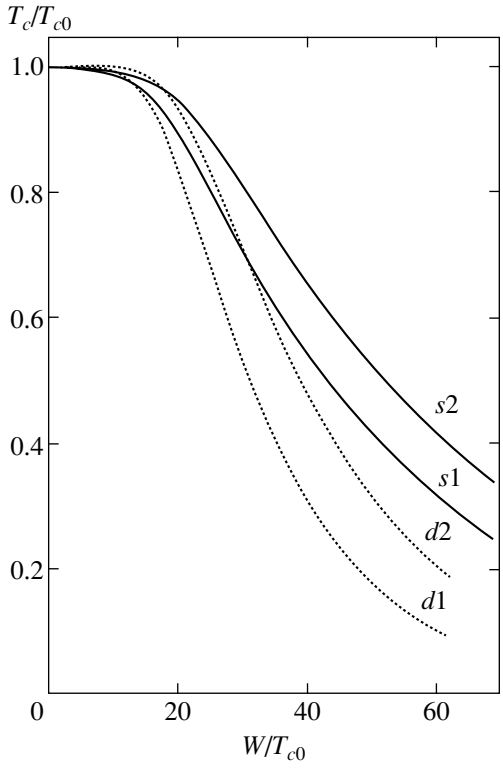


Fig. 4. Dependence of the superconducting transition temperature T_c/T_{c0} on effective pseudogap width W/T_{c0} for the *s*-type pairing and scattering from charge (CDW) fluctuations (curves *s*1 and *s*2) and for the $d_{x^2-y^2}$-type pairing and scattering from spin (AFM (SDW)) fluctuations (curves *d*1 and *d*2). The data are given for the following values of reciprocal correlation length: $\kappa a = 0.2$ (*s*1 and *d*1) and $\kappa a = 0.5$ (*s*2 and *d*2).

To determine T_c, we must consider the vertex for $\mathbf{q} = 0$. In this case, $\bar{G}_k = G_k^*$ and vertices Γ_k^+ and Γ_k^- become real-valued, which considerably simplifies procedures (12) For $\mathrm{Im}G_k$ and $\mathrm{Re}G_k$, we have the system of recurrence equations

$$\mathrm{Im}G_k = -\frac{\varepsilon_n + kv_k\kappa - W^2 s(k+1)\mathrm{Im}G_{k+1}}{D_k},$$
$$\mathrm{Re}G_k = -\frac{\xi_k(\mathbf{p}) + W^2 s(k+1)\mathrm{Re}G_{k+1}}{D_k}, \quad (14)$$

where $D_k = (\xi_k(\mathbf{p}) + W^2 s(k+1)\mathrm{Re}G_{k+1})^2 + (\varepsilon_n + kv_k\kappa - W^2 s(k+1)\mathrm{Im}G_{k+1})^2$ and the vertex part for $\mathbf{q} = 0$ can be determined from the equation

$$\Gamma_{k-1}^{\pm} = 1 \mp W^2 s(k)\frac{\mathrm{Im}G_k}{\varepsilon_n - W^2 s(k+1)\mathrm{Im}G_{k+1}}\Gamma_k^{\pm}. \quad (15)$$

Passing to numerical calculations, it is convenient to set the characteristic scale of energies (temperatures), which characterizes the superconducting state in our model in the absence of pseudogap fluctuations ($W = 0$). In this case, the equation for the corresponding superconducting transition temperature T_{c0} has the standard form for the BCS theory (in the general case of anisotropic pairing) and can be written as

$$1 = \frac{2VT}{\pi^2}\sum_{n=0}^{\bar{m}}\int_0^{\pi}dp_x\int_0^{\pi}dp_y\frac{e^2(\mathbf{p})}{\xi_{\mathbf{p}}^2 + \varepsilon_n^2}, \quad (16)$$

where $\bar{m} = [\omega_c/2\pi T_{c0}]$ is the dimensionless cutoff parameter for the sum over Matsubara frequencies. All calculations were made for a typical quasiparticle spectrum (2) in HTSC with $\mu = -1.3t$ and $t'/t = -0.4$. Choosing (quite arbitrarily) $\omega_c = 0.4t$ and $T_{c0} = 0.01t$, we can easily select the value of pairing parameter V in relation (16), which gives the same value of T_{c0} for various types of pairing enumerated in (8). In particular, we obtain $V/ta^2 = 1$ for the conventional isotropic *s*-type pairing and $V/ta^2 = 0.55$ for the $d_{x^2-y^2}$-type pairing. For the remaining types of pairing from relation (8), the values of the pairing constant for such a choice of parameters are found to be unrealistically high and we do not give the results of the corresponding calculations.[3]

Figures 4 and 5 show typical results of numerical calculations of the superconducting transition temperature T_c for a system with a pseudogap, which were obtained directly from relation (9) using the recurrence equations described above. It can be seen that pseudogap (dielectric) fluctuations considerably reduce the superconducting transition temperature in all cases. The $d_{x^2-y^2}$ pairing is suppressed much more rapidly than the isotropic *s* pairing. At the same time, a decrease in correlation length ξ (an increase in parameter κ) of pseudogap fluctuations facilitates an increase in T_c. These results are quite analogous to those obtained earlier in the model of hot regions [8, 10]. However, considerable differences also arise. It can be seen from Fig. 4 that the curve describing the dependence of T_c on pseudogap width W has a characteristic plateau in the region of $W < 10T_{c0}$ for *s* pairing and scattering from charge (CDW) fluctuations as well as for $d_{x^2-y^2}$ pairing and scattering from spin (AFM (SDW)) fluctuations[4] (i.e., in the cases when the upper sign in formulas (12) and (15) "operates", leading to sign-constant recurrence procedure for a vertex), while a consid-

[3] Of course, such a description on the basis of equations in the BCS theory with weak binding does not claim to be realistic in the cases of *s* and $d_{x^2-y^2}$ pairing considered here as well. We must just preset the characteristic scale of T_{c0} to express all temperatures in subsequent calculations in units of this temperature, assuming that a certain universality relative to this scale exists in the problem considered here.

[4] The latter case is realized, in all probability, in actual HTSC materials based on copper oxides.

erable suppression of T_c takes place on a scale of $W \sim 50T_{c0}$. Qualitative differences appear in the case of s pairing and scattering from spin (AFM (SDW)) fluctuations and in the case of $d_{x^2-y^2}$ pairing and scattering from charge fluctuations. Figure 5 shows that, in the latter case (when the lower sign in formulas (12) and (15) operates; i.e., an alternating procedure arises for a vertex), the rate of suppression of T_c is an order of magnitude higher. In the case of $d_{x^2-y^2}$ pairing, in the range of W/T_{c0} values corresponding to almost complete suppression of superconductivity, the accuracy of our calculations becomes considerably worse in view of the alternating nature of the recurrence procedure for the vertex part. In particular, a typical ambiguity of T_c may appear, which corresponds to possible existence of a narrow region of "recurrent" superconductivity on the phase diagram.[5] Such a behavior of T_c slightly resembles similar peculiarities emerging in superconductors with Kondo impurities [20]. Our calculations show, however, that the most probable scenario is the emergence of the critical value of parameter W/T_{c0}, for which superconductivity is completely suppressed. In this case, a region may appear, in which the transition to the superconducting state becomes a first-order phase transition analogously to the known situation in superconductors with a strong paramagnetic effect in an external magnetic field [21]. In any case, the effects arising in this case deserve a separate analysis. All results considered below correspond to the region of unambiguous behavior of T_c.

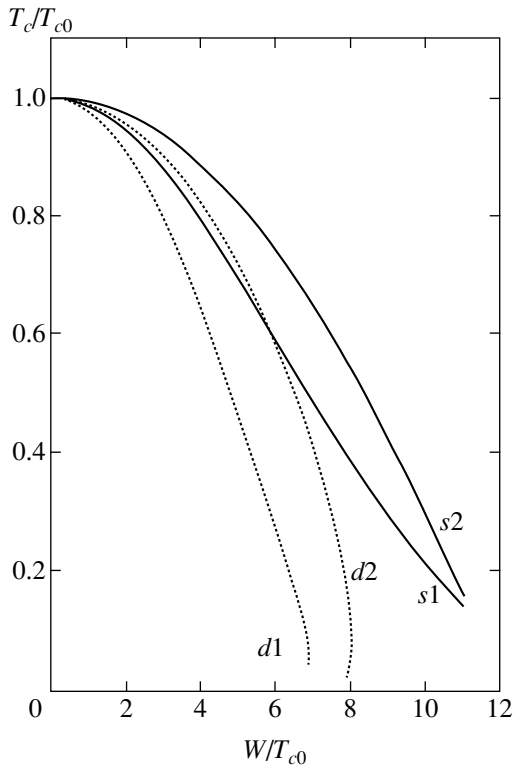


Fig. 5. Dependence of the superconducting transition temperature T_c/T_{c0} on effective pseudogap width W/T_{c0} for the s-type pairing and scattering from spin (AFM (SDW)) fluctuations (curves $s1$ and $s2$) and for the $d_{x^2-y^2}$-type pairing and scattering from charge (CDW) fluctuations (curves $d1$ and $d2$). The data are given for the following values of reciprocal correlation length: $\kappa a = 0.2$ ($s1$ and $d1$) and $\kappa a = 1.0$ ($s2$ and $d2$).

4. GINZBURG–LANDAU EXPANSION

In our earlier publication [8], the Ginzburg–Landau expansion was constructed in the exactly solvable model of a pseudogap with an infinitely large correlation length for short-range order fluctuations. Subsequently [10], these results were extended to the case of finite correlation lengths. In these publications, we considered, in fact, only charge fluctuations and used a simple model of the pseudogap state, which was based on the concept of hot (plane) regions existing at the Fermi surface. In this model, the sign of the superconducting gap remained unchanged upon a transfer over vector $\mathbf{Q}$ both for s and d pairing [10]. Here, we carry out the generalization to a more realistic case of the model of hot spots at the Fermi surface.

The Ginzburg–Landau expansion for the difference in the free energies of the superconducting and normal states can be written in the standard form

$$F_s - F_n = A|\Delta_{\mathbf{q}}|^2 + q^2 C|\Delta_{\mathbf{q}}|^2 + \frac{B}{2}|\Delta_{\mathbf{q}}|^4, \qquad (17)$$

where $\Delta_{\mathbf{q}}$ is the amplitude of the Fourier component of the order parameter, which can be written for various types of pairing in the form $\Delta(\mathbf{p}, \mathbf{q}) = \Delta_q e(\mathbf{p})$. Expansion (17) is determined by the graphs of the loop expansion for free energy in the field of order parameter fluctuations (shown by dashed lines) with a small wave vector $\mathbf{q}$ [8], which are represented in Fig. 6.

It is convenient to write the Ginzburg–Landau coefficients in the form

$$A = A_0 K_A, \quad C = C_0 K_C, \quad B = B_0 K_B, \qquad (18)$$

where A_0, C_0, and B_0 stand for the expressions for these coefficients in the absence of pseudogap fluctuations ($W = 0$), which are derived in the Appendix for an arbitrary spectrum ξ_p and various types of pairing,

$$A_0 = N_0(0)\frac{T - T_c}{T_c}\langle e^2(\mathbf{p})\rangle,$$

$$C_0 = N_0(0)\frac{7\zeta(3)}{32\pi^2 T_c^2}\langle |\mathbf{v}(\mathbf{p})|^2 e^2(\mathbf{p})\rangle, \qquad (19)$$

$$B_0 = N_0(0)\frac{7\zeta(3)}{8\pi^2 T_c^2}\langle e^4(\mathbf{p})\rangle,$$

[5] Such a peculiar behavior of T_c is manifested more strongly in the case of scattering from incommensurate pseudogap fluctuations.

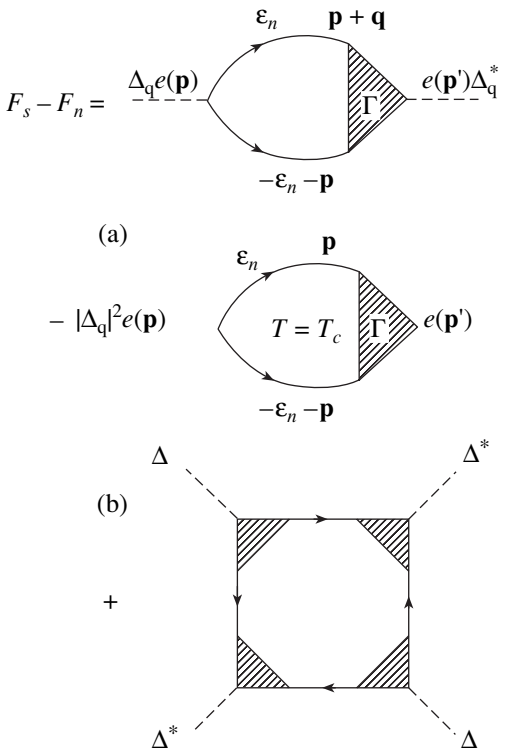


Fig. 6. Graphical form of the Ginzburg–Landau expansion.

angle brackets denote conventional averaging over the Fermi surface,

$$\langle \ldots \rangle = \frac{1}{N_0(0)} \sum_p \delta(\xi_\mathbf{p}) \ldots,$$

and $N_0(0)$ is the density of states for free electrons at the Fermi surface.

All peculiarities of the model in question, which are associated with the emergence of a pseudogap, are contained in dimensionless coefficients K_A, K_C, and K_B. In the absence of pseudogap fluctuations, all these coefficients are equal to unity.

It can be seen from Fig. 6a that coefficients K_A and K_C are completely determined by the generalized Cooper susceptibility [8, 10] $\chi(\mathbf{q}; T)$ depicted in Fig. 2:

$$K_A = \frac{\chi(0; T) - \chi(0; T_c)}{A_0}, \tag{20}$$

$$K_C = \lim_{q \to 0} \frac{\chi(\mathbf{q}; T_c) - \chi(0; T_c)}{q^2 C_0}. \tag{21}$$

It was shown above that the generalized susceptibility can be found from relation (11), where the triangular vertices are determined by recurrence procedures (12); this allows us to directly calculate coefficients K_A and K_C numerically.

The situation with coefficient B is more complicated in the general case. Calculations can be significantly simplified if we confine our analysis, as usual, to the case of $q = 0$ in the order of $|\Delta_q|^4$ and define coefficient B by the diagram show in Fig. 6b.Then we obtain the following expression for coefficient K_B:

$$K_B = \frac{T_c}{B_0} \sum_{\varepsilon_n} \sum_\mathbf{p} e^4(\mathbf{p}) (G(\varepsilon_n \mathbf{p}) G(-\varepsilon_n, -\mathbf{p}))^2 \times (\Gamma^{\pm}(\varepsilon_n, -\varepsilon_n, 0))^4. \tag{22}$$

It should be noted from the very outset that this expression leads to a positive definite coefficient B. This follows from the fact that $G(-\varepsilon_n, -\mathbf{p}) = G^*(\varepsilon_n \mathbf{p})$ so that $G(\varepsilon_n \mathbf{p}) G(-\varepsilon_n, -\mathbf{p})$ is real-valued; accordingly, vertex part $\Gamma\pm(\varepsilon_n, -\varepsilon_n, 0)$ defined by recurrence procedure (15) is also real.

5. PHYSICAL CHARACTERISTICS OF SUPERCONDUCTORS WITH A PSEUDOGAP

It is well known that the Ginzburg–Landau equations define two characteristic lengths of superconductor, viz., the coherence length and the magnetic field penetration depth.

For a given temperature, coherence length $\xi(T)$ gives the characteristic scale of inhomogeneities of order parameter Δ:

$$\xi^2(T) = -\frac{C}{A}. \tag{23}$$

In the absence of a pseudogap, we can write

$$\xi_{BCS}^2(T) = -\frac{C_0}{A_0}. \tag{24}$$

In our model, we have

$$\frac{\xi^2(T)}{\xi_{BCS}^2(T)} = \frac{K_C}{K_A}. \tag{25}$$

For the magnetic field penetration depth, we have

$$\lambda^2(T) = -\frac{c^2}{32\pi e^2} \frac{B}{AC}. \tag{26}$$

Analogously to relation (25), in the given model, we can write

$$\frac{\lambda(T)}{\lambda_{BCS}(T)} = \left(\frac{K_B}{K_A K_C}\right)^{1/2}. \tag{27}$$

In the vicinity of T_c, the upper critical field H_{c2} is

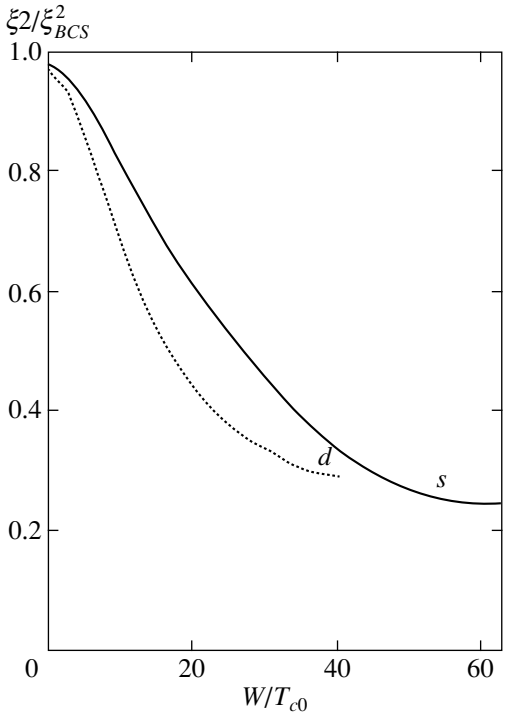


Fig. 7. Dependence of the squared coherence length ξ^2/ξ^2_{BCS} on effective pseudogap width W/T_{c0} for the *s*-type pairing and scattering from charge (CDW) fluctuations (solid curve) and for the $d_{x^2-y^2}$-type pairing and scattering from spin (AFM (SDW)) fluctuations (dotted curve). The data are given for the reciprocal correlation length $\kappa a = 0.2$.

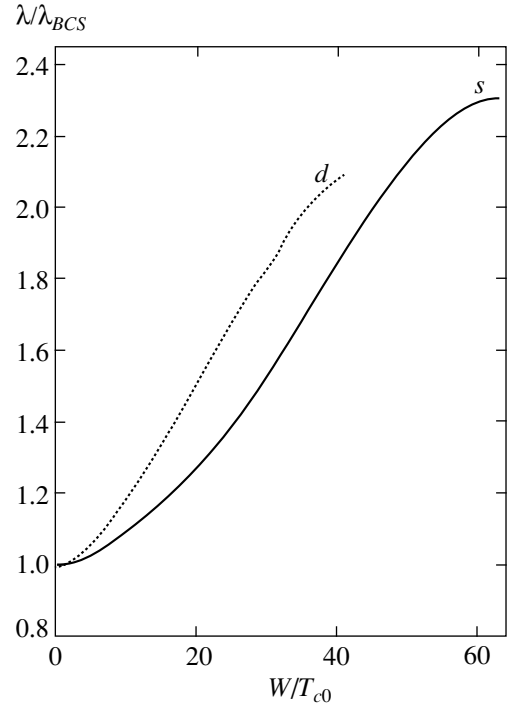


Fig. 8. Dependence of the penetration depth λ/λ_{BCS} on effective pseudogap width W/T_{c0} for the *s*-type pairing and scattering from charge (CDW) fluctuations (solid curve) and for the $d_{x^2-y^2}$-type pairing and scattering from spin (AFM (SDW)) fluctuations (dotted curve). The data are given for the reciprocal correlation length $\kappa a = 0.2$.

defined in terms of the Ginzburg–Landau coefficients as

$$H_{c2} = \frac{\phi_0}{2\pi\xi^2(T)} = -\frac{\phi_0}{2\pi}\frac{A}{C}, \tag{28}$$

where $\phi_0 = c\pi/|e|$ is the magnetic flux quantum. Then the slope of the curve describing the upper critical field near T_c is given by

$$\left|\frac{dH_{c2}}{dT}\right|_{T_c} = \frac{16\pi\phi_0\langle e^2(\mathbf{p})\rangle}{7\zeta(3)\langle|\mathbf{v}(\mathbf{p})|^2 e^2(\mathbf{p})\rangle}T_c\frac{K_A}{K_C}. \tag{29}$$

The specific heat discontinuity at the transition point has the form

$$(C_s - C_n)_{T_c} = \frac{T_c}{B}\left(\frac{A}{T - T_c}\right)^2, \tag{30}$$

where C_s and C_n are the specific heats of the superconducting and normal states, respectively. At temperature T_{c0} (in the absence of a pseudogap, $W = 0$), we have

$$(C_s - C_n)_{T_{c0}} = N(0)\frac{8\pi^2 T_{c0}\langle e^2(\mathbf{p})\rangle^2}{7\zeta(3)\langle e^4(\mathbf{p})\rangle}. \tag{31}$$

Then the relative specific heat discontinuity in the given model can be written as

$$\Delta C \equiv \frac{(C_s - C_n)_{T_c}}{(C_s - C_n)_{T_{c0}}} = \frac{T_c}{T_{c0}}\frac{K_A^2}{K_B}. \tag{32}$$

Coefficients K_A, K_B, and K_C were calculated numerically for the same typical parameters of the model as in the calculations of T_c described above. The numerical values of these coefficients as such are not very interesting and are not given here.[6] Figures 7–12 show the W/T_{c0} dependences of the corresponding physical quantities, defined by relations (23)–(32). In accordance with the situation with T_c described above, two qualitatively different modes of the behavior are also observed in this case depending on whether the behavior of the vertex part in the recurrence equations is sign-constant or alternating (the upper and lower signs in relation (12) and spin or charge fluctuations). The results of calculations of physical quantities for the first case (the *s*-type pairing and scattering from charge (CDW) fluctuations as well as the $d_{x^2-y^2}$-type pairing and scattering from spin (AFM (SDW)) fluctuations) are shown in Figs. 7–

[6] The typical dependences of these coefficients on parameter W/T_{c0} are functions rapidly decreasing from unity in the superconductivity range.

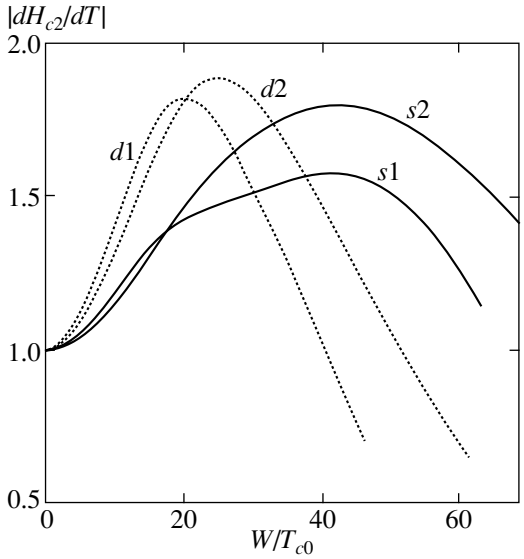


Fig. 9. Dependence of the derivative (slope) of the upper critical field on effective pseudogap width W/T_{c0} for the *s*-type pairing and scattering from charge (CDW) fluctuations (curves *s*1 and *s*2) and for the $d_{x^2-y^2}$-type pairing and scattering from spin (AFM (SDW)) fluctuations (curves *d*1 and *d*2). The data are given for the values of reciprocal correlation length $\kappa a = 0.2$ (*s*1 and *d*1) and $\kappa a = 0.5$ (*s*2 and *d*2) and are normalized to the value of the derivative in the absence of a pseudogap.

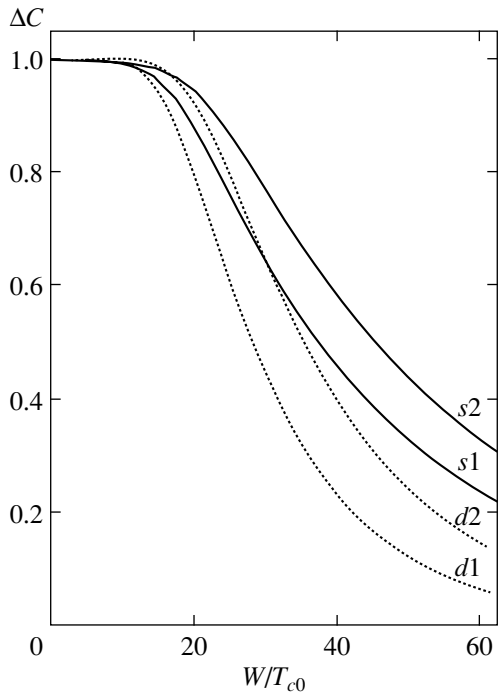


Fig. 10. Dependence of the specific heat discontinuity at the transition point on effective pseudogap width W/T_{c0} for the *s*-type pairing and scattering from charge (CDW) fluctuations (curves *s*1 and *s*2) and for the $d_{x^2-y^2}$-type pairing and scattering from spin (AFM (SDW)) fluctuations (curves *d*1 and *d*2). The data are given for the values of reciprocal correlation length $\kappa a = 0.2$ (*s*1 and *d*1) and $\kappa a = 0.5$ (*s*2 and *d*2).

10. It can be seen that, with increasing pseudogap width *W*, coherence length $\xi(T)$ decreases, while penetration depth $\lambda(T)$ increases as compared to the corresponding values in the BCS theory. Both these characteristic lengths exhibit a very weak dependence on parameter κ; for this reason, the results in Figs. 7 and 8 are given only for $\kappa a = 0.2$. The slope (derivative) of the upper critical field at $T = T_c$ first increases and then begins to decrease. The most typical is the decrease in the specific heat discontinuity as compared to the BCS value (see Fig. 10), which is in direct qualitative agreement with experimental data [22]. It should be noted that the specific heat discontinuity in our model also has a characteristic plateau in the region of $W/T_{c0} < 10$, which is similar to that noted above in the corresponding dependence of T_c.

The behavior of physical quantities in the case of the *s*-type pairing and scattering from spin (AFM (SDW)) fluctuations and the $d_{x^2-y^2}$-type pairing and scattering from charge (CDW) fluctuations is illustrated in Figs. 11 and 12. Data on the characteristic lengths are not shown since both coherence length $\xi(T)$ and penetration depth $\lambda(T)$ are virtually the same as the corresponding values in the BCS theory everywhere in the superconductivity range (except a small neighborhood of the region of ambiguity and vanishing of T_c, in which

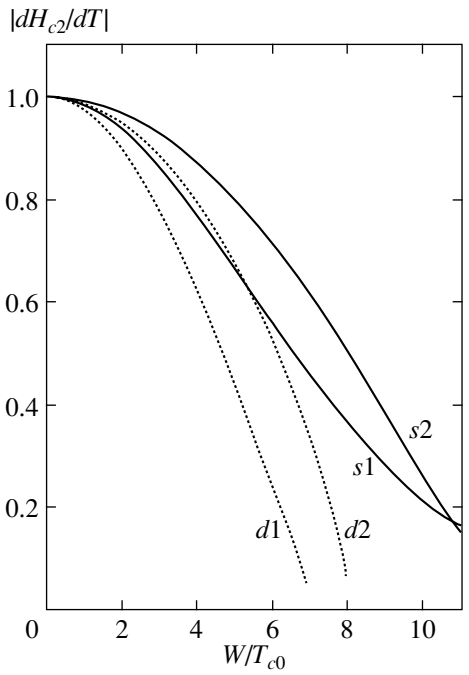


Fig. 11. Dependence of the derivative (slope) of the upper critical field on effective pseudogap width W/T_{c0} for the *s*-type pairing and scattering from spin (AFM (SDW)) fluctuations (curves *s*1 and *s*2) and for the $d_{x^2-y^2}$-type pairing and scattering from charge (CDW) fluctuations (curves *d*1 and *d*2). The data are given for the values of reciprocal correlation length $\kappa a = 0.2$ (*s*1 and *d*1) and $\kappa a = 1.0$ (*s*2 and *d*2) and normalized to the value of the derivative in the absence of a pseudogap.

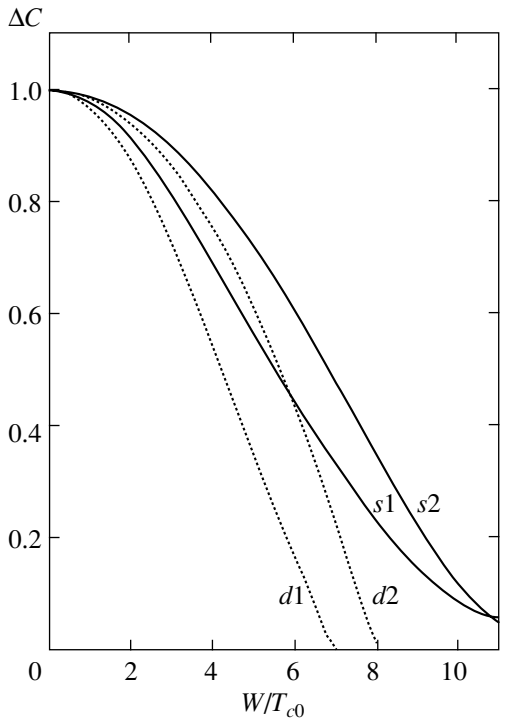


Fig. 12. Dependence of specific heat discontinuity at the transition point on the effective pseudogap width W/T_{c0} for *s*-type pairing and scattering from spin (AFM (CDW)) fluctuations (curves *s*1 and *s*2) and for $d_{x^2-y^2}$-type pairing and scattering from charge (CDW) fluctuations (curves *d*1 and *d*2). The data are given for the values of reciprocal correlation length $\kappa a = 0.2$ (*s*1 and *d*1) and $\kappa a = 1.0$ (*s*2 and *d*2).

these lengths sharply increase). As regards the derivative of the upper critical field and the specific heat discontinuity at the superconducting transition point, the values of these quantities decrease quite rapidly with increasing parameter W/T_{c0} apparently up to its critical value at which T_c is completely suppressed (or to the value at which a narrow region of the first-order transition is formed).

6. CONCLUSIONS

We have considered the peculiarities of the superconducting state emerging in the pseudogap state due to scattering of electrons from dielectric short-range order fluctuations in the model of hot spots at the Fermi surface. Our analysis was based on the microscopic derivation of the Ginzburg–Landau expansion taking into account all orders of perturbation theory in scattering from pseudogap fluctuations. The condensed phase of such a superconductor can be described on the basis of the corresponding analysis of the Gor'kov equations for a superconductor with a pseudogap (see [10]) and is the subject of special analysis.

The main result of this study is the demonstration of superconductivity suppression by pseudogap fluctuations of the CDW or AFM (SDW) type and the separation of two classes of qualitatively different models of such suppression depending on the sign-constant or alternating behavior of the vertex part in the recurrence equations (the upper or lower signs in expression (12) and spin or charge fluctuations). The version with scattering from spin fluctuations and pairing with the $d_{x^2-y^2}$-type symmetry is observed in high-T_c superconductors based on copper oxides; however, we are not aware of systems in which the peculiar behavior obtained above for the *s*-type pairing and scattering from spin (AFM (SDW)) fluctuations as well as for $d_{x^2-y^2}$-type pairing and scattering from charge (CDW) fluctuations is realized. The search for such systems is of considerable interest.

The most important question in the description of the pseudogap state of actual HTSC systems is the behavior of physical parameters upon a change in the carrier concentration. In our model, the concentration dependence must be expressed in terms of the corresponding dependence of effective width W of the pseudogap and correlation length ξ. Unfortunately, such dependences can be determined from experiment only indirectly and have been studied insufficiently [1, 2].[7] In a very rough approximation, we can state that correlation length ξ in a wide concentration range does not vary very strongly, while pseudogap width W linearly decreases with increasing charge carrier concentration from values on the order of 10^3 K in the vicinity of the dielectric phase region to values on the order of the superconducting transition temperature as we approach the optimal doping level, vanishing at slightly higher carrier concentrations (see Fig. 6 in review [2], which is based on Fig. 4 in [3], where the corresponding set of data is given for the YBCO system). Using this regularity, one can easily recalculate the above dependences on W to the corresponding dependences on the charge carrier concentration. In the extremely simplified version of our model with an infinitely large correlation length and the Fermi surface with complete nesting, such an analysis was carried out in a recent publication [23] under the assumption that the value of T_{c0} is also a linear function of the concentration. The typical form of the phase diagram for HTSC cuprates was completely reproduced qualitatively. At the same time, the obvious roughness of the model and the absence of reliable experimental data on the concentration dependences of W, ξ, and T_{c0} do not make it possible to treat the attempts at "improving" these qualitative conclusions very seriously.

In addition to the repeatedly mentioned disregard of the dynamics of short-range order fluctuations and the confinement to Gaussian fluctuations alone, it should be noted once again that the disadvantages of the model

[7] In addition, an analogous dependence of the value of T_{c0}, which is completely unknown, may turn out to be significant.

considered here also include the simplified analysis of the spin structure of interactions, which presumes that these interactions are of the Ising type. It would be interesting to also carry out a similar analysis for the general case of an interaction of the Heisenberg type.

ACKNOWLEDGMENTS

This study was partly supported by the Russian Foundation for Basic Research (project no. 02-02-16031); by the programs of fundamental studies in "Quantum Macrophysics" (the Presidium of the Russian Academy of Sciences) and "Strongly Correlated Electrons in Semiconductors, Metals, Superconductors, and Magnetic Materials" (Department of Physical Sciences, Russian Academy of Sciences); and by the project "Investigation of Collective and Quantum Effects in Condensed Media" (Ministry of Industry and Science of the Russian Federation).

APPENDIX

Ginzburg–Landau Coefficients for Anisotropic Pairing in the Absence of a Pseudogap

In the absence of fluctuations ($W = 0$), the generalized Cooper susceptibility, which is defined by the diagram in Fig. 2, assumes the form

$$\chi_0(\mathbf{q}; T) = -T\sum_{\varepsilon_n}\sum_{\mathbf{p}} e^2(\mathbf{p})\frac{1}{i\varepsilon_n - \xi_{\mathbf{p}+\mathbf{q}}}\frac{1}{-i\varepsilon_n - \xi_{\mathbf{p}}}. \quad \text{(A.1)}$$

For the susceptibility at $\mathbf{q} = 0$, which determines coefficient A_0, we obtain the expression

$$\begin{aligned}\chi_0(0; T) &= -T\sum_{\varepsilon_n}\sum_{\mathbf{p}} e^2(\mathbf{p})\frac{1}{\varepsilon_n^2 + \xi_{\mathbf{p}}^2}\\ &= -T\sum_{\varepsilon_n}\int_{-\infty}^{\infty} d\xi \frac{1}{\varepsilon_n^2 + \xi^2}\sum_{\mathbf{p}}\delta(\xi - \xi_{\mathbf{p}})e^2(\mathbf{p})\\ &\approx -N_0(0)T\sum_{\varepsilon_n}\int_{-\infty}^{\infty} d\xi \frac{1}{\varepsilon_n^2 + \xi^2}\frac{\Sigma_{\mathbf{p}}\delta(\xi_{\mathbf{p}})e^2(\mathbf{p})}{N_0(0)}\\ &= \chi_{BCS}(0; T)\langle e^2(\mathbf{p})\rangle,\end{aligned} \quad \text{(A.2)}$$

where the angle brackets denote averaging over the Fermi surface and the standard susceptibility $\chi_{BCS}(0; T)$ in the BCS model for isotropic *s* pairing is introduced.

As a result, coefficient A_0 assumes the form

$$A_0 = \chi_0(0; T) - \chi_0(0; T_c) = A_{BCS}\langle e^2(\mathbf{p})\rangle, \quad \text{(A.3)}$$

where

$$\begin{aligned}A_{BCS} &= \chi_{BCS}(0; T) - \chi_{BCS}(0; T_c)\\ &= N_0(0)\frac{T - T_c}{T_c}\end{aligned} \quad \text{(A.4)}$$

is the standard expression for coefficient A in the case of isotropic *s* pairing.

Coefficient C_0 of the Ginzburg–Landau expansion is defined by generalized susceptibility (A.1) for small values of $\mathbf{q}$:

$$C_0 = \lim_{q\to 0}\frac{\chi_0(\mathbf{q}; T_c) - \chi_0(0; T_c)}{q^2}. \quad \text{(A.5)}$$

Expanding expression (A.1) for $\chi_0(\mathbf{q}; T_c)$ into a series in small q, we obtain

$$\begin{aligned}\chi_0(\mathbf{q}; T_c) &= \chi_0(0; T_c)\\ &+ T_c\sum_{\varepsilon_n}\sum_{\mathbf{p}}\frac{3\varepsilon_n^2 - \xi_{\mathbf{p}}^2}{4(\varepsilon_n^2 + \xi_{\mathbf{p}}^2)^3}e^2(\mathbf{p})(\mathbf{v}(\mathbf{p})\mathbf{q})^2,\end{aligned} \quad \text{(A.6)}$$

so that we have for coefficient C_0 the expression

$$\begin{aligned}C_0 &= T_c\sum_{\varepsilon_n}\int_{-\infty}^{\infty} d\xi \frac{3\varepsilon_n^2 - \xi^2}{4(\varepsilon_n^2 + \xi^2)^3}\\ &\times\sum_{\mathbf{p}}\delta(\xi - \xi_{\mathbf{p}})e^2(\mathbf{p})|\mathbf{v}(\mathbf{p})|^2\cos^2\phi\\ &\approx T_c\sum_{\varepsilon_n}\int_{-\infty}^{\infty} d\xi \frac{3\varepsilon_n^2 - \xi^2}{4(\varepsilon_n^2 + \xi^2)^3}\\ &\times\sum_{\mathbf{p}}\delta(\xi_{\mathbf{p}})e^2(\mathbf{p})|\mathbf{v}(\mathbf{p})|^2\cos^2\phi\\ &= N_0(0)\frac{7\zeta(3)}{16\pi^2 T_c^2}\langle e^2(\mathbf{p})|\mathbf{v}(\mathbf{p})|^2\cos^2\phi\rangle,\end{aligned} \quad \text{(A.7)}$$

where ϕ is the angle between vectors $\mathbf{v}(\mathbf{p})$ and $\mathbf{q}$ and

$$\zeta(3) = \sum_{n=1}^{\infty}\frac{1}{n^3} \approx 1.202.$$

For a square lattice, the Fermi surface and, hence, $|\mathbf{v}(\mathbf{p})|$ also possess a symmetry relative to rotation through angle $\pi/2$; the same symmetry is also inherent

in $e^2(\mathbf{p})$ for the types of pairing considered here. Consequently, we can easily find that

$$\langle e^2(\mathbf{p})|\mathbf{v}(\mathbf{p})|^2\cos^2\phi\rangle$$
$$= \frac{1}{2}\langle e^2(\mathbf{p})|\mathbf{v}(\mathbf{p})|^2(1+\cos 2\phi)\rangle \qquad \text{(A.8)}$$
$$= \frac{1}{2}\langle e^2(\mathbf{p})|\mathbf{v}(\mathbf{p})|^2\rangle,$$

since quantity $\cos 2\phi$ reverses its sign when vector $\mathbf{p}$ rotates through angle $\pi/2$. Indeed, the direction of velocity $\mathbf{v}(\mathbf{p})$ upon this rotation changes to the perpendicular direction; accordingly, $\cos 2\phi \longrightarrow -\cos 2\phi$. As a result, we obtain the isotropic expression for coefficient C_0,

$$C_0 = N_0(0)\frac{7\zeta(3)}{32\pi^2 T_c^2}\langle|\mathbf{v}(\mathbf{p})|^2 e^2(\mathbf{p})\rangle; \qquad \text{(A.9)}$$

in the case of isotropic *s* pairing and a spherical Fermi surface, this expression acquires the standard form

$$C_{BCS} = N_0(0)\frac{7\zeta(3)v_F^2}{32\pi^2 T_c^2}. \qquad \text{(A.10)}$$

In the absence of pseudogap fluctuations ($W = 0$) and for $\mathbf{q} = 0$, coefficient B defined by the diagram in Fig. 6b has the form

$$B_0 = T_c\sum_{\varepsilon_n}\sum_{\mathbf{p}}\frac{1}{(\varepsilon_n^2+\xi_{\mathbf{p}}^2)^2}e^4(\mathbf{p})$$
$$= T_c\sum_{\varepsilon_n}\int_{-\infty}^{\infty}d\xi\frac{1}{(\varepsilon_n^2+\xi^2)^2}\sum_{\mathbf{p}}\delta(\xi-\xi_{\mathbf{p}})e^4(\mathbf{p})$$
$$\approx N_0(0)T_c\sum_{\varepsilon_n}\int_{-\infty}^{\infty}d\xi\frac{1}{(\varepsilon_n^2+\xi^2)^2}\frac{\Sigma_{\mathbf{p}}\delta(\xi_{\mathbf{p}})e^4(\mathbf{p})}{N_0(0)} \qquad \text{(A.11)}$$
$$= B_{BCS}\langle e^4(\mathbf{p})\rangle,$$

where

$$B_{BCS} = N_0(0)\frac{7\zeta(3)}{8\pi^2 T_c^2} \qquad \text{(A.12)}$$

is the standard expression for coefficient B in the case of isotropic *s* pairing.

REFERENCES

1. T. Timusk and B. Statt, Rep. Prog. Phys. **62**, 61 (1999).
2. M. V. Sadovskiĭ, Usp. Fiz. Nauk **171**, 539 (2001) [Phys. Usp. **44**, 515 (2001)].
3. J. L. Tallon and J. W. Loram, Physica C (Amsterdam) **349**, 53 (2000).
4. V. M. Krasnov, A. Yurgens, D. Winkler, *et al.*, Phys. Rev. Lett. **84**, 5860 (2000).
5. N. P. Armitage, D. H. Lu, C. Kim, *et al.*, Phys. Rev. Lett. **87**, 147003 (2001).
6. J. Schmalian, D. Pines, and B. Stojkovic, Phys. Rev. Lett. **80**, 3839 (1998); Phys. Rev. B **60**, 667 (1999).
7. É. Z. Kuchinskiĭ and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **115**, 1765 (1999) [JETP **88**, 968 (1999)].
8. A. I. Posazhennikova and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **115**, 632 (1999) [JETP **88**, 347 (1999)].
9. É. Z. Kuchinskiĭ and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **117**, 613 (2000) [JETP **90**, 535 (2000)].
10. É. Z. Kuchinskiĭ and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **119**, 553 (2001) [JETP **92**, 480 (2001)].
11. É. Z. Kuchinskiĭ and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **121**, 758 (2002) [JETP **94**, 654 (2002)].
12. P. Monthoux, A. V. Balatsky, and D. Pines, Phys. Rev. B **46**, 14803 (1992).
13. P. Monthoux and D. Pines, Phys. Rev. B **47**, 6069 (1993); **49**, 4261 (1994).
14. M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **66**, 1720 (1974) [Sov. Phys. JETP **39**, 845 (1974)]; Fiz. Tverd. Tela (Leningrad) **16**, 2504 (1974) [Sov. Phys. Solid State **16**, 1632 (1974)].
15. M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **77**, 2070 (1979) [Sov. Phys. JETP **50**, 989 (1979)].
16. L. P. Gor'kov, Zh. Éksp. Teor. Fiz. **37**, 1407 (1959) [Sov. Phys. JETP **10**, 998 (1959)].
17. P. G. de Gennes, *Superconductivity of Metals and Alloys* (Benjamin, New York, 1966; Mir, Moscow, 1968).
18. M. V. Sadovskiĭ and A. A. Timofeev, Sverkhprovodimost: Fiz. Khim. Tekh. **4**, 11 (1991); J. Mosc. Phys. Soc. **1**, 391 (1991).
19. M. V. Sadovskiĭ and N. A. Strigina, Zh. Éksp. Teor. Fiz. **122**, 610 (2002) [JETP **95**, 526 (2002)].
20. E. Müller-Hartmann and J. Zittartz, Phys. Rev. Lett. **26**, 428 (1971).
21. D. St. James, G. Sarma, and E. J. Thomas, *Type II Superconductivity* (Pergamon Press, Oxford, 1969; Mir, Moscow, 1970).
22. J. W. Loram, K. A. Mirza, J. R. Cooper, *et al.*, J. Supercond. **7**, 243 (1994).
23. A. Posazhennikova and P. Coleman, Phys. Rev. B **67**, 165109 (2003).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 760–769.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 868–878.

SOLIDS
Electronic Properties

Short-Range Order, Large-Scale Potential Fluctuations, and Photoluminescence in Amorphous SiN_x

V. A. Gritsenko[a,*], D. V. Gritsenko[a], Yu. N. Novikov[a], R. W. M. Kwok[b], and I. Bello[c]

[a]*Institute of Semiconductor Physics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 630090 Russia*
[b]*Department of Chemistry, Chinese University of Hong Kong, Shatin, Hong Kong, China*
[c]*Department of Physics and Materials Science, City University of Hong Kong, Tat Chee Avenue, Hong Kong, China*
**e-mail: grits@isp.nsc.ru*
Received September 12, 2003

Abstract—The short-range order and electron structure of amorphous silicon nitride SiN_x ($x < 4/3$) have been studied by a combination of methods including high-resolution X-ray photoelectron spectroscopy. Neither random bonding nor random mixture models can adequately describe the structure of this compound. An intermediate model is proposed, which assumes giant potential fluctuations for electrons and holes, caused by inhomogeneities in the local chemical composition. The characteristic scale of these fluctuations for both electrons and holes is about 1.5 eV. The photoluminescence in SiN_x is interpreted in terms of the optical transitions between quantum states of amorphous silicon clusters. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Amorphous silicon nitrides SiN_x, together with silicon dioxide SiO_2, are the main dielectrics used in modern silicon-based electronic devices. Silicon nitride exhibits a unique memory effect, being capable of localizing and capturing injected electrons and holes with a giant time of localized carrier trapping (about 10 years at 300 K) [1]. In recent years, the memory effect in silicon nitride has been used for developing electrically rewriteable ROM devices of Giga- and Terabit capacity [2].

There are two alternative models describing the structure of amorphous layers of nonstoichiometric tetrahedral silicon compounds (SiO_x, SiO_xN_y, SiN_x): the random mixture (RM) model and the random bonding (RB) model [3–17]. The RM model assumes that SiN_x comprises a mixture of two phases, amorphous silicon (a-Si) and silicon nitride (Si_3N_4), and is composed of $SiSi_4$ and SiN_4 tetrahedra [8, 10]. According to the RB model, the structure of SiN_x represents a network composed of $SiN_\nu Si_{4-\nu}$ tetrahedra of five types with $\nu = 0$–4 [3, 4, 11, 12]. Since amorphous SiN_x is synthesized under thermodynamically nonequilibrium conditions, the product structure depends on the method of synthesis. In particular, it was established that the structure of SiN_x obtained by plasma deposition is described by the RM model [8]. Silicon-based devices also widely employ SiN_x synthesized by high-temperature pyrolysis of silicon- and nitrogen-containing gas mixtures. The silicon-containing component is typically silane SiH_4, silicon tetrachloride $SiCl_4$, or dichlorosilane SiH_2Cl_2; the nitrogen-containing gas is ammonium NH_3. The process is carried out at a temperature of 700–800°C. The structure (short-range order) of SiN_x obtained by pyrolysis still remains unstudied.

Although the memory effect in SiN_x has been studied for more than a quarter of century, the nature of traps responsible for the localization of electrons and holes is still unknown [1, 3]. It was suggested [1] that the role of traps for electrons and holes in SiN_x can be played by silicon clusters. Recently, Park *et al.* [17] observed photoluminescence (PL) from quantum dots in plasma deposited SiN_x. Therefore, we may suggest that amorphous silicon quantum dots can exist in pyrolytic SiN_x as well and can be detected by PL measurements.

This paper reports on the results of investigations into the short-range order, electron structure, and PL in amorphous SiN_x synthesized by pyrolysis. Based on the structural data, we propose a model assuming large-scale potential fluctuations caused by inhomogeneities of the local chemical composition of SiN_x. The PL observed in SiN_x ($x \approx 4/3$) is interpreted in terms of the optical transitions between quantum states of amorphous silicon clusters.

2. SAMPLE PREPARATION AND EXPERIMENTAL TECHNIQUES

The experiments were performed with SiN_x samples synthesized in a low-pressure reactor by chemical vapor deposition (CVD) at 760°C from a SiH_2Cl_2–NH_3

gas mixture. The samples of SiN_x with various compositions were obtained by changing the ratio of SiH_2Cl_2 and NH_3 in the gas phase. The SiN_x layers were deposited onto *p*-Si(100) substrates with a resistivity of $\rho \approx 10\ \Omega$ cm. The luminescence was studied in the samples of SiN_x with $x \approx 4/3$ synthesized by decomposition of a SiH_4–NH_3–H_2 mixture at 890°C.

High-resolution X-ray photoelectron spectroscopy (XPS) measurements were performed in a Kratos AXIS-HS system using a source of monochromated AlK_α X-ray radiation with $\hbar\nu = 1486.6$ eV. The natural oxide film from the samples was removed by 2-min etching in an HF–methanol mixture (1 : 30), followed by rinsing in methanol. Prior to XPS measurements, the samples were washed in cyclohexane and blown with flow of dry nitrogen. The binding energies were measured relative to the 1*s* peak of carbon in cyclohexane with a binding energy of 285.0 eV. In cases of significant positive charging of a sample, the charge was compensated using a beam of low-energy electrons. All the XPS measurements (except for the angle-resolved ones) were performed for the sample surface oriented perpendicularly to the electron energy analyzer axis (zero polar angle).

In order to check that the XPS spectra reflect the bulk properties of SiN_x, we performed angle-resolved measurements on a Phi Quantum 2000 spectrometer. Figure 1 shows the XPS spectra of Si 2*p* levels in $SiN_{0.51}$ measured for the photoelectron take-off angles of 20°, 45°, and 90°. Weak angular dependence of the signal shape shows evidence that the bulk of a sample is probed, so that the XPS data obtained reflect the bulk properties of SiN_x.

In studying SiN_x layers with different compositions (i.e., with variable *x*), we used a nearly stoichiometric silicon nitride (Si_3N_4) as a reference sample for determining the relative sensitivity factors with respect to Si 2*p* and N 1*s* photoelectron lines. The reference sample synthesized at 800°C from a $SiCl_4$–NH_3 mixture with a 1 : 10 ratio of components had a refractive index of 1.96 and exhibited a characteristic IR absorption band at 3300 cm^{-1} due to the stretching vibrations of Si_2N–H bonds. The calculated concentration of N–H bonds in this material was 2.1×10^{21} cm^{-3}. Thus, it was established that the reference sample had a composition of $SiN_{1.41}H_{0.05}$.

The Raman spectra were measured on a Renishaw Ramanscope spectrometer using He–Ne laser radiation ($\lambda = 6328$ Å). The IR absorption spectra were recorded on a Nicolet 550 spectrometer with a resolution of 4 cm^{-1}.

The PL measurements for SiN_x ($x \approx 4/3$) samples were performed at room temperature. The emission spectra were normalized with respect to the sensitivity of the detection system. The sample film thicknesses determined using a laser ellipsometer was about 800 Å. The PL excitation spectra were measured using a deuterium lamp of the DDS-400 type.

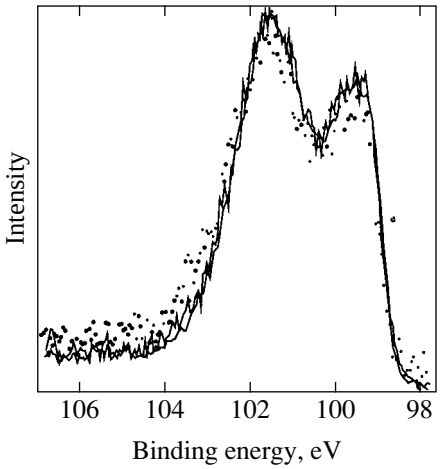


Fig. 1. The XPS spectra of Si 2*p* levels in $SiN_{0.51}$ measured for a photoelectron take-off angle of 20° (points) and 45° and 90° (solid curves).

3. SHORT-RANGE ORDER IN SiN_x BY XPS DATA

Figure 2 shows the experimental XPS spectra (depicted by symbols) of Si 2*p* levels in SiN_x samples of various compositions. All these curves exhibit either two peaks or one peak with a shoulder and are analogous to the spectra reported previously [7, 9, 13]. Applicability of the RB and RM models to description of the structure of SiN_x with variable composition was checked by comparing the experimentally measured Si 2*p* spectra to the results of calculations based on these models.

Dashed curves in Fig. 2 show the XPS spectra calculated in terms of the RB model. This model assumes that the structure consists of $SiN_\nu Si_{4-\nu}$ tetrahedra of five types corresponding to $\nu = 0$–4. The probability of finding the tetrahedron with a given ν obeys a binomial distribution [6, 8]

$$W(\nu, x) = \left(\frac{3x}{4}\right)^\nu \left(1 - \frac{3x}{4}\right)^{4-\nu} \frac{4!}{\nu!(4-\nu)!}. \qquad (1)$$

The idea of using five types of tetrahedra for modeling the XPS spectrum of Si 2*p* levels is based on the assumption that only nearest-neighbor silicon and/or nitrogen atoms contribute to the chemical shift of the Si 2*p* electron state. Equation (1) also assumes the absence of point defects such as =N–N= bonds, dangling bonds (≡Si• and =N•), and hydrogen bonds (Si–H, N–H) in SiN_x (here, symbols "–" and "•" denote a covalent bond and an unpaired electron, respectively).

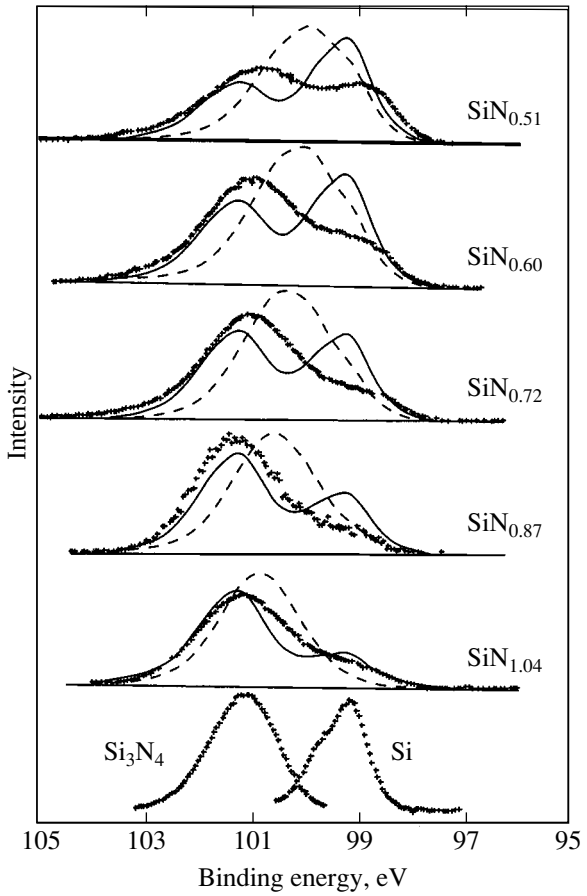


Fig. 2. The XPS spectra of Si 2*p* levels in SiN_x samples of various compositions: symbols represent the experimental spectra; the results of theoretical calculations are depicted by solid (RM model) and dashed (RB model) curves.

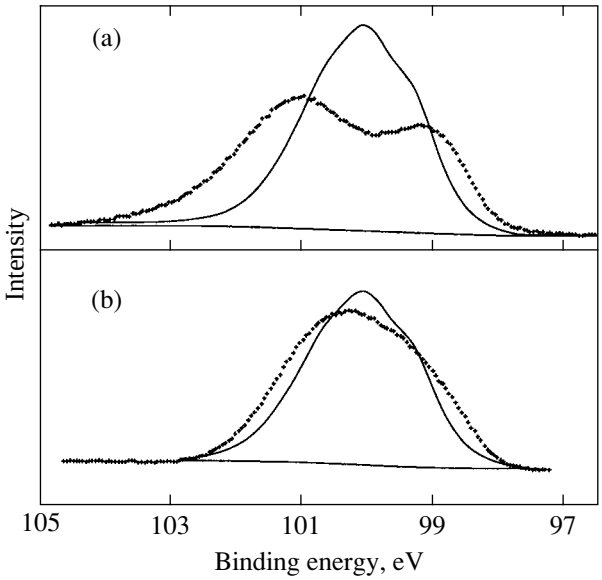


Fig. 3. The XPS spectra (points) of Si 2*p* levels in $SiN_{0.51}$ measured (a) before and (b) after irradiation with 4-keV Ar^+ ions. Solid curves show the results of calculations using the RB model.

The results of electron paramagnetic resonance (EPR) measurements showed that the concentration of dangling bonds ≡Si• and =N• in our samples did not exceed 10^{19} cm^{-3}.

Theoretical convolutions of the Si 2*p* spectra were obtained using the XPSPEAK 4.1 program package [14]. The theoretical spectra were calculated assuming the peaks corresponding to different SiN_vSi_{4-v} tetrahedra to be equidistant (equally spaced on the energy scale). The peak halfwidth (defined as the full width at half maximum, FWHM) was taken either the same for all peaks or linearly increasing as determined by extrapolation between the peak halfwidth for Si and Si_3N_4. The results of calculations according to the RB model showed that the XPS spectrum of the Si 2*p* level in SiN_x must contain a single maximum (Fig. 2, dashed curves) with the peak position being shifted toward higher binding energies with increasing nitrogen content in SiN_x.

For the sake of generality, we also checked for the applicability of the RB model to description of the structure of a disordered SiN_x sample. The disorder was produced by irradiating a $SiN_{0.51}$ sample with a beam of 4-keV Ar^+ ions (Fig. 3). As can be seen from Fig. 3a, the structure of the initial $SiN_{0.51}$ sample is not adequately described by the RB model: the experimental XPS spectrum exhibits two peaks corresponding (in the first approximation) to the $SiSi_4$ and SiN_4 tetrahedra, whereas the theoretical model predicts a single peak with a maximum positioned at a Si 2*p* binding energy of the SiN_2Si_2 tetrahedron (Fig. 3b). After irradiation, nitrogen and silicon atoms are mixed and the sample exhibits a tendency to form a substitution solid solution (RB model). The XPS spectrum of the ion-bombarded $SiN_{0.51}$ sample displays a single peak with a binding energy of the maximum close to that calculated within the RB model (Fig. 3b).

The results of simulation of the XPS spectra of SiN_x within the framework of the RM model are depicted by solid curves in Fig. 2. The spectra calculated using this model show, in agreement with experiment, a tendency to decrease in the fraction of a silicon phase in SiN_x with decreasing silicon content. However, the RM model somewhat overstates the silicon phase fraction and, in addition, predicts a dip approximately in the middle of the spectrum presented in Fig. 2. In experiments, however, such a dip is observed only for $SiN_{0.51}$ and is absent in the XPS spectra of other samples.

Thus, neither RB nor RM models can quantitatively describe the structure of an SiN_x compound. For this reason, we propose an intermediate model illustrated in Fig. 4. This model suggests the presence of separate silicon ($SiSi_4$ tetrahedra) and Si_3N_4 (SiN_4 tetrahedra) phases, as well as subnitrides (composed of SiN_3Si, SiN_2Si_2, and $SiNSi_3$ tetrahedra) in the SiN_x structure.

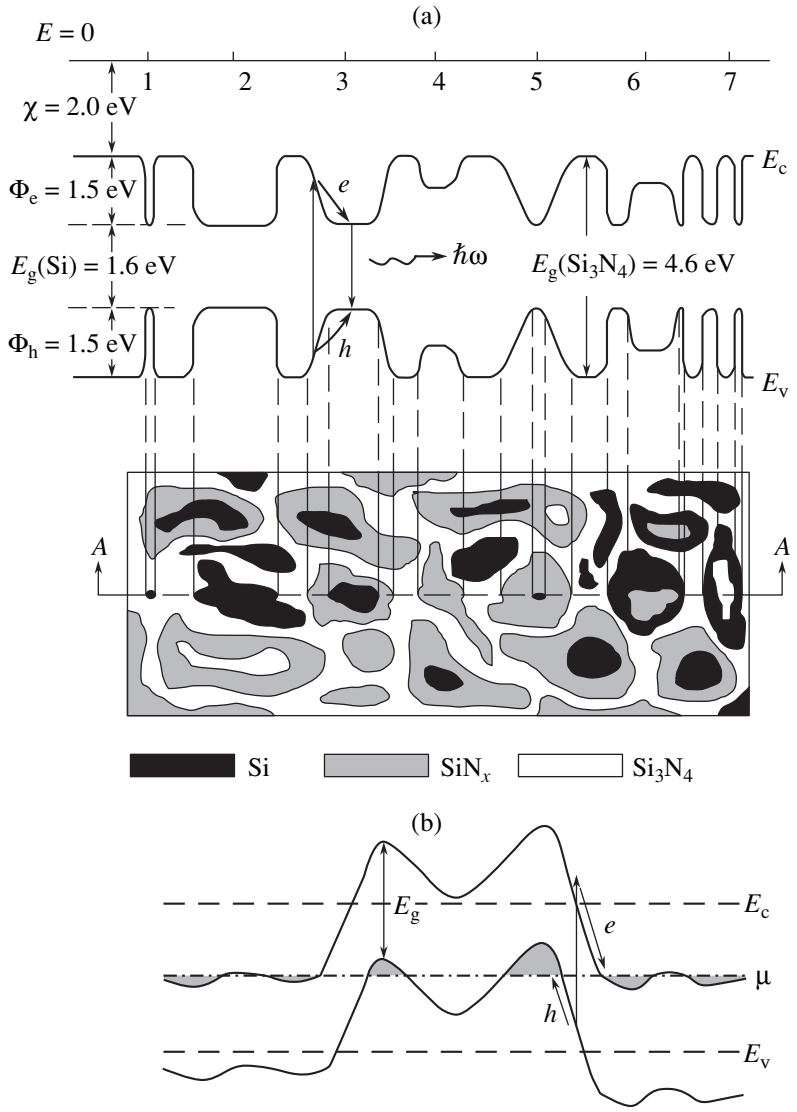


Fig. 4. Schematic diagrams illustrating the proposed intermediate model of SiN_x: (a) a two-dimensional diagram of SiN_x structure showing (bottom) the regions of a silicon phase, stoichiometric silicon nitride, and subnitrides and (top) the energy band profile of SiN_x in the *A*–*A* section (E_c is the conduction band bottom; E_v is the top of the valence band; Φ_e and Φ_h are the energy barriers for electrons and holes at the a-Si–Si_3N_4 interfaces, respectively; E_g is the bandgap width; and χ is the electron affinity); (b) fluctuations of the Shklovskii–Efros potential in a strongly doped compensated semiconductor (μ is the Fermi level).

The proposed intermediate model assumes local spatial fluctuations of the chemical composition of amorphous silicon nitride. This model will be considered in more detail in Section 6.

4. EXCESS SILICON IN SiN_x ACCORDING TO RAMAN SCATTERING DATA

The XPS measurements do not provide information on the spatial distribution of silicon in the SiN_x structure. The chemical shift of the Si 2*p* level is sensitive to the local chemical environment, but not to the long-range order. Being nonpolar, Si–Si bonds do not contribute to the IR absorption spectra either. At the same time, the method of Raman scattering allows excess silicon in SiN_x to be detected.

Figure 5 shows the Raman spectra of SiN_x samples of various compositions grown on silicon substrates. Here, the intense peak at 520 cm^{-1} is due to the longitudinal optical phonons in the silicon substrate. For SiN_x samples with a low nitrogen content ($x \le 0.72$), there is an additional weak signal in the region of 460–480 cm^{-1}, that is, at a frequency coinciding with the position of the peak of Raman scattering in amorphous silicon [18]. Previously, the Raman scattering from silicon in SiN_x was studied in [15]. Thus, the Raman spectra provide

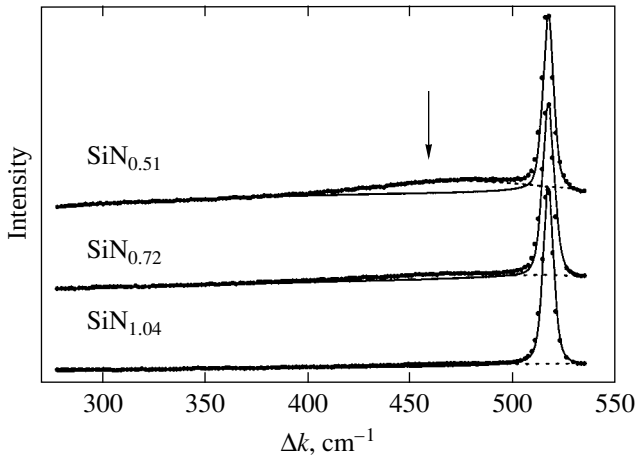


Fig. 5. The Raman spectra of $SiN_{1.04}$, $SiN_{0.72}$, and $SiN_{0.51}$ samples on silicon substrates. The peak at 520 cm^{-1} corresponds to scattering in silicon substrate; the arrow indicates the signal due to scattering from amorphous silicon clusters.

unambiguous evidence for the presence of amorphous silicon clusters in SiN_x.

The bandgap width in Si_3N_4 is $E_g = 4.6$ eV. However, experiments reveal the optical absorption in Si_3N_x at photon energies below this value [1]. This signal can be attributed to the absorption of light by silicon clusters in a Si_3N_4 matrix. The existence of silicon clusters in SiN_x is also confirmed by data on the fundamental absorption edge, according to which long-wavelength absorption takes place at 1–2 eV [16]. Silicon clusters with dimensions of 12–24 Å in hydrogenated silicon nitride (SiN_x:H) were observed in a high-resolution electron microscope [17, 19]. The Raman spectra of SiN_x with $x \geq 1.04$ exhibit no signal related to the scattering from silicon clusters. However, this result does not exclude the existence of such clusters: the lack of the signal can be explained by insufficient sensitivity of this method, related to small cluster size, low cluster density, and small thicknesses of sample films (about 1000 Å).

5. DETERMINING ENERGY BARRIERS FOR HOLES AT THE Si–Si_3N_4 INTERFACE FROM XPS DATA

For a comparative study of the valence bands of Si_3N_4, SiN_x, and a-Si, we have measured the corresponding XPS spectra. The samples of a-Si were prepared by irradiating crystalline silicon with 4-keV Ar^+ ions. The valence band of Si_3N_4 consists of two subbands separated by an ion gap (Fig. 6a). The narrow lower subband is formed by N 2*s* orbitals with an admixture of Si 3*s* and 3*p* orbitals [3]. The broad upper subband is formed by the nonbonding $2p_\pi$ orbitals of nitrogen and the bonding 3*s*, 3*p*, and 3*d* orbitals of silicon and 2*p* orbitals of nitrogen [20]. In addition, Fig. 6b presents the high-resolution XPS spectra of Si for the valence bands of amorphous silicon nitrides Si_3N_4, $SiN_{1.04}$, and Au measured relative to the top of the valence band of gold.

As can be seen, features of the valence band spectrum of $SiN_{1.04}$ are generally analogous to those for Si_3N_4, except for the region at the top of the valence band (Fig. 6b). The high-resolution XPS spectra show that the top of the valence band of a-Si coincides with that of gold. The electron work function of gold is 5.1 eV. Therefore, the top of the valence band of a-Si is spaced 5.1 eV from the electron energy level in vacuum. The top of the valence band of Si_3N_4 is spaced 1.5 eV from the top of the valence band of a-Si (Fig. 6b). Therefore, the energy barrier for holes at the a-Si–Si_3N_4 interface also amounts to 1.5 eV.

6. LARGE-SCALE POTENTIAL FLUCTUATIONS IN SiN_x CAUSED BY SPATIAL INHOMOGENEITIES IN THE CHEMICAL COMPOSITION

According to the XPS data, SiN_x comprises a mixture of Si_3N_4, silicon subnitrides, and amorphous silicon. The silicon nitride phase is composed of SiN_4 tetrahedra; subnitrides are composed of SiN_3Si, SiN_2Si_2, and $SiNSi_3$ tetrahedra; and the amorphous silicon clusters are composed of $SiSi_4$ tetrahedra. The bandgap width of compound Si_3N_4 is 4.6 eV, while that of amorphous silicon is 1.6 eV [19, 21] and that of silicon subnitrides varies within 1.6–4.6 eV. Therefore, the bandgap width in compound SiN_x also varies from 1.6 to 4.6 eV. According to the data presented in Section 5, the maximum scale of potential fluctuations for holes is 1.5 eV. Since the bandgap width of a-Si is 1.6 eV, the energy barrier for electrons at the a-Si–Si_3N_4 interface amounts to 1.5 eV. Thus, the maximum scale of potential fluctuations for electrons in SiN_x is also 1.5 eV.

Figure 4a presents the proposed model of large-scale potential fluctuations caused by variations in the local chemical composition of SiN_x, as illustrated by a two-dimensional diagram showing all possible variants of the local (spatial) structure of silicon nitride. The energy band diagram refers to the *A*–*A* section; the straight line indicates the level from which the electron energies are measured (vacuum level). A decrease in the bandgap width E_g is evidence of the presence of subnitrides in the silicon nitride matrix. The minimum bandgap width ($E_g = 1.6$ eV) corresponds to the silicon phase. This model assumes smooth variation of the chemical composition at the boundaries between silicon clusters and the Si_3N_4 matrix. Our experimental data do not allow the size of this transition region to be estimated. We reckon that this size may be on the order of several dozens of ångströms.

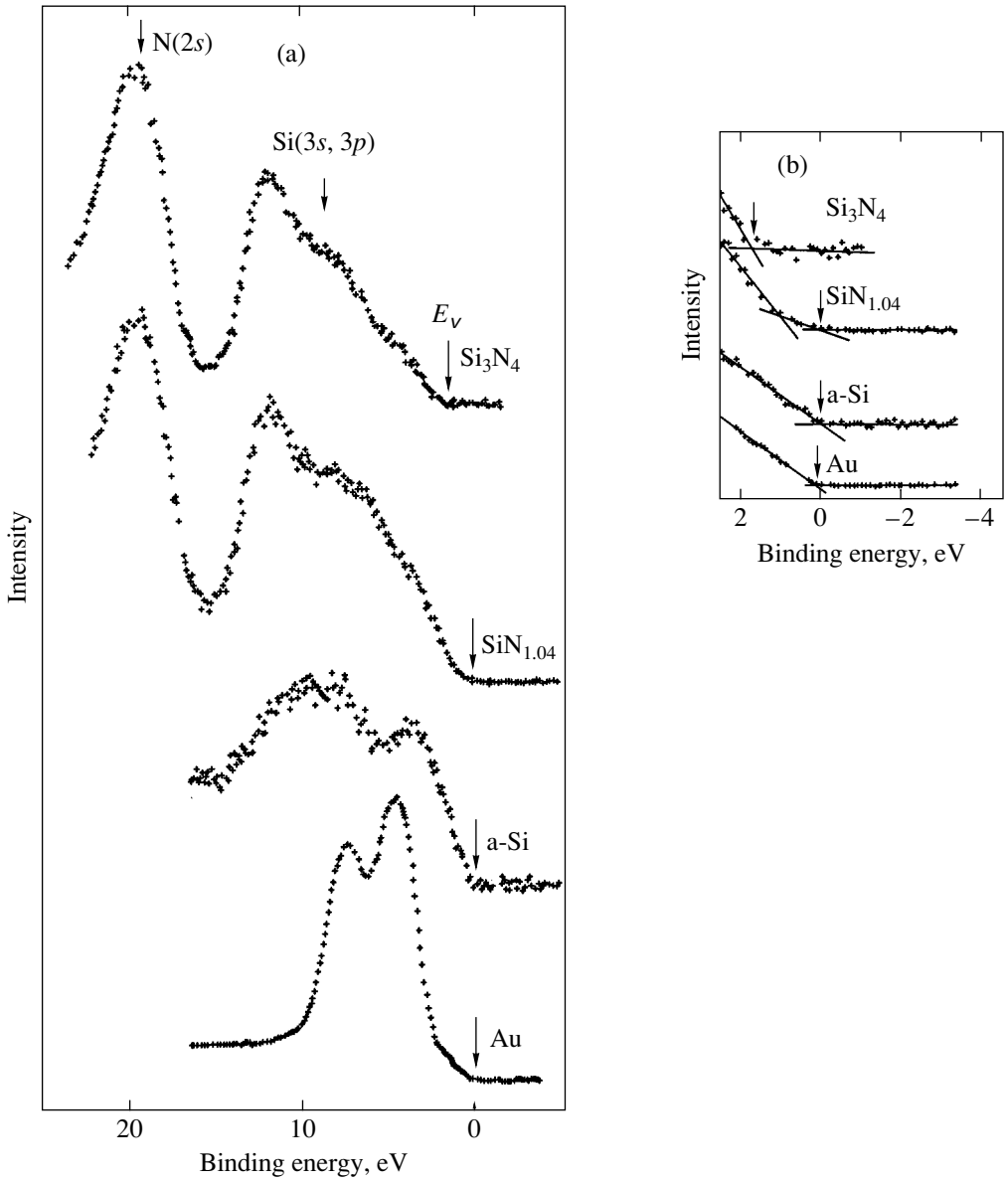


Fig. 6. (a) XPS spectra of the valence band of Si_3N_4, $SiN_{1.04}$, a-Si, and Au samples (arrows indicate the energy position of the top of the valence band); (b) the same spectra in the top of the valence band region recorded at a higher resolution.

Region 1 in Fig. 4a corresponds to a "quantum" cluster (with dimensions L on the order of the de Broglie wavelength of quasi-free electrons in a silicon cluster) incorporated into the Si_3N_4 matrix. The ground state energy in this cluster is $E = \hbar^2/2mL^2$, where m is the effective electron mass. Region 2 represents large silicon clusters surrounded by Si_3N_4. In this case, there is no transition layer of silicon subnitrides and the energy band diagram reveals a sharp Si–Si_3N_4 interface. Large clusters do not feature quantization of the energy levels of electrons and holes. Region 3 is a macroscopic silicon cluster surrounded by a silicon subnitride phase. In this situation, the transition from silicon to Si_3N_4 in the energy band diagram is smooth. Note that, here and below, we assume that the size of the transition region occupied by silicon subnitrides is significantly greater than the length of Si–N and Si–Si bonds (amounting to 1.72 and 2.35 Å, respectively). Region 4 corresponds to a silicon subnitride cluster in the silicon nitride matrix. Region 5 is a "quantum" silicon cluster incorporated into the subnitride phase, and regions 6 and 7 represent subnitride and nitride clusters, respectively, surrounded by silicon.

Thus, fluctuations of the local chemical composition of SiN_x lead to large-scale spatial fluctuations of the potential for electrons and holes. Previously, similar models of large-scale potential fluctuations have been developed for Si:H [22], SiC:H [23], SiC_xO_z:H [24],

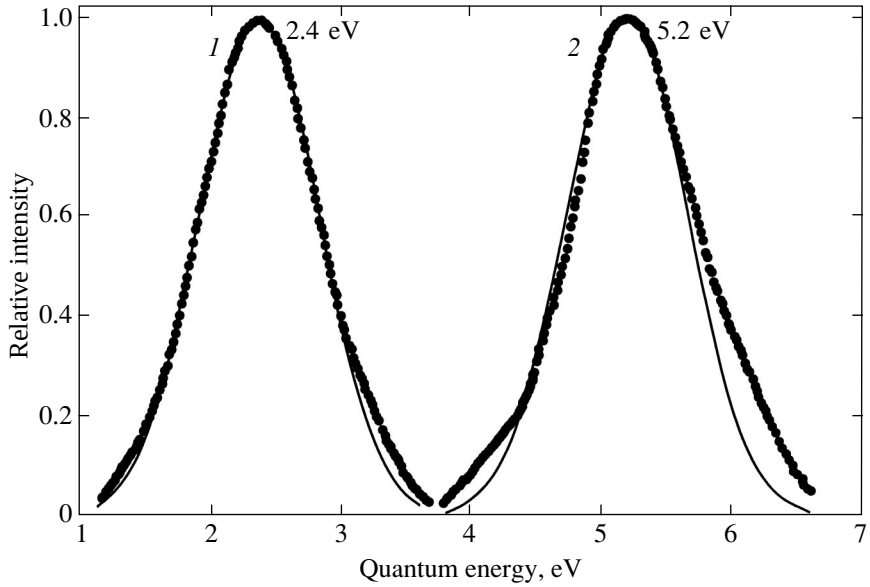


Fig. 7. The spectra of (*1*) PL and (*2*) PL excitation in SiN_x ($x \approx 4.3$) at room temperature. Points represent the experimental data, solid curves show the results of approximation by the Gauss function.

and SiO_x [25]. When an electron–hole pair is generated in silicon subnitride, the electric field is directed in the same direction for both electron and hole, thus favoring their recombination (Fig. 4a). In the case of a radiative recombination mechanism, SiN_x is an effective radiative medium. Figure 4b illustrates the Shklovskii–Efros model of large-scale potential fluctuations in a strongly doped compensated semiconductor [26]. According to this model, the bandgap width is constant and the potential fluctuations are caused by the inhomogeneous spatial distribution of charged (ionized) donors and acceptors. Here, the electron–hole pair production is accompanied by spatial separation of electrons and holes, thus not favoring their recombination.

7. PHOTOLUMINESCENCE OF SILICON NITRIDE SiN_x

Figure 7 shows the room-temperature PL spectrum of SiN_x ($x \approx 4/3$) with nearly stoichiometric composition excited by quanta with an energy of 5.2 eV (curve *1*). The emission has a maximum intensity at an energy of $E_{lum} = 2.4$ eV. Approximated by a Gauss function, the PL peak has an FWHM of 0.92 eV. The spectrum of excitation of the PL line at 2.4 eV has a maximum at $E_{ex} = 5.2$ eV. The PL excitation spectrum approximated by a Gauss function has the same FWHM (0.92 eV) as that of the emission spectrum. Deviations of the shape of the emission and excitation spectra from the Gauss function are probably caused by experimental errors.

In recent years, the PL spectra of silicon nitride (SiN_x) and oxynitride (SiO_xN_y) of variable compositions have been observed in an energy interval of 2.2–2.8 eV [27–32]. The PL excitation peak at 5.2 eV was reported in [27, 30]. Park *et al.* [17, 19] studied the optical absorption and PL spectra of amorphous silicon clusters in the Si_3N_4 matrix, observed quantum confinement of electrons and holes at these quantum dots, and determined [17] the PL peak energy as a function of the average size of amorphous clusters. As the cluster size decreased from 2.9 to 13 Å, the PL peak shifted from 1.8 to 2.7 eV. According to [17, Fig. 1], the PL quantum energy of 2.4 eV observed in our experiments corresponds to a silicon cluster size of 17 Å. Figure 8 shows a configuration diagram of a defect for a strong electron–phonon coupling constructed using the available PL and PL excitation data. The Franck–Condon shift

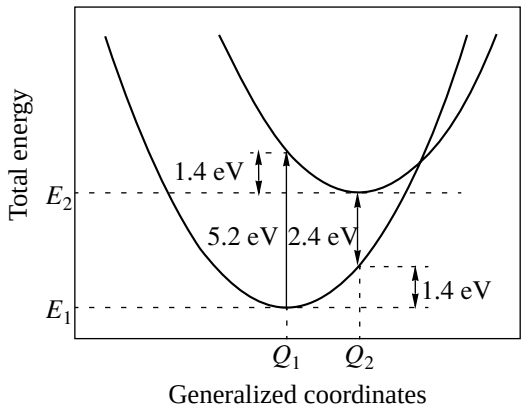


Fig. 8. A configuration diagram of a defect according to PL data: the lower and upper terms refer to the energies of the ground and excited state, respectively. The optical transitions at 5.2 and 2.4 eV correspond to excitation and emission; 1.4 eV (polaron energy) corresponds to the Franck–Condon shift.

(polaron energy) W_t estimated from this diagram equals to half of the Stokes shift: $W_t = (E_{ex} - E_{lum})/2 = 1.4$ eV.

The width Δ (FWHM) of the PL spectrum is related to the phonon energy W_{ph} in a single-mode approximation as

$$\Delta = 8W_t W_{ph} \ln 2.$$

The phonon energy determined from this relation amounts to $W_{ph} = 109$ eV and the corresponding Hung–Rice ratio is $W_t/W_{ph} = 12.7$. This W_{ph} value is significantly greater than the phonon energy (60 meV) reported for amorphous silicon [33]. The experimentally determined phonon energy coincides with the energy of Si–N bond oscillations in amorphous Si_3N_4 (900 cm^{-1} or 110 meV) [3]. The results can be explained by a large surface to volume ratio for the silicon clusters studied. Thus, the interaction in the excited electron–hole pair is related to local oscillations of the Si–N bonds at the boundary between a silicon cluster and the Si_3N_4 matrix. Previously, a strong interaction in the excited electron–hole pairs in silicon nanoclusters occurring in a SiO_2 matrix was studied by monitoring Si–O bond oscillations at low temperatures [34, 35]. Thus, according to the proposed interpretation, the emission at 2.4 eV is related to the optical transitions between quantum states of amorphous silicon clusters with an average size of about 1.7 Å.

8. DISCUSSION OF RESULTS

As was mentioned above, amorphous SiN_x can be synthesized only under thermodynamically nonequilibrium conditions. The structure and properties of this compound depend on the conditions of synthesis (the temperature and gas pressure) and subsequent high-temperature annealing [1]. We have studied the samples of SiN_x prepared at relatively high temperatures. The structure of this material is adequately described by the proposed intermediate model. Ion irradiation of the samples modifies the structure so that it approaches that described by the RB model. When the deposition temperature is decreased (plasma deposition), the structure of the synthesized SiN_x compound is described by the RM model [8].

Previously, it was demonstrated that the intermediate model describes the structure of an amorphous silicon oxide SiO_x [25], which is analogous to that of SiN_x and also depends on the conditions of synthesis. In particular, provided that the chemical composition is the same, the optical bandgap width in $SiO_{1.94}$ can be varied from 5.0 to 7.5 eV [36]. Note that the structure of silicon oxynitrides with variable composition SiO_xN_y is quantitatively described (in contrast to the structures of SiN_x and SiO_x) within the framework of the RB model [37, 38].

Basic differences between the proposed model of large-scale potential fluctuations in SiN_x and the Shklovskii–Éfros model for compensated semiconductors are as follows.

(i) Large-scale potential fluctuations in compensated semiconductors are of electrostatic nature, being related to the spatial fluctuations in the density of charged donors and acceptors, while the bandgap width is constant (Fig. 4b). The electric field caused by spatial fluctuations of the potential favors separation of electrons and holes. In SiN_x, the potential fluctuations are caused by inhomogeneities of the local chemical composition. In the proposed intermediate model (Fig. 4a), no space charge is formed (unlike the Shklovskii–Efros model) and the potential fluctuations favor the recombination of electrons and holes.

(ii) The low-frequency dielectric permittivities of Si_3N_4 and Si are 7.0 and 11.8, respectively. Therefore, SiN_x features spatial fluctuations of the permittivity.

(iii) According to the proposed model assuming potential fluctuations in SiN_x, this material is capable of localizing electrons and holes in potential wells, as experimentally observed in [1], with a giant time of localized carrier trapping (about 10 years at 300 K). The proposed intermediate model predicts the possibility of electron and hole percolation in the large-scale potential [25, 26].

The presence of silicon clusters (i.e., regions of significant excess of silicon) in SiN_x is confirmed by Raman scattering data. We believe that the excess silicon is not detected by Raman spectroscopy in nearly stoichiometric silicon nitride (SiN_x with $x \approx 4/3$) because of insufficient sensitivity of this technique. The existence of silicon clusters in SiN_x is confirmed by the following experimental data:

(i) The EPR spectrum of SiN_x with $x \approx 4/3$ displays a signal with $g = 2.0055$ belonging to a Si atom with an unpaired electron, bound to three other silicon atoms ($\equiv Si_3Si\bullet$) [39].

(ii) The low-energy electron loss spectrum of SiN_x with $x \approx 4/3$ exhibits peaks at the energies of 3.2 and 5.0 eV [40], which coincide with the energies of direct electron transitions in silicon [41].

(iii) The fundamental absorption edge in nearly stoichiometric silicon nitride SiN_x ($x \approx 4/3$) is about 4.6 eV [1, 42]. However, experiments on the photothermal absorption [16] showed the presence of absorption in the range from 1.7 to 3.9 eV. This result provides unambiguous evidence of the presence of excess silicon in SiN_x with $x \approx 4/3$.

(iv) The transport of electrons and holes in silicon nitride is conventionally interpreted within the framework of the Frenkel mechanism, according to which the Coulomb potential decreases in a strong electric field [43, 44]. However, our recent results showed that using this model at low temperatures leads to unreasonably

small values of the frequency factor ($\nu \approx 10^9$ s^{-1}) and an anomalously large tunneling mass of an electron ($m^* = 5.0m_e$) [45]. It was shown that the charge transfer in silicon nitride in a broad range of temperatures and fields can be quantitatively described using the theory of multiphonon ionization [45].

The thermal energy of trap ionization amounts to 1.4 eV [46], which coincides with the value of the Franck–Condon shift estimated in this study. The energy of local oscillations (60 meV) determined in [46] corresponds to the frequency of oscillations of silicon atoms in silicon clusters. These data provide independent evidence that amorphous silicon clusters act as traps for electrons and holes in silicon nitride.

Nevertheless, this study does not provide straightforward proof of the existence of amorphous silicon clusters with dimensions below 17 Å in nearly stoichiometric SiN_x with $x \approx 4/3$. Previously [47, 48], we formulated a hypothesis that the role of traps in silicon nitride can be played by silicon clusters of minimal size, namely, by Si–Si bonds. Indeed, Gee and Kastner [49] observed the PL at 4.4 eV with an excitation energy of 7.6 eV related to transitions on the Si–Si bonds. Our results presented above do not exclude that Si–Si bonds are the centers responsible for the luminescence at 2.4 eV and for the trapping of electrons and holes in silicon nitride. Thus, further investigations are necessary for judging between the models of amorphous silicon clusters and Si–Si bonds.

9. CONCLUSIONS

We have used high-resolution X-ray photoelectron spectroscopy and Raman scattering to study the short-range order in the layers of silicon nitride of variable composition SiN_x enriched with silicon. It has been established that neither random bonding (RB) nor random Si + Si_3N_4 mixture (RM) models can adequately describe the structure of this compound. An intermediate model has been proposed, according to which the SiN_x structure comprises five types of tetrahedral units, but the probability of finding a given tetrahedron type is not described by the RB model. It is suggested that fluctuations in the local chemical composition lead to large-scale potential fluctuations.

The PL spectra and the photoluminescence excitation spectra of SiN_x have been measured and interpreted in terms of a model assuming optical transitions between quantum states of amorphous silicon clusters. The Franck–Condon shift is 1.4 eV, which coincides with the thermal ionization energy of traps in SiN_x with $x = 4/3$. This result is evidence that either amorphous silicon clusters or Si–Si bonds (minimal silicon clusters) play the role of traps for electron and holes in nearly stoichiometric silicon nitride.

ACKNOWLEDGMENTS

The authors are grateful to V.V. Vasil'ev for kindly providing experimental data on the photoluminescence and for fruitful discussions.

This study was supported by the "Integration" Program of the Siberian Division of the Russian Academy of Sciences (project no. 116) and by the Program "Low-Dimensional Quantum Structures" of the Presidium of the Russian Academy of Sciences.

REFERENCES

1. V. A. Gritsenko, in *Silicon Nitride in Electronics*, Ed. by V. I. Belyi *et al.* (Elsevier, Amsterdam, 1988).
2. V. A. Gritsenko, K. A. Nasyrov, Yu. N. Novikov, *et al.*, Solid-State Electron. **47**, 1651 (2003).
3. V. A. Gritsenko, *Atomic and Electronic Structure of Amorphous Dielectrics in Silicon Based MIS Devices* (Nauka, Novosibirsk, 1993).
4. R. Karcher, L. Ley, and R. L. Johnson, Phys. Rev. B **30**, 1896 (1984).
5. W. R. Knolle and J. W. Osenbach, J. Appl. Phys. **58**, 1248 (1985).
6. V. P. Bolotin, I. A. Brytov, V. A. Gritsenko, *et al.*, Dokl. Akad. Nauk SSSR **310**, 114 (1990).
7. E. Bustarret, M. Bensouda, M. C. Habrard, *et al.*, Phys. Rev. B **38**, 8171 (1988).
8. S. Hasegawa, L. He, T. Inokuma, and Y. Kurata, Phys. Rev. B **46**, 12478 (1992).
9. L. Kubler, R. Haung, E. K. Hill, *et al.*, J. Vac. Sci. Technol. A **4**, 2323 (1986).
10. G. Wiech and A. Simunek, Phys. Rev. B **49**, 5398 (1994).
11. H. R. Philipp, J. Non-Cryst. Solids **8**–**10**, 627 (1972).
12. Z. Yin and F. W. Smith, Phys. Rev. B **42**, 3658 (1990).
13. G. M. Ingo, N. Zacchetti, D. Della Sala, and C. Coluzza, J. Vac. Sci. Technol. A **7**, 3048 (1989).
14. XPSPEAK 4.1, http://www.phy.cuhk.edu.hk/~surface.
15. V. A. Volodin, M. D. Efremov, and V. A. Gritsenko, Appl. Phys. Lett. **73**, 1212 (1998).
16. C. H. Seager and J. A. Knapp, Appl. Phys. Lett. **45**, 1060 (1984).
17. N.-M. Park, T.-S. Kim, and S.-J. Park, Appl. Phys. Lett. **78**, 2575 (2001).
18. F. Giorgis, Appl. Phys. Lett. **77**, 522 (2000).
19. N.-M. Park, C.-J. Choi, T.-Y. Seong, and S.-J. Park, Phys. Rev. Lett. **86**, 1355 (2001).
20. V. A. Gritsenko, Yu. N. Novikov, A. V. Shaposhnikov, and Yu. N. Morokov, Fiz. Tekh. Poluprovodn. (St. Petersburg) **35**, 1041 (2001) [Semiconductors **35**, 997 (2001)].
21. D. J. Lockwood, Z. H. Lu, and J.-M. Baribean, Phys. Rev. Lett. **76**, 539 (1996).
22. M. H. Brodsky, Solid State Commun. **36**, 55 (1980).
23. W.-J. Sah, H.-K. Tsai, and S.-C. Lee, Appl. Phys. Lett. **54**, 617 (1989).
24. R. Martins, G. Willeke, E. Fortunato, *et al.*, J. Non-Cryst. Solids **114**, 486 (1989).

25. V. A. Gritsenko, Yu. P. Kostikov, and N. A. Romanov, Pis'ma Zh. Éksp. Teor. Fiz. **34**, 1 (1981) [JETP Lett. **34**, 3 (1981)].
26. B. I. Shklovskiĭ and A. L. Éfros, *Electronic Properties of Doped Semiconductors* (Nauka, Moscow, 1979; Springer, New York, 1984).
27. V. V. Vasilev, I. P. Mikhailovskii, and K. K. Svitashev, Phys. Status Solidi A **95**, K37 (1986).
28. K. J. Price, L. E. McNeil, A. Suvkanov, *et al.*, J. Appl. Phys. **86**, 2628 (1999).
29. K. S. Seol, T. Futami, T. Watanabe, *et al.*, J. Appl. Phys. **85**, 6746 (1999).
30. K. S. Seol, Phys. Rev. B **62**, 1532 (2000).
31. T. Noma, K. S. Seol, M. Fujimaki, *et al.*, J. Appl. Phys. **39**, 6587 (2000).
32. H. Kato, N. Kashio, Y. Ohki, *et al.*, J. Appl. Phys. **93**, 239 (2003).
33. F. Giorgis, C. Vinegoni, and L. Pavesi, Phys. Rev. B **61**, 4693 (2000).
34. Y. Kanemitsu, N. Shimitzu, T. Komoda, *et al.*, Phys. Rev. B **54**, 14329 (1996).
35. Y. Kanemitsu and S. Okamoto, Phys. Rev. B **58**, 9652 (1998).
36. T. W. Hickmott and J. E. Baglin, J. Appl. Phys. **50**, 317 (1979).
37. V. A. Gritsenko, J. B. Xu, R. W. M. Kwok, *et al.*, Phys. Rev. Lett. **81**, 1054 (1998).
38. V. A. Gritsenko, R. W. M. Kwok, H. Wong, and J. B. Xu, J. Non-Cryst. Solids **297**, 96 (2002).
39. Y. Kamigaki, S. Minami, and H. Kato, J. Appl. Phys. **68**, 2211 (1990).
40. R. Hezel and N. Lieske, J. Appl. Phys. **53**, 1671 (1982).
41. P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors* (Springer, Berlin, 1996; Fizmatlit, Moscow, 2001).
42. V. A. Gritsenko, E. E. Meerson, and Yu. N. Morokov, Phys. Rev. B **57**, R2081 (1998).
43. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (Wiley, New York, 1981; Mir, Moscow, 1984).
44. V. A. Gritsenko, E. E. Meerson, I. V. Travkov, and Yu. V. Goltvyanskiĭ, Mikroélektronika **16**, 42 (1987).
45. K. A. Nasyrov, V. A. Gritsenko, M. K. Kim, and H. S. Chae, IEEE Electron. Device Lett. **23**, 336 (2002).
46. K. A. Nasyrov, V. A. Gritsenko, and Yu. N. Novikov, Phys. Rev. Lett. (in press).
47. P. A. Pundur, J. G. Shavalgin, and V. A. Gritsenko, Phys. Status Solidi A **94**, K107 (1986).
48. V. A. Gritsenko, H. Wong, I. P. Petrenko, *et al.*, J. Appl. Phys. **86**, 3234 (1999).
49. C. M. Gee and M. Kastner, Phys. Rev. Lett. **42**, 1765 (1979).

Translated by P. Pozdeev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 770–779.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 879–890.

SOLIDS
Electronic Properties

Peculiarities of Electron Spectrum Rearrangement for the Double-Well Heterostructure GaAs/AlGaAs with a Variable Dimensionality of Electronic States in an External Electric Field

Yu. A. Aleshchenko[a,*], A. E. Zhukov[b], V. V. Kapaev[a], Yu. V. Kopaev[a], P. S. Kop'ev[b], and V. M. Ustinov[b]

[a]*Lebedev Physical Institute, Russian Academy of Sciences, Leninskiĭ pr. 53, Moscow, 119991 Russia*
[b]*Ioffe Physicotechnical Institute, Russian Academy of Sciences, Politekhnicheskaya ul. 26, St. Petersburg, 194021 Russia*
**e-mail: yuriale@mail1.lebedev.ru*

Received September 29, 2003

Abstract—Peculiarities of the electron spectrum rearrangement for the double-well heterostructure GaAs/AlGaAs with a variable dimensionality of electronic states in an external electric field are investigated theoretically and experimentally. The structure is an important part of the active element of a quantum-well unipolar semiconductor laser proposed by the authors earlier. The possibility of controlling the dimensionality of the lower laser subband in such an active element by an external electric field is demonstrated. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The possibility of obtaining stimulated radiation between subbands of semiconductor quantum wells (QWs) was predicted by Kazarinov and Suris more than 30 years ago [1]. The ideas expounded on in this publication formed the basis of the injection semiconductor quantum-well laser developed by specialists at Bell Laboratories in 1994 [2]. It was a unipolar laser based on intersubband transitions in the conduction band of tunnel-coupled QWs. Such a design of a quantum-cascade laser has a number of advantages over traditional diode lasers in which radiative recombination occurs between the electron and hole states. In contrast to diode lasers, the wavelength emitted by a unipolar laser is determined by the quantum constraint, i.e., by the thickness of the layers in the active region, rather than by the forbidden gap of the material. The advantages of unipolar lasers also include the high temperature stability and the possibility of operating at room temperature due to suppression of Auger relaxation processes. Both factors are associated with the same sign of the effective mass in working subbands (parallel subbands) of a unipolar laser. Contemporary lasers of this type can operate at room temperature in a wavelength range of 3.57–16 μm [3, 4].

Unfortunately, parallel working subbands used in a unipolar laser are responsible for a serious drawback hampering the attainment of a noticeable population inversion in such a system. Owing to the similarity between the initial and final electronic states in unipolar lasers, a single LO phonon with a nonzero momentum is sufficient for electron relaxation between the parallel subbands irrespective of their separation. At the same time, it is impossible to increase the lifetime of electrons by decreasing the overlap of the wave functions since this reduces the optical efficiency of the laser. For this reason, the electron lifetimes for intersubband transitions in the structures of unipolar lasers lie in the picosecond range. Under these conditions, to attain considerable gain, the structure of a unipolar laser must include up to 500 periods of the active element. As a result, the structure of a cascade laser remains extremely complicated despite considerable advances made in recent years [5, 6].

In an earlier publication [7], we proposed an original design of the active element for a unipolar semiconductor laser. The structure is based on the physical idea of suppression of intersubband nonradiative relaxation by using the quasimomentum dependence of the wave function of a QW with strongly asymmetric barrier heights [8, 9]. In such structures, a localized electronic state exists in a limited range $(0, k_c)$ of wave vectors k in the direction along the layers of the structure. For $k = k_c$, the 2D–3D transformation of the dimensionality of states takes place [10]. This effect can be used for a sharp increase in the nonradiative recombination time in the active element of a unipolar laser, in

which the lower laser subband corresponds to a QW with strongly asymmetric barriers. Such a subband exists only for small momenta of longitudinal motion, which makes it possible to eliminate one-phonon intersubband transitions. Our calculations [7, 11] show that optimized structures with a variable dimensionality of states, which are used as the active element of a unipolar laser, ensure a considerable increase in the nonradiative relaxation time between the laser subbands and a significant gain in the population inversion.

Here, peculiarities of the electron spectrum rearrangement for the double-well heterostructure GaAs/$Al_xGa_{1-x}As$ with a variable dimensionality of electronic states in an external electric field are studied theoretically and experimentally. Such investigations of structures with asymmetric barriers are of considerable interest in view of the expected anticrossing of energy levels in complex systems and the passage of the electronic state in a QW with asymmetric barriers to the continuum and also due to possible realization of the situation studied here in an injection unipolar laser. The first step in this direction was made in [10] for a single-well structure with asymmetric barriers.

2. EXPERIMENT

The structure, which is a fragment of the active element of a unipolar laser with a variable dimensionality of the electronic states, was grown by the method of molecular beam epitaxy on a semi-insulating GaAs substrate. The structure included the following layers (in the direction of growth): a 250-nm-thick buffer layer of undoped GaAs; a 50-nm-thick silicon-doped n^+-GaAs layer ($N_D = 10^{18}$ cm^{-3}); a 25-nm-thick silicon-doped n-$Al_{0.09}Ga_{0.91}As$ barrier layer ($N_D = 5.3 \times 10^{16}$ cm^{-3}); an undoped 45-nm-thick i-$Al_{0.09}Ga_{0.91}As$ barrier; a 2.8-nm-wide GaAs QW; a 4-nm thick i-$Al_{0.35}Ga_{0.65}As$ separating barrier; a 5.3-nm-wide GaAs QW; a 10-nm-thick undoped i-$Al_{0.35}Ga_{0.65}As$ barrier; a 30-nm-thick silicon-doped n^+-$Al_{0.35}Ga_{0.65}As$ barrier ($N_D = 6.5 \times 10^{17}$ cm^{-3}); and a 10-nm-thick silicon-doped n^+-GaAs upper cap layer ($N_D = 10^{18}$ cm^{-3}). A simplified band diagram of the structure in zero ($E = 0$) and a nonzero electric field is shown in Fig. 1. The structure was designed so that it ensured the resonance of subband $\mathscr{E}_1$ (lower working subband) in well 2 with asymmetric barriers (QW2) and subband $\mathscr{E}_2$ in symmetric QW1. In this case, the degree of localization of the electron wave functions in QW2 increases and the effect of the existence of a subband in the QW with asymmetric barriers only for small momenta of longitudinal motion can be used most fully. The structure under study in fact played the role of a single QW with asymmetric barriers in the proposed design of the active element of a unipolar laser, but with a higher degree of localization of the lower laser subband.

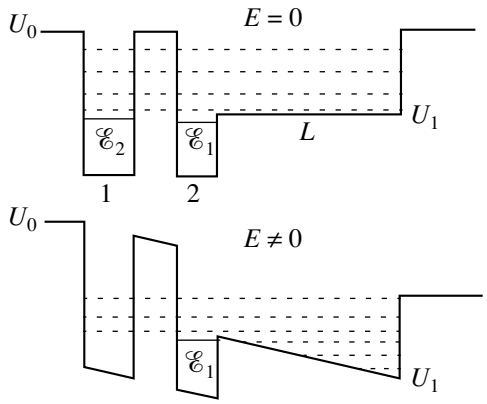


Fig. 1. Simplified band diagram of the structure in zero field ($E = 0$) and in an external electric field.

The structure was designed so that an external electric field could be applied to it. For this purpose, the upper and the buffer n^+-GaAs layers were used, in which control electrodes were formed. The doping level in the structure was selected in such a way that the fulfillment of the condition of flat bands in the active quantum region was ensured in the absence of an external bias voltage. This condition is essential when the structure is used in a unipolar laser with optical intersubband pumping (fountain laser). The upper electrode in the form of a semitransparent Ni film was deposited through a mask with a size of 3×7 mm^2. Subsequently, this electrode itself served as a mask for etching the structure around this electrode down to the n^+ lower buffer layer used as the lower electrode. Leads were soldered to the upper and lower electrodes using indium solder.

The optical properties of the structure were studied by the photoluminescence (PL) method at a temperature of 80 K and excitation by radiation emitted by an Ar^+ laser Stabilite 2017 (Spectra-Physics) with a wavelength of 5145 Å in the microprobe mode. The laser emissive power density at the sample was ~100 W/cm^2. Scattered radiation was analyzed by a monochromator Jobin Yvon T64000 and detected by a CCD matrix Spectrum One (Spex) cooled with liquid nitrogen. During measurements, a dc bias varying from +2 to –8 V was applied to the upper semitransparent electrode. The upper value of the positive bias was determined by the diode properties of the structure: for biases of +3 V and higher, a considerable current passed through the structure, which partly compensated the applied field. The minimum negative bias corresponded to the disappearance of exciton peaks in the PL spectra due to the

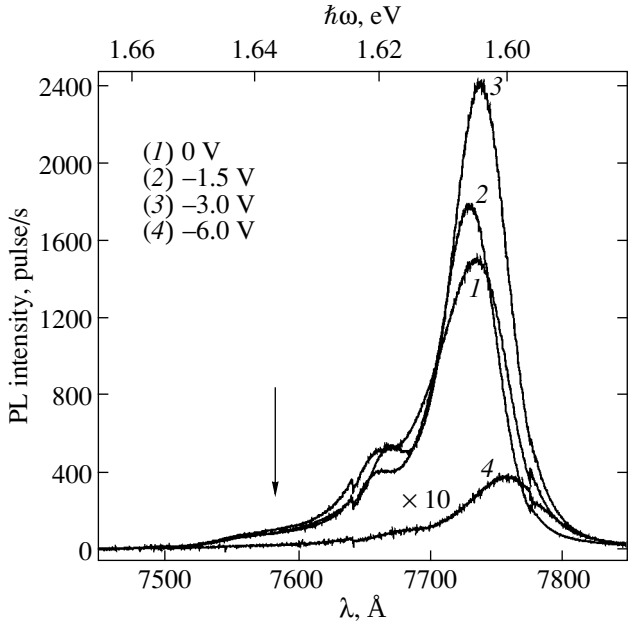


Fig. 2. PL spectra of the structure for various biases applied to the upper semitransparent electrode.

rupture of corresponding excitons by the applied electric field.

3. PHOTOLUMINESCENCE SPECTRA

Figure 2 shows the PL spectra of the structure for biases at the upper semitransparent electrode $V = 0$, –1.5, –3.0, and –6.0 V. The spectra are given on both the wavelength scale and on the energy (upper) scale. The steps on the curves in the vicinity of 7640 and 7775 Å are due to joining of different spectral regions encompassed by the multichannel matrix during one measurement. The PL spectra recorded at zero bias across the structure display four peaks at energies of 1.604, 1.619, 1.637, and 2.031 eV. The peak in the region of 1.604 eV has the highest intensity in the spectrum, while the peak at 2.031 eV (not shown in Fig. 2) has the lowest height. The intensity of the latter peak is two orders of magnitude lower than the intensity of the main peak corresponding to 1.604 eV. The peak at 2.031 eV is due to the contribution of the higher barrier to the PL spectrum. Its position corresponds to a composition of the solid solution $Al_xGa_{1-x}As$ with $x = 36\%$. Accordingly, the low-intensity peak in the energy range of 1.637 eV (marked by the arrow in Fig. 2) is due to the contribution of the lower barrier and corresponds to $x = 9.3\%$.

In accordance with our calculations, the peaks at energies of 1.604 and 1.619 eV are associated with the quantum region of the structure with layer parameters close to nominal values. It should be noted that the intensity of the peak associated with the lower electron subband in the QW with asymmetric barriers is considerably higher than the intensity of the PL spectrum of a single QW with asymmetric barriers [10] in view of a considerably stronger localization of the electron wave function in this subband in the case of its resonance with the subband in a QW with symmetric barriers. It can be seen from Fig. 2 that the position of the peaks in the energy range of 1.604 and 1.619 eV varies insignificantly upon a change in the bias across the structure from 0 to –6 V; however, the considerable change in the intensity and half-width of the peaks indicates that the structure is controlled by the external electric field.

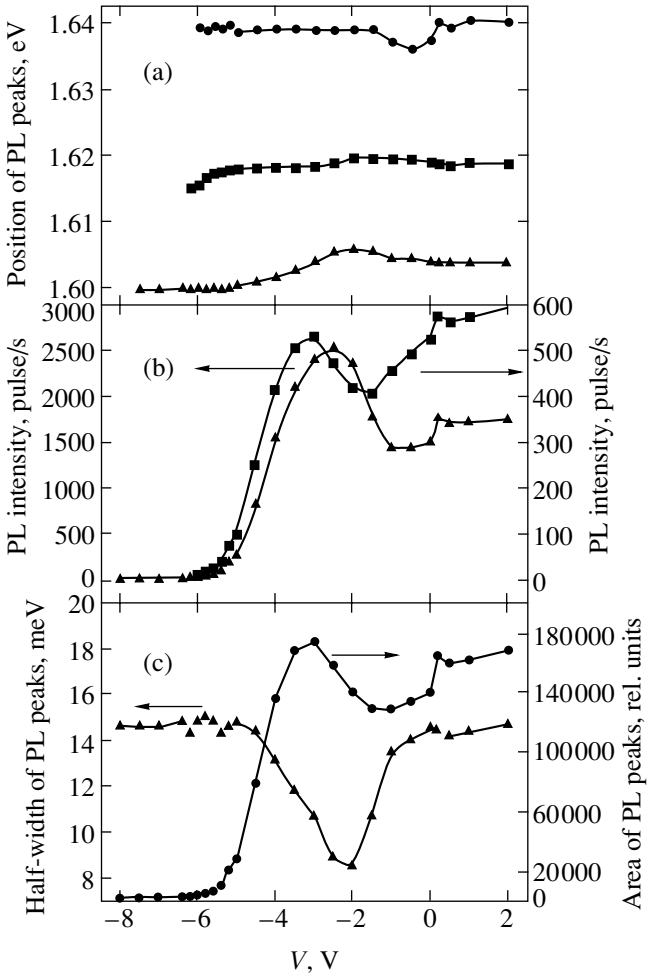


Fig. 3. Bias dependences of (a) the positions of peaks at energies 1.604, 1.637, and 1.619 eV, (b) their intensities, and (c) the half-width of the main PL peak and the total area of all peaks. The results for the main PL peak and for the high-energy peak with a lower intensity are shown by triangles and squares, respectively.

Figure 3 showing the bias dependences of the positions of the peaks at energies of 1.604 and 1.619 eV (a) and their intensity (b) as well as the half-width of the main PL peak and the total area of the peaks at energies of 1.604 and 1.619 eV (c) demonstrates these peculiarities most visually. Dark circles in Fig. 3a also show the

bias dependence of the position of the PL peak corresponding to the lower barrier. The behavior of the main PL peak (1.604 eV) can be traced down to $V = -8$ V. At lower biases, the peak cannot be detected since the corresponding exciton is torn by the strong field. The peak at an energy of 1.619 eV can be traced approximately to –6 V. At lower biases, it becomes very weak and merges with the main PL peak. Approximately at the same biases, the electric field tears the exciton of the lower barrier also. The weak dependence of the positions of the PL peaks on the applied bias in the quantum region (Fig. 3a) is worth noting. The main feature of the spectrum is the nonmonotonic behavior of all dependences depicted in Fig. 3 in the field. Indeed, as the value of V changes from 0 to –3 V, the intensity of the main PL peak in the energy range of 1.604 eV attains its maximal value (the peak height is almost doubled and its width is reduced almost by half). As bias V decreases further, the intensity of the main peak decreases and the peak becomes broader. The same tendency can also be traced for the PL peak with the maximum at 1.619 eV. In order to explain the observed peculiarities, we analyzed theoretically the spectrum of electronic states of the given structure and its modification in an external electric field.

4. THEORY

In studies of the PL spectra of the GaAs/$Al_xGa_{1-x}As$ structures with a single QW and asymmetric barriers in an external electric field [10], we demonstrated the possibility of exciton transitions involving the electron states lying above the lower barrier. Consequently, the continuous spectrum must be taken into account in theoretical analysis of relaxation processes in such structures. We developed an original method for calculating the electron spectrum of structures with a variable dimensionality of electronic states, which takes into account the contribution from the continuum, including the situation in an external electric field.

Strictly speaking, the bound states vanishes due to an indefinite decrease of the potential at infinity when an external electric field is applied to quantum wells. Nevertheless, to describe the behavior of such systems in external fields, the eigenvalue spectrum can be calculated in most cases by erecting an artificial boundary with a finite (or infinite) potential at a certain distance L to the right of the structure (see Fig. 1). As a result, instead of the actual continuous spectrum, we obtain a set of discrete states (the model of a quasi-continuous spectrum). The energies of the states and the probability of the electron location in the well proper weakly depend on L in a wide range of the values of L up to several thousands angstroms. As long as the time of tunneling through the triangular barrier remains longer than the characteristic times of the problem (e.g., the recombination time in analysis of PL spectra), such an approach ensures a quite admissible accuracy.

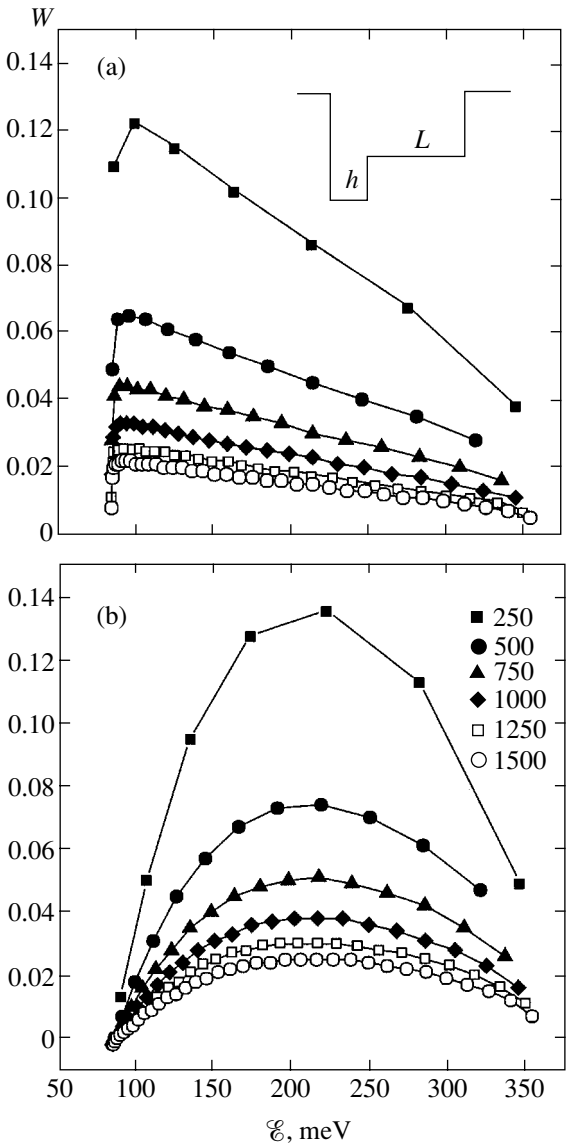


Fig. 4. (a) Dependence of probability W of finding an electron in the QW region of width $h = 24$ Å on energy $\mathscr{E}$ measured from the QW bottom; (b) the case when the potential in the QW region is equal to the height of the low barrier. The results are given for various values of L (in ångströms).

The application of this approach for QWs with asymmetric barriers under transformation of the dimensionality of states requires additional investigations. In this connection, we consider the behavior of a single GaAs QW with asymmetric $Al_xGa_{1-x}As$ barriers for a finite width L of the lower barrier in zero electric field (the level diagram of the structure is shown in the inset to Fig. 4a). In systems with asymmetric barriers and with different effective masses of carriers in the layers, the bound state is absent starting from a certain well width h_c. For $h > h_c$, this state exists for zero wave vec-

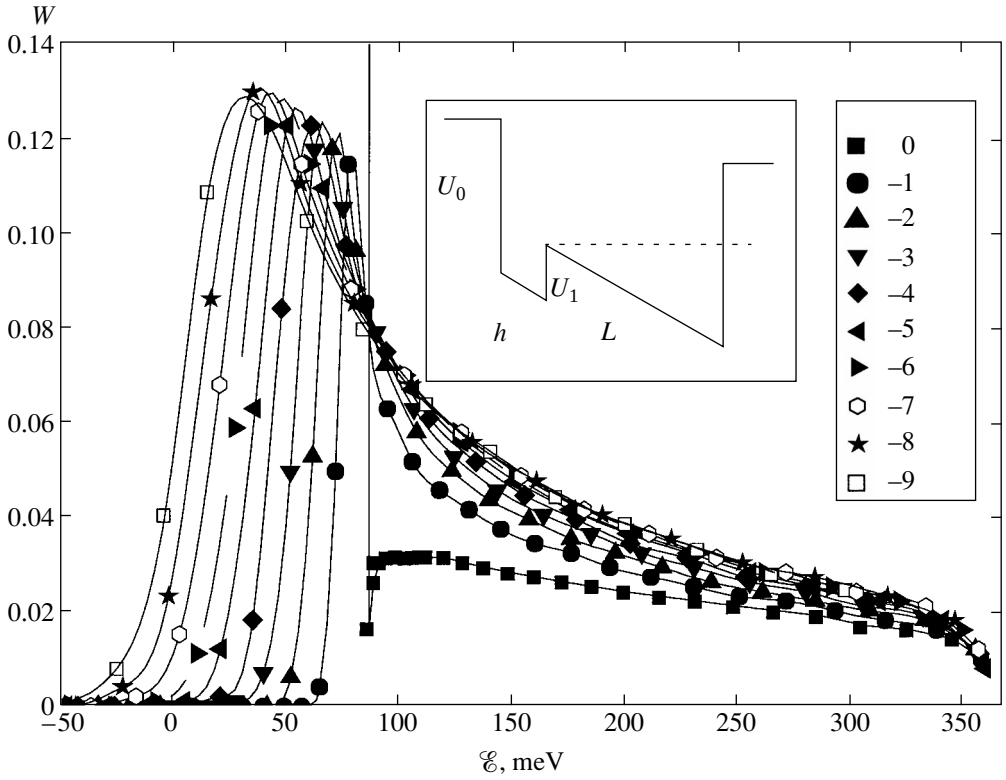


Fig. 5. Dependences of probability W of finding an electron in the QW on energy $\mathscr{E}$ for L = 1000 Å for various values of electric field E (in 10^4 V/cm).

tor k of motion along the layers of the structure and disappears for a certain finite wave vector $k = k_c$ since the "transformation of the dimensionality" of electronic states takes place. This conclusion follows from analysis of the solution of the problem with bounding barriers of an infinite width.

Figure 4a shows the dependence of probability W of finding an electron in the vicinity of QW of width 24 Å on the energy (measured from the bottom of the well), when fraction x of aluminum in the left (high) barrier is 0.36, while x = 0.093 in the right (low) barrier. The parameters of the barriers are taken close to the characteristics of the sample. The critical width for such a system is 28 Å; i.e., the bound state is absent in a well with h < 28 Å. For comparison, Fig. 4b shows the situation when the well is absent altogether (the potential in the region of the quantum well is equal to the height of the low barrier). The results of calculations are represented for various values of L (the corresponding notation is given in Fig. 4b). It can be seen from Fig. 4a that, due to the presence of a QW, probability W for energies close to the edge U_1 (about 82 meV) of the low barrier has a maximum; the dependence of the probability on L at the maximum becomes weak starting from L = 1000 Å. Since the effective mass of a hole exceeds the corresponding parameter for an electron, the hole state in the QW is strongly localized. As a result, the existence of a W peak for electrons at $\mathscr{E} \approx U_1$ may lead to recombination of this state with hole states localized in the well even in the absence of a bound electron sate. Thus, a PL signal can be observed for structures in which the bound state of electrons formally does not exist.

We have theoretically studied the behavior of states in a QW with asymmetric barrier heights in an external electric field. Proceeding from the results obtained for a QW in zero field, we can conclude that, in simulating the continuous spectrum, it is sufficient to extend the region of the low barrier to a distance of L = 1000 Å. A schematic diagram of the structure for which the calculations were made is shown in the inset to Fig. 5. The heights of the right and left barriers are chosen the same as in Fig. 4 and the QW width is h = 28 Å. For such a width, the bound state exists in zero external field.

Figure 5 shows the energy dependences of probability W of finding an electron in a QW for L = 1000 Å for various values of electric field E. For E = 0, a preferred state exists, for which W = 0.16 and the energy is lower than U_1 (bound state). For a finite value of the field, there exists a set of states with energies close to the low potential barrier height and with a high probability of electron location in the QW; i.e., in this case we can speak of the existence of quasi-bound states that can be manifested in PL spectra. The existence of relaxation in the system and tunneling in the presence of an electric

field leads to a certain ambiguity in determining the specific state from the set of quasi-bound states, which will be manifested in the PL spectrum.

Depending on the times of intersubband relaxation, tunneling, and radiative recombination, PL can be associated with the states at the peaks of $W(\mathscr{E})$ as well as with the states separated from the maximum by a certain distance. The positions of energy levels in the quasi-continuous spectrum slightly change with the value of L. To determine the characteristic energies, we approximate the discrete function $W(\mathscr{E}_n)$ by a continuous function (spline) and calculate the value of energy $\mathscr{E}_{\max}$ at the peak as well as the value $\mathscr{E}_{1/2}$ for which the probability is equal to one-half the probability $W(\mathscr{E}_{\max})$ at the maximum.

As usual in such problems, we investigate the convergence of the results upon an increase in model parameter L. Figure 6 shows the dependences of $\mathscr{E}_{\max}$ and $\mathscr{E}_{1/2}$ on electric field E for values of L equal to 750, 1000, and 1500 Å. A characteristic feature of these curves is the weak dependence on model parameter L; in other words, the use of the model of a quasi-continuous spectrum indeed makes it possible to describe the behavior of such systems in an electric field. The dashed line in Fig. 6 corresponds to $L = 0$, when the electric field exists only in the QW (the right barrier of an infinitely large width is shown by the dashed line in the inset to Fig. 5). In the field $E = E_c \approx -7.6 \times 10^4$ V/cm, the bound state in such a system disappears. In the model of a quasi-continuous spectrum, the state in the QW is localized quite strongly in fields significantly exceeding E_c as well. As a consequence, the PL line associated with recombination of such states is observed in fields considerably stronger than E_c. Another typical feature of the dependences shown in the figure is a comparatively large value of the difference $\mathscr{E}_{\max} - \mathscr{E}_{1/2}$ and its increase with electric field. This must lead to an increase in the PL linewidth with electric field. The noticeable difference between $\mathscr{E}_{\max}$ and $\mathscr{E}_{1/2}$ leads to ambiguity in the interpretation of experimental results on PL in an electric field.

Let us now describe the PL spectra for a system of tunnel-coupled QWs, one of which has barriers with asymmetric heights. The general methods of computation in this case are quite similar to the case of a single QW with asymmetric barriers. An additional difficulty in the interpretation of the PL spectra in this case emerges due to the presence of several electron and hole energy levels in the system, their formation and disappearance upon a change in the field and, finally, due to anticrossing of the levels belonging to different QWs. For convenience of interpretation, we will first calculate the transition energies, disregarding the field in the low barrier, and then solve the problem for a finite (sufficiently large) L. As in the case of a single QW, it is sufficient to take $L = 1000$ Å.

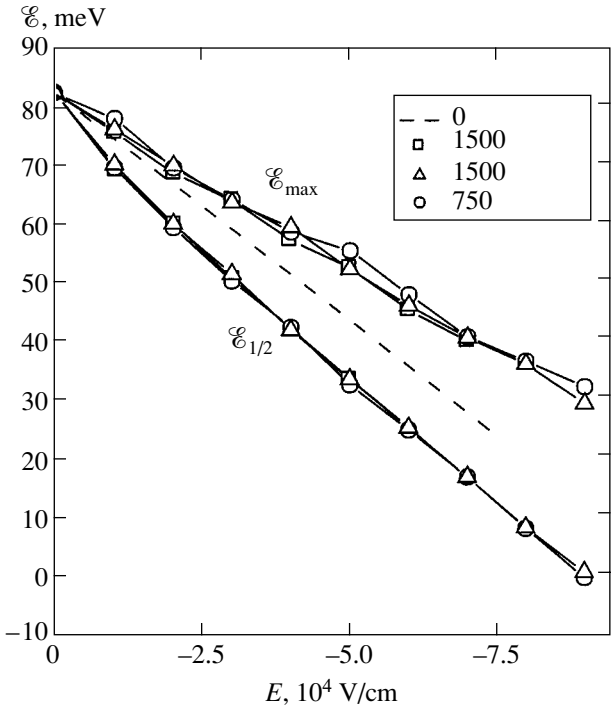


Fig. 6. Dependences of $\mathscr{E}_{\max}$ and $\mathscr{E}_{1/2}$ on electric field strength E for various values of L (in ångströms).

Figure 7 shows examples of the field dependences of the interband transition energy for double-well structures with an aluminum fraction of $x_1 = 0.36$ in the left barrier, $x_2 = 0.09$ in the right barrier, and with a separating barrier width of 40 Å. The widths of the wells are 50 and 30 Å in Fig. 7a and 45 and 35 Å in Fig. 7b. The parameters of the model structures are chosen in such a way that, in fields with a strength on the order of 3×10^4 V/cm, transitions occur with energies (including the exciton binding energy of about 10 meV) close to the experimentally observed energies in the given sample. For the first structure, the ground state for an electron is localized in the QW with symmetric barriers (50 Å), while for the second structure, it is localized in the well with asymmetric barriers (35 Å). The figure shows only the transitions for which the overlap integral for the electron and hole wave functions exceeds 0.01.

We will describe the types of transitions in the fields corresponding to characteristic points in Fig. 7. We introduce the following notation for the electron and hole states. We denote by *na* the state with level number *n*, which is localized predominantly in the well with symmetric barriers (QW1), and by *nb*, the state localized in QW2 (with asymmetric barriers). In Fig. 7a, for fields $E > E_1$, two electron subband and one hole subband are localized and transition $2b$–$1b$ takes place. This transition takes place from the second electron subband to the first hole subband; the corresponding wave functions are predominantly localized in QW2

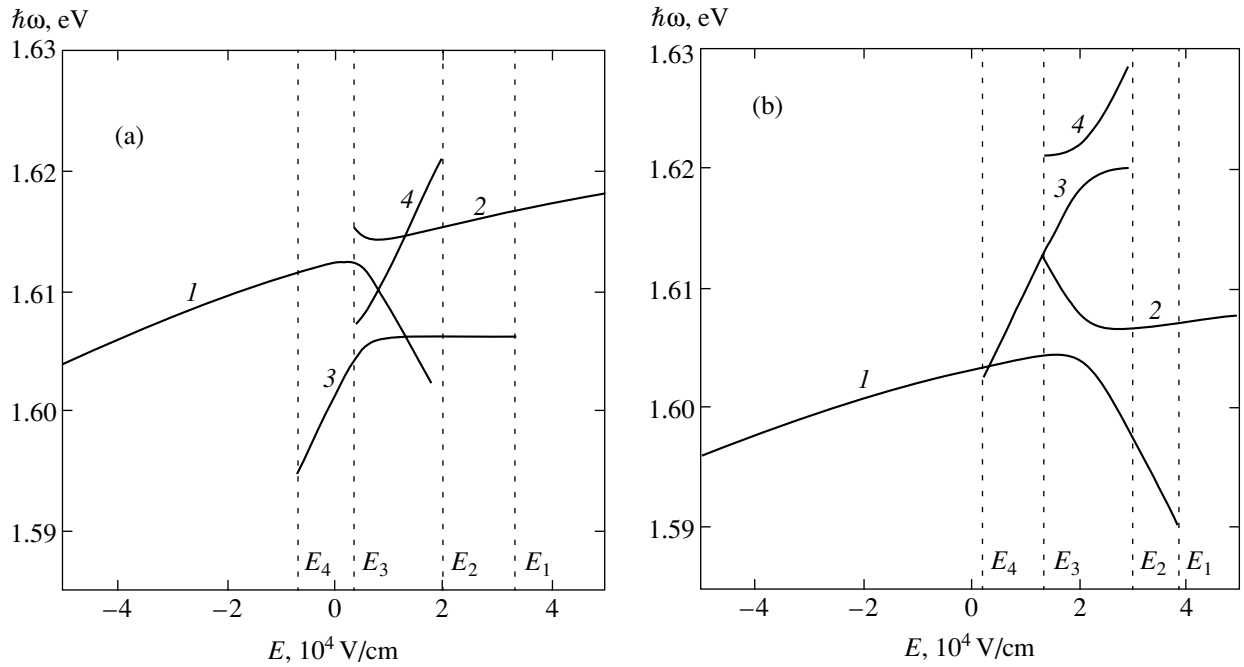


Fig. 7. Dependences of energy $\hbar\omega$ of interband transitions on electric field E for double-well structures with aluminum fraction $x_1 = 0.36$ in the left barrier and $x_2 = 0.09$ in the right barrier and with a separating barrier width of 40 Å. The widths of QWs are 50 and 30 Å (a) and 45 and 35 Å (b).

with asymmetric barriers. For the 1*a*–1*b* transition, the overlap integral is smaller than 0.01. In fields $E \sim E_1$, the second hole level localized in QW1 is formed; as a result, transition 1*a*–2*a* becomes possible. In the range $E_2 < E < E_3$, anticrossings of both electron (for a field strength of $E = E_e = 5.0 \times 10^3$ V/cm) and hole (for $E = E_h = 1.3 \times 10^4$ V/cm) subbands take place. In the anticrossing region, the probability of finding an electron (hole) in both wells becomes appreciable for each state; as a result, in addition to the above-mentioned transitions, transitions 1–1 and 2–2 begin to be manifested. In the vicinity of E_e, the upper electron subband reaches the continuum (transformation of dimensionality); as a result, only one localized electron subband remains for $E < E_3$. In this case, the PL spectrum acquires the main transition 1*b*–2*b* and the weaker 1*b*–1*b* transition. The intensity of the latter decreases smoothly with the field. Finally, for $E < E_4$, only one transition 1*b*–2*b* is left. In the field $E_5 = -5 \times 10^4$ V/cm, the first electron level reaches the continuum and the PL line disappears. On the whole, for the 50/40/30 structure, in the entire range of applied fields, we can single out curves *1* and *2* (outside the anticrossing region), which are associated with transitions in QW2, and curves *3* and *4*, for which transitions for states in QW1 play the major role.

For the 45/40/35 structure (Fig. 7b), the reverse (relative to the previous case) arrangement of curves *1*, *2* and *3*, *4* on the energy scale and reverse positions of anticrossings of electron and hole states (on the field scale) are observed. The anticrossings of electron and hole levels take place for $E_e = 2.1 \times 10^4$ V/cm and $E_h = 3.2 \times 10^3$ V/cm, respectively. For $E > E_1$, only one transition 2*b*–1*b* can be observed. For $E < E_1$, a weak nondiagonal transition 1*a*–1*b* is manifested. For $E < E_2$, the second hole subband is formed, and the electron subbands converge. As a result, the (high-intensity) 1*a*–2*a* and the (weak) 2*b*–2*a* transitions take place, and the intensity of the 1–1 transition increases. In the region of $E = E_3$, the second electron subband passes to the continuum (as in the previous case, this field is quite close to E_e) and we are left only with transitions 1*b*–1*b* (with high intensity) and 1*b*–2*a*. After attainment of anticrossing of the hole subbands ($E = E_h$), transition 1*b*–2*b* plays the major role. In field $E = E_4$, the second hole subband disappears and only one transition 1*b*–2*b* remains possible. Finally, in field $E_5 = -5 \times 10^4$ V/cm, transformation of the dimensionality for the electron subband takes place and PL disappears.

Figure 8 shows the field dependences of transition energies for the same structures, but taking into account the field in the region of the low barrier. The size of circles in Fig. 8 is proportional to the intensity of the corresponding transition. A comparison of Figs. 7 and 8 shows that the inclusion of the contribution of the continuum leads to an increase in the range of fields in which the states localized in the symmetric QW are manifested, to a decrease in the field range for transitions involving the states in the asymmetric QW, and, finally, to blurring of singularities in the regions of anti-

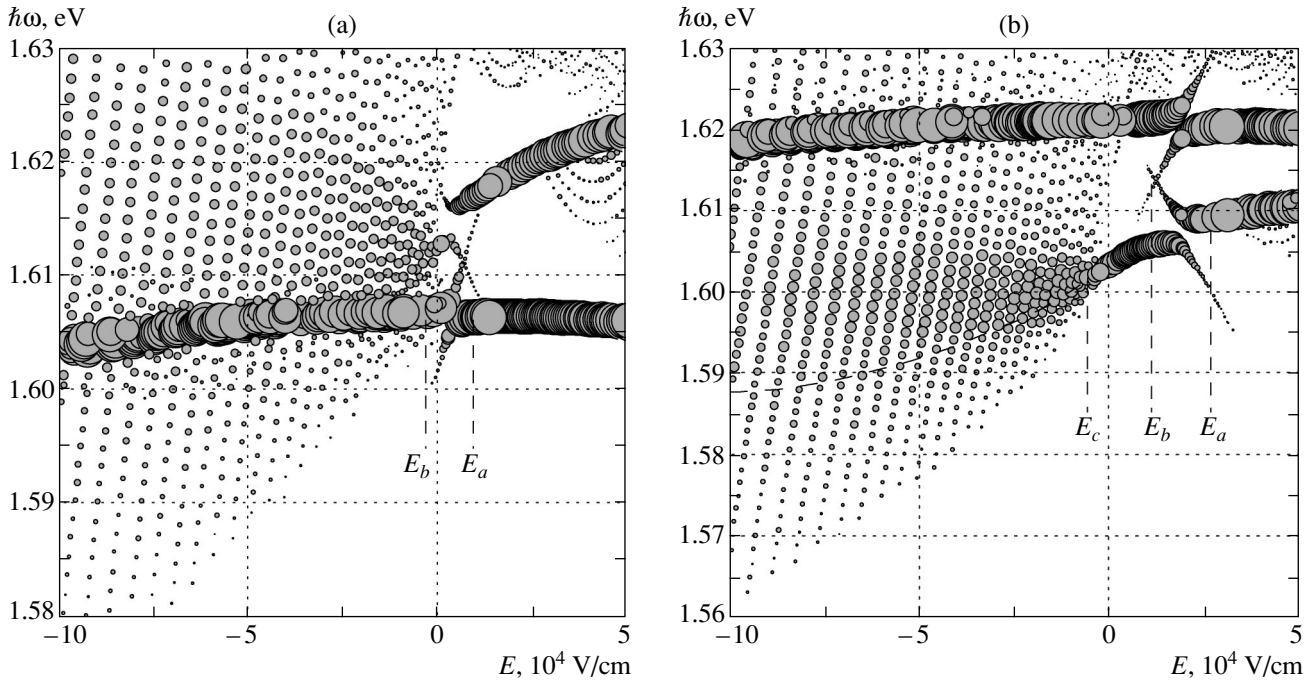


Fig. 8. Dependences of interband transition energies $\hbar\omega$ on electric field E for structures with the same geometrical parameters as in Fig. 7, but taking into account the field in the region of the low barrier. The size of the circles in the figure is proportional to the intensity of corresponding transitions.

crossing of electron and hole states. The low-energy curve exists in Fig. 8a in the entire field range under investigation, while the high-energy peak disappears for a field strength slightly exceeding E_b. For the structure corresponding to Fig. 8b, the two curves can coexist in the entire range of fields in view of what was said above about the possibility of the emergence of PL states from the continuum, which are localized in the asymmetric QW (dashed lines in the figure).

5. DISCUSSION

Taking into account the theoretical dependences of energies and intensities of transitions on the external electric field presented in Fig. 8, we return to the discussion of the experimental results. It follows from Fig. 3 that, in a wide range of biases applied to the structure, two peaks associated with the quantum region are present in the PL spectra at energies of 1.604 and 1.619 eV. For $V_2 = -2$ V, a low-intensity maximum is observed in the position of the low-energy PL peak as well as a minimum in the half-width of this peak. The intensity of the same peak and the total area of both peaks associated with the quantum region attain their minimum values for $V_1 = -1$ V, while the maximal values of these quantities are observed for $V_3 \approx -3$ V. Finally, the intensity minimum of the high-energy PL peak corresponds to lower biases as compared to V_1. The lack of exact information on the actual distribution of the doping impurity concentration and the complexity of the sample geometry did not allow us to carry out direct recalculation of the bias applied to the structure to the field in the region of quantum wells. Moreover, dependence $E(V)$ may be essentially nonlinear due to currents present in the structure. For this reason, a comparison with experiment can be carried out only on qualitative level.

A comparison of Figs. 3 and 8 shows that Fig. 8b is in better agreement with the experimental data. Indeed, the theory predicts for the 45/40/35 structure a considerable intensity of two transitions with energy values close to experiment both below and above the anticrossing region. If we assume that, for a positive bias across the sample, the field in the quantum region corresponds to E_a, the changes in the spectra in accordance with Fig. 8b must be as follows. When the external field varies in interval E_a–E_b, the system gets in the region of anticrossing of states. This field interval corresponds to the range of biases from +2 to –1 V in Fig. 3. Delocalization of electronic states in this region leads to splitting of each line into two lines, which is manifested in experiment as a slight broadening of PL peaks with a simultaneous decrease both in the line intensities and in the total area of the peaks since delocalization of the electron wave functions facilitates the departure of electrons from QW1 with symmetric barriers to the states of the continuum (i.e., the departure of electrons from the quantum region). For $E < E_b$, the system leaves the state of anticrossing of electron levels and the contribution from the continuum is still virtually

absent; this can be attributed to the strong localization of the lower electronic state in QW2 with asymmetric barriers and to the low intensity of electron tunneling from the state in QW1 in the absence of resonance between subbands due to the large width of the separating barrier. As a result, the PL peaks associated with the quantum region become narrower, while their intensity increases (bias range from –1 to –3 V). At the same time, the total area of the peaks in this bias range increases.

A peculiar feature of the structure considered here is that the bias at which the transformation of the dimensionality of the lower electronic state takes place is close to the bias corresponding to anticrossing of states. The effect of the 2D–3D transformation of the dimensionality of the subband belonging to QW2 plays a decisive role in the suppression of nonradiative relaxation to this subband. Due to the asymmetry in the barrier heights in the given structure, the localization of the wave functions of the subband corresponding to QW2 is determined by wave vector k (the state can be localized in the well for $k = 0$ or in the region of the low barrier for $k > k_c$). Let us define the "width" of the subband as $\Delta = \mathscr{E}(k_c) - \mathscr{E}(0)$. In Fig. 8b, the effect of the 2D–3D transformation of the dimensionality of the electron subband in QW2 takes place in region E_b–E_c. A decrease in Δ results in an additional decrease in the width (proportional to Δ) and in the area of the PL peak associated with the lower electronic state due to the suppression of relaxation from the second subband to the first. As a result of the competition with the effects associated with anticrossing of energy levels, the positions of the minima of the width of the PL peak, its intensity, and area for different lines do not coincide. The transformation of the dimensionality leads to a decrease in the exciton binding energy, which might be responsible for the formation of a low-intensity maximum in the position of the low-energy PL peak at –2 V in Fig. 3a. An additional argument in favor of this explanation is the coincidence of the positions of this maximum and the minimum of the linewidth on the bias scale.

A further increase in the field enhances the processes of tunneling and relaxation involving the states localized in the region of the low barrier; this first slows the variation of line widths and intensities and leads to the attainment of intensity maxima at biases $V \approx -3$ V. In the region of fields $E < E_c$, the contribution of the continuous spectrum becomes predominant, which leads to a decrease in the transition energy and to a sharp decrease in the PL peak intensities down to their complete disappearance as a result of the rupture of an exciton. The shift in the position of extrema in the intensity for the high-energy PL peak towards lower biases as compared to the low-energy peak (see Fig. 3b) can be explained by the facts that the first extremum is associated with transitions from the states of QW1 and the departure of electrons to the states of the continuum is hampered by a broad (40 Å) barrier.

It should be noted that the bias dependences of the energies of PL peaks in the quantum region (Fig. 3a) are more gently sloping than those predicted by the theory (Fig. 8b). Indeed, the range of transition energy variation with the field in Fig. 8b is 15 meV for the low-energy PL peak and about 5 meV for the high-energy peak, while this range in Fig. 3a does not exceed 5 meV for both peaks. This is due to the fact that the decrease in the transition energy with increasing field is partly compensated by the decrease in the binding energy of the corresponding exciton.

Thus, a comparison of experimental and theoretical results for a double-well structure enclosed between strongly asymmetric (in height) barriers shows that the observed singularities in the variation of the characteristics of PL spectra in an external electric field are associated with the transformation of the dimensionality of the electronic state in the QW with asymmetric barriers and with anticrossing of this state with a state in the QW with symmetric barriers.

6. CONCLUSIONS

Peculiarities of rearrangement of the electron spectrum for the double-well GaAs/AlGaAs heterostructure with variable dimensionality of electronic states in an external electric field are investigated theoretically and experimentally. The structure is an important component of the active element of the quantum-well unipolar semiconductor laser proposed by us earlier. An original method developed for calculating the electron spectrum of structures with variable dimensionality takes into account the contribution from the continuum, including the situation in an external electric field. It is shown that the dimensionality transformation effect in the lower subband associated with the QW with asymmetric barriers plays a decisive role in the variation of the PL spectra in an external electric field. The possibility of controlling the dimensionality of the lower laser subband in such an active element by an external electric field is demonstrated. This makes it possible to construct the active element of a quantum-well unipolar laser with record-high characteristics on the basis of QWs with asymmetric barriers.

ACKNOWLEDGMENTS

The authors thank S.S. Shmelev for preparing electrical contacts for the structure and to V.G. Plotnichenko and V.V. Koltashev for their assistance in optical measurements.

This study was partly financed in the framework of the program "Low-Dimensional Quantum Structures" of the Presidium of the Russian Academy of Sciences, the program "Physics of Solid-State Nanostructures" of the Ministry of Industry, Science, and Technology of

the Russian Federation, and the Federal Targeted Program “Integration” (grant no. B0049).

REFERENCES

1. R. F. Kazarinov and R. A. Suris, Fiz. Tekh. Poluprovodn. (Leningrad) **5**, 797 (1971) [Sov. Phys. Semicond. **5**, 707 (1971)].
2. J. Faist, F. Capasso, D. L. Sivco, *et al.*, Science **264**, 553 (1994).
3. F. Liu, Y. Zhang, Q. Zhang, *et al.*, Semicond. Sci. Technol. **15**, L44 (2000).
4. M. Rochat, D. Hofstetter, M. Beck, and J. Faist, Appl. Phys. Lett. **79**, 4271 (2001).
5. C. Gmachl, F. Capasso, D. L. Sivco, and A. Y. Cho, Rep. Prog. Phys. **64**, 1533 (2001).
6. F. Capasso, C. Gmachl, D. L. Sivco, and A. Y. Cho, Phys. Today **55**, 34 (2002).
7. Yu. A. Aleshchenko, V. V. Kapaev, Yu. V. Kopaev, and N. V. Kornyakov, Nanotechnology **11**, 206 (2000).
8. V. V. Kapaev and Yu. V. Kopaev, Pis’ma Zh. Éksp. Teor. Fiz. **65**, 188 (1997) [JETP Lett. **65**, 202 (1997)].
9. V. F. Elesin, V. V. Kapaev, Yu. V. Kopaev, and A. V. Tsukanov, Pis’ma Zh. Éksp. Teor. Fiz. **66**, 709 (1997) [JETP Lett. **66**, 742 (1997)].
10. Yu. A. Aleshchenko, I. P. Kazakov, V. V. Kapaev, and Yu. V. Kopaev, Pis’ma Zh. Éksp. Teor. Fiz. **67**, 207 (1998) [JETP Lett. **67**, 222 (1998)].
11. V. V. Kapaev, Yu. V. Kopaev, and N. V. Kornyakov, in *Proceedings of 9th International Symposium on Nanostructures*: *Physics and Technology* (St. Petersburg, Russia, 2001), p. 522.

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 780–792.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 891–905.

SOLIDS
Electronic Properties

Peculiarities of the Electron Mechanism of Superconductivity

R. O. Zaĭtsev

Russian Research Centre Kurchatov Institute, pl. Kurchatova 1, Moscow, 123182 Russia
e-mail: agydel@veernet.iol.ru

Received October 2, 2003

Abstract—A two-particle scattering amplitude calculated within the framework of the Hubbard model has been used for establishing the region of development of the Cooper instability. Dependence of the superconducting transition temperature T_c on the electron density has been determined. The temperature dependence of the time of relaxation with electron spin flipping has been determined taking into account spin fluctuations. Factors responsible for an increase in the $2\Delta_0/T_c$ ratio as compared to the classical value have been established.

1. INTRODUCTION

There are only two independent undetermined quantities in the theory of ideal superconductors: the Bardeen–Cooper–Schrieffer (BCS) constant and the preexponential factor $\bar{\epsilon}$. In the superconductivity theory based on the Hamiltonian of the electron–phonon interaction, the $\bar{\epsilon}$ value is on the order of the Debye frequency, whereas the BCS constant can be expressed via the slope of the temperature dependence of resistivity in the region above the Debye temperature. In these terms, the superconducting transition temperature is inversely proportional to the square root of the total mass of atoms in a unit cell and hence decreases with increasing mass of the introduced isotope. There is a large group of simple metals (Hg, Pb, Sn, Tl, Zn) possessing a correct sign and order of the isotope effect. However, the power of the isotope effect in complex compounds exhibiting a rather high transition temperature ($T_c > 20$ K) is always below 0.4 and keeps decreasing with increasing T_c. Therefore, we may suggest that there exists a nonphonon mechanism of the Cooper electron pairing in high-T_c superconductors.

The experiments with tunneling in high-T_c superconductors showed that the $2\Delta_0/T_c$ ratio in these materials is almost two times that according to the BCS theory. However, the most intriguing peculiarity is an extremely sharp dependence of the transition temperature on the dopant concentration. For example, the superconductivity in $La_{2-x}Sr_xCuO_4$ exists for $0.05 < x < 0.34$ and the maximum T_c corresponds to $x = 0.15$–0.16. Compounds of the $Nd_{2-x}Ce_xCuO_4$ system can be superconducting within a rather narrow interval of $0.14 < x < 0.18$, where the maximum T_c corresponds to $x = 0.15$. In the $YBa_2Cu_3O_{7-\delta}$ system, the maximum transition temperature is observed for $\delta = 0$, while the compound with $\delta = 1/2$ exhibits no superconductivity.

This paper is devoted to the so-called kinematic mechanism of superconductivity, which is capable of explaining the aforementioned peculiarities in the behavior of high-T_c superconductors.

The notions that superconductivity is possible in the Hubbard model with strong repulsion were criticized, first because the anomalous Gor'kov mean values [1] appearing below the point of development of the Cooper instability have to obey an additional sum rule, the existence of which implies infinite Hubbard energy. However, researchers formulating these objections (N.M. Plakida, Yu.A. Izyumov, V.Yu. Yushankhai, *et al.*) had no doubts about the existence of the kinematic interaction discovered by Dyson [2] and about the correctness of the equation derived by the author [3] for determining the superconducting transition temperature. The objection concerning the anomalous Gor'kov means was based on intuitive (so-called "physical") considerations and, until now, has not been confirmed by rigorous mathematical calculations.

A solution to this problem was suggested by Val'kov *et al.* [4]. It was demonstrated that "the inclusion of a singular contribution to the spectral intensity of the anomalous correlation function is shown to regain the sum rule and remove the unjustified forbidding of the *s*-symmetry order parameter in superconductors with strong correlations."

Another critical remark is related to possible ferromagnetic ordering in the region where the superconductivity appears. As will be shown below, a tendency toward ferromagnetism is manifested only in the region of small concentrations, where superconductivity is impossible. It will be demonstrated that the presence of paramagnetic fluctuations in the superconducting region leads to the appearance of a temperature-depen-

dent finite time of relaxation with electron spin flipping. In the region of ultimately low temperatures, the effect of paramagnetic fluctuations is insignificant, whereas a significant decrease in the effective BCS constant at a finite temperature leads to a decrease in the superconducting transition temperature T_c and to a corresponding increase in the $2\Delta_0/T_c$ value.

The third problem under dispute is the behavior of the concentration dependence of the superconducting transition temperature, $T_c(n)$, in the limit as $n \longrightarrow 1$. It will be shown below that the transition temperature in this limit tends to zero for any finite Hubbard energy. With allowance for a finite value V of the Coulomb interaction between adjacent cells, the superconductivity disappears within a finite interval of concentrations in the vicinity of $n = 1$, beginning with a certain critical value $V = V_c$.

Although the integral equations obtained admit solutions of different symmetry, we will consider solutions of the *s* type symmetry, which are free of nodes and lead to a phase diagram with a maximum temperature of the superconducting transition.

2. CALCULATION OF THE SCATTERING AMPLITUDE FOR INFINITE HUBBARD ENERGY

Our task here is to calculate the BCS constant directly as a function of the dopant concentration n_d. For definiteness, let us consider the lower Hubbard subband for $n_d < 1$ and, for simplicity, assume that the Hubbard energy is infinite ($U \longrightarrow \infty$). The spectrum of excitations is expressed via the product of the Fourier component of the hopping integral $t_\mathbf{p}$ and the so-called end factor f equal to the sum of the occupation numbers of the initial and final states,

$$\xi_\mathbf{p} = f t_\mathbf{p} - \mu, \quad f = n_0 + n_\sigma = 1 - n_{-\sigma}. \tag{1}$$

The equation of state in zero external field is as follows:

$$n_d = 2f\sum_\mathbf{p} n_F(\xi_\mathbf{p}), \quad f = 1 - \frac{n_d}{2}. \tag{2}$$

The one-particle Green function is conventionally defined via the excitation spectrum as

$$G_{\omega_n}(\mathbf{p}) = \frac{1}{i\omega - \xi_\mathbf{p}}, \quad \omega_n = \pi T(2n+1). \tag{3}$$

The Cooper instability develops when the two-particle vertex part acquires a singularity for zero total momentum, spin, and energy. The condition for the appearance of this singularity can be formulated as the condition of solvability of the corresponding homogeneous system [1]. For the model under consideration featuring only the interaction between electrons having opposite spins, we obtain the following ladder equation,

$$\begin{aligned}\Gamma_s(\mathbf{p}_1, \mathbf{p}_2|\mathbf{p}_3, \mathbf{p}_4) &= \Gamma_s^{(0)}(\mathbf{p}_1, \mathbf{p}_2|\mathbf{p}_3, \mathbf{p}_4)\\ &- T\sum_{\omega, \mathbf{p}} \Gamma_s^{(0)}(\mathbf{p}_1, \mathbf{p}_2|\mathbf{p}, s-\mathbf{p}) G_\omega(\mathbf{p})\\ &\times G_{\Omega-\omega}(\mathbf{s}-\mathbf{p})\Gamma_s(\mathbf{p}, \mathbf{s}-\mathbf{p}|\mathbf{p}_3, \mathbf{p}_4),\end{aligned} \tag{4}$$

where $\mathbf{s} = \mathbf{p}_1 + \mathbf{p}_2 = \mathbf{p}_3 + \mathbf{p}_4$, $\mathbf{p}$ are the operators of momenta, and $\Gamma_s^{(0)}(\mathbf{p}_1, \mathbf{p}_2|\mathbf{p}_3, \mathbf{p}_4)$ is the vertex part irreducible for cutoffs with respect to parallel momenta. This vertex part will be calculated using the method developed by Dyson [2] for calculating the spin wave scattering amplitude.

In the limiting case of infinite Hubbard energy, it is sufficient to use a Hamiltonian corresponding to the lower Hubbard subband,

$$\hat{H} = \sum_{\mathbf{r}, \mathbf{r}'(\mathbf{r}\neq\mathbf{r}'), \sigma} t(\mathbf{r}-\mathbf{r}')\hat{X}_\mathbf{r}^{\sigma,0}\hat{X}_{\mathbf{r}'}^{0,\sigma}, \tag{5}$$

where X operators obey the commutation relations

$$\begin{aligned}\{X_\mathbf{r}^{+,0}, X_{\mathbf{r}'}^{0,-}\} &= X_\mathbf{r}^{+,-}\delta_{\mathbf{r},\mathbf{r}'},\\ \{X_\mathbf{r}^{\sigma,0}, X_{\mathbf{r}'}^{0,\sigma}\} &= (X_\mathbf{r}^{0,0} + X^{\sigma,\sigma})\delta_{\mathbf{r},\mathbf{r}'},\end{aligned} \tag{6}$$

and the other anticommutators are zero.

Now let us determine the ground ($|0\rangle$), one-particle ($X_\mathbf{r}^{\sigma,0}|0\rangle$), and two-particle ($X_\mathbf{r}^{+,0}X_{\mathbf{r}'}^{-,0}|0\rangle$) states and calculate their energies. Denoting by E_0 the energy of the ground state, $\hat{H}|0\rangle = E_0|0\rangle$, we define the one-particle excitation energy (measured from the ground state energy level) via the commutator $[\hat{H}, X_\mathbf{r}^{\sigma,0}]$:

$$\begin{aligned}\hat{H}X_\mathbf{r}^{\sigma,0}|0\rangle - X_\mathbf{r}^{\sigma,0}\hat{H}|0\rangle &= (E_1 - E_0)X_\mathbf{r}^{\sigma,0}|0\rangle\\ &= \epsilon_\mathbf{p}^{(\sigma)}X_{\mathbf{r}'}^{\sigma,0}|0\rangle.\end{aligned} \tag{7}$$

By the same token, the two-particle excitation energy is defined via the commutator $[\hat{H}, X_\mathbf{r}^{+,0}X_{\mathbf{r}'}^{-,0}]$:

$$\begin{aligned}\hat{H}X_\mathbf{r}^{+,0}X_{\mathbf{r}'}^{-,0}|0\rangle - X_\mathbf{r}^{+,0}X_{\mathbf{r}'}^{-,0}\hat{H}|0\rangle\\ = (E_2 - E_0)X_\mathbf{r}^{+,0}X_{\mathbf{r}'}^{-,0}|0\rangle.\end{aligned} \tag{8}$$

The two-particle commutator $[\hat{H}, X_\mathbf{r}^{+,0}X_{\mathbf{r}'}^{-,0}]$ is calculated using the operator identity relation

$$\begin{aligned}[\hat{H}, X_\mathbf{r}^{+,0}X_{\mathbf{r}'}^{-,0}] &= X_\mathbf{r}^{+,0}[\hat{H}, X_{\mathbf{r}'}^{-,0}]\\ &- X_{\mathbf{r}'}^{-,0}[\hat{H}, X_\mathbf{r}^{+,0}] + \{[\hat{H}, X_\mathbf{r}^{+,0}]X_{\mathbf{r}'}^{-,0}\},\end{aligned} \tag{9}$$

where the first two terms can be transformed using

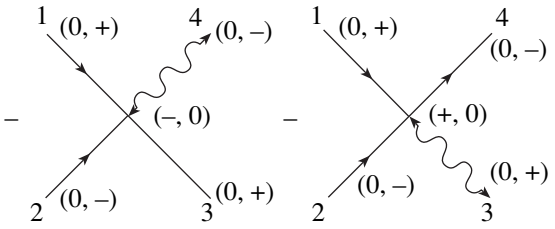


Fig. 1. The Born amplitudes of kinematic interaction (14).

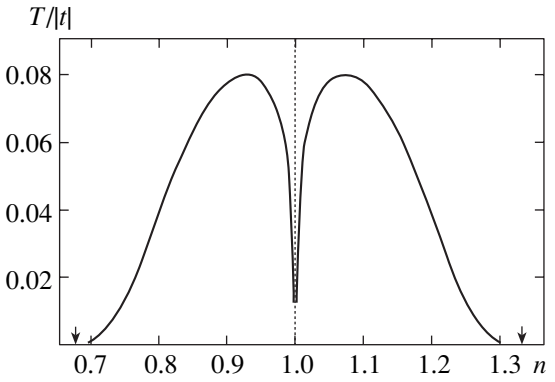


Fig. 2. The curve of transition temperature T_c versus electron density n calculated using Eq. (16). The arrows indicate critical electron densities.

definition (7) as

$$X_{\mathbf{r}}^{+,0}[\hat{H}, X_{\mathbf{r}'}^{-,0}] = \hat{\epsilon}_{\mathbf{p}}^{(-)} X_{\mathbf{r}}^{+,0} X_{\mathbf{r}'}^{-,0},$$
$$X_{\mathbf{r}'}^{-,0}[\hat{H}, X_{\mathbf{r}}^{+,0}] = \hat{\epsilon}_{\mathbf{p}}^{(+)} X_{\mathbf{r}'}^{-,0} X_{\mathbf{r}}^{+,0} \quad (10)$$

(here and above, $\mathbf{p}$ and $\mathbf{p}'$ are the operators of momenta acting upon $\mathbf{r}$ and $\mathbf{r}'$, respectively), and the third term is directly calculated as

$$\{[\hat{H}, X_{\mathbf{r}}^{+,0}]X_{\mathbf{r}'}^{-,0}\}$$
$$= -\sum_{\mathbf{r}_1} t(\mathbf{r}_1 - \mathbf{r})(X_{\mathbf{r}_1}^{+,0} X_{\mathbf{r}'}^{-,0} + X_{\mathbf{r}}^{+,0} X_{\mathbf{r}_1}^{-,0})\delta_{\mathbf{r},\mathbf{r}'}. \quad (11)$$

The action of the two-particle commutator on the ground state gives the following equation:

$$(E_2 - E_0)X_{\mathbf{r}}^{+,0} X_{\mathbf{r}'}^{-,0}|0\rangle = [\hat{\epsilon}_{\mathbf{p}}^{(+)} + \hat{\epsilon}_{\mathbf{p}}^{(-)}]X_{\mathbf{r}}^{+,0} X_{\mathbf{r}'}^{-,0}|0\rangle$$
$$+ \{[\hat{H}, X_{\mathbf{r}}^{+,0}]X_{\mathbf{r}'}^{-,0}\}|0\rangle. \quad (12)$$

An approximate expression for the one-particle energy operator $\hat{\epsilon}(\mathbf{p})$ is provided by Eq. (1). The effect of scattering is determined by the double commutator (11). Passing to the momentum representation in Eqs. (11) and (12),

$$X_{\mathbf{r}}^{+,0} X_{\mathbf{r}'}^{-,0}|0\rangle = \sum_{\mathbf{p},\mathbf{q}} \psi(\mathbf{p}, \mathbf{q}) \exp(i\mathbf{p}\cdot\mathbf{r} + i\mathbf{q}\cdot\mathbf{r}'),$$

we obtain an explicit expression for the two-particle interaction energy

$$(E_2 - E_0)\psi(\mathbf{p}, \mathbf{q}) = [\hat{\epsilon}_{\mathbf{p}}^{(+)} + \hat{\epsilon}_{\mathbf{p}}^{(-)}]\psi(\mathbf{p}, \mathbf{q})$$
$$- \sum_{\mathbf{k}} (t(-\mathbf{k}) + t(\mathbf{k} - \mathbf{p} - \mathbf{q}))\psi(\mathbf{k}, \mathbf{p} + \mathbf{q} - \mathbf{k}). \quad (13)$$

Thus, the scattering amplitude in the Born approximation depends only on the momenta of scattered particles [2, 3]:

$$\Gamma^{(0)}(\mathbf{p}_1, \mathbf{p}_2 | \mathbf{p}_3, \mathbf{p}_4) = -t(\mathbf{p}_3) - t(\mathbf{p}_4). \quad (14)$$

The Born amplitudes corresponding to this interaction are depicted in Fig. 1.

3. SUPERCONDUCTING TRANSITION TEMPERATURE

Substituting expression (14) into the homogeneous part of Eq. (4), we obtain an equation for determining the point of appearance of the Cooper instability

$$\Gamma_{\mathbf{s}} = T\sum_{\omega,\mathbf{p}} [t_{\mathbf{p}} + t_{\mathbf{s}-\mathbf{p}}] G_{\omega}(\mathbf{p}) G_{-\omega}(\mathbf{s} - \mathbf{p})\Gamma_{\mathbf{s}}. \quad (15)$$

Using Green functions (3) and summing over frequencies $\omega = \pi T(2n + 1)$ for $\mathbf{s} = 0$, we arrive at the equation for determining the temperature T_c of the superconducting transition,

$$\sum_{\mathbf{p}} \frac{t(\mathbf{p})}{\xi_{\mathbf{p}}} \tanh\left(\frac{\xi_{\mathbf{p}}}{2T_c}\right) = 1, \quad (16)$$

where $\xi_{\mathbf{p}}$ is the excitation energy defined by Eq. (1).

Equation (16) together with the equation of state (2) and the condition of electroneutrality determine the dependence of the superconducting transition temperature T_c on the dopant concentration. Figure 2 shows a plot of the transition temperature versus the electron density n, which was calculated using Eq. (16) (see [3–5]).

It should be noted that the integration in Eq. (16) is performed predominantly in the vicinity of the Fermi surface, where $\xi_{\mathbf{p}} = 0$ or $t_{\mathbf{p}} \approx \mu/f$. From this it follows that the Cooper instability is not developed for negative values of the chemical potential. This result appears as quite natural because small values of the chemical potential correspond to small occupation numbers. The scattering amplitude will have the same sign as that for

two particles possessing a small relative velocity and exhibiting infinitely strong repulsion at small distances.

As the energy of the relative motion increases, the scattering amplitude decreases and changes sign for a wavevector on the order of the inverse radius of the interaction potential. These properties are characteristic of the scattering amplitude (14) for zero total momentum $\mathbf{s} = \mathbf{p}_1 + \mathbf{p}_2 = \mathbf{p}_3 + \mathbf{p}_4 = 0$.

According to Eq. (16), the scattering amplitude calculated on the Fermi surface changes sign at zero chemical potential. Using the equation of state for $T = 0$, we obtain a critical value of $n_d = 2/3$. Beginning with this dopant concentration, the scattering amplitude is negative and the transition temperature is finite within the entire interval of $2/3 < n_d < 1$. The analysis of filling of the upper Hubbard subband reveals the symmetry with respect to the particle–hole transformation: $n_d \longrightarrow 2 - n_d$. For this reason, the Cooper pairing according to this simplest model also takes place in the region of $1 < n_d < 4/3$.

Thus, even the Born approximation stipulates the possibility of a change in the sign of the scattering amplitude over the entire Fermi surface. This provides explanation of a rapid change in the superconducting transition temperature and the existence of narrow regions of high-temperature superconductivity with respect to the dopant concentration.

As the electron density increases so that $n \longrightarrow 1$ and the lower Hubbard subband is completely filled, the system passes to a semiconductor state. The radius of screening of the Coulomb potential rapidly grows to become infinite for $n = 1$ and zero temperature. From this we infer that, in the region of $n \approx 1$, it is necessary to take into account the direct intercell Coulomb interaction,

$$\hat{V} = \frac{1}{2}\sum_{\mathbf{r},\,\mathbf{r}'(\mathbf{r}\neq\mathbf{r}')} \hat{n}_{\mathbf{r}}\hat{n}_{\mathbf{r}'}\varphi(|\mathbf{r}-\mathbf{r}'|),$$
$$\hat{n}_{\mathbf{r}} = \hat{X}^{+,+}_{\mathbf{r}} + \hat{X}^{-,-}_{\mathbf{r}} + 2\hat{X}^{2,2}_{\mathbf{r}}, \qquad (17)$$

where $\hat{n}_{\mathbf{r}}$ is the electron density operator expressed in terms of the Hubbard operators X. Using the commutation relations,

$$[\hat{X}^{0,\sigma}, \hat{n}] = \hat{X}^{0,\sigma}, \quad \hat{n} = \hat{X}^{+,+} + \hat{X}^{-,-} + 2\hat{X}^{2,2},$$

we obtain an expression for the Born scattering amplitude,

$$\Gamma^{(0)}(\mathbf{p}_1, \mathbf{p}_2|\mathbf{p}_3, \mathbf{p}_4) = -t(\mathbf{p}_3) - t(\mathbf{p}_4) + \varphi(\mathbf{p}_3 - \mathbf{p}_1), \quad (18)$$

which is a generalization of relation (14). Figure 3 shows the Born amplitudes of the Coulomb scattering, which represent only the first term corresponding to scattering in the lower Hubbard subband.

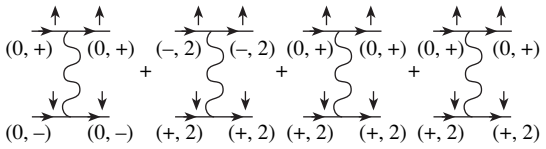


Fig. 3. The Born amplitudes of scattering on the Coulomb potential in terms of X operators.

In order to determine the superconducting transition temperature, let us consider the simplest model in which the influence of the direct Coulomb potential is bounded by the given potential of interaction V between nearest neighbors,

$$\varphi(\mathbf{q}) = V\sum_{\text{n.n.}} \exp(-i\mathbf{q}\cdot\mathbf{r}).$$

Using the Coulomb potential in this form, it is possible to solve the homogeneous equation for T_c by means of separating variables.

Assuming that the total momentum is zero, we obtain a solution $\psi(\mathbf{p})$ of the homogeneous equation depending only on the relative momentum $\mathbf{p} = \mathbf{p}_1 = -\mathbf{p}_2$ of colliding particles,

$$\psi(\mathbf{p}) = -T\sum_{\omega,\,\mathbf{q}} \Gamma^{(0)}(\mathbf{p}|\mathbf{q})G_{\omega}(\mathbf{q})G_{-\omega}(-\mathbf{q})\psi(\mathbf{q}), \quad (19)$$

where the scattering amplitude for a simple cubic or square lattice is

$$\Gamma^{(0)}(\mathbf{p}|\mathbf{q}) = -2t_{\mathbf{q}} + \sum_k 2V\cos(p_k - q_k)$$
$$= -2t_{\mathbf{q}} + \sum_k 2V\cos p_k\cos q_k \qquad (20)$$
$$+ \sum_k 2V\sin p_k\sin q_k.$$

Under the assumption that the unknown function $\psi(\mathbf{p})$ is even with respect to a change in the sign of momentum, integration of the sum of sines yields zero. By virtue of the cubic symmetry, integration of the sum of cosines yields equal values for all summands. Therefore, with respect to the integration over $\cos q_k$, the vertex part in Eq. (20) is equivalent to

$$\tilde{\Gamma}^{(0)}(\mathbf{p}|\mathbf{q}) = -2t_{\mathbf{q}} + \beta t_{\mathbf{p}}t_{\mathbf{q}}, \qquad (21)$$

where $\beta = V/Dt^2$. For a bcc lattice, we obtain an analogous expression for β without the factor $1/D$, that is, not divided by the number of measurements D.

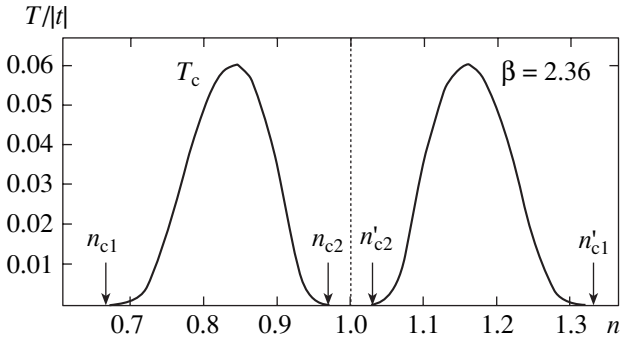


Fig. 4. The curve of transition temperature T_c versus electron density n calculated using Eq. (23). The arrows indicate critical electron densities $n_{c1} = 2/3$, $n_{c2} = 0.96$, $n'_{c1} = 4/3$, and $n'_{c2} = 1.04$.

Thus, we arrive at the following equation for determining T_c:

$$\psi(\mathbf{p}) = 2T\sum_{\omega,\,\mathbf{q}} t_{\mathbf{q}} G_{\omega}(\mathbf{q}) G_{-\omega}(-\mathbf{q})\psi(\mathbf{q}) - \beta t_{\mathbf{p}} T\sum_{\omega,\,\mathbf{p}} t_{\mathbf{q}} G_{\omega}(\mathbf{q}) G_{-\omega}(-\mathbf{q})\psi(\mathbf{q}). \quad (22)$$

This equation can be readily solved by separating variables as $\psi(\mathbf{p}) = A + Bt_{\mathbf{p}}$. As a result, we obtain the condition of solvability that generalizes relation (16),

$$\sum_{\mathbf{p}} \frac{t(\mathbf{p})}{\xi_{\mathbf{p}}} \tanh\left(\frac{\xi_{\mathbf{p}}}{2T_c}\right) = 1 + \frac{\beta}{2}\sum_{\mathbf{p}} \frac{t^2(\mathbf{p})}{\xi_{\mathbf{p}}} \tanh\left(\frac{\xi_{\mathbf{p}}}{2T_c}\right), \quad (23)$$

where $\xi_{\mathbf{p}} = ft_{\mathbf{p}} - \mu$. In the limit of $T_c = 0$, this integral is finite provided that either of the two conditions is satisfied,

$$\mu = 0 \quad \text{or} \quad ft_{\mathbf{p}} - \mu \equiv C\left(1 - \frac{\beta}{2} t_{\mathbf{p}}\right). \quad (24)$$

The first condition is independent of the Coulomb potential and determines a lower critical concentration ($n_d = 2/3$) above which the superconductivity appears. The second condition reduces to the relation $\beta = (2 - n)/\mu$ determining the upper critical dopant concentration which depends on the ratio $\beta w \sim V/w$ of the Coulomb potential V and the seeding energy width w of the electron subband.

For $T = 0$ and $n = 1$, the maximum value of the chemical potential is $\mu = 1/2$ and, hence, the upper critical electron density for $\beta < 2$ remains equal to unity. In other words, the superconductivity under these conditions exists inside the fixed interval $2/3 < n < 1$. The transition temperature significantly decreases when β increases from zero to two. However, as the Coulomb potential increases further, the second critical concentration determined from the system of equations appears,

$$\beta = \frac{2f}{\mu}, \quad n = 2f\int_{-1}^{\mu/f} \rho_0(\epsilon)\,d\epsilon. \quad (25)$$

Since $f = 1 - n/2$, the function $n_{c2}(\beta)$ can be set in the parametric form,

$$\beta = \frac{2}{\zeta}, \quad n_{c2}(\zeta) = \frac{2K(\zeta)}{1 + K(\zeta)}, \qquad K(\zeta) = \int_{-1}^{\zeta} \rho_0(\epsilon)\,d\epsilon, \quad \zeta = \frac{\mu}{f}, \quad 0 < \zeta < 1, \quad (26)$$

where $\rho_0(\epsilon)$ is the dimensionless seeding density of states.

In the case of a constant density of states, we obtain $n_{c2} = 2(2 + \beta)/(2 + 3\beta)$. When the parameter β increases, n_{c2} decreases and, in the limit of $\beta \longrightarrow \infty$, $n_{c2} \longrightarrow 2/3$ so that the region of existence of the superconducting state disappears. As can be checked, this behavior is retained in the general case of an arbitrary seeding density of states.

Figure 4 shows the plot of the transition temperature versus the electron density n, which was calculated for a constant density of states. The value of $\beta = 2.36$ corresponds to the experimentally observed critical concentration $n'_{c2} = 1.04$. A comparison of the $T_c(n)$ curves presented in Figs. 2 and 4 shows that the inclusion of the direct Coulomb interaction leads to a significant shift of the position of maximum toward lower (and higher) electron densities: from $n_m = 0.93$ to $n_m = 0.8$ (and accordingly, from $n'_m = 1.07$ to $n'_m = 1.16$).

Thus, by selecting the magnitude of the direct Coulomb repulsion, it is possible to fit the critical electron density corresponding to the disappearance of superconductivity. However, the value of the transition temperature turns out to be overstated because the calculation was performed using the Born approximation for the scattering amplitude. As will be shown below, the inclusion of scattering on the spin fluctuations provides for a decrease in T_c.

4. ALLOWANCE FOR RELAXATION PROCESSES

As is known from the original Hubbard papers, electron excitations in the normal phase are characterized by a finite free path length determined by scattering on the charge and spin fluctuations. In the limit of infinite Hubbard energy, the fluctuations of one-particle charge and spin states are determined by the longitudinal and

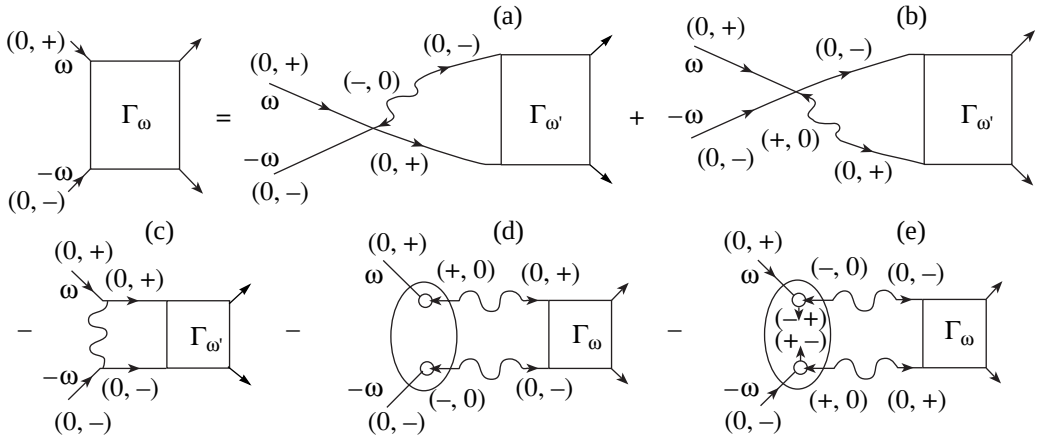


Fig. 5. A homogeneous equation for the vertex part with allowance for (a, b) kinematic and (c) Coulomb interaction and the scattering on (d) longitudinal, and (e) transverse spin fluctuations.

transverse parts of spin correlators calculated in the same unit cell,

$$K_\parallel = \langle \Delta\{\hat{X}^{+,0}, \hat{X}^{0,+}\}_{\mathbf{r}} \Delta\{\hat{X}^{+,0}, \hat{X}^{0,+}\}_{\mathbf{r}} \rangle,$$
$$K_\perp = \frac{1}{2}\langle \{\hat{X}^{+,-}_{\mathbf{r}}, \hat{X}^{-,+}_{\mathbf{r}}\} \rangle, \tag{27}$$

where

$$\Delta\{\hat{X}^{+,0}, \hat{X}^{0,+}\}_{\mathbf{r}} = \hat{X}^{+,+}_{\mathbf{r}} + \hat{X}^{0,0}_{\mathbf{r}} - \langle (\hat{X}^{+,+}_{\mathbf{r}} + \hat{X}^{0,0}_{\mathbf{r}}) \rangle.$$

In the one-loop approximation,

$$[G^{(\sigma)}_\omega(\mathbf{p})]^{-1} = [G^{(\sigma),(0)}_\omega(\mathbf{p})]^{-1} - \Sigma^{(\sigma)}_\omega(\mathbf{p}),$$

$$\Sigma^{(\sigma)}_\omega(\mathbf{p}) = K_\parallel t(\mathbf{p}) \sum_{\mathbf{p}'} G^{(\sigma)}_\omega(\mathbf{p}') t(\mathbf{p}') \tag{28}$$

$$+ K_\perp t(\mathbf{p}) \sum_{\mathbf{p}'} G^{(-\sigma)}_\omega(\mathbf{p}') t(\mathbf{p}').$$

For the normal paramagnetic phase, Eqs. (28) are rewritten as

$$[G^{(\sigma)}_\omega(\mathbf{p})]^{(-1)} = i\omega - (f + K_1\sigma_\omega) t_{\mathbf{p}} + \mu, \tag{29}$$

where

$$f = \frac{1+x}{2}, \quad \sigma_\omega = \sum_{\mathbf{p}} G_\omega(\mathbf{p}) t(\mathbf{p}),$$
$$x = 1 - n, \quad K_1 = K_\parallel + K_\perp. \tag{30}$$

Equations for the transition temperature with allowance for the scattering on the charge and spin fluctuations have a graphical representation as in Fig. 5 and can be analytically represented as

$$\Gamma_\omega(\mathbf{p}) + K_2 \sum_{\mathbf{p}'} G_\omega(\mathbf{p}') G_{-\omega}(-\mathbf{p}') t^2_{\mathbf{p}'} \Gamma_\omega(\mathbf{p}')$$
$$= 2T \sum_{\omega', \mathbf{p}'} \Gamma_{\omega'}(\mathbf{p}') G_{\omega'}(\mathbf{p}') G_{-\omega'}(-\mathbf{p}') t_{\mathbf{p}'}, \tag{31}$$

where

$$K_2 = K_\perp - \langle \Delta\{\hat{X}^{+,0}, \hat{X}^{0,+}\}_{\mathbf{r}} \Delta\{\hat{X}^{-,0}, \hat{X}^{0,-}\}_{\mathbf{r}} \rangle.$$

After the separation of variables, the conditions of solvability of Eqs. (31) can be represented in the following form:

$$2T \sum_\omega \frac{S^{(1)}_\omega}{1 + K_2 S^{(2)}_\omega} = 1, \tag{32}$$

where the sums over momenta are expressed via the function σ_ω determined from the self-consistency condition (30),

$$S^{(1)}_\omega = \sum_{\mathbf{p}} t(\mathbf{p}) G_\omega(\mathbf{p}) G_{-\omega}(-\mathbf{p})$$
$$= \frac{\sigma_r [K_1\sigma_s + f]}{\mu K_1 \sigma_r - 2i\omega f - i\omega K_1 \sigma_s},$$
$$S^{(2)}_\omega = \sum_{\mathbf{p}} t^2(\mathbf{p}) G_\omega(\mathbf{p}) G_{-\omega}(-\mathbf{p}) \tag{33}$$
$$= \frac{\mu\sigma_r + i\omega\sigma_s}{\mu K_1 \sigma_r - 2i\omega f - i\omega K_1 \sigma_s},$$

and $\sigma_{s,r} = \sigma_\omega \pm \sigma_{-\omega}$. Substituting these expressions

into (32), we obtain an equation

$$2T\sum_{\omega}\frac{\sigma_r(f+K_s\sigma_s)}{\mu K_s\sigma_r - 2i\omega f - i\omega K_r\sigma_s} = 1, \qquad (34)$$

where $K_{s,r} = K_1 \pm K_2$. In the low-temperature limit where $K_{1,2} \longrightarrow 0$, we obtain Eq. (16) for the superconducting transition temperature in the ideal model.

The quantities $\sigma_{r,s}$ can be expressed via the density of states $\rho = \sum_{\mathbf{p}}\delta(\xi_{\mathbf{p}})$ on the Fermi surface:

$$\begin{aligned}\sigma_s &= \sum t_{\mathbf{p}}\frac{1}{i\omega_n - \xi_{\mathbf{p}}} + \sum t_{\mathbf{p}}\frac{1}{-i\omega_n - \xi_{\mathbf{p}}}\\ &= -\sum t_{\mathbf{p}}\frac{2\xi_{\mathbf{p}}}{\omega_n^2 + \xi_{\mathbf{p}}^2} = -2t^*\rho\int_{-\infty}^{\infty}d\xi\frac{2\xi}{\omega_n^2+\xi^2} = 0,\end{aligned} \qquad (35a)$$

$$\begin{aligned}\sigma_r &= -2\sum t_{\mathbf{p}}\frac{i\omega}{\omega_n^2+\xi_{\mathbf{p}}^2} = -2i\omega t^*\rho\int_{-\infty}^{\infty}d\xi\frac{1}{\omega_n^2+\xi^2}\\ &= -2i\omega t^*\rho\frac{\pi}{|\omega_n|} = -2it^*\pi\rho\,\mathrm{sgn}(\omega_n),\end{aligned} \qquad (35b)$$

where the t^* value is determined from the condition $\xi(t^*) = 0$. The density of states on the Fermi surface is expressed via the seeding density of states $\rho_0(\epsilon) = \sum\delta(\epsilon - t_{\mathbf{p}})$ as

$$\begin{aligned}\xi(t^*) &= ft^* - \mu = 0,\\ \rho &= \sum_{\mathbf{p}}\delta(ft_{\mathbf{p}} - \mu) = \int\rho_0(\epsilon)\delta(f\epsilon - \mu)d\epsilon = \frac{1}{f}\rho_0\left(\frac{\mu}{f}\right).\end{aligned} \qquad (36)$$

As can be seen, Eq. (34) in fact contains only the combination μK_s, which has the meaning of the inverse time of relaxation with spin flipping.

In order to determine the temperature dependence of the time of relaxation with spin flipping, let us express K_s via mean-square spin fluctuations. To this end, the initial expression for K_s can be transformed as

$$\begin{aligned}K_s &= 2K_\perp + \left\langle\Delta\{\hat{X}^{+,0},\hat{X}^{0,+}\}_{\mathbf{r}}\Delta\{\hat{X}^{+,0},\hat{X}^{0,+}\}_{\mathbf{r}}\right\rangle\\ &\quad - \left\langle\Delta\{\hat{X}^{+,0},\hat{X}^{0,+}\}_{\mathbf{r}}\Delta\{\hat{X}^{-,0},\hat{X}^{0,-}\}_{\mathbf{r}}\right\rangle\\ &= 2K_\perp + \frac{1}{2}\left\langle(\Delta\hat{X}_{\mathbf{r}}^{+,+} + \Delta\hat{X}_{\mathbf{r}}^{0,0})(\Delta\hat{X}_{\mathbf{r}}^{+,+} - \Delta\hat{X}_{\mathbf{r}}^{-,-})\right\rangle\\ &\quad + \frac{1}{2}\left\langle(\Delta\hat{X}_{\mathbf{r}}^{-,-} + \Delta\hat{X}_{\mathbf{r}}^{0,0})(\Delta\hat{X}_{\mathbf{r}}^{-,-} - \Delta\hat{X}_{\mathbf{r}}^{+,+})\right\rangle\\ &= 2K_\perp + \frac{1}{2}\left\langle(\Delta\hat{X}_{\mathbf{r}}^{+,+} - \Delta\hat{X}_{\mathbf{r}}^{-,-})(\Delta\hat{X}_{\mathbf{r}}^{+,+} - \Delta\hat{X}_{\mathbf{r}}^{-,-})\right\rangle\\ &= 2K_\perp + \frac{1}{2}\left\langle(\Delta\hat{X}_{\mathbf{r}}^{+,+} - \Delta\hat{X}_{\mathbf{r}}^{-,-})^2\right\rangle.\end{aligned} \qquad (37)$$

Introducing the electron spin operators

$$\begin{aligned}\hat{S}^z &= \frac{1}{2}(\hat{X}^{+,+} - \hat{X}^{-,-}),\\ \hat{S}^+ &= \hat{S}^x + i\hat{S}^y = \hat{X}^{+,-},\\ \hat{S}^- &= \hat{S}^x - i\hat{S}^y = \hat{X}^{-,+},\end{aligned} \qquad (38)$$

we obtain an expression for K_s via a sum of the mean-square spin fluctuations:

$$\begin{aligned}K_s &= 2\langle\hat{S}_{\mathbf{r}}^+\hat{S}_{\mathbf{r}}^-\rangle + 2\left\langle(\Delta\hat{S}_{\mathbf{r}}^z)^2\right\rangle\\ &= 2\sum_{k=x,y,z}\left\langle(\Delta\hat{S}_{\mathbf{r}}^k)^2\right\rangle.\end{aligned} \qquad (39)$$

Based on the isotropy considerations, this sum is expressed via the mean-square fluctuation $\left\langle(\Delta S_{\mathbf{r}}^z)^2\right\rangle$ which, in turn, can be written in terms of the static spin susceptibility as

$$K_s = 6\left\langle(\Delta\hat{S}_{\mathbf{r}}^z)^2\right\rangle = 6T\frac{\partial\langle S^z\rangle}{\partial H} = 6T\chi(T), \qquad (40)$$

where $\chi(T)$ is the static susceptibility of the normal phase.

Recently, the author demonstrated [6] that the spin magnetic susceptibility in the Hubbard model with infinite repulsion is given by the formula

$$\begin{aligned}\chi(n,T) &= 2\frac{\partial n^\sigma}{\partial h}\\ &= \frac{-2fD_0}{1-K-fD_1-(1-K)D_1-f(D_0D_2-D_1^2)}.\end{aligned} \qquad (41)$$

By virtue of the equation of state, the K value can be expressed via the electron density and the other coefficients, via the integrals of a derivative of the Fermi function:

$$\begin{aligned}1-K &= \frac{2(1-n)}{(2-n)}, \quad f = 1-\frac{n}{2},\\ D_k &= \sum_{\mathbf{p}}t_{\mathbf{p}}^k n_F'(\xi_{\mathbf{p}}).\end{aligned} \qquad (42)$$

For $T \longrightarrow 0$, these coefficients can be expressed via the seeding density of states $\rho_0(\epsilon)$ and, hence, via the density of states on the Fermi level:

$$\begin{aligned}\rho_0(\epsilon) &= \sum_{\mathbf{p}}\delta(\epsilon - t_{\mathbf{p}}), \quad K = \int\rho_0(\epsilon)\theta(\mu - f\epsilon)d\epsilon,\\ D_s &= -\int\epsilon^s\rho_0(\epsilon)\delta(f\epsilon - \mu)d\epsilon.\end{aligned} \qquad (43)$$

In the limit of $T = 0$, the coefficients obey the relation $D_0D_2 = D_1^2$ and, hence, the susceptibility can be

expressed as a function of the electron density with only two coefficients (D_0 and D_1):

$$\chi(n, 0) = \frac{-2fD_0}{1-K-(1-K+f)D_1} = \frac{2\rho_0(\mu/f)}{1-K+(1-K+f)(\mu/f^2)\rho_0(\mu/f)}. \quad (44)$$

Since both the density of states $\rho_0(\epsilon)$ and the factor $1 - K + f$ are always positive, we may ascertain that the system with small electron densities ($\mu < 0$) always exhibits a tendency toward ferromagnetism.

In the case of a constant seeding density of states $\rho_0(\epsilon) = \theta(1-\epsilon^2)/2w$ at $T = 0$, the magnetic susceptibility goes to infinity only for $n = 0$ (see Fig. 6):

$$\chi(n, T=0) = \frac{1}{w}\frac{2(2-n)^3}{n(8-6n-n^2)}, \quad (45)$$

where n is the electron density and w is the energy half-width of the lower Hubbard subband.

For a square or bcc lattices, the density of states logarithmically tends to infinity on the zero energy level. However, the influence of the singularity in the denominator is compensated by the energy factor μ/f so that the magnetic susceptibility remains constant for all electron densities.

At a finite temperature and a constant density of states, all integrals are explicitly calculated and the chemical potential and, hence, the magnetic susceptibility can be determined (Fig. 6). At a small electron density, we obtain the Curie–Weiss law corresponding to the gas phase. In this limit, the mean-square correlator is temperature-independent and acquires the classical value $S(S+1) = 3/4$. For intermediate densities, the Curie–Weiss law is valid only at a sufficiently high temperature. At low temperatures ($T \ll t$), the system exhibits the Pauli spin susceptibility with a Stoner factor significantly above unity (Fig. 6). Here, the susceptibility never goes to infinity and the system remains paramagnetic at all electron densities (see [7, 8]).

Thus, it can be ascertained that, for $T \ll t$ and $n \gg T/t$, the value of K_s linearly tends to zero with decreasing temperature, $K_s \longrightarrow 0$, with a proportionality factor equal to the limiting value of the magnetic susceptibility at $T = 0$. From this we infer that the spin correlators are linear functions of the temperature and strongly depend on the degree of doping $x = 1 - n$.

Performing calculations for zero magnitude of spin fluctuations and subtracting the corresponding sum over frequencies from the left-hand part of Eq. (34), we can obtain a simpler form of the equation for the superconducting transition temperature T_c (for analogous calculations, see [9]). It can noticed that, after such subtraction, the sums over frequencies are taken in the region of $|\omega_n| \approx T_c$ and, hence, it is sufficient to take the limit of $\mathrm{sgn}(\omega)\sigma_r(0+)$ instead of the function $\sigma_r(\omega)$. Then, the right-hand part is calculated with a logarithmic accuracy, while the differential sum in the left-hand part can be expressed in terms of the digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$ as

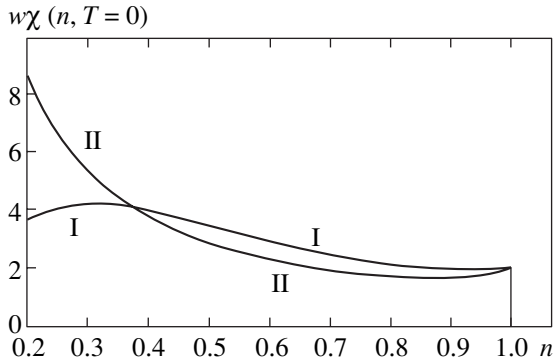


Fig. 6. Plots of the magnetic susceptibility versus electron density at $T = 0$ for (I) a semielliptic density of states and (II) a flat band model.

$$\frac{1}{g} + \psi\left[\frac{1}{2} + \frac{\mu\sigma^* K_s}{4\pi f T_c}\right] - \psi\left(\frac{1}{2}\right) = \int_0^{\bar{\omega}} \frac{\tanh(\xi/2T_c)}{\xi} d\xi = \ln\frac{2\gamma\bar{\omega}}{\pi T_c}. \quad (46)$$

Here, all values are determined on the Fermi level $\epsilon^* = \mu/f$ and expressed via the seeding density of states $\rho_0(\epsilon) = \sum_{\mathbf{p}} \delta(\epsilon - t_{\mathbf{p}})$:

$$\sigma^* = 2\pi\epsilon^*\rho_0(\epsilon^*)/f, \quad g = 2\epsilon^*\rho_0(\epsilon^*)/f. \quad (47)$$

The correlation function K_s goes to zero in the limit as $T \longrightarrow 0$ and is a linear function of the temperature at $T \gg w$. For this reason, the influence of spin fluctuations reduces to renormalization of the BCS constant:

$$\frac{1}{g} \longrightarrow \frac{1}{g^*} = \frac{1}{g} + \psi\left[\frac{1}{2} + \frac{\mu\sigma^* K_s}{4\pi f T_c}\right] - \psi\left(\frac{1}{2}\right), \quad \pi T_c \approx \gamma\bar{\omega}\exp\left(-\frac{1}{g^*}\right). \quad (48)$$

This relation is supplemented by an equation for determining the energy gap at $T = 0$:

$$\frac{1}{g} = \ln\frac{2\bar{\omega}}{\Delta_0}. \quad (49)$$

Thus, the ratio $2\Delta_0/T_c$ is independent of the parameter $\zeta = \mu\sigma^* K_s/4\pi f T$ calculated at $T = T_c$:

$$\frac{2\Delta_0}{T_c} = 8\pi\exp\left[\psi\left(\frac{1}{2} + \zeta\right)\right]. \quad (50)$$

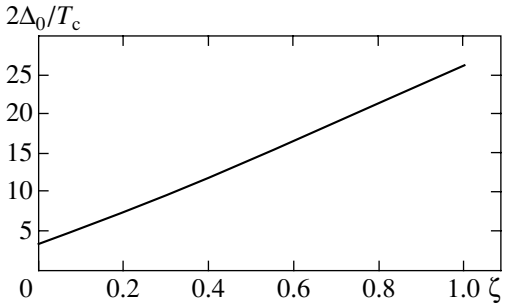


Fig. 7. A plot of the dimensionless $2\Delta_0/T_c$ ratio versus parameter ζ constructed using relation (50).

Since the value of K_s is proportional to the ratio of the temperature to the absolute value of the hopping integral, the parameter ζ is eventually on the order of unity. Accordingly, the ratio $2\Delta_0/T_c$ is overstated as compared to the well-known value approximately equal to 3.53 according to the classical BCS theory. This can be observed for most of high-T_c superconductors.

The measured values of $2\Delta_0/3.53T_c$ for various high-T_c superconductors fall within the interval from 1.27 to 3.12. As can be seen from Fig. 7, these values correspond to the parameters $\zeta \approx 0.05$–0.55. For example, the experimental values for $YBa_2Cu_3O_7$ are $2\Delta_0/3.53T_c \approx 1.98$–$2.27$ and $\zeta \approx 0.2$.

5. FINITE HUBBARD ENERGIES

In order to write the equation of state, let us obtain the Green functions in the simplest zero-loop approximation. For zero external field and a given spin projec-

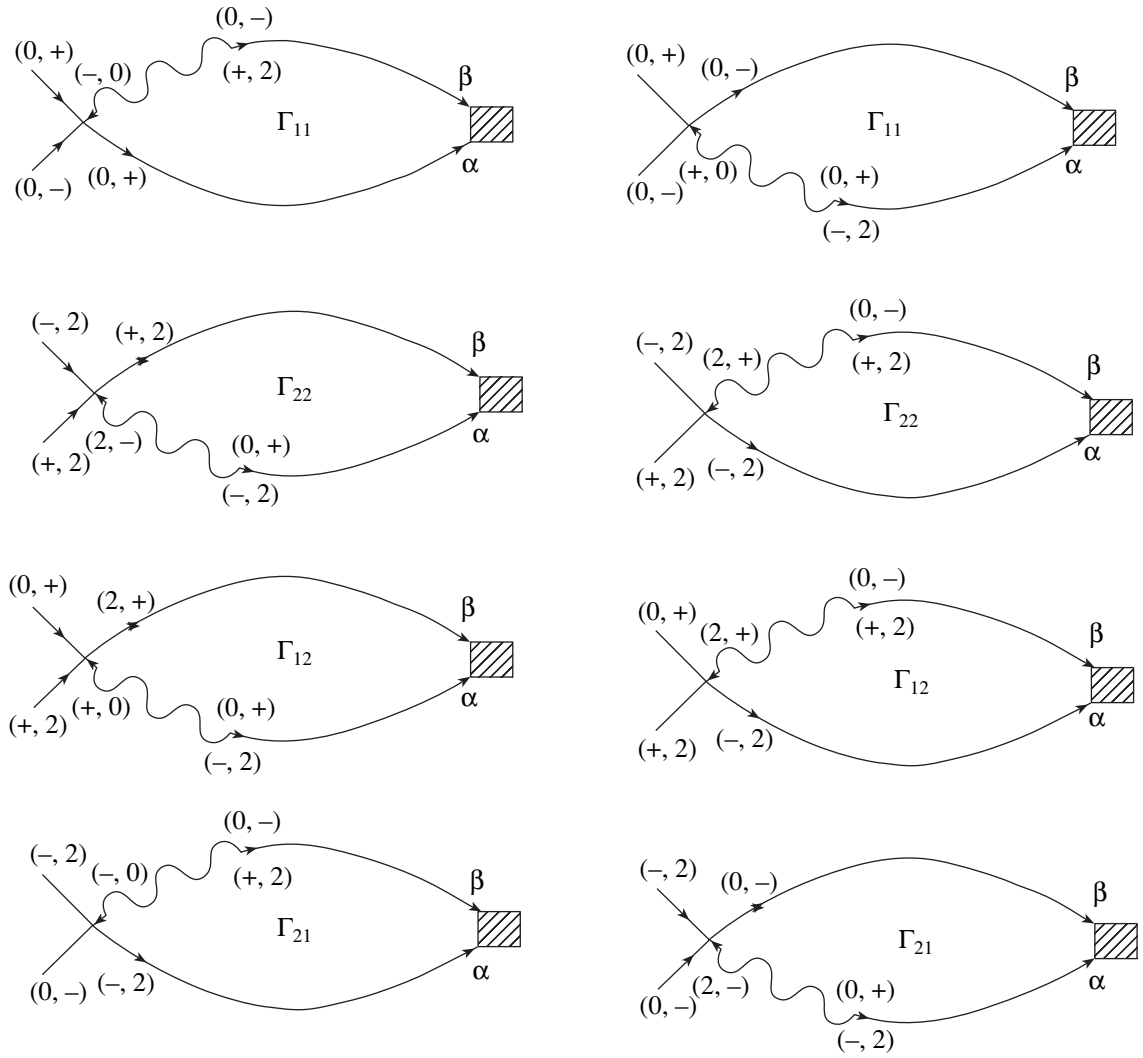


Fig. 8. The right-hand part of a homogeneous system of equation for four vertex parts of $\Gamma_{i,k}$. Cross-hatched squares indicate the vertex parts of $\Gamma_{\alpha,\beta}$. Indices "α" and "β" denote the transitions (0, +), (–, 2) and (0, –, (+, 2), respectively.

tion, the inverse Green function is represented by the matrix

$$G_\omega^{-1}\mathbf{p} = \begin{pmatrix} i\omega - \epsilon_1 - f_1 t_\mathbf{p}, & -\sigma f_1 t_\mathbf{p} \\ -\sigma f_2 t_\mathbf{p}, & i\omega - \epsilon_2 - f_2 t_\mathbf{p} \end{pmatrix}, \quad (51)$$

where $\epsilon_1 = -\mu$ and $\epsilon_2 = U - \mu$ are the energies of transitions between the empty and one-particle and the one-particle and two-particle levels, respectively (differing by the Hubbard energy U); $f_1 = 1 - n/2$ and $f_2 = n/2$ are the end factors expressed via the electron density n.

Within the framework of the zero-loop approximation, the equation of state is written using the sum of all possible products of the components of one-particle Green function and the end factors:

$$\langle \hat{a}_\mathbf{r}^\sigma + \hat{a}_\mathbf{r}^\sigma \rangle = \left\langle (\hat{X}_\mathbf{r}^{\sigma,0} + \sigma \hat{X}_\mathbf{r}^{2,-\sigma})(\hat{X}_\mathbf{r}^{0,\sigma} + \sigma \hat{X}_\mathbf{r}^{-\sigma,2}) \right\rangle$$

$$= T\sum_{\omega,\mathbf{p}} \exp(i\omega\delta) \quad (52)$$

$$\times \{(G_\omega^{11}(\mathbf{p}) + \sigma G_\omega^{21}(\mathbf{p}))f_1 + (G_\omega^{22}(\mathbf{p}) + \sigma G_\omega^{12}(\mathbf{p}))f_2\}.$$

Upon calculation and substitution of the matrix elements and summation over the spin projections, we arrive at the relation

$$n = 2T\sum_{\omega,\mathbf{p}} \exp(i\omega\delta)\left\{\frac{(i\omega - \epsilon_2)f_1 + (i\omega - \epsilon_1)f_2}{(i\omega - \xi_\mathbf{p}^+)(i\omega - \xi_\mathbf{p}^-)}\right\}, \quad (53)$$

where

$$\xi_\mathbf{p}^\pm = \frac{U}{2} + \frac{t_\mathbf{p}}{2} \pm \frac{1}{2}\sqrt{U^2 + t_\mathbf{p}^2 - 2U(1-n)t_\mathbf{p}} - \mu$$

are the two branches of the energy spectrum of one-particle excitations. Expanding the summand into simple multipliers, we obtain

$$n = 2T\sum_{\mathbf{p},\lambda=\pm} A^\lambda(\mathbf{p}) n_F(\xi_\mathbf{p}^\lambda),$$
$$A^\pm = \frac{1}{2}\left[1 \pm \frac{t_\mathbf{p} - U(1-n)}{\sqrt{U^2 + t_\mathbf{p}^2 - 2U(1-n)t_\mathbf{p}}}\right]. \quad (54)$$

The conditions of appearance of the Cooper instability at a finite Hubbard energy consist of two pairs of equations for the two types of excitations, $(0, \sigma)$ and $(-\sigma, 2)$, differing by the direction of spin projection and the sign of the transition energy $\epsilon_1 = \epsilon_\sigma - \epsilon_0$ and $\epsilon_2 = \epsilon_2 - \epsilon_{-\sigma}$ (see Fig. 8).

Taking into account the commutation relations for the electron density operators,

$$[\hat{X}^{0,\sigma}, \hat{n}] = \hat{X}^{0,\sigma}, \quad [\hat{X}^{-\sigma,2}, \hat{n}] = \hat{X}^{-\sigma,2},$$
$$\hat{n} = \hat{X}^{+,+} + \hat{X}^{-,-} + 2\hat{X}^{2,2}, \quad (55)$$

we readily obtain the Coulomb corrections to four scattering amplitudes (see Fig. 3).

Using the double commutation relations between four operators X, $(\hat{X}^{0,\sigma}, \hat{X}^{-\sigma,2})$, and $(\hat{X}^{\sigma,0}, \hat{X}^{2,-\sigma})$, we obtain a system of homogeneous equations for four vertex parts:

$$\Gamma_{11} = \Gamma(0+|0-), \quad \Gamma_{22} = \Gamma(-2|+2),$$
$$\Gamma_{12} = \Gamma(0+|+2), \quad \Gamma_{21} = \Gamma(-2|0-),$$
$$\Gamma_{11} = A_{\alpha',\beta'}\Gamma_{\alpha',\beta'}$$
$$-\beta t(\mathbf{p})T\sum_{\omega,\mathbf{p}'} G_\omega^{0+|\alpha'}(\mathbf{p}')G_{-\omega}^{0-|\beta'}(-\mathbf{p}')t(\mathbf{p}')\Gamma_{\alpha',\beta'},$$
$$\Gamma_{22} = B_{\alpha',\beta'}\Gamma_{\alpha',\beta'}$$
$$-\beta t(\mathbf{p})T\sum_{\omega,\mathbf{p}'} G_\omega^{-2|\alpha'}(\mathbf{p}')G_{-\omega}^{+2|\beta'}(-\mathbf{p}')t(\mathbf{p}')\Gamma_{\alpha',\beta'}, \quad (56)$$
$$\Gamma_{12} = \frac{1}{2}(A_{\alpha',\beta'} - B_{\alpha',\beta'})\Gamma_{\alpha',\beta'}$$
$$-\beta t(\mathbf{p})T\sum_{\omega,\mathbf{p}'} G_\omega^{0+|\alpha'}(\mathbf{p}')G_{-\omega}^{+2|\beta'}(-\mathbf{p}')t(\mathbf{p}')\Gamma_{\alpha',\beta'},$$
$$\Gamma_{21} = \frac{1}{2}(B_{\alpha',\beta'} - A_{\alpha',\beta'})\Gamma_{\alpha',\beta'}$$
$$-\beta t(\mathbf{p})T\sum_{\omega,\mathbf{p}'} G_\omega^{-2|\alpha'}(\mathbf{p}')G_{-\omega}^{0+|\beta'}(-\mathbf{p}')t(\mathbf{p}')\Gamma_{\alpha',\beta'},$$

where indices "α" and "β" independently run over two couples (0, +), (–, 2) and (0, –), (+, 2), respectively. Direct calculation of the matrix elements $A_{\alpha,\beta}$ and $B_{\alpha,\beta}$ leads to the following system of relations:

$$A_{11} = T\sum_{\omega,\mathbf{p}} \Phi t_\mathbf{p}\{(i\omega_1 - \epsilon_2)(i\omega_2 - \epsilon_2 - f_2 t_\mathbf{p}) + (i\omega_1 - \epsilon_2)(i\omega - \epsilon_2 - f_2 t_\mathbf{p})\},$$

$$A_{22} = -T\sum_{\omega,\mathbf{p}} \Phi t_\mathbf{p}^2 f_1\{(i\omega_1 - \epsilon_1) + (i\omega_2 - \epsilon_1)\},$$

$$A_{12} = -T\sum_{\omega,\mathbf{p}} \Phi t_\mathbf{p}\{(i\omega_1 - \epsilon_2)f_1 t_\mathbf{p} + (i\omega_1 - \epsilon_2 - f_2 t_\mathbf{p})(i\omega_2 - \epsilon_1)\},$$

$$
\begin{aligned}
A_{21} &= T\sum_{\omega,\mathbf{p}}\Phi t_{\mathbf{p}}\{(i\omega_1-\epsilon_1)(i\omega_2-\epsilon_2-f_2t_{\mathbf{p}})\\
&\quad+(i\omega_2-\epsilon_2)f_1t_{\mathbf{p}}\},\\
B_{11} &= T\sum_{\omega,\mathbf{p}}\Phi t_{\mathbf{p}}^2 f_2\{(i\omega_1-\epsilon_2)+(i\omega_2-\epsilon_2)\},\\
B_{22} &= -T\sum_{\omega,\mathbf{p}}\Phi t_{\mathbf{p}}\{(i\omega_1-\epsilon_1)(i\omega_2-\epsilon_1-f_1t_{\mathbf{p}})\\
&\quad+(i\omega_1-\epsilon_1)(i\omega_2-\epsilon_1-f_1t_{\mathbf{p}})\},\\
B_{12} &= -T\sum_{\omega,\mathbf{p}}\Phi t_{\mathbf{p}}\{(i\omega_1-\epsilon_2)(i\omega_2-\epsilon_1-f_1t_{\mathbf{p}})\\
&\quad+(i\omega_2-\epsilon_1)f_2t_{\mathbf{p}}\},\\
B_{21} &= T\sum_{\omega,\mathbf{p}}\Phi t_{\mathbf{p}}\{(i\omega_2-\epsilon_2)(i\omega_1-\epsilon_1-f_1t_{\mathbf{p}})\\
&\quad+(i\omega_2-\epsilon_1)f_2t_{\mathbf{p}}\},
\end{aligned}
\tag{57}
$$

$$
\begin{aligned}
\Phi &= [(\omega_n^2+\xi_{(+)}^2)(\omega_n^2+\xi_{(-)}^2)]^{-1},\\
\xi_\pm &= \frac{U}{2}+\frac{t_{\mathbf{p}}}{2}-\mu\pm\sqrt{U^2+t_{\mathbf{p}}^2-2U(1-n)t_{\mathbf{p}}},
\end{aligned}
\tag{58}
$$

where

$$
\begin{aligned}
i\omega_1 &= -i\omega_2 = i\omega_n = i\pi T(2n+1),\\
\epsilon_1 &= -\mu,\quad \epsilon_2 = U-\mu,\\
f_1 &= f(\sigma,0) = \left\langle(\hat{X}^{(\sigma,\sigma)}+\hat{X}^{(0,0)})\right\rangle,\\
f_2 &= f(2,-\sigma) = \left\langle(\hat{X}^{(-\sigma,-\sigma)}+\hat{X}^{(2,2)})\right\rangle.
\end{aligned}
$$

The coefficients $A_{\alpha,\beta}$ and $B_{\alpha,\beta}$ are symmetric: the substitution $\epsilon_1 \leftrightarrow \epsilon_2$, $f_1 \leftrightarrow f_2$ corresponds to $A_{11} \leftrightarrow -B_{22}$, $B_{11} \leftrightarrow -A_{22}$, $A_{12} \leftrightarrow -B_{21}$, and $A_{21} \leftrightarrow -B_{12}$.

Separating the variables, we obtain the following equations:

$$
\begin{aligned}
x_i &= A_{ik}^{(1)}x_k + A_{ik}^{(2)}y_k,\quad A_{ik}^{(\alpha)} = -A_{ki}^{(\alpha)},\\
y_i &= B_{ik}^{(1)}x_k + B_{ik}^{(2)}y_k,\quad i,k = 1,2,3.
\end{aligned}
\tag{59}
$$

Here, the first index corresponds to the scattering of (0, +) on the (0, –) excitation; the second index corresponds to the scattering of (–, 2) on the (+, 2) excitation; and the third index corresponds to a mixed scattering of (0, +) on the (–, 2) excitation (the fourth index is missing because $x_4 = -x_3$ and $y_4 = -y_3$). It can be also noted that $x_3 = (x_1 - x_2)/2$; this relation follows from the explicit form of $A^{(\alpha)}$ matrices. Upon summation over the frequencies, it turns out that summands depend on the momentum only via the function $t_{\mathbf{p}}$, which makes expedient the introduction of the seeding density of states $\rho_0(\epsilon) = \sum_{\mathbf{p}}\delta(\epsilon - t_{\mathbf{p}})$. In addition, it can be noted that integration for determining the pole part of the singular integrals is performed near the Fermi surface determined from the relation $\xi_{\mathbf{p}}^- = 0$. Using this condition, we can fix the hopping integral $t_{\mathbf{p}} = t^*$ (instead of the chemical potential μ) determined from the condition

$$
\mu = \frac{U}{2}+\frac{t^*}{2}-\frac{1}{2}\sqrt{U^2+(t^*)^2-2U(1-n)t^*}. \tag{60}
$$

As a result, all integrals can be expressed via t^*, $\epsilon_1 = -\mu$, and $\epsilon_2 = -\mu + U$:

$$
\hat{A}^{(1)} = L\begin{pmatrix}
-2t_*^2 f_1\epsilon_2^2\epsilon_1^{-1} & 2t_*^2\epsilon_1 f_1 & 2t_*^2\epsilon_2 f_1\\
-2t_*^2 f_2\epsilon_2 & 2t_*^2\epsilon_1^2 f_2\epsilon_2^{-1} & 2t_*^2\epsilon_1 f_2\\
-t_*^2\epsilon_2\epsilon_1^{-1}(f_1\epsilon_2 - f_2\epsilon_1) & t_*^2\epsilon_1\epsilon_2^{-1}(f_1\epsilon_2 - f_2\epsilon_1) & t_*^2(f_1\epsilon_2 - f_2\epsilon_1)
\end{pmatrix},
\tag{61}
$$

$$
\hat{A}^{(2)} = t_*\hat{A}^{(1)},\quad t_* = -\frac{\epsilon_1\epsilon_2}{f_1\epsilon_2 + f_2\epsilon_1},\quad L = T\sum_{\omega,\mathbf{p}}\frac{1}{(\omega_n^2+\xi_{(+)}^2)(\omega_n^2+\xi_{(-)}^2)}.
$$

Matrices $B^{(2)} = t_*B^{(1)}$ also exhibit a strong degeneracy,

$$
\hat{B}^{(2)} = -\beta L\begin{pmatrix}
t_*^4 f_1^2\epsilon_2^2\epsilon_1^{-2} & -t_*^4 f_1^2 & -t_*^4\epsilon_2\epsilon_1^{-1} f_1^2\\
-t_*^4 f_2^2 & t_*^4\epsilon_1^2 f_2^2\epsilon_2^{-2} & t_*^4\epsilon_1 f_2^2\epsilon_2^{-1}\\
-t_*^4 f_1 f_2\epsilon_2\epsilon_1^{-1} & t_*^4 f_1 f_2\epsilon_1\epsilon_2^{-1} & t_*^4 f_1 f_2
\end{pmatrix}.
\tag{62}
$$

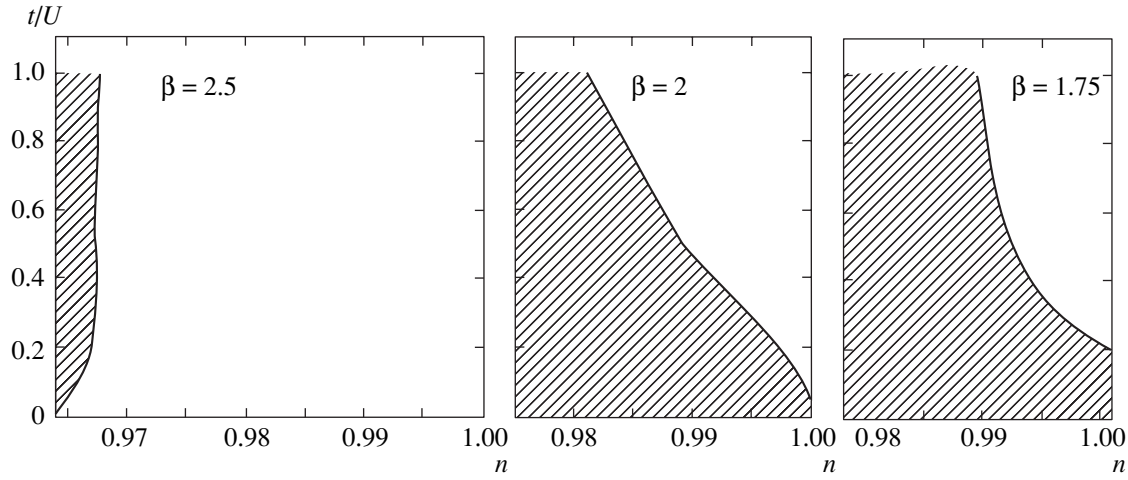


Fig. 9. Phase diagrams showing the regions of existence of the superconducting phase at $T = 0$ for the various values of the Coulomb potential β and the rectangular density of states.

The variables $x_{1,2}$ and $y_{1,2,3,4}$ in Eq. (59) are expressed via each other using the relations following from the form of A and B matrices:

$$x_1 = x_2\frac{f_1\epsilon_2}{f_2\epsilon_1}, \quad x_3 = -x_4 = \frac{(x_1 - x_2)}{2},$$
$$y_1 = y_2\frac{f_1^2\epsilon_2^2}{f_2^2\epsilon_1^2}, \quad y_3 = -y_4 = -y_1\frac{f_2\epsilon_1}{f_1\epsilon_2}. \tag{63}$$

Substituting these relations into Eqs. (59), we obtain the following conditions of solvability:

$$\Lambda\int_0^{\bar\omega}\frac{\tanh(\xi/2T)}{\xi}d\xi = 1,$$
$$\Lambda = -\left[\frac{2U\epsilon_1\epsilon_2}{f_2\epsilon_1^2 + f_1\epsilon_2^2} + \beta t_*^2\right]\rho_0(t_*)\left|\frac{\partial\xi_{(-)}}{\partial t_{\mathbf{p}}}\right|^{-1}. \tag{64}$$

Here, the calculations were performed for the lower Hubbard subband, so that the partial derivative is calculated for $\xi^{(-)}(\mathbf{p}) = 0$ or $t(\mathbf{p}) = t^*$.

The superconductivity takes place for the positive values of Λ, which has the meaning of the effective BCS constant. The coefficients in Eqs. (64) can be expressed via the degree of underfilling $x = 1 - n$ and the chemical potential $\bar\mu = \mu - U/2$. Eventually, we obtain

$$\epsilon_1 = -\bar\mu - \frac{U}{2}, \quad \epsilon_2 = -\bar\mu + \frac{U}{2},$$
$$f_1 = \frac{1+x}{2}, \quad f_2 = \frac{1-x}{2}, \quad t_* = \frac{U^2 - 4\bar\mu^2}{2(Ux - 2\bar\mu)},$$
$$-\frac{2U\epsilon_1\epsilon_2}{f_2\epsilon_1^2 + f_1\epsilon_2^2} - \beta t_*^2 = (U^2 - 4\bar\mu^2)$$
$$\times\left\{\frac{2U}{U^2 + 4\bar\mu^2 + 4xU\bar\mu} - \beta\frac{(U^2 - 4\bar\mu^2)}{4(Ux - 2\bar\mu)^2}\right\}. \tag{65}$$

Substitution of these expressions into Eqs. (64) shows that, for a given finite value of the interstitial Coulomb repulsion between electrons (parameter β), the phase diagram is symmetric relative to the simultaneous substitution $\bar\mu \longrightarrow -\bar\mu$ and $x \longrightarrow -x$. It can be also shown that $\bar\mu(x)$ is antisymmetric if the seeding density of states $\rho_0(\epsilon) = \sum_{\mathbf{p}}\delta(\epsilon - t_{(\mathbf{p})})$ is an even function. For this reason, the phase diagram describing the existence of superconductivity in terms of the variables $(x, t/U)$ is symmetric with respect to the particle–hole transformation $x \longrightarrow -x$.

For the sake of illustration, let us perform further calculations for a constant seeding density of states $\rho_0(\epsilon)$ with a unity halfwidth. In this case, the equation of state (53) can be represented in the form of a full derivative,

$$n = 2\sum_{\mathbf{p}}\frac{\delta\xi^-(t_{\mathbf{p}})}{\delta t_{\mathbf{p}}}\theta(-\xi^-(t_{\mathbf{p}})) = \xi^-(t^*) - \xi^-(-1). \tag{66}$$

For $T = 0$, this yields the following explicit function $t^*(n)$ and the energies:

$$t^* = [2n^2 + u(1-n) - 2n + 1 + (1-2n)\sqrt{(1+u)^2 - 2nu}]$$
$$\times[2n - u(1-n) - 1 - \sqrt{(1+u)^2 - 2nu}]^{-1}, \quad u = \frac{U}{t}, \tag{67}$$
$$\epsilon_{1,2} = \mp\frac{u}{2} - n + \frac{1}{2} + \frac{1}{2}\sqrt{(1+u)^2 - 2nu}.$$

These relations, together with the effective BCS constant determined above, allow the phase diagram describing the existence of superconductivity can be

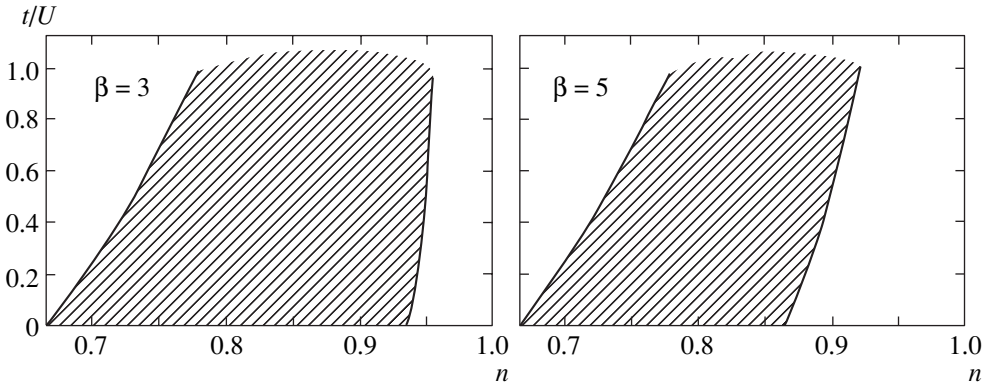


Fig. 10. Phase diagrams showing the regions of existence of the superconducting phase at $T = 0$ for the various values of the Coulomb potential β and the rectangular density of states.

constructed for all values of the dimensionless Hubbard energy U/t, Coulomb parameter β, and electron density n.

For any strength of the Coulomb interaction, there exists a critical electron density, corresponding to zero chemical potential, above which the Cooper instability can arise. In the flat band model, the corresponding value can be determined from the condition $\epsilon_1(n, u) = 0$, which yields

$$n_{c1} = \frac{1}{2} - \frac{3u}{4} + \frac{1}{4}\sqrt{9u^2 + 4u + 4}. \quad (68)$$

As the intercell Coulomb repulsion increases, the BCS constant decreases and the region of existence of the superconducting state diminishes. Beginning with a certain critical value $\beta_c = 2$ in a system with arbitrarily large Hubbard repulsion, the superconductivity appears when there is a finite underfilling n_{c2} of the Hubbard subband (cf. Figs. 9 and 10). The temperature of the superconducting transition exhibits two zeros, corresponding to the electron densities n_{c1} and n_{c2}. This very form of the phase diagram is observed for high-T_c superconductors.

6. CONCLUSIONS

It was demonstrated that the phase space of the Hubbard model with strong repulsion contains a finite region in which the scattering amplitude is negative. It was also found that the Fermi level can occur entirely in the region where the scattering amplitude corresponds to the Cooper pairing. In this system, the superconductivity can exist only in a limited interval of electron densities.

In this study, various phase diagrams were obtained for the most symmetric case of superconductivity of the Cooper type with *s*-pairing. As for the phase diagram in a system with *d*-pairing, it can be obtained proceeding from the same Eqs. (56–59) written for a finite Hubbard energy. In the limit of infinite Hubbard energy, only the region of *s*-pairing is retained.

In the nonsuperconducting part of the phase diagram, the effect of magnetic fluctuations reduces to the appearance of relaxation with spin flipping, the rate of which is proportional to the first power of the temperature. This leads to a decrease in the superconducting transition temperature. In the superconducting part of the phase diagram, the relaxation rate exhibits an additional decrease caused by a rapid drop of the superconducting gap and the corresponding decrease in the density of state on the Fermi level. This corresponds to an increase in the $2\Delta_0/T_c$ ratio.

ACKNOWLEDGMENTS

This study was supported by the Scientific Council on Superconductivity within the framework of the Federal Scientific-Technological Program “Basic Directions in the Physics of Condensed Media.”

REFERENCES

1. L. P. Gor’kov, Zh. Éksp. Teor. Fiz. **34**, 735 (1958) [Sov. Phys. JETP **7**, 505 (1958)].
2. F. Dyson, Phys. Rev. **102**, 1217, 1230 (1956).
3. R. O. Zaitsev, Phys. Lett. A **134**, 199 (1988).
4. V. V. Val’kov, D. M. Dzebisashvili, and A. S. Kravtsov, Pis’ma Zh. Éksp. Teor. Fiz. **77**, 604 (2003) [JETP Lett. **77**, 505 (2003)].
5. J. Zelinski, M. Matlak, and P. Entel, Phys. Lett. A **136**, 441 (1989).
6. R. O. Zaĭtsev, Zh. Éksp. Teor. Fiz. **123**, 325 (2003) [JETP **96**, 286 (2003)].
7. J. Hubbard and K. R. Jain, J. Phys. C **1**, 1650 (1968).
8. J. W. Schweitzer, Phys. Rev. B **3**, 2357 (1971).
9. L. P. Gor’kov and A. I. Rusinov, Zh. Éksp. Teor. Fiz. **46**, 1361 (1964) [Sov. Phys. JETP **19**, 920 (1964)].

Translated by P. Pozdeev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 793–810.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 906–926.

SOLIDS
Electronic Properties

Genesis of the Anomalous Hall Effect in $CeAl_2$

N. E. Sluchanko[a,b,*], A. V. Bogach[a,b], V. V. Glushkov[a,b], S. V. Demishev[a,b], M. I. Ignatov[a,b], N. A. Samarin[a], G. S. Burkhanov[c], and O. D. Chistyakov[c]

[a] *Institute of General Physics, Russian Academy of Sciences, ul. Vavilova 38, Moscow, 119991 Russia*
[b] *Moscow Physicotechnical Institute, Institutskiĭ proezd 9, Dolgoprudnyĭ, Moscow oblast, 141700 Russia*
[c] *Baĭkov Institute of Metallurgy and Materials Science, Russian Academy of Sciences, Leninskiĭ pr. 49, Moscow, 119991 Russia*
**e-mail: nes@lt.gpi.ru*

Received October 6, 2003

Abstract—The Hall coefficient R_H, resistivity ρ, and Seebeck coefficient S of the $CeAl_2$ compound with fast electron density fluctuations were studied in a wide temperature range (from 1.8 to 300 K). Detailed measurements of the angular dependences $R_H(\varphi, T, H \le 70$ kOe) were performed to determine contributions to the anomalous Hall effect and study the behavior of the anomalous magnetic R_H^{am} and main R_H^a components of the Hall signal of this compound with strong electron correlation. The special features of the behavior of the anomalous magnetic component R_H^{am} were used to analyze the complex magnetic phase diagram H–T determined by magnetic ordering in the presence of strong spin fluctuations. An analysis of changes in the main contribution $R_H^a(H, T)$ to the Hall effect made it possible to determine the complex activation behavior of this anomalous component in the $CeAl_2$ intermetallic compound. The results led us to conclude that taking into account spin–polaron effects was necessary and that the Kondo lattice and skew-scattering models were of very limited applicability as methods for describing the low-temperature transport of charge carriers in cerium-based intermetallic compounds. The effective masses and localization radii of manybody states in the $CeAl_2$ matrix were estimated to be (55–90)m_0 and 6–10 Å, respectively. The behaviors of the parameters R_H, S, and ρ were jointly analyzed. The results allowed us to consistently describe the transport coefficients of $CeAl_2$. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

One of the most interesting and complicated properties of rare-earth metal-based compounds with intermediate valence and heavy fermions is the Hall coefficient R_H [1–4]. In particular, in the overwhelming majority of cases, the Hall effect in cerium-based intermetallic compounds is anomalous both in the magnitude and sign of R_H. Indeed, the R_H value for conducting cerium compounds is dozens of times larger than the Hall coefficient of their nonmagnetic analogues (La, Y, Lu, etc., compounds) and positive at temperatures comparable to the characteristic temperature of spin fluctuations T_{sf} [1, 2]. Studies performed by various authors for various Ce-based intermetallic compounds ($CeAl_3$ [5], $CeCu_2Si_2$ [6], $CeCu_6$ [7], $CePd_3$ [8], CeNiSn [9], $CeOs_2$ [10], $CePb_3$ [11], etc.) showed that their $R_H(T)$ temperature dependences contained large-amplitude maxima at $T_{\max}^{R_H}$ in the neighborhood of the T_{sf} temperature. According to [1, 2], the most correct explanation of this behavior of the $R_H(T)$ parameter can be obtained using the model of skew-scattering of charge carriers by localized magnetic moments of rare-earth metal ions. However, our preliminary study of the Hall coefficient performed comparatively recently [12] for a typical representative of this class of compounds, the so-called magnetic Kondo lattice $CeAl_2$, revealed a complex activation behavior of the $R_H(T)$ parameter, which did not fit in with the concept [1, 2] of the determining role played by the scattering effect in the formation of anomalies of the R_H coefficient in this intermetallic compound.

In this work, we thoroughly studied the Hall effect in $CeAl_2$ in a wide temperature range from 1.8 to 300 K and in magnetic fields of up to 70 kOe. Our goal was to experimentally examine if the existing theoretical approaches could be used to describe the anomalous Hall effect in rare-earth metal compounds with long-range magnetic order and heavy fermions. To elucidate the special features of the low-temperature transport of charge carriers in $CeAl_2$, we also measured the temperature dependences of the Seebeck coefficient $S(T)$ and resistivity $\rho(H_0, T)$ at fixed magnetic field values.

2. EXPERIMENTAL

The transport characteristics were measured in this work for polycrystalline $CeAl_2$ samples with a cubic Laves phase structure. The samples were synthesized from stoichiometric amounts of high-purity components in an electric arc furnace with a nonconsumable tungsten electrode on a water-cooled copper hearth in an atmosphere of purified helium. Composition homogeneity in the bulk was attained by repeatedly remelting the samples with subsequent homogenizing annealing in evacuated quartz ampules. X-ray (DRON-3) and microstructure (optical microscopy) analyses showed the products to be single-phase.

The Seebeck coefficient was measured by the four-probe method using a setup of an original design similar to that described in [13]. The temperature gradients ΔT on the samples were varied in the region of a linear response of the thermal electromotive force voltage $U_{1,2} \propto \Delta T$ measured for various pairs of sample contacts [13]. The relative maximum superheating value $\Delta T/T$ between the "cold" and "warm" sample ends was no more than 5%.

An experimental setup of an original design was used to measure the Hall coefficient; its block diagram is shown in Fig. 1. Measurements were taken in cryostat *1* with superconducting magnet *2* while the sample was step-by-step rotated in a magnetic field. Sample *3* prepared for direct-current measurements in the four-contact configuration (Fig. 1, inset a) was placed on copper plate *4* (Fig. 1, inset b) of a rotary device together with Hall probe *5* and CERNOX 1050 standard resistance thermometer *6* (Lake Shore Cryotronics, USA). Hall probe *5* measured the magnetic field vector component normal to the surface of the sample. The assemblage on copper plate *4* was rotated in the magnetic field of the superconducting magnet step-by-step (steps of $\Delta\varphi = 1.8°$) by step motor *7*. After each rotation through 2–5 steps, the position of the assemblage was fixed (see insets in Fig. 1), and signals from the sample Hall contacts, Hall probe, and resistance thermometer were measured. A model 2182 (Keithley, USA) nanovoltmeter was used to directly perform precision measurements of the

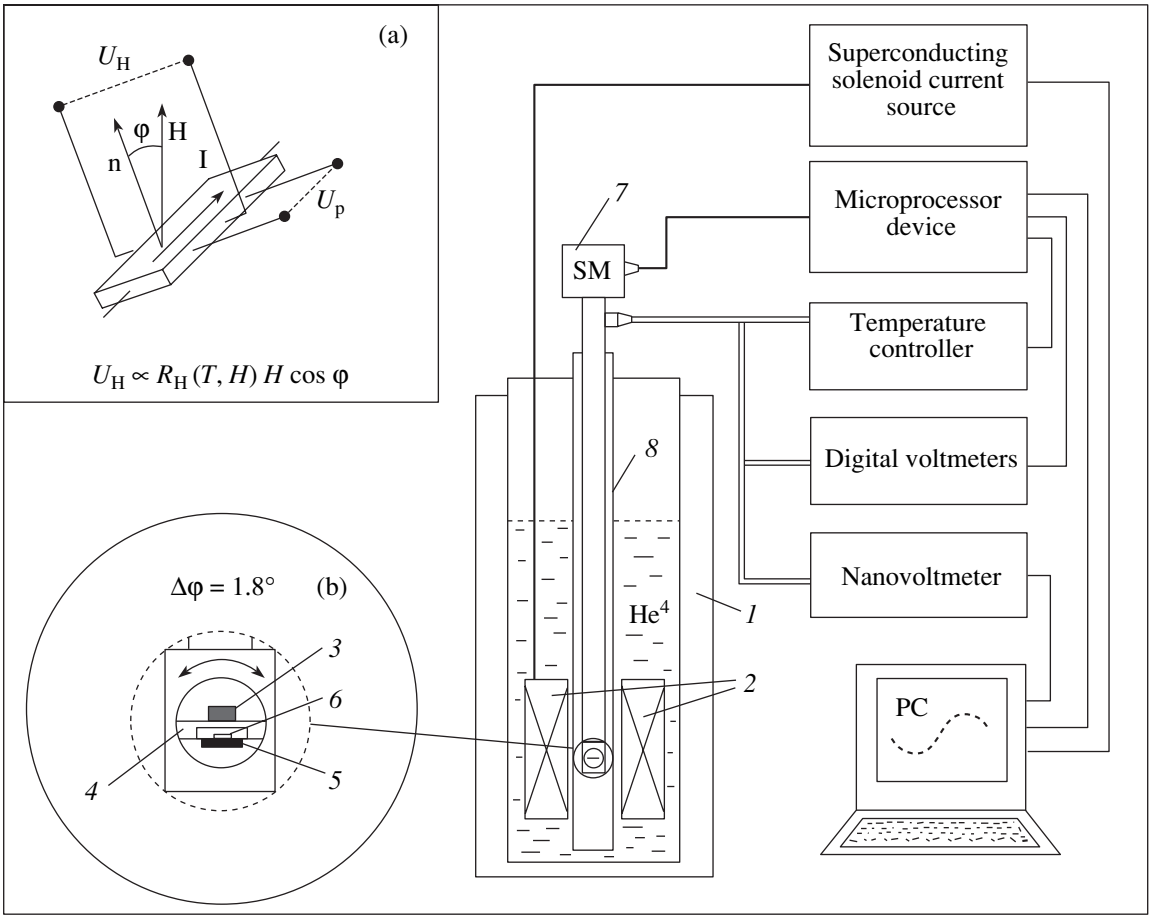


Fig. 1. Block diagram of setup for measuring transport characteristics: (*1*) cryostat, (*2*) superconducting magnet, (*3*) sample, (*4*) copper plate, (*5*) Hall probe, (*6*) resistance thermometer, (*7*) step motor, and (*8*) double-walled ampule. The geometry of contacts to the sample and the position of the sample on the plate are shown in insets a and b, respectively.

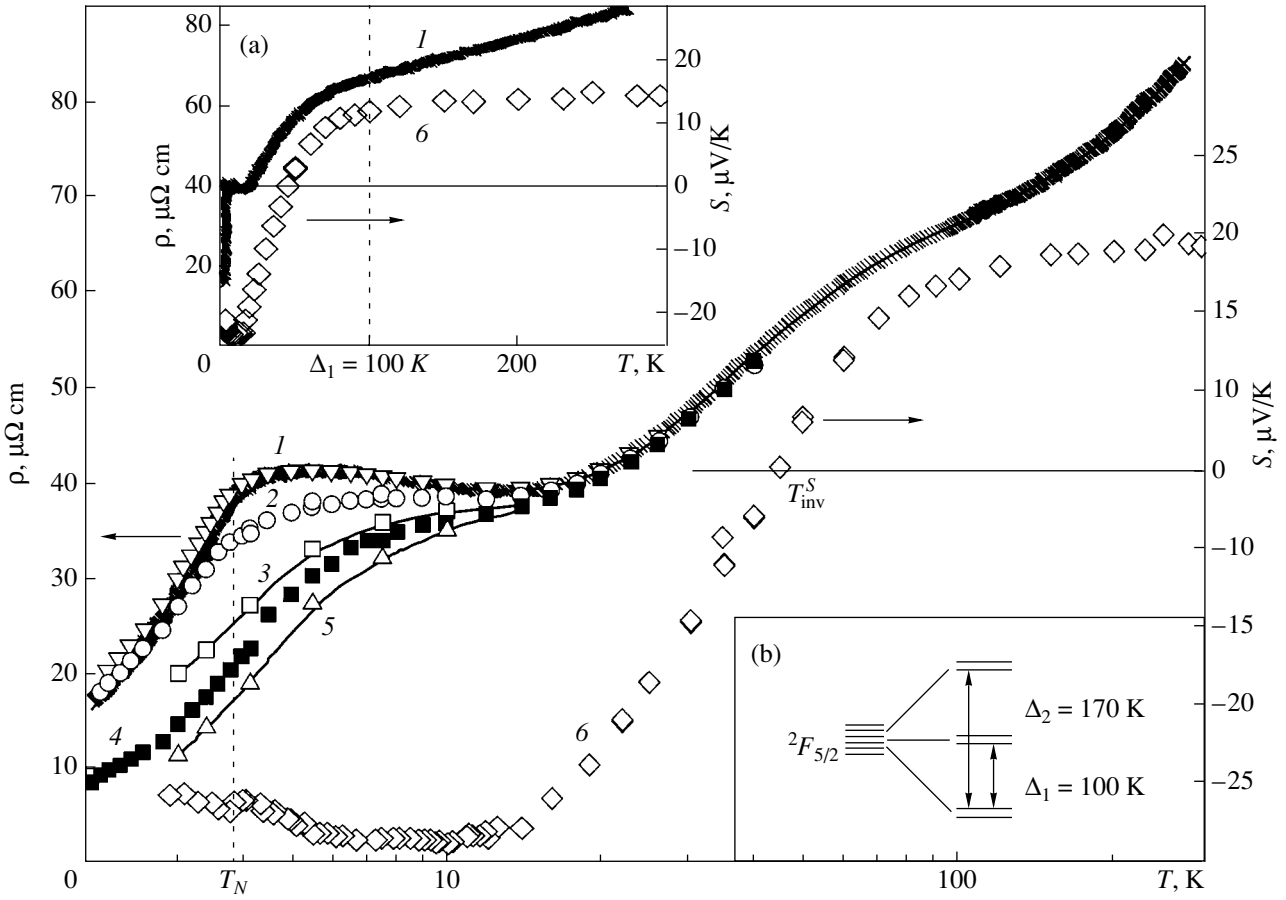


Fig. 2. Temperature dependences of the resistivity of $CeAl_2$ in various magnetic fields: H_0 = 0 (*1*), 32 (*2*), 50.5 (*3*), 60 (*4*), and 70.7 kOe (*5*). Curve *6* is the temperature dependence of the Seebeck coefficient. Curves *1* and *6* are shown in inset a on a linear temperature scale; a scheme of crystal-field splitting of the $^2F_{5/2}$ Ce ground state in $CeAl_2$ is shown in inset b.

Hall voltage from the sample. The temperature of the measuring cell with the sample, which was placed in double-walled ampule *8*, was stabilized and controlled by a temperature controller of an original design. The controller provided a 0.01 K accuracy of temperature stabilization. The system for recording and controlling low-temperature experiments was connected to a PC (Fig. 1) through a microprocessor device. The PC was used to accumulate and process experimental information and set the required parameters and operating conditions for the electronic components of the setup.

The resistance of $CeAl_2$ was measured by the four-probe direct-current method. In magnetic fields, the behavior of transverse (**I** ⊥ **H**, Fig. 1, inset a) magnetoresistance was studied.

3. RESULTS

The results of measurements of the resistivity $\rho(T, H_0)$ in a fixed magnetic field $H_0 \le 70$ kOe and of the Seebeck coefficient $S(T)$ of $CeAl_2$ are shown in Fig. 2 (curves *1–5* and *6*, respectively). The temperature dependence $\rho(T)$ in the absence of an external magnetic field (Fig. 2, curve *1*) is close to linear in the temperature range 100–300 K (Fig. 2, inset a). At $T < 100$ K, the resistance decreases more sharply as temperature lowers. After the $\rho(T)$ curve passes through a minimum near 13 K (Fig. 2, inset a), ρ increases almost logarithmically as temperature decreases and passes through a maximum at $T^{\rho}_{max} \approx 5.5$ K. Below this maximum, the $\rho(T)$ curve has a kink at $T = T_N \approx 3.8$ K (marked by a dashed line in Fig. 2), which corresponds to long-range magnetic ordering (antiferromagnetic modulated structure [14]) of the system of magnetic moments localized on cerium centers in the $CeAl_2$ matrix.

An external magnetic field of up to 70 kOe noticeably changes the form of the $\rho(T)$ curve. The appearance of substantial (up to 50%) negative magnetoresistance in the specified range of field values at $T < 20$ K (Fig. 2, curves *1–5*) is accompanied by a broadening of the low-temperature $\rho(T)$ maximum and its shift upward along the temperature axis. At the same time, the suppression of the low-temperature magnetic con-

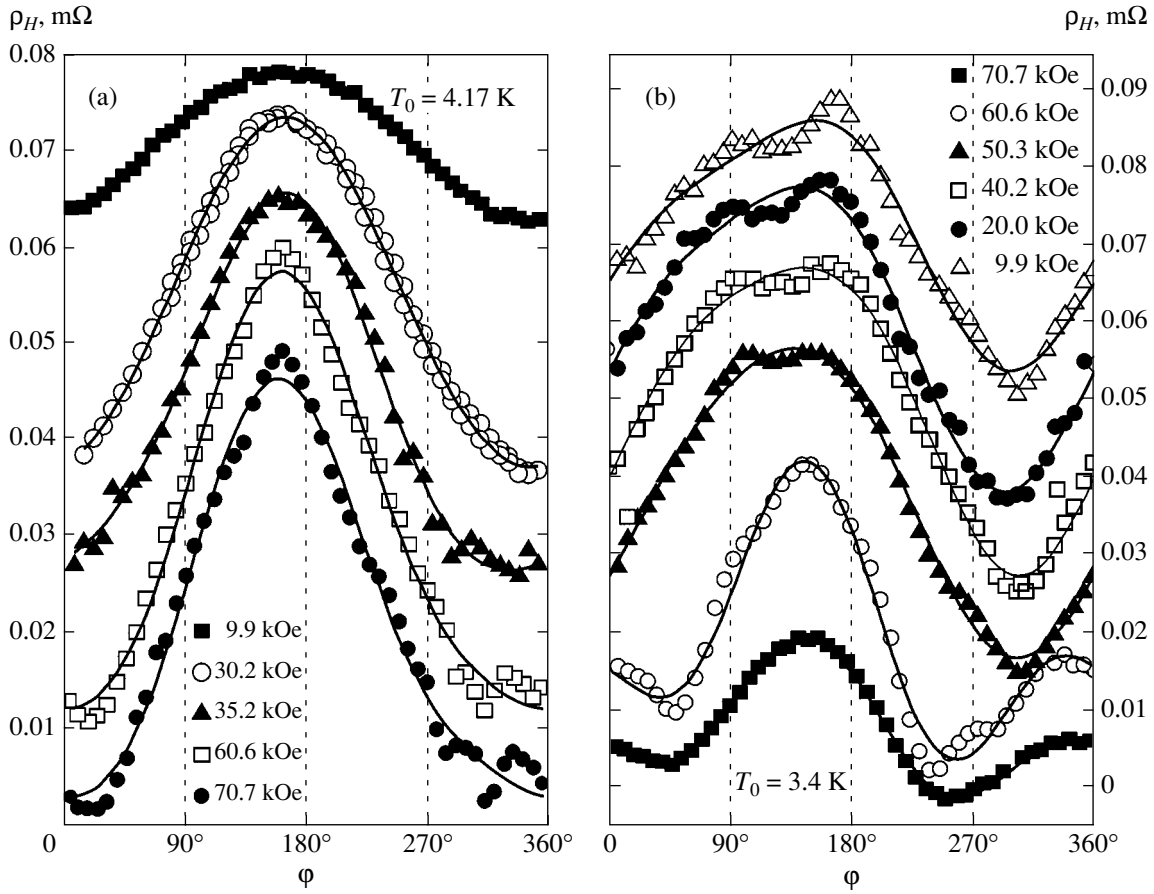


Fig. 3. Angular dependences of the Hall resistance of $CeAl_2$ in magnetic fields up to 70 kOe at temperatures T_0 = 4.17 (a) and 3.4 K (b).

tribution to $\rho(T)$ in magnetic fields $H_0 > 30$ kOe is much more effective at helium temperatures $T < 5$ K. This in turn transforms the low-temperature $\rho(T, H_0)$ maximum into a step (Fig. 2, curves *1–5*). Note that the experimental data presented in Fig. 2 (curves *1–5*) are in close agreement with the results of measurements performed in [15] for polycrystalline $CeAl_2$ samples at low and superlow temperatures in magnetic fields up to 200 kOe.

The temperature dependence of the Seebeck coefficient $S(T)$ of $CeAl_2$ measured in this work is shown in Fig. 2 (curve *6*). In the temperature range 100–300 K, the Seebeck coefficient is a slowly varying function of temperature, which takes on positive values; $S(T)$, however, sharply decreases as temperature lowers ($T <$ 100 K) and changes sign at $T = T_{inv} \approx 45$ K (Fig. 2, see also inset a). In addition to a negative $S(T)$ minimum of a considerable amplitude at $T_{min} \approx 10$ K, our precision Seebeck coefficient measurements revealed the presence of a singularity in the vicinity of the Néel temperature $T_N \approx 3.85$ K (the dashed line in Fig. 2); this singularity corresponded to the transition to the magnetically ordered state of localized cerium magnetic moments. Note that, as far as we know, this result is the first experimental observation of the $S(T)$ curve singularity that corresponds to the transition of $CeAl_2$ to the magnetically ordered phase. The small amplitude of $S(T)$ changes at $T \approx T_N$ ($\Delta S \leq 1$ μV/K at $|S| > 20$ μV/K) allows us to exclude the formation of a magnetically ordered state of $CeAl_2$ as a factor responsible for the deep negative Seebeck coefficient minimum.

As mentioned above, we measured the Hall resistance component $\rho_H(H, T)$ on the setup shown in Fig. 1 by rotating the sample through a certain angle and then by fixing its position in a magnetic field. Typical families of the angular dependences $\rho_H(\varphi, H_0, T_0)$ obtained for $CeAl_2$ in various magnetic fields $H \leq 70$ kOe both at temperatures above the Néel temperature (for instance, at $T_0 = 4.17$ K $> T_N \approx 3.8$ K) and in the magnetically ordered state ($T_0 = 3.4$ K $< T_N$) are shown in Figs. 3a and 3b, respectively. Measurements with sample rotations in a magnetic field, when the amplitude of the normal component of the external magnetic field vector $\mathbf{H}_\perp \parallel \mathbf{n}$ changes by the harmonic law $H_\perp = H_0\cos\varphi$, should usually be expected to give a sine Hall voltage dependence of the form $U_H \propto R_H(T, H_0)H_0\cos\varphi$ (Fig. 1,

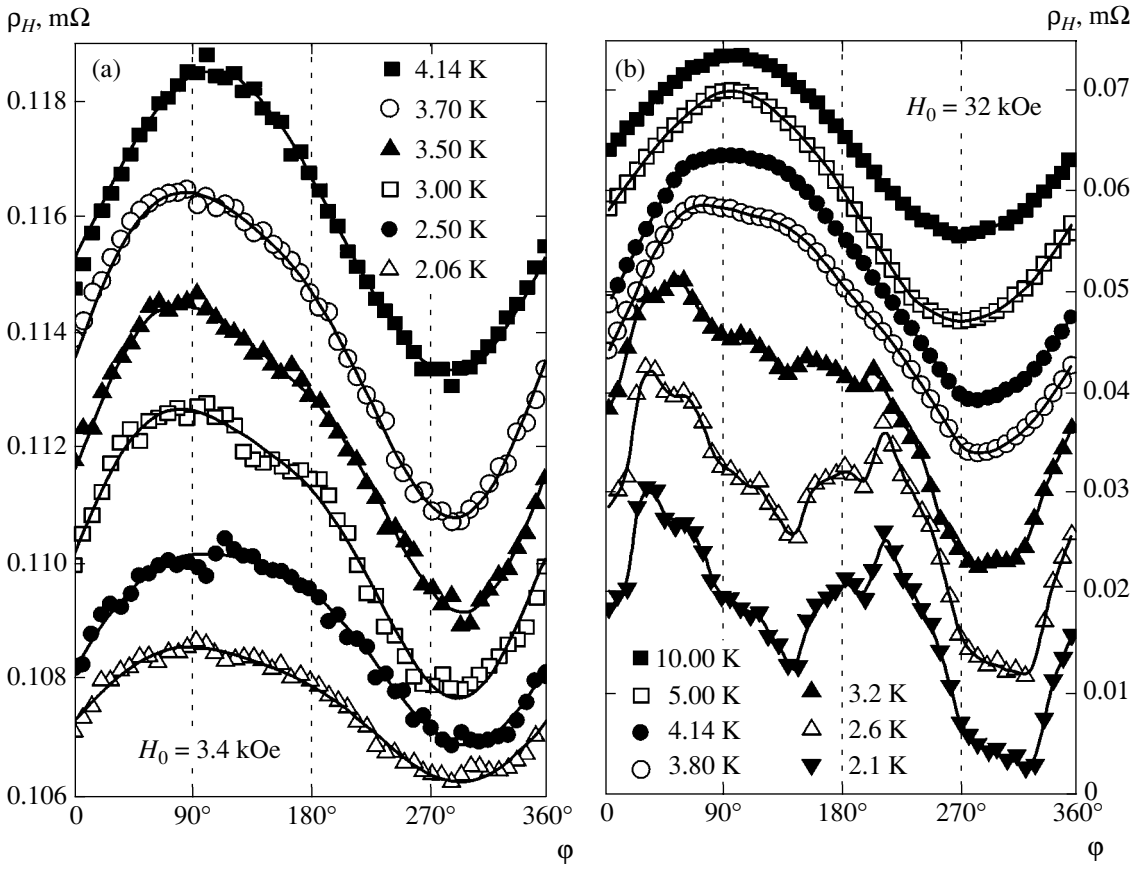


Fig. 4. Angular dependences of the Hall resistance of $CeAl_2$ at various temperatures in magnetic fields $H_0 = 3.4$ (a) and 32 kOe (b).

inset a). However, with $CeAl_2$ samples, such a form of $\rho_H(\varphi, H_0, T_0)$ curves is only observed in limited temperature and magnetic field ranges. In particular, at helium temperatures, an angular dependence of the Hall signal close to sinusoidal is only observed in fields up to 35 kOe (Fig. 3a). At temperatures below T_N, the shape of the $\rho_H(\varphi)$ curves becomes much more complex and a contribution of even harmonics is added to the main signal constituent $\rho_H(\varphi) \propto \cos\varphi$ over the whole range of magnetic fields used in this work (Fig. 3b). Changes in the form of the $\rho_H(\varphi)$ curves as temperature lowers from $T > T_N$ to $T < T_N$ are most visually shown in Figs. 4a and 4b, where the angular dependences of the Hall resistance measured at various temperatures in magnetic fields $H_0 \approx 3.4$ and 32 kOe, respectively, are plotted.

Another special feature of the $\rho_H(\varphi)$ dependences in $CeAl_2$ is the appearance of a contribution of even harmonics to the Hall effect in high magnetic fields $H \geq 40$ kOe. This contribution is observed both in the temperature region $4 \leq T \leq 7.5$ K (above T_N) and at $T < T_N$ for T and H values outside the region of the low-temperature antiferromagnetic modulated phase in the H–T magnetic phase diagram of $CeAl_2$. This anomalous Hall resistance behavior most clearly manifests itself in field $H_0 \approx 60$ kOe [Fig. 5a, all experimental curves were obtained at temperatures $T > T_N(H)$]. Another visual example of the appearance of the specified additional contribution of even harmonics is provided by the family of $\rho_H(T)$ curves recorded at temperature $T_0 \approx 5.5$ K > T_N (Fig. 6a, curves in the field range 40–80 kOe). At the same time, the angular dependences of the Hall resistance again become sinusoidal already at $T_0 \geq 8$ K (e.g., see the family of curves shown in Fig. 6b) over the whole filed range $H \leq 70$ kOe studied in this work.

To conclude this section, we stress that the anomalous component observed in this work and caused by the appearance of even harmonics in the Hall signal cannot be explained by asymmetry in the arrangement of Hall contacts on the sample of the intermetallic compound and, therefore, cannot be related to the addition of a contribution of $CeAl_2$ magnetoresistance, which is an even magnetic field function, to $U_H(T, H, \varphi)$. The angular dependences of magnetoresistance $\rho(\varphi, H, T)$ measured simultaneously with Hall effect dependences allow us to exclude such an influence of the ordinary resistive component, which arises because of "nonequivalence" in the arrangement of Hall contacts to the sample, as a factor determining the shape and character of variations of $U_H(T, H, \varphi)$ for all $CeAl_2$ samples studied in this work.

4. DISCUSSION

4.1. Separation of Contributions to the Hall Effect

We analyzed the results obtained for the angular dependence of the Hall resistance (Figs. 3–6) of the $CeAl_2$ intermetallic compound using the representation

$$\rho_H(\varphi, T_0, H_0) = \rho_{H0} + \rho_{H1}\sin(\varphi - \varphi_{01}) + \rho_{H2}\sin[2(\varphi - \varphi_{02})], \quad (1)$$

which includes not only the main component ρ_{H1} (odd with respect to the magnetic field) and the constant shift ρ_{H0}, but also the second harmonic contribution. The procedure for separating the contributions to (1) is most clearly shown in Fig. 5b, where the contributions with the amplitudes ρ_{H1} and ρ_{H2} found for the family of curves recorded at $T \approx 3.8$ K in various magnetic fields $H \le 70$ kOe are plotted along with the experimental ρ_H^{expt} curves. The accuracy of approximating the experimental data by (1) can be estimated from the differences

$$\rho_H^{\text{expt}}(\varphi, T_0, H_0) - \rho_{H0} - \rho_{H1}\sin(\varphi - \varphi_{01}) - \rho_{H2}\sin[2(\varphi - \varphi_{02})],$$

which are also plotted in Fig. 5b. Note that (1) is a fairly good approximation everywhere except the region of high magnetic fields $H > 50$ kOe, where a noticeable additional contribution of higher even harmonics appears in the $\rho_H^{\text{expt}}(\varphi)$ curves (Fig. 5b, curves *D*).

We applied the approach described above to determine the temperature and magnetic field dependences of the amplitudes of both the main contribution to the Hall effect $\rho_{H1}[\text{m}\Omega]d[\text{cm}] = \rho_H^a[\text{m}\Omega\ \text{cm}]$, where d is the sample thickness and ρ_H^a is the anomalous Hall resistance component (according to the classification suggested in [1–4], this is the anomalous contribution of skew scattering) and the anomalous magnetic component $\rho_{H2}d = \rho_H^{am}$. The results are plotted in Figs. 7 and 8, respectively. Figure 7 shows that the field depen-

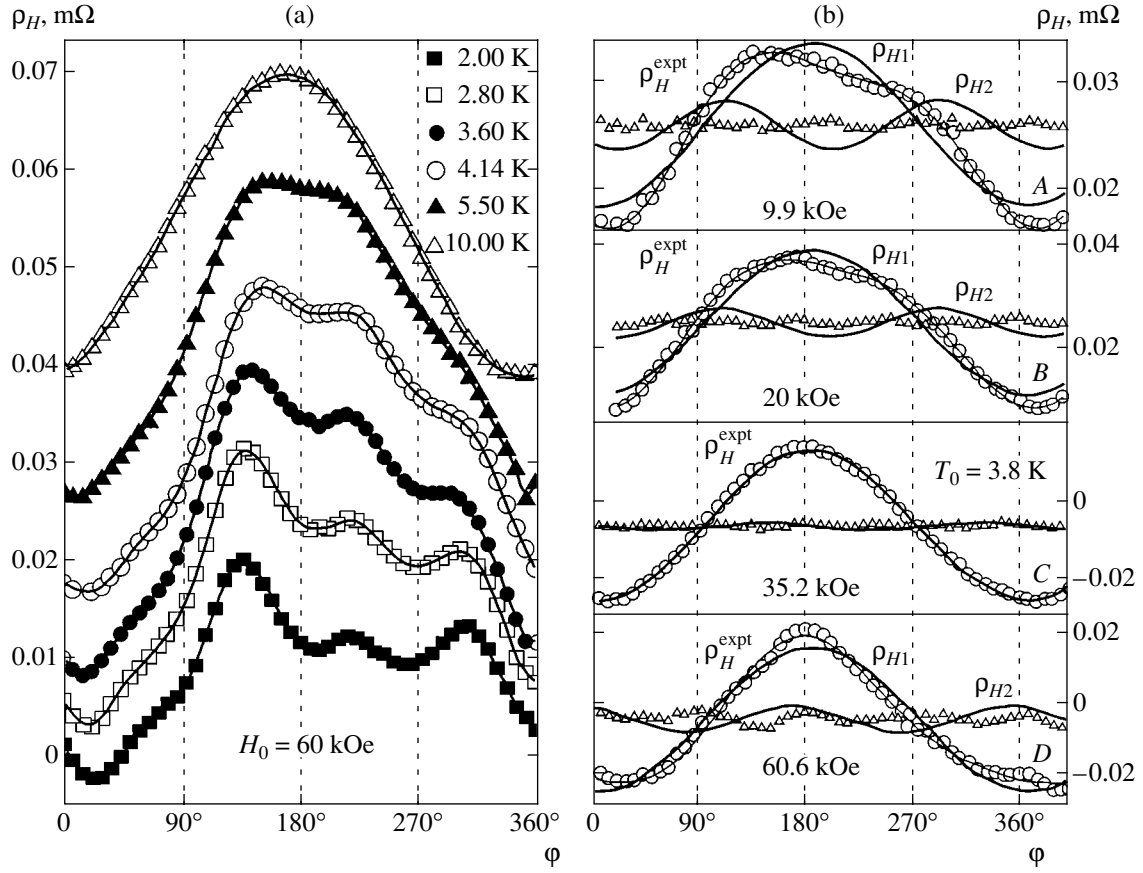


Fig. 5. (a) Angular dependences of the Hall resistance of $CeAl_2$ at various temperatures in magnetic field $H_0 = 60$ kOe and (b) separation of contributions to (1) (see text) at various H values and $T_0 = 3.8$ K; ρ_H^{expt} are experimental data, ρ_{H1} is the main component contribution, ρ_{H2} is the second harmonic contribution, and $\rho_H - \rho_{H1} - \rho_{H2}$ is the difference signal (triangles).

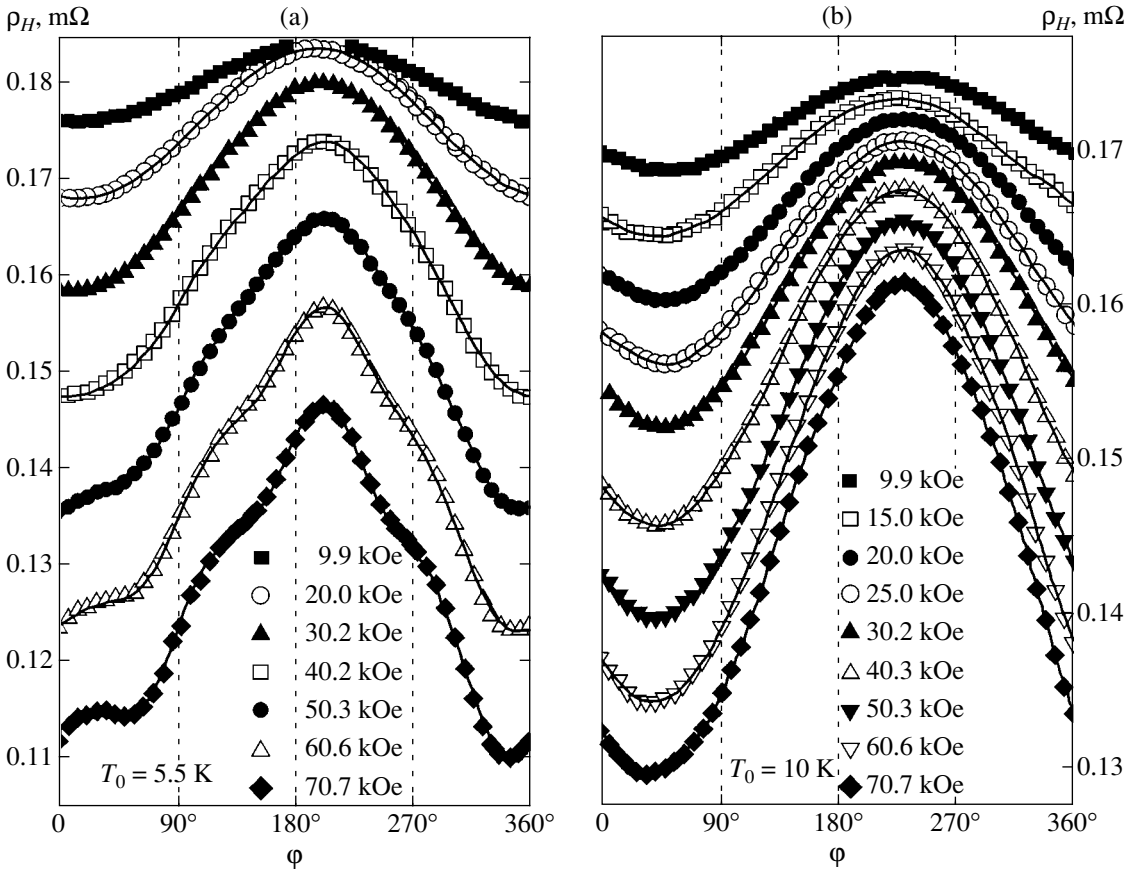


Fig. 6. Angular dependences of the Hall resistance of $CeAl_2$ in magnetic fields up to 70 kOe at temperatures $T = 5.5$ (a) and 10 K (b).

dences of the Hall resistivity $\rho_H^a(H, T_0)$ are essentially nonlinear not only for the antiferromagnetic modulated phase of $CeAl_2$ but also in the immediate vicinity of the Néel temperature (Fig. 7a). In magnetic fields $H \leq 70$ kOe, they remain nonlinear up to temperatures above 10 K, and, only at $T \geq 30$ K do the field dependences $\rho_H^a(H)$ become linear (Fig. 7b).

4.2. The Anomalous Magnetic Hall Effect Component

The temperature dependences of the anomalous magnetic contribution to the Hall resistivity $\rho_H^{am}(H, T_0)$ were studied at various fixed external magnetic field values; the results are shown in Fig. 8. At comparatively low fields $H \leq 30$ kOe, the appearance of the anomalous magnetic contribution can be unambiguously related to the transition to the antiferromagnetic modulated phase at temperatures below $T_N \approx 3.85$ K (Fig. 8). The appearance of the anomalous magnetic contribution ρ_H^{am} becomes noticeable also at $T > T_N$ as field H increases, and, in field $H_0 \approx 60$ kOe, the presence of a small ρ_H^{am} component is observed at temperatures up to $T \approx 8$–10 K (Fig. 8). Such a behavior of $\rho_H^{am}(T, H)$ is, on the whole, in agreement with the conclusions drawn in [16–18], where quasi-elastic neutron scattering by $CeAl_2$ was studied, and with the H–T magnetic phase diagram of this compound [19] (see also inset in Fig. 8). In particular, our data lend support to the conclusion [16] of the existence in $CeAl_2$ of a region of strong magnetic fluctuations with the coherence length $\xi \geq 20$ Å at temperatures up to 12 K. Note that, in the series of Laves phases $LnAl_2$ (Ln = Ce, Nd, Tb, Dy, etc.), all trivalent rare-earth metal dialuminides except $CeAl_2$ are ferromagnets. This leads us to suggest that the main factor responsible for the complex long-range antiferromagnetic ordering in the magnetic Kondo lattice of $CeAl_2$ is precisely the competition between magnetic RKKY exchange interaction and the mechanism of spin-flip scattering. The latter mechanism is responsible for the compensation (commonly attributed to the Kondo effect) of the localized rare-earth metal magnetic moment. This reduction of the magnetic moments of rare-earth metal ions determines the trend toward the formation of a nonmagnetic ground state and results in a substantial instability of the ferromagnetic structure. It appears that, in this situation, the presence of strong ferromagnetic fluctuations in the $CeAl_2$ matrix in high

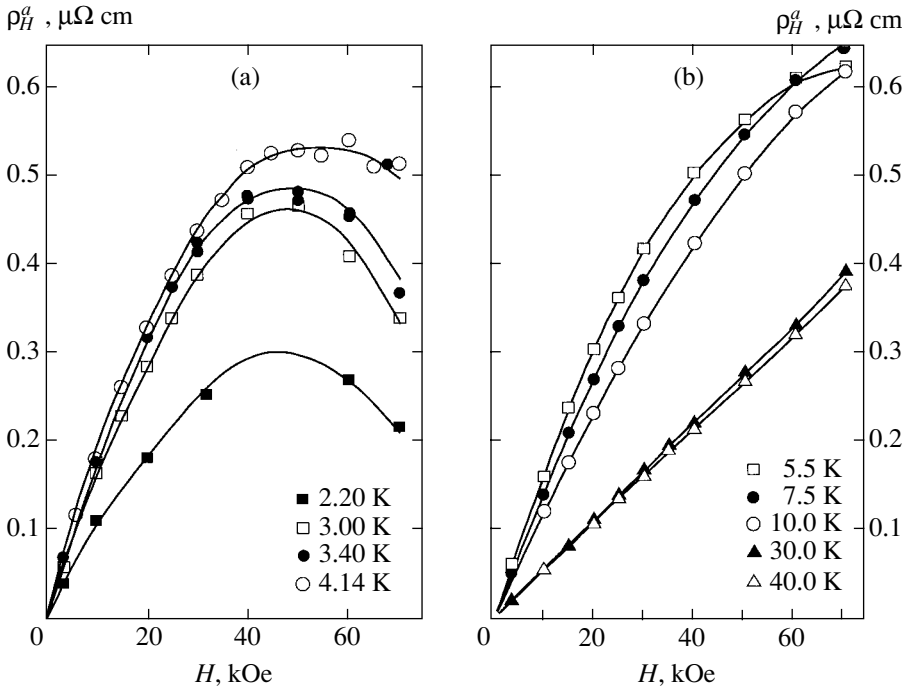


Fig. 7. Field dependences of the main anomalous component ρ_H^a of the Hall resistivity of $CeAl_2$ at various temperatures.

magnetic fields $H \geq 60$ kOe at both helium and intermediate (5–10 K) temperatures can result from the suppression of Kondo magnetic moment fluctuations on cerium centers by an external magnetic field $H \sim H_K \approx k_B T_K/\mu_B \approx 70$ kOe ($T_K^{CeAl_2} \approx 5$ K [16]). Another consequence of magnetic ordering in $CeAl_2$ at low temperatures under the conditions of the competition between various mechanisms (see above) appears to be the formation of new phases in the *H–T* magnetic diagram of $CeAl_2$. In particular, the phase transition from the antiferromagnetic modulated phase to a noncollinear magnetic structure was observed in [19] (Fig. 8, inset).

Yet another interesting feature of the anomalous magnetic component of the Hall resistivity is the nonmonotonic magnetic field dependence of the $\rho_H^{am}(H)$ amplitude at temperatures $T < T_N$ (Fig. 8). These results can more conveniently be discussed using the R_H^a and R_H^{am} Hall coefficients, which can be directly obtained from ρ_H^a and ρ_H^{am} and the magnetic field value. The field dependences of R_H^a and R_H^{am} found from the data shown in Figs. 3–6 are plotted in Fig. 9. According to Fig. 9a, the R_H^a and R_H^{am} Hall coefficients are comparable in order of magnitude at helium temperatures, and the anomalous magnetic contribution $R_H^{am}(H, T_0)$ is indeed essentially nonmonotonic. The $R_H^{am}(H)$ curves have a singularity (a maximum) in the vicinity of $H_{max} \approx 15$ kOe (Fig. 9a), which increases in amplitude as temperature decreases below liquid helium tempera-

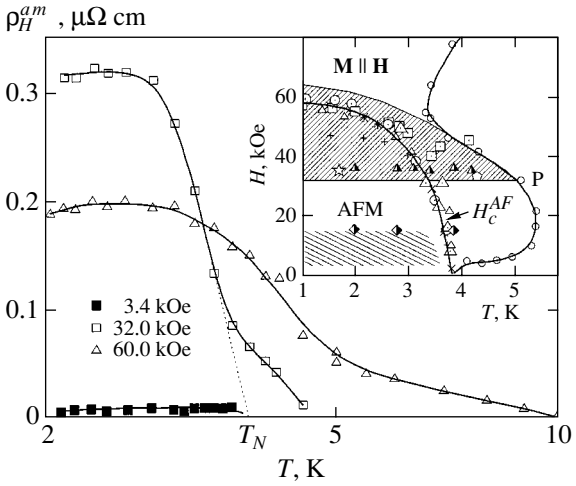


Fig. 8. Temperature dependences of the anomalous magnetic component ρ_H^{am} (see text) of the Hall resistivity in magnetic fields $H_0 = 3.4$, 32.0, and 60.0 kOe; the *H–T* phase diagram of $CeAl_2$ is shown in the inset: AFM is the antiferromagnetic modulated phase, P is the paramagnetic phase, and H_c^{AF} is the phase boundary of the AFM phase; the lower hatched region corresponds to the mode of the reorientation of antiferromagnetic domains by an external field (see text), and the upper hatched region corresponds to a noncollinear magnetic structure (according to [19]).

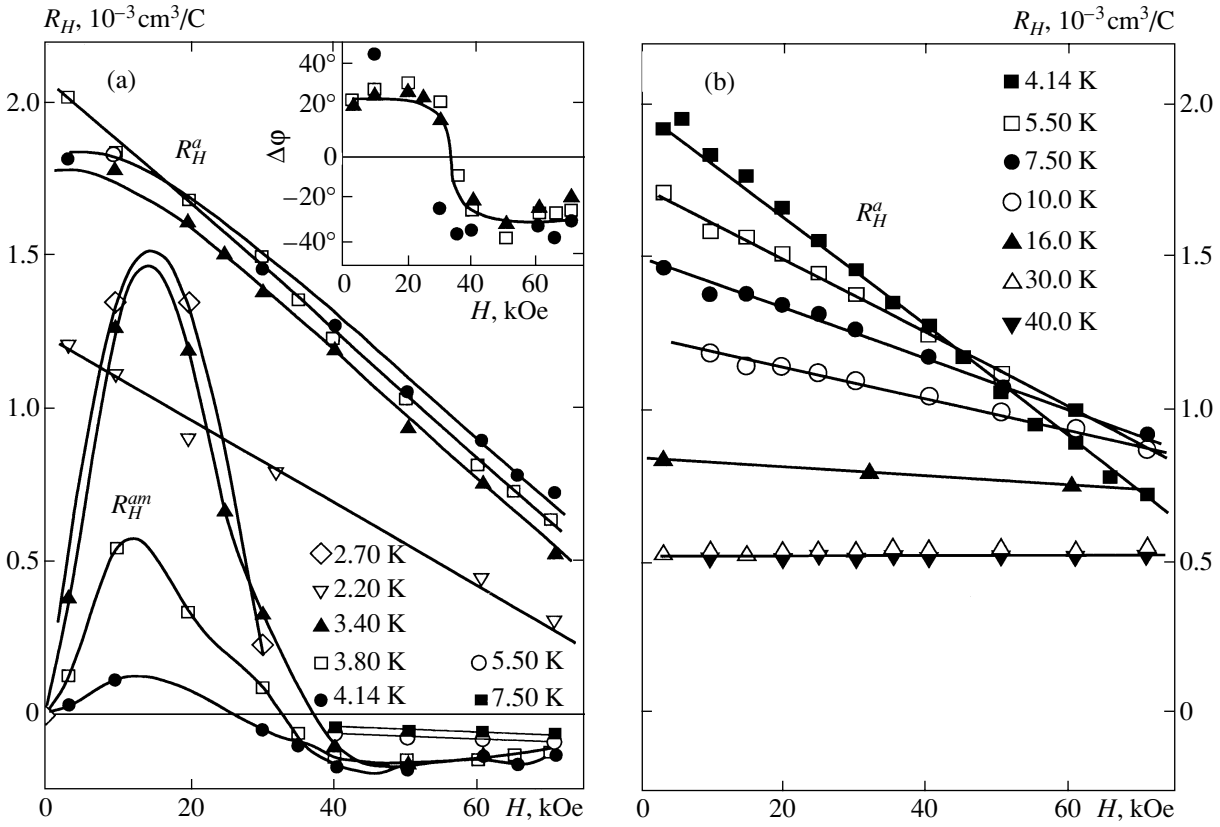


Fig. 9. Field dependences of the anomalous R_H^a and anomalous magnetic R_H^{am} components of the Hall coefficient of $CeAl_2$ at various temperatures. Field dependences of the phase shift between harmonics (see text) are shown in the inset.

ture and becomes equal to R_H^a at $T \le 3.4$ K. An analysis based on (1) also makes it possible to determine the phase shift $\Delta\varphi = \varphi_{01} - \varphi_{02}$ between the main and even Hall signal harmonics, which offers an additional possibility of quantitative characterization of the change in the anomalous magnetic contribution and related scattering of charge carriers by $CeAl_2$ magnetic structure features as the temperature and magnetic field vary. The inset in Fig. 9a shows that, to within the accuracy of measurements, the phase shift $\Delta\varphi$ takes on fixed values at $H \le 30$ kOe ($\Delta\varphi \approx 25°$) and $H \ge 40$ kOe ($\Delta\varphi \approx -35°$). A sharp change in the sign and value of $\Delta\varphi$ occurs in a fairly narrow neighborhood of $H^* \approx 35$ kOe, and this H^* value remains virtually unchanged both at $T < T_N \approx 3.85$ K (this corresponds to the antiferromagnetic modulated phase in the $CeAl_2$ matrix) and in the region from 4 to 8 K characterized by strong magnetic fluctuations in the $CeAl_2$ matrix (Fig. 9a). For a more visual and convenient representation, the change by 60° in the phase shift $\Delta\varphi = \varphi_{01} - \varphi_{02}$ of the second harmonic with respect to the main signal in the vicinity of $H^* \approx 35$ kOe is shown in Fig. 9a as a reversal of sign of the R_H^{am} component. In our view, the observed behavior of the $\Delta\varphi$ parameter is an additional argument in favor of the rearrangement of the magnetic structure of $CeAl_2$. It also lends support to the conclusion [19] on the existence of a "horizontal" (H_c = const) phase boundary in the H–T magnetic phase diagram of this compound (see inset in Fig. 8).

While discussing the nature of the $R_H^{am}(H, T_0)$ maxima in the vicinity of $H_{max} \approx 15$ kOe (Fig. 9a), we must mention the results obtained in [20–22] for the thermal expansion and magnetostriction of $CeAl_2$. According to [20, 21], the reorientation of antiferromagnetic domains occurs in the antiferromagnetic modulated phase of $CeAl_2$ as the magnetic field value increases to 15 kOe. The antiferromagnetic domains, initially oriented chaotically (uniformly), assume the orientation in which the antiferromagnetic polarization direction is transverse with respect to the external magnetic field. As $CeAl_2$ magnetization reversal at $H \le 20$ kOe virtually does not change the T_N temperature of the magnetic transition in this compound (the lower hatched phase diagram region in the inset in Fig. 8), the authors of [20, 21] arrived at the conclusion of the reorientation of antiferromagnetic domains by an external magnetic field without noticeable changes in their size and antiferromagnetic polarization. These findings lead us to conclude that the appearance of the anomalous magnetic scattering of charge carriers and, accordingly,

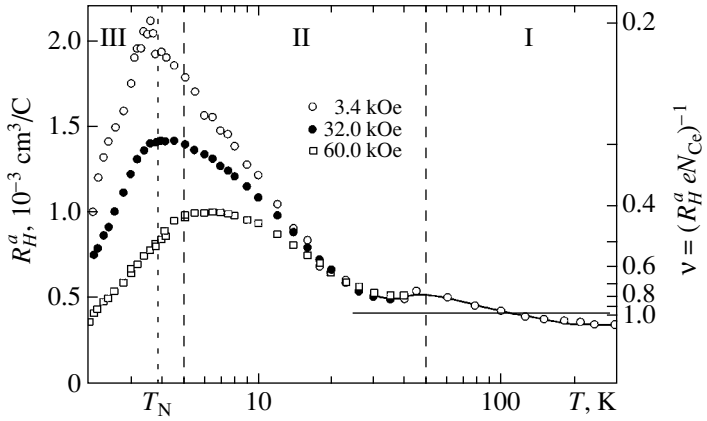


Fig. 10. Temperature dependences of the main anomalous component of the Hall coefficient of $CeAl_2$ in magnetic fields H_0 = 3.4, 32.0, and 60.0 kOe.

essentially nonmonotonic behavior of the anomalous magnetic component of the Hall coefficient R_H^{am} at $H <$ 30 kOe are likely caused by magnetic domain magnetization reversal in the magnetically ordered state of $CeAl_2$.

4.3. The Main Hall Effect Component in $CeAl_2$

The temperature dependences of the main (odd in magnetic field) anomalous component of the Hall coefficient $R_H^a(T, H_0)$ of $CeAl_2$ measured in this work in a wide temperature range from 1.8 to 300 K at several fixed magnetic field values are shown in Fig. 10. In agreement with the results obtained in [5–10] for other cerium-based intermetallic compounds with charge and spin fluctuations, a positive large-amplitude $R_H^a(T)$ maximum was observed in the vicinity of the characteristic temperature of spin fluctuations T_{sf} equal to $T_{sf} = T_K \approx 5$ K for the compound under study [16]. As the condition $\mu_B H \approx k_B T_{sf}$ is satisfied in fields $H \le 70$ kOe used in this work (see above), the amplitude of the $R_H^a(T)$ maximum essentially depends on the external magnetic field value. The magnetic field-induced suppression of spin fluctuations that are caused by spin-flip scattering of charge carriers results in nonlinear field dependences of the Hall resistivity $\rho_H^a(T)$ (see Fig. 7) and, as a consequence, a sharp decrease in the amplitude of the $R_H^a(T)$ maximum (Fig. 10). Note that the anomalous positive component $R_H^a(T)$ cannot be related to the magnetic (antiferromagnetic) ordering of the $CeAl_2$ matrix at temperatures $T < T_N \approx 3.85$ K. Indeed, the width of the Hall coefficient $R_H^a(T)$ maximum is fairly large compared with T_N ($\Delta T \sim 10$ K, Fig. 10), and the suppression of the R_H^a amplitude in magnetic fields occurs almost equally effectively at both $T_0 \ge T_N$ (e.g., see Fig. 9a, curves for $T_0 \approx 4.14$ and 3.8 K) and $T_0 < T_N$ (e.g., $T_0 \approx 3.4$ K in Fig. 9a). It follows that the nature of the appearance of the anomalous positive Hall effect in the vicinity of T_{sf} may be common to $CeAl_2$, which is a magnetic Kondo lattice according to the classification suggested in [23] ($T_{sf} = T_K$), and other cerium-based intermetallic compounds with strong spin (nonmagnetic Kondo lattices [23]) and charge (intermediate-valence compounds) fluctuations, in which nonmagnetic ground states are formed as temperature decreases because of manybody effects. Similarly, the resistivity maximum at $T_{\max}^{\rho} \approx 5.5$ K (Fig. 2, curve *1*) and the negative Seebeck coefficient minimum (Fig. 2, curve *6*) should in our view be treated as special features of low-temperature transport in compounds with electron density fluctuations. In the $CeAl_2$ matrix, these fluctuations are caused by the formation of manybody states. The problem of a consistent interpretation of $\rho(T)$, $R_H^a(T)$, and $S(T)$ singularities will be considered in the next section in more detail.

The suppression of the $R_H^a(T)$ maximum in magnetic fields is accompanied by its noticeable shift upward along the temperature axis (see Fig. 10). For instance, the $R_H^a(T_{\max}^{R_H})$ Hall coefficient decreases more than twofold in magnitude in magnetic field $H \approx 60$ kOe (Fig. 10), whereas the $T_{\max}^{R_H}$ value increases to $T_{\max}^{R_H}(60\text{ kOe}) \approx 6.5$ K. Such a noticeable shift of the low-temperature singularity of $R_H^a(T)$ in high magnetic fields upward along the temperature axis correlates well with the behavior of the low-temperature resistivity maximum in magnetic fields (see Fig. 2 and [15]). As a result, while the absolute $R_H^a(4.2\text{ K})$ and $\rho(4.2\text{ K})$ values decrease substantially and in parallel to each other

(both parameters decrease approximately threefold in magnetic field $H \approx 70$ kOe), their ratio $\mu_H = R_H^a/\rho$ is a slowly varying magnetic field function (Fig. 11).

We must stress that this noticeable shift of the low-temperature singularities of $R_H^a(T)$ and $\rho(T)$ upward along the temperature axis, which depends on the external magnetic field strength, cannot be given a simple explanation within the traditional Kondo lattice model. Indeed, in this approach, the specified anomalies of the galvanomagnetic properties of the system with strong electron correlation appear to be related to the formation of a manybody resonance in the density of electronic states in the neighborhood of Fermi energy E_F having a width about k_BT_K. It is expected in this situation that the resonance spin-flip scattering of conduction electrons by localized cerium magnetic moments, which determines both the renormalization of the density of states mentioned above and the appearance of the anomalies of $R_H^a(T)$ and $\rho(T)$, should be almost fully suppressed by magnetic field $H \sim k_BT_K/\mu_B$, and that the $R_H^a(T_{\max}^{R_H})$ and $\rho(T_{\max}^{\rho})$ parameters should decrease without noticeable changes in the positions of the low-temperature galvanomagnetic singularities.

Another special feature of the behavior of the anomalous Hall coefficient in $CeAl_2$, which also does not fit in with the traditionally used approach, is the complex activation dependence of $R_H^a(T)$ in this intermetallic compound; this dependence was for the first time studied in [12]. The temperature dependence of the anomalous Hall coefficient is plotted in the coordinates $\log(R_H^a - R_H^{LaAl_2}) = f(1/T)$ in Fig. 12. In these coordinates, we easily see three characteristic temperature intervals of the $R_H^a(T)$ dependence and, accordingly, three asymptotic behaviors. In the temperature intervals 50–300 K (I) and 10–40 K (II), an anomalous activation growth is observed as the temperature decreases,

$$R_H^a(T) \propto \exp(E_{a1,2}/k_BT) \qquad (2)$$

(see Fig. 12); the activation energies in these intervals are $E_{a1}/k_B \approx 12.0 \pm 0.5$ K and $E_{a2}/k_B \approx 7.6 \pm 0.2$ K, respectively. Note that, as distinct from the results reported in [12], these E_{a1} and E_{a2} values were obtained with the Hall coefficient of the $LaAl_2$ nonmagnetic analog of cerium dialuminate used as the R_H^a anomalous component of the $CeAl_2$ compound; for $LaAl_2$, $R_H^{LaAl_2} \approx -6 \times 10^{-4}$ cm^3/C [24]. In temperature interval III ($T \leq 5$ K), the behavior of $R_H^a(T)$ below its maximum is fairly well described by the dependence

$$R_H^a(T) \propto \exp(-E_{a3}/k_BT). \qquad (3)$$

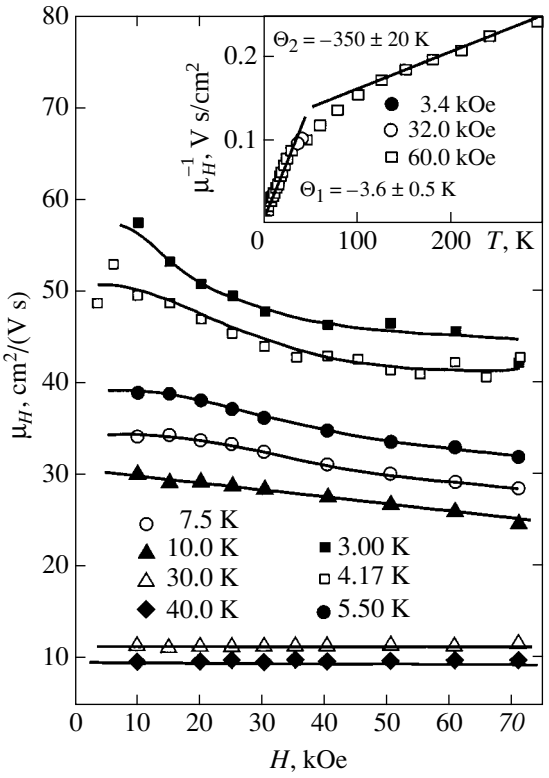


Fig. 11. Field dependences of the $\mu_H = R_H^a/\rho$ parameter for $CeAl_2$ at various temperatures. Temperature dependences $\mu_H^{-1}(T)$ in magnetic fields $H_0 = 3.4$, 32.0, and 60.0 kOe are shown in the inset.

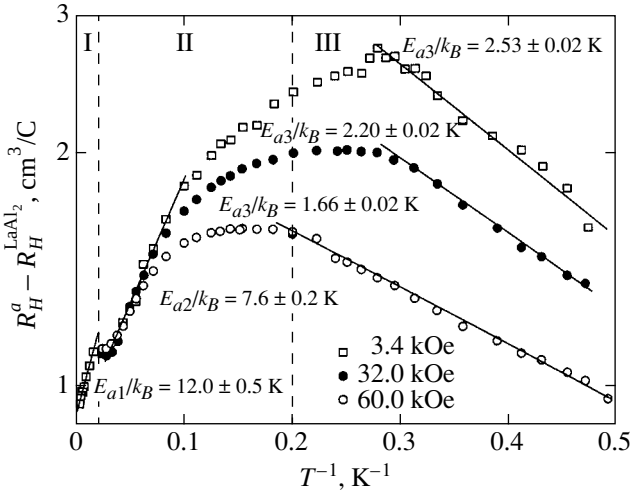


Fig. 12. Temperature dependences of the Hall coefficient $R_H^a - R_H^{LaAl_2}$ (see text) of $CeAl_2$ in reciprocal logarithmical coordinates at various magnetic field values.

The E_{a3}/k_B value, which lies in the interval 1.5–2.6 K, depends on the external magnetic field (see Fig. 12).

As has been mentioned above, such a behavior of the Hall coefficient $R_H^a(T)$, very unusual for metallic

systems, not only does not fit in with the Kondo lattice model, but also cannot be given a simple explanation in terms of the model of skew-scattering of charge carriers [1–4]. Indeed, both approaches under consideration are based on the assumption of a prevailing role played by resonance spin-flip scattering of conduction electrons by localized magnetic moments of rare-earth metal ions. From the point of view of the authors of [1–4], both the anomalous positive Hall effect in compounds with heavy fermions (including cerium-based compounds) and resistivity anomalies are consequences exclusively of the special features of scattering effects. In particular, the contribution of skew-scattering at temperatures above the $R_H^a(T)$ maximum (that is, at $T > T_K$) is estimated in [1–4] by the approximate equation

$$R_H^a(T) \propto \rho(T)\chi(T), \tag{4}$$

where $\chi(T)$ is the reduced magnetic susceptibility of the system. On the whole, applying (4) to analyze the data on the $\mu_H = R_H^a(T)/\rho(T) \propto \chi(T)$ parameter (Figs. 2, 10), which characterizes the scattering of charge carriers, leads to qualitative agreement between the experimental results obtained in this work (see inset in Fig. 11) and the conclusions drawn in [1–4]. The Curie–Weiss behavior of $\mu_H^{-1}(T) \propto (T - \Theta_{1,2}) \propto \chi^{-1}(T)$ is characterized by the paramagnetic Curie temperatures $\Theta_1 = -350 \pm 20$ K and $\Theta_2 = -3.6 \pm 0.5$ K. However, as has been mentioned above, the whole set of manybody effects in the low-temperature transport of charge carriers, namely, the activation behavior of the Hall coefficient, shifts of $R_H^a(T)$ and $\rho(T)$ singularities in magnetic fields, etc., fall outside the scope of this approach. In addition, the influence of crystal-field splitting of the $^2F_{5/2}$ cerium state (see Fig. 2, inset b) on the behavior of the Hall mobility $\mu_H = R_H^a(T)/\rho(T)$ was not consistently taken into account in [1–4], which impedes a quantitative analysis of these results.

The use of the approach based on the formation of spin-polaron states in Hubbard bands to interpret the anomalies of the low-temperature transport in the compound with heavy fermions that we are studying is in our view much more preferable (e.g., see monograph [25] and [26, 27]). Such states arise as a result of fast spin fluctuations in the immediate vicinity of localized cerium magnetic moments in the $CeAl_2$ matrix. This approach provides a natural explanation of the activation behavior of the Hall coefficient in $CeAl_2$ (see Fig. 12). In our view, the activation energies $E_{a1,2}$ should be put in correspondence to the characteristics of spin–polaron complexes formed in the vicinity of Ce sites. The transition from $T \geq \Delta_{1,2}$ (interval I in Fig. 12) to $T < \Delta_1 \approx 100$ K determines the change in the mode of fast spin $4f$–$5d$ fluctuations (here, $\Delta_{1,2}$ are the crystal-field splitting parameters of the $^2F_{5/2}$ cerium state, $\Delta_1 = 100$ K and $\Delta_2 = 170$ K [16–18, 28–30], see Fig. 2, inset b). As a consequence, the parameters of spin-polaron states in $CeAl_2$ change. The localization radii $a_{p1,2}^*$ and effective masses $m_{1,2}^*$ of manybody states with the binding energies $E_{a1}/k_B \approx 12$ K (Fig. 12, interval I) and $E_{a2}/k_B \approx 7.6$ K (Fig. 12, interval II) can be estimated by the equations

$$m_{1,2}^* = e\tau_{\text{eff}}/\mu_H, \tag{5}$$

$$a_{p1,2}^* = \hbar/\sqrt{2E_{a1,2}m_{1,2}^*}. \tag{6}$$

The relaxation time τ_{eff} at various temperatures between 4 and 300 K can be determined from the half-width $\Gamma/2$ of the quasi-elastic peak in the neutron scattering spectra of $CeAl_2$ (e.g., see [16]). The equation

$$\Gamma/2 = \hbar/\tau_{\text{eff}}, \tag{7}$$

gives 1.3×10^{-12} and 4.1×10^{-13} s for τ_{eff} at 5 and 60 K, respectively. Using the experimental data on $\mu_H(T)$ presented in Figs. 2, 10, and 11 and Eqs. (5) and (6), we obtain the following effective masses of heavy charge carriers and the corresponding localization radii of spin-polaron states:

$$m_1^*(60\text{ K}) \approx 90m_0, \quad a_{p1}^* = 6.4\text{ Å},$$

$$m_2^*(5\text{ K}) \approx 57m_0, \quad a_{p2}^* = 10\text{ Å}$$

(m_0 is the mass of the electron). Note that these estimates of $m_{1,2}^*$ and $a_{p1,2}^*$ are comparable in order of magnitude with the $m_e^* \approx 30m_0$ and $a_{ep}^* \approx 6$ Å parameters of manybody exciton-polaron states found in [31] for the classic compound with fast charge and spin fluctuations, samarium hexaboride SmB_6.

The magnetic field-induced shift of the anomalies of the galvanomagnetic characteristics upward along the temperature axis observed in this work for $CeAl_2$ (see Figs. 2, 10) also finds a natural explanation within the framework of the suggested approach. The formation of spin polarons as a result of fast spin fluctuations on Ce sites is accompanied by the appearance of exchange field H_{ex}, which is to a great extent responsible for the anomalous increase in the Hall coefficient and the appearance of $R_H^a(T)$ singularities in the vicinity of $k_BT_{sf} \approx \mu_BH_{ex}$. In addition to the suppression of fast spin fluctuations on Ce sites, the summation of the components $H + H_{ex}$ substantially shifts the anomalies of the transport characteristics upward along the temperature axis as the external magnetic field increases. Note that the $H_{ex} \approx 75$ kOe value in $CeAl_2$ at low temperatures was estimated in [14, 32, 33] by analyzing the polarized

neutron diffraction spectra. Close H_{ex} values were also obtained in [22] from $CeAl_2$ magnetostriction measurements at helium temperatures, $H_{ex} \approx 79 \pm 2$ kOe. An additional "estimate from below" of the exchange field $\mu_B H_{ex}/k_B$ is provided by the $\Theta_2 \approx -3.6 \pm 0.5$ K value found in this work (Fig. 11, inset), which reproduces well the paramagnetic Curie temperature $\Theta \approx -3.9$ K obtained in [34] from $CeAl_2$ magnetic susceptibility measurements.

Another no less important argument in favor of the spin-polaron approach that we propose for interpreting the low-temperature properties of $CeAl_2$ is, in our view, the Schottky anomaly of the low-temperature heat capacity at about $T \approx 6$ K observed in [35], where measurements were performed in magnetic field $H \approx 50$ kOe. Recall that, according to the result obtained in this work (see Fig. 10), the Zeeman splitting of the ground state of the system in the effective field $H_{\text{eff}} = H + H_{ex}$ (external magnetic field $H \approx 60$ kOe) causes the appearance of an $R_H^a(T)$ maximum at about $T_{\max}^{R_H}(60\text{ kOe}) \approx 6.5$ K. The calculations of two-level system parameters in the resonance level model performed in [35] using the results of low-temperature $CeAl_2$ heat capacity measurements then acquire special significance. In [35], the activation energy $E_a/k_B \approx 9.6$ K was found from the $\Gamma/2 \approx \hbar/\tau_{\text{eff}} = k_B T_K \approx 0.5$ meV value as an estimate of the width of two-level system levels. To within calculation errors [35] and taking into account the contribution of the external magnetic field $H = 50$ kOe to the Zeeman splitting of the Γ_7 doublet, this value fairly closely agrees with that found in the present work, $E_a/k_B \approx 7.6 \pm 0.2$ K.

The suggested approach predicts that temperature lowering in the region $T < 50$ K (Fig. 10, region II) will not cause only an increase in the manybody resonance amplitude in the vicinity of the Fermi energy E_F but also an essential rearrangement of the magnetic structure of spin polarons. By analogy with the result obtained for the FeSi compound with strong electron correlation [36, 37], the transition to coherent spin fluctuations in the vicinity of Ce sites should be accompanied by the formation of ferromagnetic microregions (ferrons) from spin polarons in the $CeAl_2$ matrix. According to the study performed in [36, 37] for iron monosilicide, an additional special feature of such a "phase transition" in a system of nanosized magnetic regions is the retention of virtually unchanged activation characteristics (band structure parameters) both for spin polarons and for ferromagnetic nanoclusters formed from them. It appears that the situation with $CeAl_2$ is similar. At $T < 20$ K, a strong dispersion of elastic constants and the related substantial anomaly in the absorption of ultrasound was observed in this compound [38]. Also note the result obtained in [39, 40] in studying NMR spin-lattice relaxation in the antiferromagnetic modulated phase of $CeAl_2$. In these works, an "energy gap in the excitation spectrum of magnons" with $E_g = 0.87 \pm 0.08$ meV [39] ($E_g = 11 \pm 3$ K [40]) was observed. In addition, the inelastic neutron scattering spectra of the magnetically ordered $CeAl_2$ phase contained two absorption singularities, or "magnon peaks," at $E_{a1} = 1.2 \pm 0.8$ meV and $E_{a2} = 0.7 \pm 0.4$ meV [17]. To within the error of measurements performed in [17], these values correspond to the binding energies of spin polarons $E_{a1,2}$ found in this work (see Fig. 12). At the same time, it is necessary to stress important differences between the formation of manybody states in the narrow-band FeSi semiconductor with a fairly low concentration of spin polarons at low temperatures $T <$ 40 K (10^{17}–10^{18} cm^{-3} [36, 37]) and in the $CeAl_2$ intermetallic compound with a fairly broad conduction band. In addition to substantial screening effects, the strong magnetic interaction of ferromagnetic microregions through RKKY electron density oscillations (indirect exchange) arises in $CeAl_2$. The related special features of magnetic ordering in $CeAl_2$ were discussed in [19] comparatively recently. It appears that so complex a magnetic structure in $CeAl_2$ at medium- and long-range magnetic order distances (antiferromagnetic ordering in the system of ferrons of submicron dimensions) is the main reason why the identification of magnetically ordered phases in this compound encounters serious difficulties (e.g., see [41–47]).

Let us return to the experimental results shown in Figs. 10 and 12. Note that the $R_H^a(T) \propto \exp(-E_{a3}/k_B T)$ dependence observed in this work for the anomalous contribution to the Hall coefficient of $CeAl_2$ at $T < 10$ K is similar to that predicted in [48] on the basis of calculations of the behavior of the Hall coefficient in a system with topologically nontrivial spin configurations (Berry phases). According to [48, 49], the Hall effect is modified in this situation because of the appearance of the internal magnetic field $H_{\text{int}} = \langle h_z \rangle \propto (1/k_B T)\exp(-E_a/k_B T)$, which adds to the external field H.

To conclude this section, let us fairly roughly estimate the localization radius of manybody states from the results presented in Fig. 10. The R_H^a parameter will be put in correspondence to the effective reduced concentration of carriers per cerium atom $\nu = (R_H e N_{\text{Ce}})^{-1}$ (the right-hand axis in Fig. 10). In this situation, an increase in ν in the interval $0 < \nu \leq 1$ can be treated as an increase in the effective volume per carrier; we believe this increase to be caused by manybody effects in $CeAl_2$. The characteristic Ce–Ce distance in the crystal structure of the $CeAl_2$ Laves phase is $a_{\text{Ce–Ce}} \approx 3.5$ Å [50]. Using this value, we obtain the crude estimate $a_p^* = 6$–16 Å for $\nu = N/N_{\text{Ce}} = 0.2$–0.3 in the vicinity of the R_H^a maximum.

4.4. The Separation of Contributions to the ρ(T), S(T), and $R_H^a(T)$ Transport Coefficients

As has been mentioned in Section 3, the resistivity ρ(*T*) (Fig. 2, curve *1*), Seebeck coefficient *S(T)* (Fig. 2, curve *6*), and Hall coefficient $R_H^a(T)$ (Figs. 10, 12, H = 3.4 kOe) dependences contain several anomalies in the temperature range 1.8–300 K. These anomalies are evidence of concerted changes in the specified $CeAl_2$ parameters in the temperature ranges 50–300 K (I), 5–50 K (II), and $T < 5$ K (III). The distinguishing feature of the transport of charge carriers in region I is a determining role played by the contribution of inelastic scattering related to the transitions between the ground (Γ_7) and two excited doublets (Fig. 2, inset b) of the $^2F_{5/2}$ cerium state. The transport of carriers in mode I at 100–300 K is characterized by a linear ρ(*T*) dependence combined with slow variations in the positive (12–15 μV/K) Seebeck coefficient and an almost activation behavior of the Hall coefficient with $E_{a1}/k_B \approx$ 12 K. The transition from mode I to II (see Figs. 2, 10, 12, 13) is accompanied by sharp changes in the measured ρ, *S*, and R_H^a values. In addition to substantial deviations of the ρ(*T*) dependence from linearity, *S*(*T*) sign reversal (Fig. 2, curve *6*) and a kink in the $R_H^a(T)$ dependence (Fig. 12) are observed at $T \approx 50$ K. The low-temperature features of the behavior of ρ, *S*, and R_H^a in the transition region 4–12 K between intervals II and III (Figs. 2, 10, 12, 13) also provide clear evidence of a change in the asymptotic behaviors of the specified $CeAl_2$ characteristics.

Currently, neither reliable and consistent explanation can be suggested for the transport characteristics of cerium-based compounds with heavy fermions, including the nature of various contributions to the conductivity, Seebeck coefficient, and Hall coefficient nor can these contributions be identified. It is therefore of interest to perform a comparative analysis based on the results of measurements performed in this work for high-quality polycrystalline $CeAl_2$ samples. Among the few investigations in which resistance and thermoelectric data on cerium intermetallic compounds were considered jointly, we must mention work [51], where the Nordheim equation

$$\rho S = \rho_0 S_0 + \rho_{mag} S_{mag}$$

was used to analyze the impurity and magnetic contributions to ρ(*T*) and *S*(*T*) of $CeNi_2Sn_2$. Such a simplified representation of the sum of the low-temperature transport components for a tetragonal compound with $\Delta \approx$ 20 K and $T_{sf} = T_K \approx 1.6$ K is at least, insufficiently accurate. In addition, according to [52], both the Gorter–Nordheim equation and the Matissen rule $\rho = \sum_i \rho_i$ are not good approximations for compounds with heavy fermions, which should be treated taking into account a qualitative rearrangement of the density of electronic states in the neighborhood of E_F.

For this reason, it is more correct to use the standard equations for σ, *S*, and R_H^a in the form [53]

$$\sigma = \sum_{i=1}^{3} \sigma_i, \tag{8}$$

$$S\sigma = \sum_{i=1}^{3} \sigma_i S_i, \tag{9}$$

$$R_H^a \sum_{i=1}^{3} \sigma_i^2 = \sum_{i=1}^{3} \sigma_i^2 R_{Hi}^a. \tag{10}$$

We could not analyze the experimental data of this work using the approach suggested in [52] and based on summing the nonmagnetic S_0, positive Kondo $S_d^{(1)}(T)$, and negative resonance $S_d^{(2)}(T)$ terms (in [52], the inversion temperature of the Seebeck coefficient was found to be $T_{inv}^S < 0.6T_K \approx 3$ K, which is obviously at variance with the value for $CeAl_2$, $T_{inv}^S \approx 46$ K). For this reason, we applied a phenomenological approach to separate the contributions to σ, *S*, and R_H^a in this compound.

Also note that, as far as we know, no self-consistent analysis of contributions to the σ, *S*, and R_H^a transport coefficients for rare-earth metal-based compounds with heavy fermions has been performed as yet.

In region I, of the greatest importance is inelastic scattering of carriers accompanied by transitions between cerium $^2F_{5/2}$ state doublets spaced $\Delta_1 \approx 100$ K and $\Delta_2 \approx 170$ K apart from the ground-state doublet Γ_7. For this reason, the contributions σ_1, S_1, and R_{H1}^a (see Fig. 13) were approximated by the analytic equations

$$\begin{gathered}
\sigma_1 = \sigma_0(T)\exp(-\Delta_1/k_BT), \\
\sigma_0(T) = 1.03/T^{\alpha}, \quad \alpha = 0.73, \\
S_1 = S_0^{(I)} + BT, \quad S_0^{(I)} = 11\ \mu V/K, \\
B = 0.0085\ \mu V/K^2, \\
R_{H1}^a = R_{H1}^{a(0)}\exp\left(\frac{E_{a1}^{R_H}}{k_BT}\right) - R_H^{LaAl_2}, \\
R_{H1}^{(0)} \approx 0.89 \times 10^{-3}\ cm^3/C, \\
E_{a1}^{R_H}/k_B \approx 12\ K.
\end{gathered} \tag{11}$$

The preexponential factor $\sigma_0(T)$ was found by fitting the experimental dependence with the use of the opti-

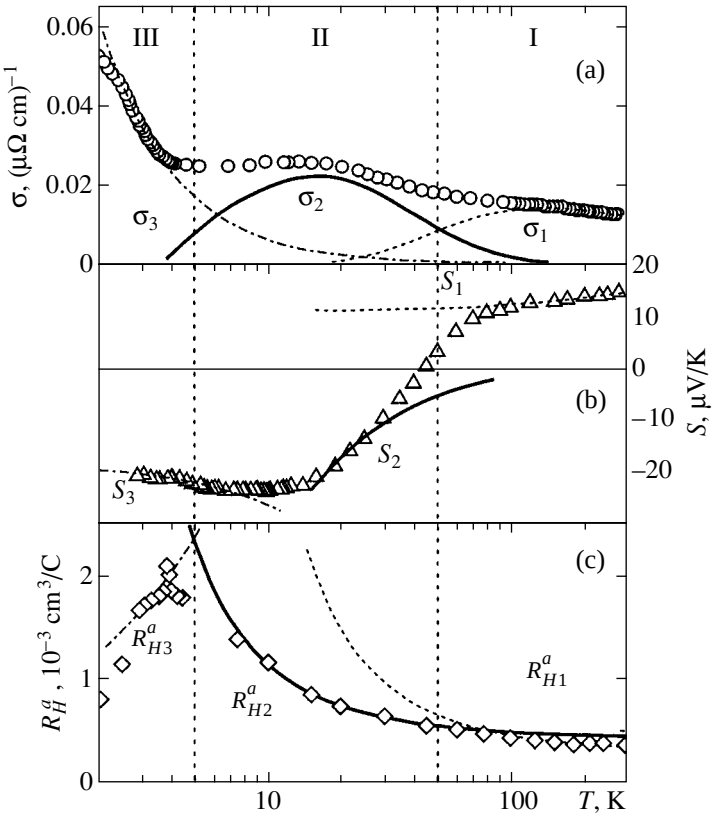


Fig. 13. Decomposition of the temperature dependences of the transport characteristics (a) $\sigma(T)$, (b) $S(T)$, and (c) $R_H^a(T)$ of $CeAl_2$ into contributions in temperature intervals I, II, and III.

mization procedures implemented in the ORIGIN 6.1 program.

The "inelastic" contribution σ_1 is described in (11) by a very simplified equation, which gives an activation dependence of conductivity in the transition region at $T < 100$ K. The main requirement imposed on (11) is the vanishing of the σ_1 component at $T \ll 100$ K. Clearly, the high-temperature contribution also contains the σ_{Γ_7} component determined by the scattering of carriers by the ground-state doublet Γ_7. Unlike the inelastic contribution, this component is also retained at low temperatures. Comparative estimates of its relative value in region I, however, give $\sigma_{\Gamma_7} \leq 0.1\sigma_1$, which in our view justifies the approximation that we use.

In addition, (11) contains a large constant contribution to the Seebeck coefficient (S_1) alongside a small linear term. We can therefore use the Heikes' formula to describe the Seebeck coefficient under strong Hubbard correlation conditions,

$$S_0^{\mathrm{I}} = \frac{k_B}{e}\ln\left(\frac{1-\nu}{\nu}\right), \qquad (12)$$

where, as previously, we use the notation $\nu = N/N_{\mathrm{Ce}}$, and $N_{\mathrm{Ce}} \approx 1.5 \times 10^{22}$ cm^{-3} in $CeAl_2$. It follows from (12) that the concentration of carriers in the temperature range 100–300 K can be estimated at $\nu_I \approx 0.53$ and, therefore, $N_I \approx 8 \times 10^{21}$ cm^{-3}. Note that (12) describes the asymptotic behavior of S, which, in our case, corresponds to the formation of a band of spin-polaron states of width $E_{a2}/k_{\mathrm{B}} \approx 12$ K in the vicinity of E_F.

According to [36], an analogous activation dependence of the Hall coefficient of spin-polaron transport in (11) can also be used to approximately estimate the concentration of carriers for the spin-polaron states of the FeSi matrix,

$$N_I^{\mathrm{CeAl_2}} = \frac{1}{eR_{H1}^{a(0)}} \approx 7 \times 10^{21}\ \mathrm{cm}^{-3}.$$

This estimate agrees satisfactorily with the results obtained above.

The asymptotic behaviors of all charge carrier transport parameters of $CeAl_2$ change as temperature decreases in the interval 50–100 K (see Figs. 13, 14). The corresponding contributions to σ, S, and R_H^a are comparable in magnitude. The situation in the transition temperature region 5–10 K between intervals II

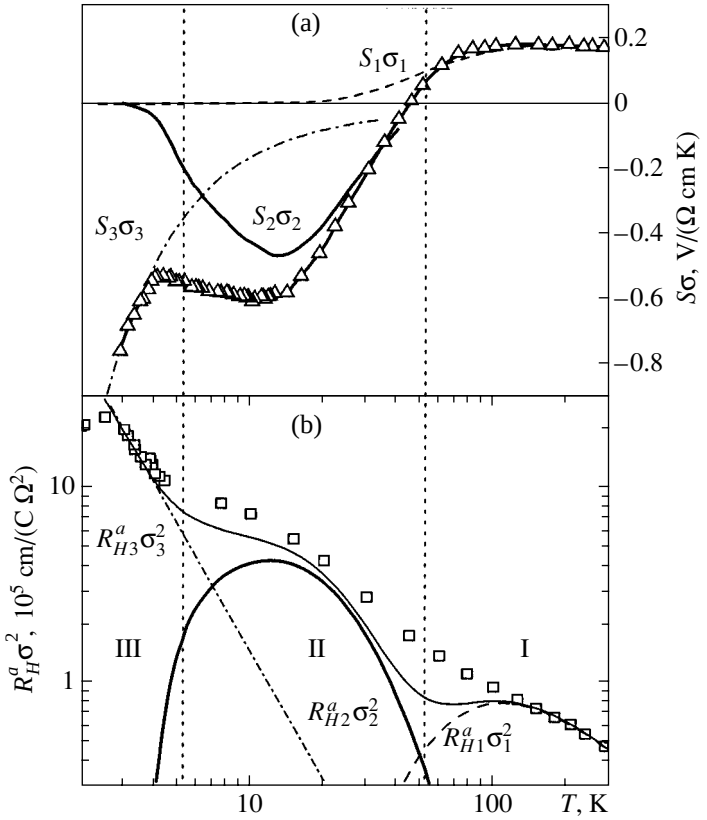


Fig. 14. Decomposition of (a) $S\sigma$ and (b) $R_H^a\sigma^2$ into partial contributions in temperature intervals I, II, and III for $CeAl_2$.

and III is similar. In this region, the σ_2 and σ_3, S_2 and S_3, and R_{H2}^a and R_{H3}^a components have comparable values (see Figs. 13, 14). The procedure used within the phenomenological approach to separating the contributions to low-temperature transport was as follows: In interval III, the data given in Figs. 2 and 10 were approximated by the analytic dependences

$$\sigma_3 = AT^{-\beta}, \quad \beta = 1.44,$$

$$S_3 = S_0^{\mathrm{III}} + CT, \quad S_0^{\mathrm{III}} = -18\ \mu\mathrm{V/K}, \quad C = -0.8\ \mu\mathrm{V/K}^2, \tag{13}$$

$$R_{H3}^a = DT^{0.7}, \quad D = 0.76386.$$

Equations (13) fairly accurately describe the σ^{expt}, S^{expt}, and R_H^{expt} curves at $T < 4$ K.

Next, additivity condition (8) for σ was used to obtain the σ_2 component by subtracting the sum $\sigma_1 + \sigma_3$ from the experimental σ^{expt} curve. The S_2 and R_{H2}^a contributions were analyzed using the equations that described the activation behavior of the thermoelectric and Hall coefficients in interval II,

$$S_2 \propto \frac{k_B}{e}\frac{E_{a2}^S}{k_BT},$$

$$R_{H2}^a = R_{H2}^{a(0)}\exp\left(\frac{E_{a2}^{R_H}}{k_BT}\right) - R_H^{\mathrm{LaAl_2}}. \tag{14}$$

Figures 13, where the components of σ, S, and R_H are plotted, and 14, where their additivity is verified [Eqs. (9) and (10)], show that, on the whole, the suggested procedure for separating contributions, in spite of its approximate character, gives a quantitative description of the behavior of the transport coefficients of $CeAl_2$. It can therefore be used to estimate certain microscopic parameters that characterize the electronic structure of this compound. Within this procedure, the activation energy of the Seebeck coefficient in interval II is estimated at $E_{a2}^S/k_B \approx 3.6$ K, which is noticeably smaller than the value for the Hall coefficient $E_{a2}^H/k_B \approx 7$ K. The use of $R_{H2}^{a(0)} \approx 1.03 \times 10^{-3}$ cm^3/C in (14) for estimating the concentration of carriers in interval II gives $N_{\mathrm{II}} = 6 \times 10^{21}$ cm^{-3} or $\nu = N_{\mathrm{II}}/N_{\mathrm{Ce}} \approx 0.4$. The

$S_0^{III} = -18$ μV/K value can also be formally used in (13) within the framework of the spin-polaron approach to estimate the reduced concentration of carriers in the transition temperature region 5–10 K. The use of the high-temperature asymptotic behavior of the Seebeck coefficient [Eq. (12)] is justified in this temperature region, because the formation of a narrow band of manybody states in the vicinity of E_F with the activation energy $E_{a2}/k_B \approx 7.6$ K results in the appearance of ferromagnetic nanoclusters based on spin-polaron states at $T < 20$ K (see the argumentation given above). As a result, we have

$$\nu_{III} \approx 0.45, \quad N_{III} \approx 6.8 \times 10^{21}\ \text{cm}^{-3}.$$

It should also be stressed that a comparison of the activation energies of the thermoelectric ($E_{a2}^S/k_B \approx 3.6$ K) and Hall ($E_{a2}^{R_H}/k_B \approx 7$ K) coefficients and the paramagnetic Curie temperatures found in this work ($\Theta_p \approx 3.6$ K) and measured in [34] ($\Theta_p \approx 3.9$ K) lead us to suggest that there are two approximately equal components that determine the formation of manybody states in $CeAl_2$ at low temperatures. It can be expected that the spin-polaron (magnetic) contribution to E_{a2}, although it does not manifest itself in the temperature dependence of the Seebeck coefficient [53], at the same time plays a key role in determining the Θ_p and $H_{ex} \approx 75$ kOe magnetic exchange parameters. This suggestion allows us to expect that the sum of the contributions of the exciton ($4f^+$–$5d^-$) and spin-polaron components should determine the $E_{a2}^{R_H}$ value, which characterizes the low-temperature behavior of the Hall coefficient. At the same time, the experimental results presented in this work are obviously insufficient as reliable evidence of exciton–polaron nature of manybody states in $CeAl_2$ with fast electron density fluctuations.

Note in conclusion that, within the framework of the approach that we use to separate the contributions to the low-temperature charge transport, the most complex problem is, in our view, a quantitative analysis of the σ_3, S_3, and R_{H3}^a components at $T < 5$ K. In this region, we must take into account not only the establishment of coherence (the formation of heavy carrier bands), but also effects related to complex magnetic ordering of the cerium-based intermetallic compounds under consideration.

5. CONCLUSIONS

The detailed measurements of the Hall effect in $CeAl_2$ with fast electron density fluctuations performed in this work allowed us to separate and classify the contributions to the anomalous Hall effect in this compound with heavy fermions. The appearance of an anomalous magnetic contribution "even in magnetic field" to the Hall resistance observed in this work at $T <$ 10 K is caused by the special features of medium- and long-range magnetic ordering and the complex H–T magnetic phase diagram of $CeAl_2$ at low temperatures. This result and the estimates of the $\Theta_p \approx 3.6$ K and $H_{ex} \approx 75$ kOe magnetic exchange parameters appear to provide evidence in favor of the formation of nanosized ferromagnetic regions in the $CeAl_2$ matrix at temperatures substantially higher than the Néel temperature $T_N \approx 3.85$ K of this compound. It was shown that the temperature dependence of the main anomalous component R_H^a in this compound with heavy fermions has a complex activation character. The behavior of $R_H^a(T)$ observed in $CeAl_2$ does not fit in with the interpretation within the framework of the skew-scattering model, according to which scattering effects play a determining role in the formation of Hall coefficient anomalies in concentrated Kondo systems.

The special features of the suppression of the anomalous Hall effect in high magnetic fields studied in this work are likely to be evidence that spin-polaron effects should be taken into account to interpret the behavior of the transport characteristics of cerium-based intermetallic compounds. We estimated the parameters characteristic of manybody states that arise in the $CeAl_2$ matrix at low and intermediate temperatures (their effective masses and localization radii). The nontrivial analysis of contributions to the transport characteristics of $CeAl_2$ performed in this work based on the results of Hall effect measurements combined with resistivity and Seebeck coefficient data led us to conclude that the approach that uses the Kondo lattice model has serious limitations as applied to the totality of properties of cerium-based concentrated Kondo systems.

ACKNOWLEDGMENTS

This work was financially supported by the Russian Foundation for Basic research (project nos. 01-02-16601 and 03-02-06531); the "New Materials" project of the Ministry of Education of Russian Federation (no. 202.07.01.023); the program "Strongly Correlated Electrons in Semiconductors, Metals, Superconductors, and Magnetic Materials" of the Russian Academy of Sciences (Division of Physical Sciences); the program for the development of the instrumental base of scientific organizations of RF Ministry of Industry and Science; INTAS project no. 00-807; and the program of the Russian Academy of Sciences for support of young scientists. Special thanks for individual support are due to the Foundation for Promoting Science in this country (V.V.G. and S.V.D.) and Government of Moscow and Soros Foundation (A.V.B. and M.I.I.).

REFERENCES

1. P. Coleman, P. W. Anderson, and T. V. Ramakrishnan, Phys. Rev. Lett. **55**, 414 (1985).

2. A. Fert and P. M. Levy, Phys. Rev. B **36**, 1907 (1987).
3. P. M. Levy and A. Fert, Phys. Rev. B **39**, 12224 (1989).
4. P. M. Levy, Phys. Rev. B **38**, 6779 (1988).
5. N. B. Brandt, V. V. Moshchalkov, N. E. Sluchanko, *et al.*, Solid State Commun. **53**, 645 (1985).
6. V. V. Moshchalkov, F. G. Aliev, N. E. Sluchanko, *et al.*, J. Less-Common Met. **127**, 321 (1986).
7. T. Penney, F. P. Milliken, S. von Molnar, *et al.*, Phys. Rev. B **34**, 5959 (1986).
8. A. Fert, P. Pureur, A. Hamzic, and J. P. Kappler, Phys. Rev. B **32**, 7003 (1985).
9. T. Hiraoka, E. Kinoshita, T. Takabatake, *et al.*, Physica B (Amsterdam) **199**–**200**, 440 (1994).
10. H. Sugawara, H. R. Sato, Y. Aoki, and H. Sato, J. Phys. Soc. Jpn. **66**, 174 (1997).
11. U. Welp, P. Haen, G. Bruls, *et al.*, J. Magn. Magn. Mater. **63**–**64**, 28 (1987).
12. N. E. Sluchanko, A. V. Bogach, V. V. Glushkov, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **76**, 31 (2002) [JETP Lett. **76**, 26 (2002)].
13. N. E. Sluchanko, V. V. Glushkov, S. V. Demishev, *et al.*, Zh. Éksp. Teor. Fiz. **113**, 339 (1998) [JETP **86**, 190 (1998)].
14. B. Barbara, J. X. Boucherle, J. L. Buevoz, *et al.*, Solid State Commun. **24**, 481 (1977).
15. F. Lapierre, P. Haen, A. Briggs, and M. Sera, J. Magn. Magn. Mater. **63**–**64**, 76 (1987).
16. F. Steglich, C. D. Bredl, M. Loewenhaupt, and K. D. Schotte, J. Phys. Colloq. **40** (C5), 301 (1979).
17. S. Osborn, M. Loewenhaupt, B. D. Rainford, and W. G. Stirling, J. Magn. Magn. Mater. **63**–**64**, 70 (1987).
18. M. Loewenhaupt, W. Reichardt, R. Pynn, and E. Lindley, J. Magn. Magn. Mater. **63**–**64**, 73 (1987).
19. N. E. Sluchanko, A. V. Bogach, I. B. Voskoboĭnikov, *et al.*, Fiz. Tverd. Tela (St. Petersburg) **45**, 1046 (2003) [Phys. Solid State **45**, 1096 (2003)].
20. M. Croft, I. Zoric, and R. D. Parks, Phys. Rev. B **18**, 345 (1978).
21. M. Croft, I. Zoric, and R. D. Parks, Phys. Rev. B **18**, 5065 (1978).
22. E. Fawcett, V. Pluzhnikov, and H. Klimker, Phys. Rev. B **43**, 8531 (1991).
23. N. B. Brandt and V. V. Moshchalkov, Adv. Phys. **33**, 373 (1984).
24. M. Christen and M. Godet, Phys. Lett. A **63A**, 125 (1977).
25. N. F. Mott, *Metal–Insulator Transitions* (Taylor and Francis, London, 1974; Nauka, Moscow, 1979).
26. S. H. Liu, Phys. Rev. B **37**, 3542 (1988).
27. T. Portengen, Th. Ostreich, and L. J. Sham, Phys. Rev. B **54**, 17452 (1996).
28. M. Loewenhaupt, B. D. Rainford, and F. Steglich, Phys. Rev. Lett. **42**, 1709 (1979); M. Loewenhaupt and U. Witte, J. Phys.: Condens. Matter **15**, S519 (2003).
29. P. Thalmeier and P. Fulde, Phys. Rev. Lett. **49**, 1588 (1982).
30. G. Guntherodt, A. Jayaraman, G. Batlogg, *et al.*, Phys. Rev. Lett. **51**, 2330 (1983).
31. N. E. Sluchanko, V. V. Glushkov, B. P. Gorshunov, *et al.*, Phys. Rev. B **61**, 9906 (2000).
32. B. Barbara, M. F. Rossignol, J. X. Boucherle, *et al.*, Phys. Rev. Lett. **45**, 938 (1980).
33. A. Benoit, J. X. Boucherle, J. Flouquet, *et al.*, in *Valence Fluctuations in Solids*, Ed. by L. M. Falicov, W. Hanke, and M. B. Maple (North-Holland, Amsterdam, 1981), p. 197.
34. M. C. Croft, R. P. Guertin, L. C. Kupferberg, and R. D. Parks, Phys. Rev. B **20**, 2073 (1979).
35. C. D. Bredl, F. Steglich, and K. D. Schotte, Z. Phys. B **29**, 327 (1978).
36. N. E. Sluchanko, V. V. Glushkov, S. V. Demishev, *et al.*, Zh. Éksp. Teor. Fiz. **119**, 359 (2001) [JETP **92**, 312 (2001)].
37. N. E. Sluchanko, V. V. Glushkov, S. V. Demishev, *et al.*, Phys. Rev. B **65**, 064404 (2002).
38. G. Hampel and R. H. Blick, J. Low Temp. Phys. **99**, 71 (1995).
39. D. E. MacLaughlin, O. Peca, and M. Lysak, Phys. Rev. B **23**, 1039 (1981).
40. J. L. Gavilano, J. Hunziker, O. Hudak, *et al.*, Phys. Rev. B **47**, 3438 (1993).
41. S. M. Schapiro, E. Gurewitz, R. D. Parks, and L. C. Kupferberg, Phys. Rev. Lett. **43**, 1748 (1979).
42. A. Schenk, D. Andreica, M. Pinkpank, *et al.*, Physica B (Amsterdam) **259**–**261**, 14 (1999).
43. A. Schenk, D. Andreica, F. N. Gygax, and H. R. Ott, Phys. Rev. B **65**, 024444 (2002).
44. A. Amato, Rev. Mod. Phys. **69**, 1119 (1997).
45. E. M. Forgan, B. D. Rainford, S. L. Lee, *et al.*, J. Phys.: Condens. Matter **2**, 10211 (1990).
46. F. Giford, J. Schweizer, and F. Tasset, Physica B (Amsterdam) **234**–**236**, 685 (1997).
47. T. Chattopadhyay and G. J. McIntyre, Physica B (Amsterdam) **234**–**236**, 682 (1997).
48. J. Ye, Y. B. Kim, A. J. Millis, *et al.*, Phys. Rev. Lett. **83**, 3737 (1999).
49. Y. B. Kim, P. Majumdar, A. J. Millis, and B. I. Shraiman, cond-mat/9803350 (1998).
50. E. Walker, H. G. Purwins, M. Landolt, and F. Hulliger, J. Less-Common Met. **33**, 203 (1973).
51. J. Sakurai, H. Takagi, T. Kuwai, and Y. Isikawa, J. Magn. Magn. Mater. **177**–**181**, 407 (1998).
52. K. H. Fisher, Z. Phys. B **76**, 315 (1989).
53. P. M. Chaikin, in *Organic Superconductivity*, Ed. by V. Z. Kresin and W. A. Little (Plenum, New York, 1990), p. 101.

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 811–819.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 927–937.

NONLINEAR PHYSICS

Numerical Modeling of Shock-Wave Instability in Thermodynamically Nonideal Media

A. V. Konyukhov[a,b], A. P. Likhachev[a], A. M. Oparin[b], S. I. Anisimov[c], and V. E. Fortov[a]

[a]*Institute of Thermophysics of Extreme States, Joint Institute of High Temperatures, Russian Academy of Sciences, Moscow, 125412 Russia*

[b]*Institute for Computer-Aided Design, Russian Academy of Sciences, Vtoraya Brestskaya ul. 19/18, Moscow, 123056 Russia*

[c]*Landau Institute for Theoretical Physics, Russian Academy of Sciences, Chernogolovka, Moscow oblast, 142432 Russia*

e-mail: a.oparin@icad.org.ru

Received August 27, 2003

Abstract—A numerical analysis of the nonlinear instability of shock waves is presented for solid deuterium and for a model medium described by a properly constructed equation of state. The splitting of an unstable shock wave into an absolutely stable shock and a shock that emits acoustic waves is simulated for the first time. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

For a variety of media (in particular, for media with phase transitions), the shock Hugoniot curves include segments corresponding to shock waves that are unstable with respect to small disturbances of the wave front. The linear analysis developed in [1, 2] predicts two types of shock-wave evolution: small periodic shock-front disturbances may either grow exponentially with time or persist for an indefinitely long time without being either damped or amplified. In the latter case, the shock front emits acoustic and entropy waves propagating downstream. The D'yakov–Kontorovich linear theory was the first to provide criteria for shock-wave instability. In terms of the stability parameter

$$L = j^2\left(\frac{\partial V}{\partial p}\right)_H,$$

where $j^2 = (p_2 - p_1)/(V_1 - V_2)$ is the mass flux across the shock front, $V = 1/\rho$ is specific volume, p is pressure, and the derivative is taken along the Hugoniot curve, the following regions of qualitatively different behavior are identified: absolute stability ($-1 < L < L_0$), acoustic emission ($L_0 < L < 1 + 2M$), and exponential growth ($L < -1$, $L > 1 + 2M$). We use the following notation here:

$$L_0 = \frac{1 - \theta M^2 - M^2}{1 + \theta M^2 - M^2}, \quad \theta = \frac{V_0}{V}, \quad M = \frac{u}{c}.$$

Subsequently, it was shown that the Hugoniot segment corresponding to exponential growth lies within a region of shock-wave nonuniqueness, where a single shock wave splits into multiple stable elementary waves (shock waves, isentropic rarefaction/compression waves, and/or contact discontinuities) [3]. These findings led to the hypothesis that exponential growth cannot be observed because an unstable initial shock wave must evolve into an admissible split-wave configuration [3].

To facilitate further analysis, we recall here some facts concerning the problem of shock-wave stability. It was shown in [4] that the point in the p–V diagram where the shock Hugoniot curve is tangent to a line passing through the initial state is a sonic point; i.e., the Mach number behind the shock front is unity in a reference frame tied to the front:

$$M = \frac{|u - D|}{c} = \frac{D - u}{c} = 1.$$

The slope of the tangent line to the corresponding isentrope satisfies the relation

$$\left(\frac{\partial p}{\partial V}\right)_S = -\frac{j^2}{M^2} = -j^2.$$

By definition,

$$\left(\frac{\partial p}{\partial V}\right)_H = -j^2$$

at the tangency point. Therefore, a point where the Rayleigh line is tangent to the Hugoniot curve is a point of tangency of the Hugoniot with the isentrope:

$$\left(\frac{\partial p}{\partial V}\right)_S = -j^2 = \left(\frac{\partial p}{\partial V}\right)_H.$$

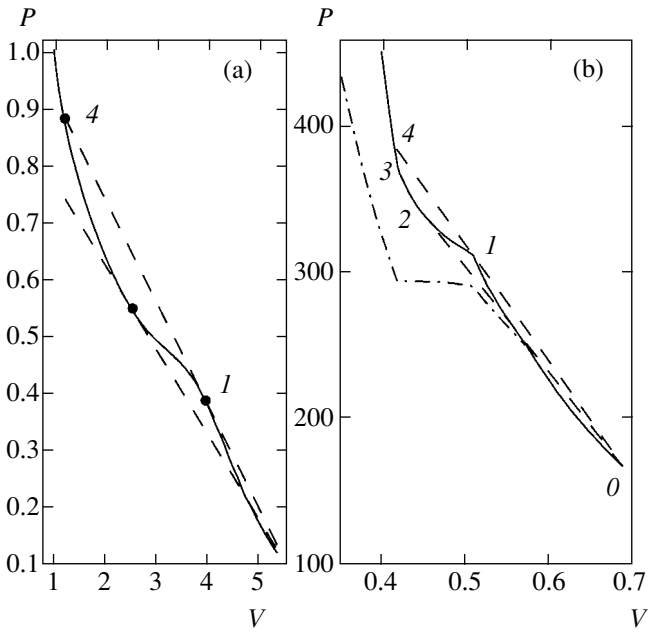


Fig. 1. Typical Hugoniots (a) without and (b) with a kink corresponding to the model equation of state and the SESAME tabular equation of state for deuterium, respectively. The intervals of stability (*0–1*), exponential growth (*1–2*), acoustic emission (*2–3*), and shock-wave splitting (*1–4*) are indicated. The dot–dash curve is an isotherm.

At such points, the D'yakov–Kontorovich stability parameter is

$$L = j^2\left(\frac{\partial V}{\partial p}\right)_H = -1.$$

Thus, the points where the Rayleigh line is tangent to the Hugoniot curve correspond to the boundaries of exponential growth. Between these boundaries, $L < -1$.

If the Hugoniot is smooth, then the convexity condition for equation of state (EOS) is violated below the lower tangency point:

$$\left(\frac{\partial^2 p}{\partial V^2}\right)_S < 0.$$

Above this point, a single initial shock wave evolves into the following split wave configuration:

$$\vec{\mathrm{S}} \longrightarrow \overleftarrow{\mathrm{W}}\mathrm{T}\vec{\mathrm{C}}\vec{\mathrm{S}},$$

where S is a shock wave, W is a shock or rarefaction wave, T is a tangential discontinuity, and C is an isentropic compression wave [3]. This is explained by the fact that $u + c < D$ below the point where

$$M = \frac{D-u}{c} = 1,$$

i.e., acoustic disturbances (with velocity $u + c$) lag behind the shock, transforming into an isentropic compression wave [3].

The lower point of tangency of the Rayleigh line with the shock Hugoniot is the lower boundary of acoustic emission:

$$L_0 = \frac{1-\theta M^2 - M^2}{1+\theta M^2 - M^2} = -1 = L.$$

If the Hugoniot curve is smooth (see Fig. 1a), then the $L(P)$ and $L_0(P)$ curves are either mutually tangent or they intersect at this point. If the Hugoniot curve has a kink (Fig. 1b), then the values of L and L_0 are not defined at this point, whereas $L - L_0$ either reverses or retains its sign across the kink (as in the case of intersection or tangency, respectively). However, the $L(P)$ and $L_0(P)$ curves must intersect at the upper point of Rayleigh-line tangency with a Hugoniot of any shape. This point separates Hugoniot segments associated with exponential growth and acoustic emission. Shock waves with parameters corresponding to a portion of the latter segment adjoining the former one emit acoustic waves at angles close to the shock propagation direction. As $L \longrightarrow -1$ and $M \longrightarrow 1$, the equation for the cosine of this angle tends to a limit form [4] that has a unique solution: $\cos\alpha = -1$. When $\cos\alpha < 0$, the acoustic wave vector points in the direction of shock propagation. The wave is "outgoing" because of advection by the flow.

The upper boundary of shock-wave splitting is determined by equating the velocities of the leading wave and the downstream shock. The pressure at this boundary can be found from the equation

$$\frac{p_4 - p_1}{V_4 - V_1} = -\frac{c_1^2}{V_1^2}$$

as a tangency condition for a Rayleigh line and an isentrope. (Here, subscripts correspond to points in Figs. 1a and 1b.) This relation also means that the downstream-shock velocity relative to the medium equals the speed of sound:

$$D_{\mathrm{II}} - u_1 = V_1\left(\frac{p_4 - p_1}{V_1 - V_4}\right)^{1/2} = c_1.$$

A detailed analysis of the splitting wave configurations corresponding to smooth and kinked Hugoniot curves was presented in [3].

Even though shock-wave stability has been analyzed in numerous theoretical studies for almost half a century, experimental observations supporting the theoretical predictions mentioned above remain scarce to this day. On the one hand, the corresponding thermodynamic conditions are difficult to implement experimen-

tally. Another reason is the lack of data required to specify the conditions of an appropriate physical experiment and the expected manifestations of shock-wave instability. To obtain data of this kind, a parametric numerical analysis of the behavior of a shock wave must be performed by varying its intensity under suitable choice of a properly constructed or realistic equation of state. Only a few numerical simulations of shock-wave splitting [5, 6] and acoustic emission [7] have been performed to date, and their results are far from complete. Thus, systematic numerical analysis of nonlinear shock-wave instability remains a challenging task. This problem is addressed in the present study.

In this paper, we analyze the stability of a shock wave in solid deuterium (using an EOS borrowed from the SESAME library [8]) and in a model medium described by the equation [5]

$$e(p, \rho) = (1 - e^{-p^2})(4 - e^{-(4-1/\rho)^2}).$$

This equation of state is thermodynamically consistent, and the corresponding shock Hugoniot curves include segments associated with instabilities of all known types. The realistic three-phase SESAME equation of state (given in tabular form) describes the first-order molecular-to-metallic phase transition in solid deuterium [8].

The shock Hugoniot curve based on the model equation of state characterizes shock compression processes involving phase transitions or endothermic chemical reactions. This curve allows for both shock splitting and acoustic emission. Figure 1a shows a low-pressure portion of this curve.

Figure 1b shows the shock Hugoniot curve for low-temperature deuterium compressed from an initial state characterized by $p_0 = 1.6 \times 10^3$ GPa and $V_0 = 0.7$ cm^3/g. Unlike the curve shown in Fig. 1a, it has two kinks where adiabatic compressibility decreases and increases stepwise (points *1* and *3*, respectively). Since a kinked curve is a limit case of a smooth curve, one may say that a smooth region of EOS convexity violation reduces to a point in this limit and tangency at a kink is interpreted accordingly.

2. NUMERICAL METHOD

To solve the governing equations, we use Roe's method [9] extended to an arbitrary equation of state of the form $\varepsilon = (p, \rho)$. In this conservative method based on exact characteristic flux splitting, a stationary shock satisfying the Rankine–Hugoniot jump conditions is an exact solution to the finite-difference problem. We use a computationally efficient formulation of the method in terms of the vector $Y = (p, \rho^{1/2}, u)^T$ [10].

Following the approach developed in [10], we find the solution vector on the $(n+1)$th time layer from integral conservation laws written as expressions for the grid pressure, velocity, and $\rho^{1/2}$:

$$z^{n+1} = ((z^n)^2 + \xi_1)^{1/2},$$

$$u^{n+1} = \frac{(z^n)^2 u^n + \xi_2}{(z^{n+1})^2},$$

$$\varepsilon(p^{n+1}, \rho^{n+1}) = \varepsilon(p^n, \rho^n) + \frac{1}{2}((z^n u^n)^2 - (z^{n+1}u^{n+1})^2) + \xi_3,$$

where $z = \rho^{1/2}$, ε is the internal energy per unit volume,

$$\boldsymbol{\xi} = -\frac{\tau}{h}(\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2})$$

is the net flux of the conserved quantities across a cell, τ is the time step, and h is the mesh size. The numerical flux across a cell face can be written in a more compact form in terms of pressure, density, and velocity (as compared to the conserved quantities):

$$\mathbf{F}_{i+1/2} = \frac{1}{2}[\mathbf{F}_i + \mathbf{F}_{i+1} + R_{i+1/2}\boldsymbol{\Phi}_{i+1/2}],$$

$$\mathbf{F} = \begin{pmatrix} \rho u \\ p + \rho u^2 \\ \rho u H \end{pmatrix},$$

$$\mathbf{R} = \begin{pmatrix} 1 & \dfrac{\rho}{2c} & \dfrac{\rho}{2c} \\ u & \dfrac{\rho(u-c)}{2c} & \dfrac{\rho(u+c)}{2c} \\ \varepsilon_\rho + \dfrac{1}{2}u^2 & \dfrac{\rho(H-uc)}{2c} & \dfrac{\rho(H+uc)}{2c} \end{pmatrix},$$

$$\boldsymbol{\alpha} = \begin{pmatrix} \Delta\rho - \Delta p/c^2 \\ \Delta p/(\rho c) - \Delta u \\ \Delta p/(\rho c) + \Delta u \end{pmatrix},$$

$$\boldsymbol{\lambda} = (u, u-c, u+c)^T, \quad \boldsymbol{\lambda}^{\pm} = \frac{\boldsymbol{\lambda} \pm |\boldsymbol{\lambda}|}{2}.$$

The components ϕ^l of the vector $\boldsymbol{\Phi}_{j+1/2}$ are expressed as

$$\phi^l_{i+1/2} = g^1_i + g^1_{i+1} - \psi\left(\lambda^1_{i+1/2} + \frac{g^1_{i+1} - g^1_i}{\alpha^1_{i+1/2}}\right)\alpha^i_{i+1/2}.$$

In TVD1 (first-order accurate scheme),

$$\phi^l_{i+1/2} = -\psi(\lambda^l_{i+1/2})\alpha^1_{i+1/2}.$$

In TVD2 (Harten's second-order accurate scheme [11]),

$$g^l_i = \frac{1-|\sigma|}{2}\psi(\lambda^1_{i+1/2})m[\alpha^1_{j+1/2}, \alpha^1_{j-1/2}]$$

$$+\begin{cases}\left(\lambda^{1+}_{i+1/2}\dfrac{2-3|\sigma|+\sigma^2}{6}+\lambda^{1-}_{i+1/2}\dfrac{\sigma^2-1}{6}\right)\bar{m}[\Delta_-\alpha^1_{j-1/2}, \Delta_+\alpha^1_{j-1/2}], & \text{if } |\alpha^1_{i-1/2}| \le |\alpha^1_{i+1/2}|,\\ \left(\lambda^{1-}_{i+1/2}\dfrac{2-3|\sigma|+\sigma^2}{6}+\lambda^{1+}_{i+1/2}\dfrac{\sigma^2-1}{6}\right)\bar{m}[\Delta_-\alpha^1_{j+1/2}, \Delta_+\alpha^1_{j+1/2}], & \text{if } |\alpha^1_{i-1/2}| > |\alpha^1_{i+1/2}|,\end{cases}$$

with

$$\sigma = \lambda^1_{j+1/2}\frac{\tau}{h},$$

$$m[x,y] = \begin{cases}0, & xy \le 0,\\ \min(|x|,|y|)\operatorname{sgn}(x), & xy > 0,\end{cases}$$

$$\bar{m}[x,y] = \begin{cases}x, & |x| \le |y|,\\ y, & |x| > |y|,\end{cases} \qquad (1)$$

$$\psi(x) = \begin{cases}|x|, & |x| \ge \epsilon,\\ \dfrac{x^2+\epsilon^2}{2\epsilon}, & |x| < \epsilon,\end{cases}$$

where ϵ is the entropy correction parameter [11].

An approximate solution to the Riemann problem for two states corresponding to adjacent cells labeled "1" and "2" is given by the relations

$$\rho = z_1 z_2,$$

$$u = \frac{z_1u_1 + z_2u_2}{z_1+z_2},$$

$$H = \frac{z_1H_1 + z_2H_2}{z_1+z_2},$$

$$c^2 = \frac{h-\varepsilon_\rho}{\varepsilon_p}.$$

In the UNO3 scheme relying on the third-order ENO interpolation [12],

The EOS derivatives are calculated numerically by using states "1" and "2:"[1]

$$\varepsilon_\rho = \frac{0.5}{\tilde{\rho}_2-\rho_1}(\varepsilon(\tilde{p}_2,\tilde{\rho}_2)+\varepsilon(p_1,\tilde{\rho}_2) - \varepsilon(\tilde{p}_2,\rho_1)-\varepsilon(p_1,\rho_1)),$$

$$\varepsilon_p = \frac{0.5}{\tilde{\rho}_2-\rho_1}(\varepsilon(\tilde{p}_2,\tilde{\rho}_2)+\varepsilon(\tilde{p}_2,\rho_1) - \varepsilon(p_1,\tilde{\rho}_2)-\varepsilon(p_1,\rho_1)),$$

$$\tilde{p}_2 = \begin{cases}p_2, & |p_2-p_1| \ge p_1\delta,\\ p_1+p_1\delta, & |p_2-p_1| < p_1\delta,\end{cases}$$

$$\tilde{\rho}_2 = \begin{cases}\rho_2, & |\rho_2-\rho_1| \ge \rho_1\delta,\\ \rho_1+\rho_1\delta, & |\rho_2-\rho_1| < \rho_1\delta,\end{cases}$$

where δ is a small positive number.

This differentiation procedure ensures that the jump condition

$$\varepsilon_2-\varepsilon_1 = \varepsilon_\rho(\rho_2-\rho_1)+\varepsilon_p(p_2-p_1)$$

and hence the relation

$$\Delta F = A\Delta U,$$

hold across a shock. (Here, A is the Jacobian matrix of the flux vector for an averaged state.) Then, an arbitrary

[1] This approach is analogous to Glaister's solution [13], where different states were combined for the first time in calculating pressure derivatives to extend Roe's method to an arbitrary EOS.

steady shock satisfying the Hugoniot relations, i.e., defined by a grid function of the form

$$U_i^n = \begin{cases} U_1, & i \le i_0, \\ U_2, & i > i_0, \end{cases} \quad (2)$$

such that $F(U_2) - F(U_1) = 0$, is an exact solution to the finite-difference equations (i_0 is a particular value of grid index).[2]

In the region of shock-wave nonuniqueness, single shock waves are exact solutions to the finite-difference problem in a reference frame tied to its front and small disturbances introduced into initial conditions either decay (when the shock is stable) or grow exponentially. In the course of computations, single waves split and evolve into alternative scale-invariant configurations, because solutions to the finite-difference problem are not unique either. Figure 2 demonstrates that the norm of residual computed by the schemes employed in this study levels off as a split-wave solution is approached under nonuniqueness conditions.

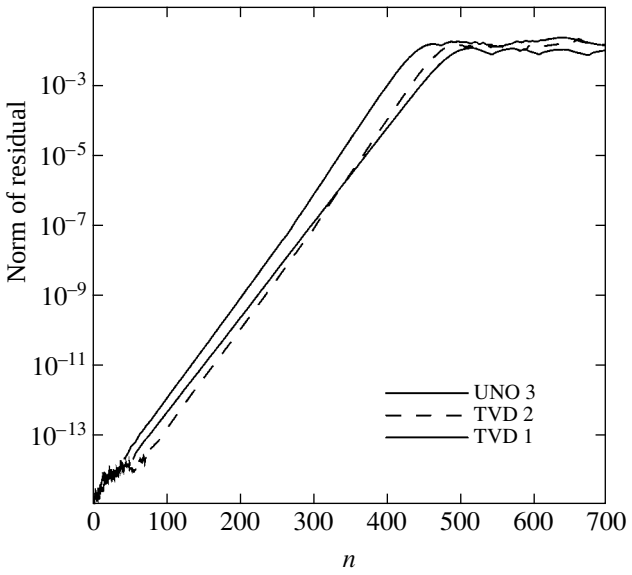


Fig. 2. Norm of residual for split-wave solutions versus time step number.

3. ANALYSIS OF SHOCK-WAVE SPLITTING

The Cauchy problem for the Euler equations supplemented with initial conditions (2) was solved in a coordinate system moving with the front; i.e., the initial shock wave was stationary on the numerical grid. Figure 3 shows the resulting scale-invariant solutions as functions of x/t for various Hugoniot points characterized by post-shock pressures P. In the right panel, *0* is the initial state, *1* is the lower boundary of exponential growth and shock-wave splitting, *2* is the upper boundary of EOS convexity violation, *3* is the upper boundary of exponential growth and the lower boundary of acoustic emission, and *4* is the upper boundary of shock-wave splitting.

The solution is depicted by constant-pressure contours in the (P, x/t) plane. The pressure values on the contours correspond to those on the P axis. Hereinafter, this representation of results obtained for various Hugoniot curves is called a splitting diagram.

Below point *1*, shock waves are absolutely stable. Between points *1* and *2*, a single shock wave splits into a configuration involving a shock and an isentropic compression wave:

$$\vec{\mathrm{S}} \longrightarrow \overleftarrow{\mathrm{W}}\mathrm{T}\vec{\mathrm{C}}\vec{\mathrm{S}}.$$

Between points *2* and *3*, the split-wave configuration involves two shocks separated by an isentropic compression wave:

$$\vec{\mathrm{S}} \longrightarrow \overleftarrow{\mathrm{W}}\mathrm{T}\vec{\mathrm{S}}\vec{\mathrm{C}}\vec{\mathrm{S}}.$$

At point *3*, $M = 1$. The difference in propagation velocity between shock waves can be determined as the distance between the corresponding fronts on the x/t axis. As P increases along the Hugoniot, the leading-wave intensity remains invariant, the downstream-shock intensity and velocity increase, and the pressure rise in the isentropic compression wave decreases. At point *4*, the leading-wave and downstream-shock velocities are equal, and the isentropic-wave pressure rise vanishes. This means that point *4* is the upper boundary of shock-wave splitting.

In contrast to the case of smooth violation of EOS convexity, the split-wave configuration corresponding to a kinked Hugoniot curve (Fig. 1b) does not involve any isentropic compression wave:

$$\vec{\mathrm{S}} \longrightarrow \overleftarrow{\mathrm{W}}\mathrm{T}\vec{\mathrm{S}}\,\vec{\mathrm{S}}.$$

Figure 4 compares the scale-invariant pressure profiles obtained for the model equation of state at $P = 0.54$ (left panel) and for deuterium at $P = 3.6 \times 10^3$ GPa (right panel). The pressure interval where the isentropic compression wave (ICW) exists is indicated here. The velocity difference between the leading wave and downstream shock resulting from shock-wave splitting in deuterium is about 2.8×10^3 m/s.

These results were verified by using various numerical techniques (other than those described above), including Eulerian and Lagrangian schemes. All of

[2] This is true only when the Harten entropy correction parameter is zero. Generally, entropy correction is applied to eliminate the rarefaction shock arising in Roe's method. In the present context, rarefaction shocks do not arise when EOS convexity holds. However, we have to deal with the "symmetric" problem of compression shocks when the EOS convexity condition is violated.

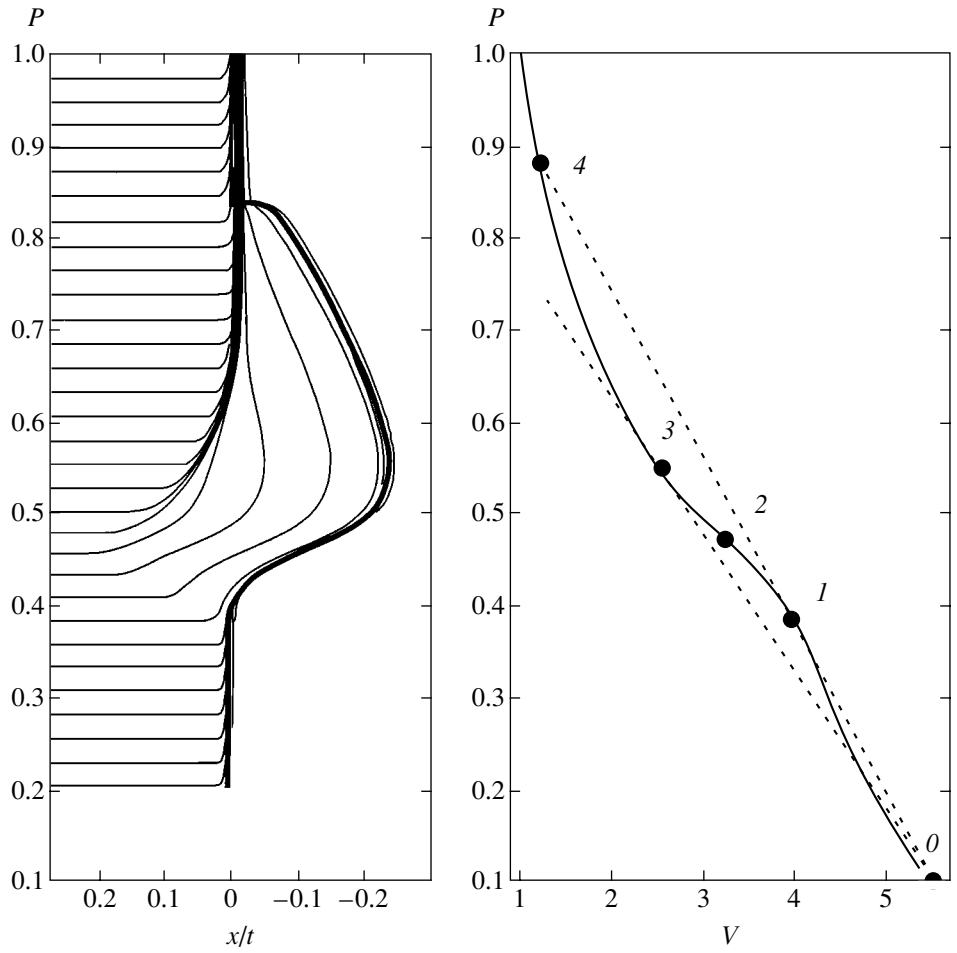


Fig. 3. Pressure versus scaling variable x/t (left panel) for one-dimensional shock waves corresponding to Hugoniot curve for model equation of state (right panel).

these results are mutually consistent and agree with previous computations [6].

4. COMPARISON OF RESULTS FOR SEVERAL FINITE-DIFFERENCE SCHEMES AND THE REQUIREMENT OF ENTROPY CORRECTION

Test computations were performed with the first- and second-order accurate TVD schemes and with fluxes calculated by using the third-order one-dimensional ENO interpolation [12]. It was shown that physical solutions satisfying the law of entropy increase under conditions of EOS convexity violation can be computed by the TVD schemes only with entropy corrections. Among the tested schemes, only UNO3 provides a correct solution without entropy correction. For example, the first-order accurate scheme without entropy correction yields the incorrect two-shock solution $\vec{S} \longrightarrow \overleftarrow{W}\vec{T}\vec{S}\vec{S}$ instead of $\vec{S} \longrightarrow \overleftarrow{W}\vec{T}\vec{S}\vec{C}\vec{S}$. Furthermore, the upper boundary of shock-wave splitting predicted by using higher order accurate schemes is lower than its experimental value. This result was

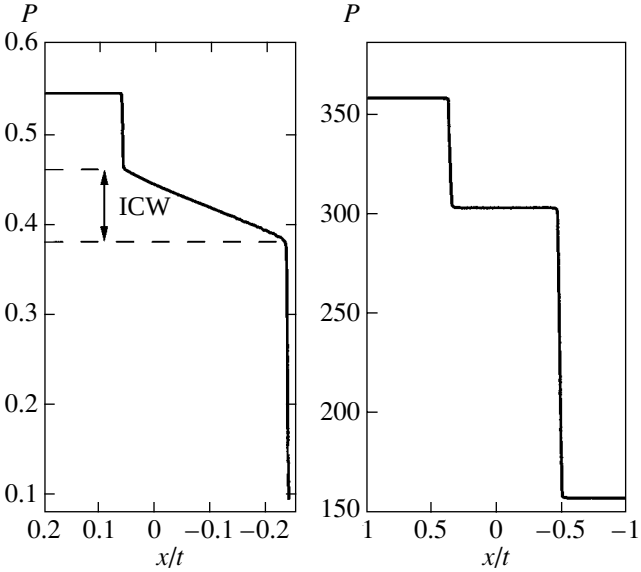


Fig. 4. Shock-wave splitting configurations: $\vec{S} \longrightarrow \overleftarrow{W}\vec{T}\vec{S}\vec{C}\vec{S}$ for Hugoniot without kinks (model equation of state, left panel); $\vec{S} \longrightarrow \overleftarrow{W}\vec{T}\vec{S}\vec{S}$ for kinked Hugoniot (SESAME equation for deuterium, right panel).

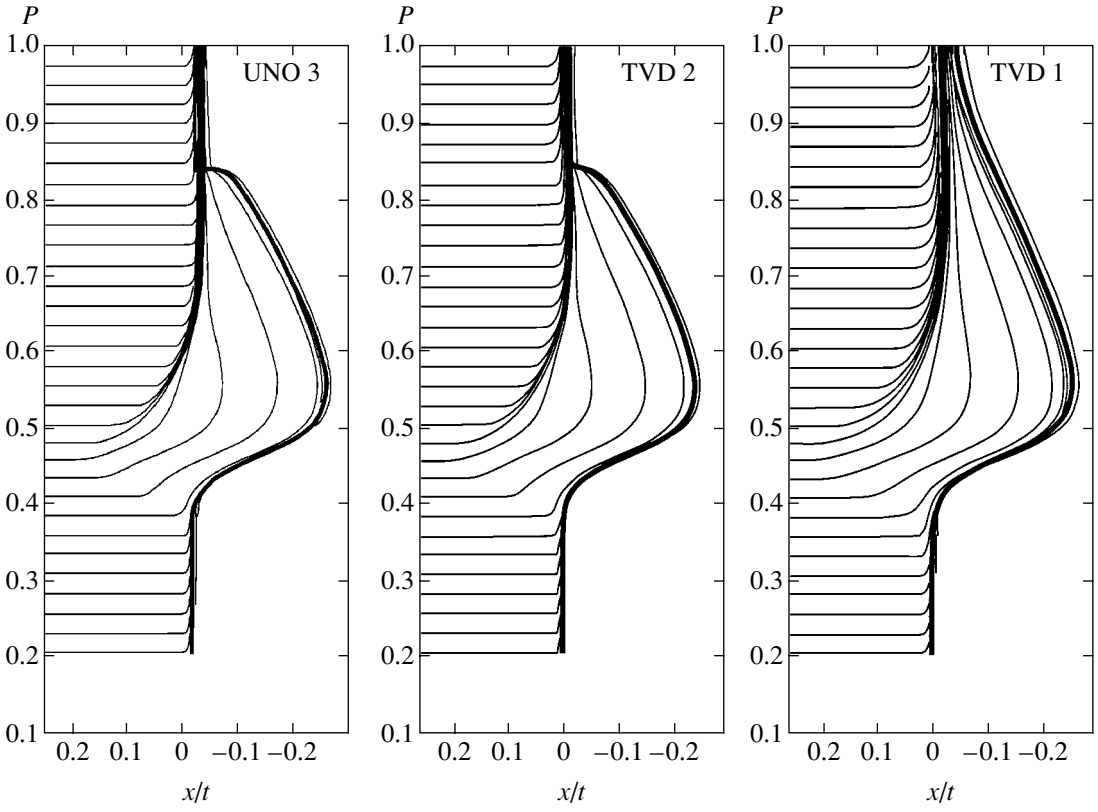


Fig. 5. Splitting diagrams obtained by using different finite-difference schemes.

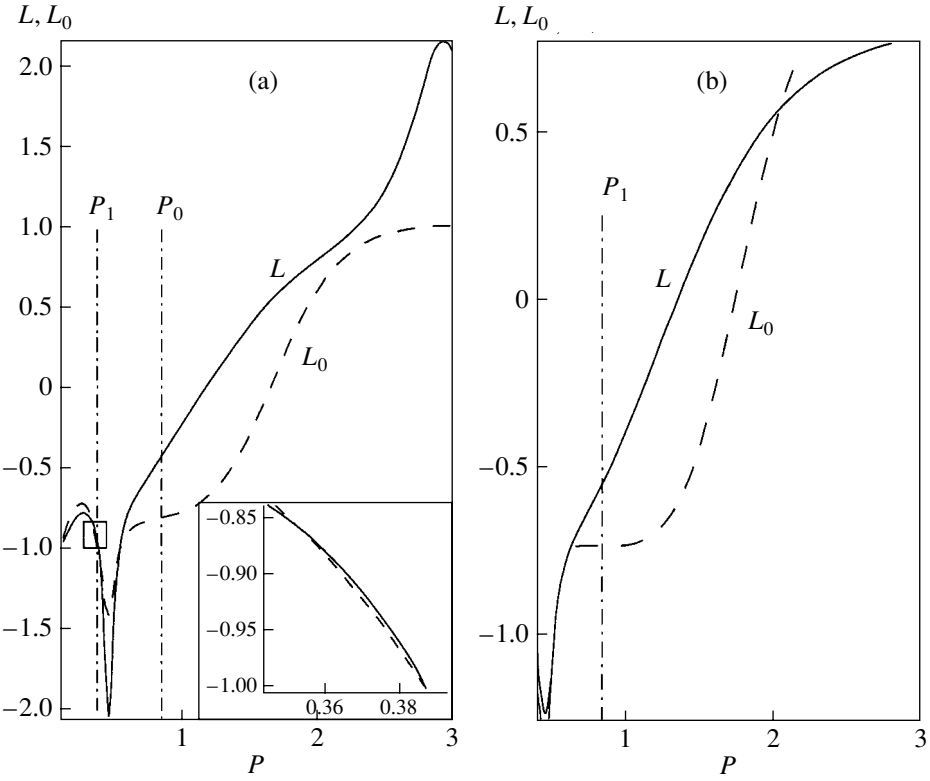


Fig. 6. D'yakov–Kontorovich stability parameter L and its critical value L_0 versus post-shock pressure P. Pressures behind (a) initial, leading, and (b) downstream shocks are denoted by P_0, P_1, and P_2, respectively.

improved by introducing additional diffusion, as in Harten entropy correction (1). A satisfactory solution was obtained near the upper shock-splitting boundary with an entropy correction parameter of 0.125ω, where ω is the spectral radius of the Jacobian matrix of the flux vector.

Figure 5 compares splitting diagrams computed for the model equation of state (see Fig. 3) by using different schemes with this value of the entropy correction parameter.

5. SIMULATION OF ACOUSTIC EMISSION UNDER SHOCK-SPLITTING CONDITIONS

In computations, acoustic emission can be observed if the post-shock state lies above point *4* on the Hugoniot curve in Fig. 1, during a limited time interval under shock-splitting conditions before the distance between split shocks exceeds the acoustic wavelength (if acoustic emission develops faster than does shock splitting), or when acoustic waves are emitted by the downstream shock in a split-wave configuration.

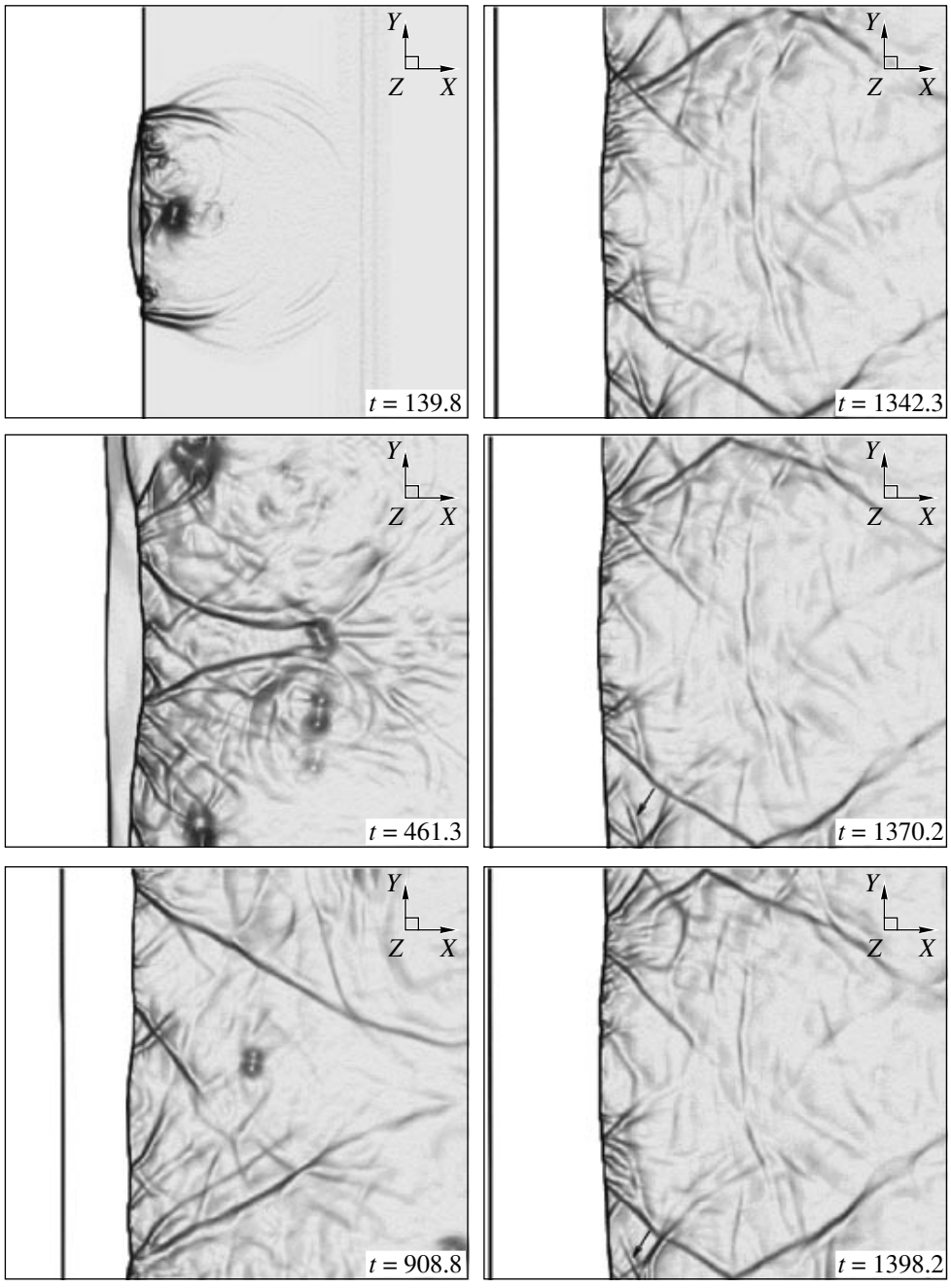


Fig. 7. Magnitude of pressure gradient. Darker areas correspond to steeper gradients.

Numerical results illustrating the last manifestation are presented here for an initial shock wave corresponding to the Hugoniot point at $P = 0.82$ in Fig. 3. At this point, the original wave simultaneously satisfies the Kontorovich criterion for acoustic emission and the shock-splitting conditions. Figure 6 shows the D'yakov–Kontorovich stability parameter L and its critical value L_0 for two Hugoniot curves: the original Hugoniot emanating from point 0 and the Hugoniot corresponding to the downstream shock. The pressures indicated by vertical lines in Fig. 6a correspond to the initial and leading shocks. The vertical line in Fig. 6b represents the pressure corresponding to the downstream shock. The figure demonstrates that both initial and downstream shocks satisfy the acoustic emission condition $L_0 < L < 1 + 2M$. The inset in the left panel shows that the equation $L(P) = L_0(P)$ has two nearly equal roots, which implies that there is a small region of acoustic emission below point *1* in Fig. 3a.

The splitting into stable and unstable shocks was simulated numerically by solving a two-dimensional Riemann problem on a 400×300 grid in a coordinate system tied to the initial shock front. Initial conditions were set to approximate an unstable shock-front fragment and adjoining uniform flow regions. The initial disturbance was a 1% drop in the velocity profile at the center of a cell contiguous to the shock front. No qualitative change in the ensuing flow pattern was induced by varying the perturbed velocity component or perturbation amplitude. The gray-scale plots of pressure-gradient magnitude shown in Fig. 7 illustrate the numerical solution obtained for $P = 0.82$ with the time step corresponding to a Courant number of 0.3 (darker areas correspond to steeper gradients). Time is measured here in arbitrary units since the system does not have any intrinsic time scale. In the three-wave configuration shown here, the cusps in the downstream shock front emit weak downstream-propagating acoustic waves and move along the front. The acoustic-emission parameters have the following values: $L = -0.59$, $M = 0.83$, and $\theta = 2.8$. The cosine of the angle between the acoustic wave vector and the positive x axis is determined by the equation [4]

$$M^2\left(\frac{4}{1+L} + \theta - 1\right)\cos^2\alpha + 2M\left(\frac{3+M^2}{1+L} - 1\right)\cos\alpha + \frac{2(1+M^2)}{1+L} - (1+\theta M^2) = 0,$$

whose roots are -1 and -0.66 in this particular case. The range of the latter root is $-M_2 < \cos^2\alpha < 1$, which corresponds to an outgoing acoustic wave with wave vector making an angle of $\pm 131°$ with the positive x axis. The snapshots shown for three successive instants in the right-hand part of Fig. 7 illustrate the propagation of acoustic wave fronts and demonstrate good agreement with the theoretical value of the angle.

6. CONCLUSIONS

The evolution of a shock wave into a split-wave configuration involving a stable shock and a shock emitting acoustic waves has been simulated for the first time under conditions of shock-wave nonuniqueness. The direction of the simulated acoustic wave propagation is consistent with predictions of the linear theory.

An analysis of scale-invariant solutions represented by splitting diagrams in the $(P, x/t)$ plane supports the theoretical results of [3] in terms of structure of solutions and boundaries for the split wave configurations $\vec{S} \longrightarrow \overleftarrow{W}\,T\vec{C}\,\vec{S}$, $\vec{S} \longrightarrow \overleftarrow{W}\,T\vec{S}\,\vec{C}\,\vec{S}$, and $\vec{S} \longrightarrow \overleftarrow{W}\,T\vec{S}\,\vec{S}$.

A numerical simulation of shock splitting performed with the use of generalized Roe's method (approximate solution of the Riemann problem for an arbitrary equation of state) has demonstrated that entropy correction is required to obtain physical solutions near the upper shock-splitting boundary and solutions satisfying the condition of entropy increase for split-wave configurations involving isentropic compression waves.

ACKNOWLEDGMENTS

This work was supported by Presidium of the Russian Academy of Sciences under the program "Mathematical Modeling, Intelligent Systems, and Control of Nonlinear Mechanical Systems."

REFERENCES

1. S. P. D'yakov, Zh. Éksp. Teor. Fiz. **27**, 288 (1954).
2. V. M. Kontorovich, Zh. Éksp. Teor. Fiz. **33**, 1525 (1957) [Sov. Phys. JETP **6**, 1179 (1957)].
3. N. M. Kuznetsov, Zh. Éksp. Teor. Fiz. **88**, 470 (1985) [Sov. Phys. JETP **61**, 275 (1985)].
4. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 6: *Fluid Mechanics*, 3rd ed. (Nauka, Moscow, 1986; Pergamon, New York, 1987).
5. A. L. Ni, S. G. Sugak, and V. E. Fortov, Teplofiz. Vys. Temp. **24**, 564 (1986).
6. V. A. Gushchin, A. P. Likhachev, N. G. Nechiporenko, and E. R. Pavlyukova, *News in Numerical Simulation: Algorithms, Computer Simulation, and Results* (Nauka, Moscow, 2000), p. 165.
7. J. W. Bates and D. C. Montgomery, Phys. Rev. Lett. **84**, 1180 (2000).
8. SESAME Report on the Los Alamos Equation-of-State Library (1983), Report LANL-83-4.
9. P. L. Roe, J. Comput. Phys. **43**, 357 (1981).
10. O. M. Belotserkovskiĭ and A. V. Konyukhov, Zh. Vychisl. Mat. Mat. Fiz. **42**, 235 (2002).
11. A. Harten, J. Comput. Phys. **49**, 357 (1983).
12. A. Harten and S. Osher, SIAM J. Numer. Anal. **24**, 279 (1987).
13. P. Glaister, J. Comput. Phys. **74**, 382 (1988).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 98, No. 4, 2004, pp. 820–836.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 125, No. 4, 2004, pp. 938–956.

MISCELLANEOUS

Dynamical Tunneling of Bound Systems through a Potential Barrier: Complex Way to the Top¶

F. Bezrukov[a] and D. Levkov[a,b]

[a]*Institute for Nuclear Research, Russian Academy of Sciences, Moscow, 117312 Russia*
[b]*Department of Physics, Moscow State University, Moscow, 119899 Russia*
e-mail: fedor@ms2.inr.ac.ru, levkov@ms2.inr.ac.ru

Received January 15, 2003

Abstract—A semiclassical method of complex trajectories for the calculation of the tunneling exponent in systems with many degrees of freedom is further developed. It is supplemented with an easily implemented technique that enables one to single out the physically relevant trajectory from the whole set of complex classical trajectories. The method is applied to semiclassical transitions of a bound system through a potential barrier. We find that the properties of physically relevant complex trajectories are qualitatively different in the cases of potential tunneling at low energy and dynamical tunneling at energies exceeding the barrier height. Namely, in the case of high energies, the physically relevant complex trajectories describe tunneling via creation of a state close to the top of the barrier. The method is checked against exact solutions of the Schrödinger equation in a quantum-mechanical system of two degrees of freedom. © *2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Semiclassical methods provide a useful tool for the study of nonperturbative processes. Tunneling phenomena represent one of the most notable cases where semiclassical techniques are used to obtain otherwise unattainable information on the dynamics of the transition. One standard semiclassical technique is the WKB approximation to tunneling in quantum mechanics of one degree of freedom. In this case, solutions $S(q)$ of the Hamilton–Jacobi equation are purely imaginary in the classically forbidden region. Therefore, the function $S(q)$ can be obtained as the action functional on a real trajectory $q(\tau)$, which is a solution to the equations to motion in the Euclidean time domain,

$$t = -i\tau,$$

with the real Euclidean action

$$S_E = -iS.$$

This simple picture of tunneling is no longer valid for systems with many degrees of freedom, where solutions $S(\mathbf{q})$ of the Hamilton–Jacobi equation are known to be generically complex in the classically forbidden region (see [1, 2] for recent discussion). This leads to the concept of "mixed" tunneling, as opposed to "pure" tunneling, where $S(\mathbf{q})$ is purely imaginary. Mixed tunneling cannot be described by any real tunneling trajectory. However, it can be related to a complex trajectory, in which case the function $S(\mathbf{q})$ (and, therefore, the exponential part of the wave function) is calculated as the action functional on this complex trajectory.

A particularly difficult situation arises when one considers transitions of a nonseparable system with a strong interaction between its degrees of freedom such that the quantum numbers of the system change considerably during the transition. Methods based on the adiabatic expansion are not applicable in this situation, while the method of complex trajectories proves to be extremely useful.

The method of complex trajectories in the form suitable for the calculation of S-matrix elements was formulated and checked by direct numerical calculations in [3–5] (see [6] for review). Further studies [7–12] showed that this method can be generalized to the calculation of the tunneling wave functions and tunneling probabilities, energy splitting in double-well potentials, and decay rates from metastable states. Similar methods were successful in the study of tunneling in high-energy collisions in field theory [13–16], where one considers systems with a definite particle number ($\mathcal{N} = 2$) in the initial state, and in the study of chemical reactions and atom ionization processes, where the initial bound systems are in definite quantum states [6, 17, 18], etc. The main advantage of the method of complex trajectories is that it can be easily generalized and numerically implemented in the cases of a large and even infinite (field theory) number of degrees of freedom, in contrast to other methods, such as the Huygens-type construction in [1, 2] and the initial value representation (IVR) in [3, 19–23].

¶This article was submitted by the authors in English.

In this paper, we develop the method of complex trajectories further. Namely, we concentrate on the following problem. It is known [3] that a physically relevant complex trajectory satisfies the classical equations of motion with certain boundary conditions. However, this boundary value problem generically has also an infinite, although discrete, set of unphysical solutions. In one-dimensional quantum mechanics, all solutions can easily be classified. In systems with many degrees of freedom, such a classification is extremely difficult, if at all possible. In the case of a small number of degrees of freedom (realistically, $N = 2$), one can scan over all solutions and find the solution giving the largest tunneling probability [3, 9, 10], but in systems with a large or infinite number of degrees of freedom, the problem of choosing the physically relevant solution becomes a formidable task.

The problem of choosing the appropriate solution becomes even more pronounced when the qualitative properties of the relevant complex trajectory are different in different energy regions. This may happen when the physically relevant classical solution "meets" an unphysical one at some energy value $E = E_1$, or in other words, when solutions of the boundary value problem, viewed as functions of the energy, bifurcate at $E = E_1$.

In this paper, we give an example of this kind, which appears to be fairly generic (see also [11, 12, 15, 16, 24]). We then develop a method that allows one to choose the physically relevant solution automatically, implement it numerically, and check this method against the numerical solution to the full Schrödinger equation.

We study inelastic transitions of a bound system through a potential barrier. To be specific, we consider a model with one internal degree of freedom in addition to the center-of-mass coordinate. We consider a situation in which the spacing between the levels of the bound system is small compared to the height of the barrier and assume a sufficiently strong coupling between the degrees of freedom to make sure that the quantum numbers of the bound system change considerably during the transition process. This is precisely the situation in which the method of complex trajectories shows its full strength.

Transitions of bound systems involve a particular energy scale, the barrier height V_0. At energies below V_0, classical overbarrier transitions are forbidden energetically; the corresponding regime is called "potential tunneling." For $E > V_0$, it is energetically allowed for the system to evolve classically to the other side of the barrier. However, overbarrier transitions may be forbidden dynamically even at $E > V_0$. Indeed, inelastic interactions of a bound system with a potential barrier generally lead to the excitation of the internal degrees of freedom with the simultaneous decrease of the center-of-mass energy, and this may prevent the system from the overbarrier transition. The tunneling regime at energies exceeding the barrier height is called "dynamical tunneling."[1]

Examples of dynamical tunneling are well known in scattering theory [4]. This type of tunneling between bound states was discovered in [25], and the generality of dynamical tunneling in large molecules was stressed in [26, 27]. Dynamical tunneling is of primary interest in our study.

We observe a novel phenomenon that dynamical tunneling at $E \gtrsim V_0$ (more precisely, at $E > E_1$, where E_1 is somewhat larger than V_0) occurs in the following way: the system jumps on top of the barrier and restarts its classical evolution from the region near the top. From the physical standpoint, this is not quite what is normally meant by "tunneling through a barrier." Yet, the transitions remain exponentially suppressed, but the reason is different: to jump above the barrier, the system has to undergo considerable rearrangement, unless the incoming state is chosen in a special way (see footnote 1). This rearrangement costs an exponentially small probability factor. We note that a similar exponential factor was argued to appear in various field theory processes with multiparticle final states [28–31].

We find that the new physical behavior of the system is related to a bifurcation of the family of complex-time classical solutions, viewed as functions of energy. This is precisely the bifurcation mentioned above. Our method for dealing with this bifurcation is to regularize the boundary value problem so that the bifurcations disappear altogether (at real energies), and the only solutions recovered after removing the regularization are physical ones.

This paper is organized as follows. The system to be discussed in what follows is introduced in Section 2.1. In Section 2.2, we formulate the boundary value problem for the calculation of the tunneling exponent. In Section 2.3, we then examine the classical overbarrier solutions and find all initial states that lead to classically allowed transitions. In Section 2.4, we present a straightforward application of the semiclassical technique outlined in Section 2.2 and find that it ceases to produce relevant complex trajectories in a certain region of initial data, namely, at $E > E_1$. In Section 3, we introduce our regularization technique and show that it indeed enables us to find all the relevant complex trajectories, including those with $E > E_1$ (Section 3.1). We check our method against the numerical solution of the full Schrödinger equation in Section 3.2. In Section 3.3 and Appendix C, we show how our regulariza-

[1] It is clear that the properties of transitions of a bound system at $E > V_0$ depend on the choice of the initial state. Namely, there always exists a certain class of states the transitions from which are not exponentially suppressed. To construct an example, one places the bound system on top of the barrier and evolves it classically backwards in time to the region where the interaction with the barrier is negligibly small. On the other hand, even at $E > V_0$, there are states the transitions from which are exponentially suppressed (dynamical tunneling).

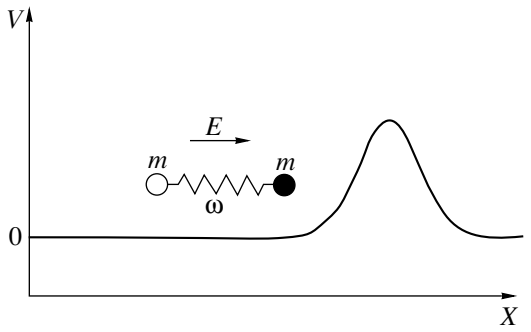


Fig. 1. An oscillator hitting a potential barrier, with only the "dark" particle interacting with the barrier.

tion technique is used to smoothly join the "classically allowed" and "classically forbidden" families of solutions in the respective cases of two- and one-dimensional quantum mechanics.

2. SEMICLASSICAL TRANSITIONS THROUGH A POTENTIAL BARRIER

2.1. Model

The situation discussed in this paper is a transition through a potential barrier of the bound system considered in [11, 12], namely, the system made of two particles of mass m, moving in one dimension and bound by a harmonic oscillator potential of frequency ω (Fig. 1). One of the particles interacts with a repulsive potential barrier. The potential barrier is assumed to be high and wide, while the spacing between the oscillator levels is much smaller than the barrier height V_0. The Hamiltonian of the model is

$$H = \frac{p_1^2}{2m} + \frac{p_2^2}{2m} + \frac{m\omega^2}{4}(x_1 - x_2)^2 + V_0 \exp\left(-\frac{x_1^2}{2\sigma^2}\right), \quad (1)$$

where the conditions on the oscillator frequency and potential barrier are

$$\hbar\omega \ll V_0, \qquad \sigma \gg \hbar/\sqrt{mV_0}. \quad (2)$$

Because the variables do not separate, this is certainly a nontrivial system.

We choose units with $\hbar = 1$, $m = 1$. It is also convenient to treat the frequency ω as a dimensionless parameter, so that all physical quantities are dimensionless. In our subsequent numerical study, we use the value $\omega = 0.5$ but keep the notation "ω" in formulas. The system is semiclassical, i.e., conditions (2) are satisfied, if we choose $\sigma = 1\sqrt{2\lambda}$ and $V_0 = 1/\lambda$, where λ is a small parameter. At the classical level, this parameter is irrelevant: after rescaling the variables[2] as

$$x_1 \longrightarrow x_1/\sqrt{\lambda}, \quad x_2 \longrightarrow x_2/\sqrt{\lambda},$$

the small parameter enters only through the overall multiplicative factor $1/\lambda$ in the Hamiltonian. Therefore, the semiclassical technique can be developed as an asymptotic expansion in λ.

The properties of the system are made clearer by replacing the variables x_1 and x_2 with the center-of-mass coordinate

$$X \equiv \frac{x_1 + x_2}{\sqrt{2}}$$

and the relative oscillator coordinate

$$y \equiv \frac{x_1 - x_2}{\sqrt{2}}.$$

In terms of these variables, the Hamiltonian becomes

$$H = \frac{p_X^2}{2} + \frac{p_y^2}{2} + \frac{\omega^2}{2}y^2 + \frac{1}{\lambda}\exp\left(-\frac{\lambda(X+y)^2}{2}\right). \quad (3)$$

The interaction potential

$$U_{\text{int}} \equiv \frac{1}{\lambda}\exp\left(-\frac{\lambda(X+y)^2}{2}\right)$$

vanishes in the asymptotic regions $X \longrightarrow \pm\infty$ and describes a potential barrier between these regions. At $X \longrightarrow \pm\infty$, Hamiltonian (3) corresponds to an oscillator of the frequency ω moving along the center-of-mass coordinate X. The oscillator asymptotic state is characterized by its excitation number N and total energy

$$E = \frac{p_X^2}{2} + \omega\left(N + \frac{1}{2}\right).$$

We are interested in the transmissions through the potential barrier of the oscillator with given initial values of E and N.

2.2. T/θ Boundary Value Problem

The probability of tunneling from a state with a fixed initial energy E and oscillator excitation number N from the asymptotic region $X \longrightarrow -\infty$ to any state in the other asymptotic region $X \longrightarrow +\infty$ is

$$\mathcal{T}(E, N) = \lim_{t_f - t_i \to \infty} \sum_f \left|\langle f|\exp(-i\hat{H}(t_f - t_i))|E, N\rangle\right|^2, \quad (4)$$

where it is implicit that the initial and final states have support only well outside the range of the potential,

[2] To keep the notation simple, we use the same symbols x_1, x_2 for the rescaled variables.

with $X < 0$ and $X > 0$, respectively. Semiclassical methods are applicable if the initial energy and excitation number are parametrically large,

$$E = \tilde{E}/\lambda, \quad N = \tilde{N}/\lambda,$$

where $\tilde{E}$ and $\tilde{N}$ are kept constant as $\lambda \longrightarrow 0$. The transition probability has the exponential form

$$\mathcal{T} = D\exp\left(-\frac{1}{\lambda}F(\tilde{E}, \tilde{N})\right), \tag{5}$$

where D is a preexponential factor, which is not considered in this paper. Our purpose is to calculate the leading semiclassical exponent $F(\tilde{E}, \tilde{N})$. The exponent for tunneling from the oscillator ground state is obtained in [11–13, 32] by taking the limit $\tilde{N} \longrightarrow 0$ in $F(\tilde{E}, \tilde{N})$.

In what follows, we rescale the variables as

$$X \longrightarrow X/\sqrt{\lambda}, \quad y \longrightarrow y/\sqrt{\lambda}$$

and omit the tilde over the rescaled quantities $\tilde{E}$ and $\tilde{N}$.

The exponent $F(E, N)$ is related to a complex trajectory that satisfies a certain complexified classical boundary value problem. We present the derivation of this problem in Appendix A. The outcome is as follows. There are two Lagrange multipliers T and θ, which are related to the parameters E and N characterizing the incoming state. The boundary value problem is conveniently formulated on the contour $ABCD$ in the complex time plane (see Fig. 2), with the imaginary part of the initial time equal to $T/2$. The coordinates $X(t)$ and $y(t)$ must satisfy the complexified equations of motion in the interior points of the contour and must be real in the asymptotic future (region D):

$$\frac{\delta S}{\delta X(t)} = \frac{\delta S}{\delta y(t)} = 0, \tag{6a}$$

$$\mathrm{Im}y(t) \longrightarrow 0, \quad \mathrm{Im}X(t) \longrightarrow 0, \quad \text{as } t \longrightarrow +\infty. \tag{6b}$$

In the asymptotic past (region A of the contour, where $t = t' + iT/2$, t' is real negative), the interaction potential U_{int} can be neglected and the oscillator decouples,

$$y = \frac{1}{\sqrt{2\omega}}(u\exp(-i\omega t') + v\exp(i\omega t')).$$

The boundary conditions in the asymptotic past, $t' \longrightarrow -\infty$, are that the center-of-mass coordinate X must be real, while the complex amplitudes of the decoupled oscillator must be linearly related,

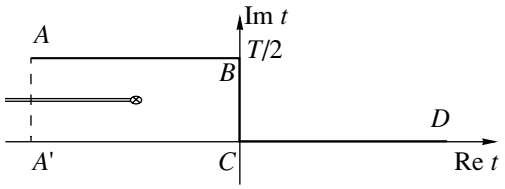


Fig. 2. Contour in the complex time plane.

$$\mathrm{Im}X \longrightarrow 0, \quad v \longrightarrow e^{\theta}u^*, \quad \text{as} \quad t \longrightarrow -\infty. \tag{6c}$$

Boundary conditions (6b) and (6c) in fact make eight real conditions (because, e.g., $\mathrm{Im}X(t') \longrightarrow 0$ implies that both $\mathrm{Im}X$ and $\mathrm{Im}\dot{X}$ tend to zero) and completely determine a solution, up to the time translation invariance (see the discussion in Appendix A).

It is shown in Appendix A that a solution of this boundary value problem is an extremum of the functional

$$F[X, y; X^*, y^*; T, \theta] = -iS[X, y] + iS[X^*, y^*] - ET - N\theta + \text{Boundary Terms}. \tag{7}$$

The value of this functional at the extremum gives the exponent for the transition probability (up to the large overall factor $1/\lambda$, see Eq. (5)),

$$F(E, N) = 2\mathrm{Im}S_0(T, \theta) - ET - N\theta, \tag{8}$$

where S_0 is the action of the solution, integrated by parts,

$$S_0 = \int dt\left(-\frac{1}{2}X\frac{d^2X}{dt^2} - \frac{1}{2}y\frac{d^2y}{dt^2} - \frac{1}{2}\omega^2y^2 - U_{\mathrm{int}}(X, y)\right). \tag{9}$$

Here, the integration runs along the contour $ABCD$. The values of the Lagrange multipliers T and θ are related to the energy and excitation number as

$$E(T, \theta) = \frac{\partial}{\partial T}2\mathrm{Im}S_0(T, \theta), \tag{10}$$

$$N(T, \theta) = \frac{\partial}{\partial \theta}2\mathrm{Im}S_0(T, \theta). \tag{11}$$

Using Eq. (8), it is also straightforward to verify the inverse Legendre transformation formulas

$$T(E, N) = -\frac{\partial}{\partial E}F(E, N), \tag{12}$$

$$\theta(E, N) = -\frac{\partial}{\partial N}F(E, N). \tag{13}$$

It can also be verified that the right-hand side of Eq. (10) coincides with the energy of the classical solution and the right-hand side of Eq. (11) is equal to the classical counterpart of the occupation number,

$$E = \frac{\dot{X}^2}{2} + \omega N, \quad N = uv. \tag{14}$$

Therefore, we can either seek the values of T and θ that correspond to given E and N or, following a computationally simpler procedure, solve boundary value problem (6) for given T and θ and then find the corresponding values of E and N from Eq. (14). We note that initial conditions (6c) complemented by Eqs. (14) are equivalent to the initial conditions in [3–5], the latter being expressed in terms of action–angle variables. The boundary conditions in the asymptotic future (6b) are different from those in [3–5], because we consider an inclusive, rather than fixed, final state.

We now discuss some subtle points of boundary value problem (6). First, we note that the asymptotic reality condition in (6b) does not always coincide with the reality condition at finite time. Of course, if the solution approaches the asymptotic region $X \longrightarrow +\infty$ on the part CD of the contour, asymptotic reality condition (6b) implies that the solution is real at any finite positive t. Indeed, the oscillator decouples as $X \longrightarrow +\infty$, and, therefore, condition (6b) means that its phase and amplitude, as well as $X(t)$, are real as $t \longrightarrow +\infty$. Due to the equations of motion, $X(t)$ and $y(t)$ are real on the entire CD part of the contour. This situation corresponds to the transition directly to the asymptotic region $X \longrightarrow +\infty$. However, the situation can be drastically different if the solution on the final part of the time contour remains in the interaction region. For example, we can imagine that the solution approaches the saddle point of the potential $X = 0$, $y = 0$ as $t \longrightarrow +\infty$. Because one of the perturbations around this point is unstable, there may exist solutions that approach this point exponentially along the unstable direction, i.e.,

$$X(t), y(t) \propto \exp(-\text{const} \cdot t)$$

with possibly complex prefactors. In this case, the solution may be complex at any finite time and become real only asymptotically as $t \longrightarrow +\infty$. Such a solution corresponds to tunneling to the saddle point of the barrier, after which the system rolls down classically towards $X \longrightarrow +\infty$ (with probability of order 1, inessential for the tunneling exponent F). We see in Section 3.1 that a situation of this sort indeed takes place for some values of the energy and excitation number.

Second, because the interaction potential disappears at large negative time (in the asymptotic region $X \longrightarrow -\infty$), it is straightforward to continue the asymptotic the solution to the real time axis. For solutions satisfying (6c), this gives

$$y(t) = \frac{1}{\sqrt{2\omega}}\left(u\exp\left(-\frac{\omega T}{2}\right)\exp(-i\omega t) + u^*\exp\left(\theta + \frac{\omega T}{2}\right)\exp(i\omega t)\right),$$

$$\mathrm{Im}X(t) = -\frac{T}{2}p_X$$

at large negative time. We see that the dynamical coordinates on the negative side of the real time axis are generally complex. For solutions approaching the asymptotic region $X \longrightarrow +\infty$ as $t \longrightarrow +\infty$ (such that X and y are exactly real at finite $t > 0$), this means that there should exist a branch point in the complex time plane: the contour $A'ABC$ in Fig. 2 winds around this point and cannot be deformed to the real time axis. This argument does not work for solutions ending in the interaction region as $t \longrightarrow +\infty$, and, hence, branch points between the AB part of the contour and the real time axis may be absent. We see in Section 3.1 that this is indeed the case in our model in a certain range of E and N.

2.3. Overbarrier Transitions: the Region of Classically Allowed Transitions and Its Boundary $E_0(N)$

Before studying the exponentially suppressed transitions, we consider the classically allowed ones. For this, we study the classical evolution (real time, real-valued coordinates) in which the system is initially located at large negative X and moves with a positive center-of-mass velocity towards the asymptotic region $X \longrightarrow +\infty$. The classical dynamics of the system is specified by four initial parameters. One of them (e.g., the initial center-of-mass coordinate) fixes the invariance under time translation, while the other three are the total energy E; the initial excitation number of the y oscillator, defined in classical theory as $N \equiv E_{\mathrm{osc}}/\omega$; and the initial oscillator phase φ_i.

Any initial quantum state of our system can be fully determined by the energy E and the initial oscillator excitation number N; we can represent each state by a point in the EN plane. There is, however, one additional classically relevant initial parameter, the oscillator phase φ_i. An initial state (E, N) leads to unsuppressed transmission if the corresponding classical overbarrier transitions[3] are possible for some value(s) of φ_i. These

[3] We note that the corresponding classical solutions obey boundary conditions (6b) and (6c) with $T = \theta = 0$; i.e., they are solutions to boundary value problem (6).

states form some region in the EN plane, which is to be found in this section.

For given N, at sufficiently large E, the system can certainly evolve to the other side of the barrier. On the other hand, if E is smaller than the barrier height, the system definitely undergoes reflection. Thus, there exists some boundary energy $E_0(N)$ such that classical transitions are possible for $E > E_0(N)$, while, for $E < E_0(N)$, they do not occur for any initial phase φ_i. The line $E_0(N)$ represents the boundary of the region of classically allowed transitions. We have calculated $E_0(N)$ numerically: the result[4] is shown in Fig. 3.

An important point of the boundary $E_0(N)$ corresponds to the static unstable classical solution $X(t) = y(t) = 0$. In the field theory context, such a solution is called "sphaleron" [33], and we keep this terminology in what follows. This solution is the saddle point of the potential

$$U(X, y) \equiv \omega^2 y^2/2 + U_{\text{int}}(X, y)$$

and has exactly one unstable direction, the negative mode (see Fig. 4). The sphaleron energy $E_S = U(0, 0) = 1$ determines the minimum value of the function $E_0(N)$. Indeed, classical overbarrier transitions with $E < E_S$ are impossible, but the overbarrier solution with a slightly higher energy can be obtained as follows: a momentum along the negative mode is added at the point $X = y = 0$, "pushing" the system towards $X \longrightarrow +\infty$. Continuing this solution backwards in time shows that the system tends to $X \longrightarrow -\infty$ for large negative time and has a certain oscillator excitation number. Solutions with the energy closer to the sphaleron energy correspond to a smaller "push" and, thus, spend a longer time near the sphaleron. In the limiting case where the energy is equal to E_S, the solution spends an infinite time in the vicinity of the sphaleron. This limiting case has a definite initial excitation number N_S, so that $E_0(N_S) = E_S$ (see Fig. 3). The value of N_S is unique because there is exactly one negative direction of the potential in the vicinity of the sphaleron.

In complete analogy to the features of the overbarrier classical solutions near the sphaleron point (E_S, N_S), we expect that, as the values of E and N approach any other boundary point $(E_0(N), N)$, the corresponding overbarrier solutions spend more and more time in the interaction region, where $U_{\text{int}} \neq 0$. This follows from a continuity argument. Namely, we first fix the initial and final times, t_i and t_f. If, within this time interval, a solution with the energy E_1 evolves to the other side of the barrier and a solution with the energy E_2 and the same oscillator excitation number is reflected, there exists an intermediate energy at which the solution ends up at $t = t_f$ in the interaction region. Taking the limit as $t_f \longrightarrow +\infty$ and $E_1 - E_2 \longrightarrow 0$, we obtain a point at the boundary $E_0(N)$ and a solution tending asymptotically to some unstable time-dependent solution that spends an infinite time in the interaction region. We call the latter solution the excited sphaleron; it describes some (in general, nonlinear) oscillations above the sphaleron along the stable direction in the coordinate space. Therefore, every point of the boundary $(E_0(N), N)$ corresponds to some excited sphaleron. In the phase space, solutions tending asymptotically to the excited sphalerons form a surface (separatrix) that separates regions of qualitatively different classical motions of the system.

In Fig. 4, we display a solution found numerically in our model that tends to an excited sphaleron. We see that the trajectory of the excited sphaleron is, roughly speaking, orthogonal to the unstable direction at the saddle point $(X = 0, y = 0)$.

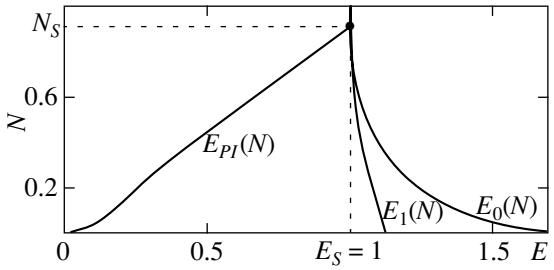


Fig. 3. The boundary $E_0(N)$ of the region of classically allowed transitions, the bifurcation line $E_1(N)$, and the line of the periodic instantons $E_{PI}(N)$.

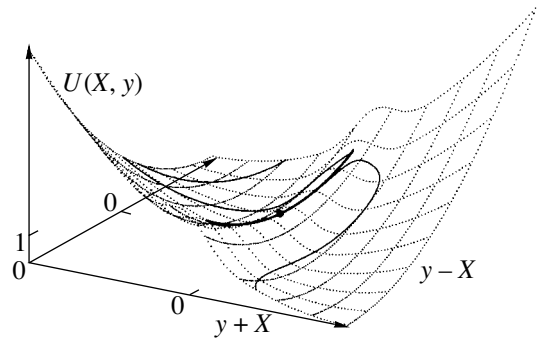


Fig. 4. The potential (dotted lines) in the vicinity of the sphaleron ($X = 0$, $y = 0$) (marked by the point), the excited sphaleron (thick line) corresponding to the point (E, N) = (1.985, 3.72) at the boundary of the region of classically allowed transitions, and the trajectory of a solution that is close to this excited sphaleron (thin line). The asymptotic regions $X \longrightarrow \pm\infty$ are along the diagonal.

[4] We note that the boundary $E_0(N)$ of the region of classically allowed transitions can be extended to $N > N_S$. Because $E = E_S$ is the absolute minimum of the energy of classically allowed transitions, the function $E_0(N)$ grows with N at $N > N_S$. In fact, it tends to the asymptotics $E_0^{as} = \omega N$ as $N \longrightarrow +\infty$. In what follows, we are not interested in transitions with $N > N_S$, and, therefore, this part of the boundary $E_0(N)$ is not shown in Fig. 3.

2.4. Suppressed Transitions: Bifurcation Line $E_1(N)$

We now turn to classically forbidden transitions and consider the boundary value problem in Eq. (6). It is relatively straightforward to obtain solutions for $\theta = 0$ numerically. In this case, boundary conditions (6b) and (6c) take the form of reality conditions in the asymptotic future and past. It can be shown [34] that the physically relevant solutions with $\theta = 0$ are real on the entire contour *ABCD* in Fig. 2 and describe nonlinear oscillations in the upside-down potential on the Euclidean part *BC* of the contour. The period of the oscillations is equal to *T*, and, hence, points *B* and *C* are two different turning points where $\dot{X} = \dot{y} = 0$. These real Euclidean solutions are called periodic instantons. A practical technique for obtaining these solutions numerically on the Euclidean part *BC* consists in minimizing the Euclidean action (for example, with the method of conjugate gradients, see [11, 12] for details). The solutions on the entire contour are then obtained by solving the Cauchy problem numerically, forward in time along the line *CD* and backward in time along the line *BA*. From the solution in the asymptotic past (region *A*), we then calculate its energy and excitation number (14). The solutions of this Cauchy problem are obviously real, and, hence, boundary conditions (6b) and (6c) are indeed satisfied for $\theta = 0$. It is worth noting that solutions with $\theta = 0$ are similar to the ones in quantum mechanics of one degree of freedom. The line of periodic instantons in the *EN* plane in our model is shown in Fig. 3.

Once the solutions with $\theta = 0$ are found, it is natural to try to cover the entire region of classically forbidden transitions in the *EN* plane with a deformation procedure, by moving in small steps in θ and *T*. The solution of the boundary value problem with $(T + \Delta T, \theta + \Delta\theta)$ may be obtained numerically by applying an iteration technique, with the known solution at (T, θ) serving as the initial approximation.[5] If the solutions end up in the correct asymptotic region at each step, i.e., $X \longrightarrow +\infty$ on part *D* of the contour, the solutions obtained by this procedure of small deformations are physically relevant. But the method of small deformations fails to produce relevant solution if there are bifurcation points in the *EN* plane, where the physical branch of solutions merges to an unphysical branch. Because there are unphysical solutions close to physical ones in the vicinity of bifurcation points, the procedure of small deformations cannot be used near these points.

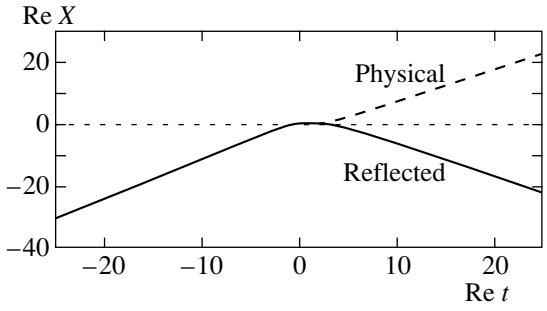


Fig. 5. The dependence of the tunneling coordinate *X* on time for two solutions with nearly the same energy and initial excitation number. The physical solution tunnels to the asymptotic region $X \longrightarrow +\infty$, while the unphysical one is reflected to $X \longrightarrow -\infty$. The physical solution has $E = 1.028$, $N = 0.44$, while the unphysical one has $E = 1.034$, $N = 0.44$. These two solutions are close to the point on the bifurcation line $E_1(N = 0.44) = 1.031$.

We have found numerically that, in our model, the method of small deformations produces correct solutions of the T/θ boundary value problem in a large region of the *EN* plane where $E < E_1(N)$. However, at sufficiently high energy $E > E_1(N)$, where $E_1(N) \gtrsim E_S$, the deformation procedure generates solutions that bounce back from the barrier (see Fig. 5), i.e., have a wrong "topology." This occurs deep inside the region of classically forbidden transitions, where the suppression is large, and one naively expects the semiclassical technique to work well. Clearly, solutions with a wrong topology do not describe the tunneling transitions of interest. Therefore, if the semiclassical method is applicable in the region $E_1(N) < E < E_0(N)$ at all, there exists another, physical, branch of solutions. In that case, the line $E_1(N)$ is the bifurcation line where the physical solutions meet the ones with a wrong topology. Walking in small steps in θ and *T* is useless in the vicinity of this bifurcation line, and a special trick is required to find the relevant solutions beyond that line. The bifurcation line $E_1(N)$ for our quantum-mechanical problem of two degrees of freedom is shown in Fig. 3.

The loss of topology beyond a certain bifurcation line in the *EN* plane is by no means a property of our model only. This phenomenon has been observed in field theory models in the context of both induced false vacuum decay [14] and baryon-number violating transitions in gauge theory [15] (in field theory models, the parameter *N* is the number of incoming particles). In all cases, the loss of topology prevented one from computing the semiclassical exponent for the transition probability in the interesting region of relatively high energies.

Returning to quantum mechanics of two degrees of freedom, we point out that the properties of tunneling solutions with different energies approaching the bifurcation line $E_1(N)$ from the left of the *EN* plane are in some sense similar to the properties of tunneling solutions in one-dimensional quantum mechanics whose energy is close to the barrier height (see Appendix C). Again by continuity, these solutions of our two-dimensional model spend a long time in the interaction region; this time tends to infinity on the line $E_1(N)$.

[5] In practice, the Newton–Raphson method is particularly convenient (see [11, 12, 14, 15]).

Hence, at any point of this line, there is a solution that starts in the asymptotic region left of the barrier and ends up on an excited sphaleron. Such behavior is indeed possible because of the existence of an unstable direction near the (excited) sphaleron, even for complex initial data. In the next section, we suggest a trick to deal with this situation—this is our regularization technique.

3. REGULARIZATION TECHNIQUE

In this section, we develop our regularization technique and find the physically relevant solutions between the lines $E_1(N)$ and $E_0(N)$. We see that all solutions from the new branch (and not only on the lines $E_0(N)$ and $E_1(N)$) correspond to tunneling onto the excited sphaleron ("tunneling on top of the barrier"). These solutions would be very difficult, if at all possible, to obtain directly by numerically solving the nonregularized classical boundary value problem (6): they are complex at finite times and become real only asymptotically as $t \longrightarrow +\infty$, whereas numerical methods require working with finite time intervals.

As an additional advantage, our regularization technique allows one to obtain a family of overbarrier solutions that covers all the region of the initial data corresponding to classically allowed transitions, including its boundary. This is of interest in models with a large number of degrees of freedom and in field theory, where finding the boundary $E_0(N)$ by direct methods is difficult (see, e.g., [35] for a discussion of this point).

3.1. Regularized Problem: Classically Forbidden Transitions

The main idea of our method is to regularize the equations of motion by adding a term proportional to a small parameter ϵ such that configurations staying near the sphaleron for an infinite time no longer exist among the solutions of the T/θ boundary value problem. After performing the regularization, we explore all the region of classically forbidden transitions without crossing the bifurcation line. Taking the limit $\epsilon \longrightarrow 0$, we then reconstruct the correct values of F, E, and N.

In formulating the regularization technique, it is more convenient to work with the functional $F[X, y; X^*, y^*; T, \theta]$, Eq. (7), itself rather than with the equations of motion. We prevent F from being extremized by configurations approaching the excited sphalerons asymptotically. To achieve this, we add a new term of the form $2\epsilon T_{\text{int}}$ to the original functional (7), where T_{int} estimates the time the solution "spends" in the interaction region. The regularization parameter ϵ is the smallest one in the problem, and, hence, any regular extremum of the functional F (the solution that spends a finite time in the region $U_{\text{int}} \neq 0$) changes only slightly after the regularization. At the same time, the excited sphaleron configuration has $T_{\text{int}} = \infty$, which leads to the infinite value of the regularized functional

$$F_\epsilon \equiv F + 2\epsilon T_{\text{int}}.$$

Hence, the excited sphalerons are not stationary points of the regularized functional.

For the problem under consideration, $U_{\text{int}} \sim 1$ in the interaction region and T_{int} can be defined as

$$T_{\text{int}} = \frac{1}{2}\left[\int dt U_{\text{int}}(X, y) + \int dt U_{\text{int}}(X^*, y^*)\right]. \quad (15)$$

We note that T_{int} is real and that the regularization is equivalent to the multiplication of the interaction potential with a complex factor,

$$U_{\text{int}} \longrightarrow (1 - i\epsilon)U_{\text{int}} = e^{-i\epsilon}U_{\text{int}} + O(\epsilon^2). \quad (16)$$

This results in the corresponding change of the classical equations of motion, while boundary conditions (6b) and (6c) remain unaltered.

We still have to understand whether solutions with $\epsilon \neq 0$ exist at all. The reason for the existence of such solutions is as follows. We consider a well-defined (for $\epsilon > 0$) matrix element

$$\mathcal{T}_\epsilon = \lim_{t_f - t_i \to \infty} \sum_f \left|\langle f| \exp[(-i\hat{H} - \epsilon U_{\text{int}})(t_f - t_i)]|E, N\rangle\right|^2,$$

where, as before, $|E, N\rangle$ denotes the incoming state with given energy and oscillator excitation number. The quantity $\mathcal{T}_\epsilon$ has a well-defined limit as $\epsilon \longrightarrow 0$, equal to tunneling probability (4). Because the saddle point of the regularized functional F_ϵ gives the semiclassical exponent for the quantity $\mathcal{T}_\epsilon$, we expect that such a saddle point indeed exists.

Therefore, the regularized T/θ boundary value problem is expected to have solutions necessarily spending a finite time in the interaction region. By continuity, these solutions do not experience reflection from the barrier if the procedure of small deformations starting from solutions with the correct topology is used. The line $E_1(N)$ is no longer a bifurcation line of the regularized system, and the procedure of small deformations therefore enables us to cover the entire region of classically forbidden transitions. The semiclassical suppression factor of the original problem is recovered in the limit $\epsilon \longrightarrow 0$.

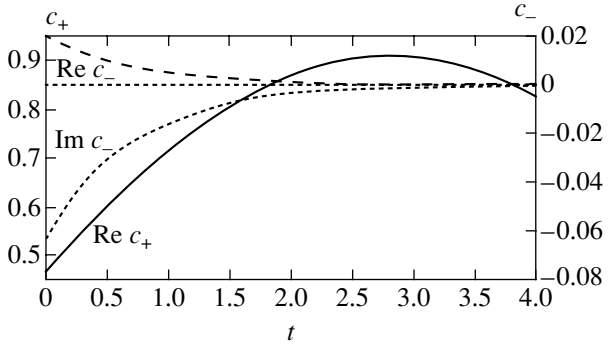


Fig. 6. Large-time behavior of a solution with $\epsilon = 0$ at ($E = 1.05$, $N = 0.43$). The coordinates X and y are decomposed on the basis of the eigenmodes near the sphaleron. We note that $\mathrm{Im} c_+ = 0$.

It is worth noting that the interaction time is Legendre conjugate to ϵ,

$$T_{\mathrm{int}} = \frac{1}{2}\frac{\partial}{\partial\epsilon}F_\epsilon(E, N, \epsilon). \tag{17}$$

This equation can be used as a check of numerical calculations.

We implemented the regularization procedure numerically. To solve the boundary value problem, we use the computational methods described in [11, 12]. To obtain the semiclassical tunneling exponent in the region between the bifurcation line $E_1(N)$ and the boundary of the region of classically allowed transitions $E_0(N)$, we began with a solution to the nonregularized problem deep in the "forbidden" region of the initial data (i.e., at $E < E_1(N)$). We then increased the value of ϵ from zero to a certain small positive number, keeping T and θ fixed. We next changed T and θ in small steps, keeping ϵ finite, and found solutions to the regularized problem in the region $E_1(N) < E < E_0(N)$. These solutions had the correct topology; i.e., they indeed ended up in the asymptotic region $X \longrightarrow +\infty$. Finally, we lowered ϵ and extrapolated F, E, and N to the limit $\epsilon \longrightarrow 0$.

We now consider the solutions in the region $E_1(N) < E < E_0(N)$, which we obtain in the limit $\epsilon \longrightarrow 0$ more carefully. They belong to a new branch and may, therefore, exhibit new physical properties. Indeed, we found that, as the value of ϵ decreases to zero, the solution at any point (E, N) with $E_1(N) < E < E_0(N)$ spends more and more time in the interaction region. The limiting solution corresponding to $\epsilon = 0$ has infinite interaction time: in other words, as $t \longrightarrow +\infty$, it tends to one of the excited sphalerons. The resulting physical picture is that, at a sufficiently large energy (i.e., at $E > E_1(N)$), the system prefers to tunnel exactly onto an unstable classical solution, excited sphaleron, that oscillates about the top of the potential barrier. To demonstrate this, we have plotted in Fig. 6 the solution $\mathbf{x}(t) \equiv (X(t), y(t))$ at large times after taking the limit $\epsilon \longrightarrow 0$ numerically. To understand this figure, we recall that the potential near the sphaleron point $X = y = 0$ has one positive mode and one negative mode. Namely, introducing new coordinates c_+, c_- as

$$X = \cos\alpha c_+ + \sin\alpha c_-,$$
$$y = -\sin\alpha c_+ + \cos\alpha c_-,$$
$$\cot 2\alpha = -\frac{\omega^2}{2},$$

we write, in the vicinity of the sphaleron,

$$H = 1 + \frac{p_+^2}{2} + \frac{p_-^2}{2} + \frac{\omega_+^2}{2}c_+^2 - \frac{\omega_-^2}{2}c_-^2,$$

where

$$\omega_\pm^2 = \pm\left(-1 + \frac{\omega^2}{2}\right) + \sqrt{1 + \frac{\omega^4}{4}} > 0.$$

Because the solutions of the T/θ boundary value problem are complex, the coordinates c_+ and c_- are also complex. In Fig. 6, we show real and imaginary parts of c_+ and c_- at a large real time t (part CD of the contour). We see that, while $\mathrm{Re} c_+$ oscillates, the unstable coordinate c_- asymptotically approaches the sphaleron value: $c_- \longrightarrow 0$ as $t \longrightarrow +\infty$. The imaginary part of c_- is nonzero at any finite time. This is the reason for the failure of straightforward numerical methods in the region $E > E_1(N)$: the solutions from the physical branch do not satisfy the reality conditions at any large but finite final time. We have pointed out in Section 2.2 that this can happen only if the solution ends up near the sphaleron, which has a negative mode. This is precisely what happens: for $\epsilon = 0$ at asymptotically large t, our solutions are real and oscillate near the sphaleron, remaining in the interaction region.

3.2. Regularization Technique versus Exact Quantum-Mechanical Solution

Quantum mechanics of two degrees of freedom is a convenient testing ground for checking the semiclassical methods and, in particular, our regularization technique. We have found solutions to the full stationary Schrödinger equation and exact tunneling probability $\mathcal{T}$ by applying the numerical technique in [11, 12]. Our numerical calculations were performed for several small values of the semiclassical parameter λ, namely, for $\lambda = 0.01$–0.1. Transitions through the barrier for these values of the semiclassical parameter are well suppressed. In particular, for $\lambda = 0.02$, the tunneling probability $\mathcal{T}$ is of the order e^{-14}. To check the semi-

classical result with better precision, we have calculated the exact suppression exponent

$$F_{QM}(\lambda) \equiv -\lambda \log \mathcal{T}$$

(cf. (5)) for $\lambda = 0.09$, 0.05, 0.03, and 0.02 and extrapolated F_{QM} to $\lambda = 0$ by polynomials of the third and fourth degree. The extrapolation results are independent of the degree (3 or 4) of polynomials with a precision of 1%. The extrapolated suppression exponent $F_{QM}(0)$ corresponds to infinite suppression and must coincide exactly (up to numerical errors) with the correct semiclassical result.

We performed this check in the region $E > E_S = 1$, which is most interesting for our purposes. The results of the full quantum-mechanical calculation of the suppression exponent F_{QM} in the limit $\lambda \longrightarrow 0$ are represented by points in Fig. 7. The lines in that figure represent the values of the semiclassical exponent $F(E, N)$ for constant N, which we obtain in the limit $\epsilon \longrightarrow 0$ of the regularization procedure. In practice, instead of taking the limit $\epsilon \longrightarrow 0$, we calculate the regularized functional

$$F_\epsilon(E, N) = F(E, N) + O(\epsilon)$$

for sufficiently small ϵ. We used the value $\epsilon = 10^{-6}$, and the value of the suppression exponent was then found with a precision of the order 10^{-5}. We see that, in the entire region of classically forbidden transitions (including the region $E > E_1(N)$), the semiclassical result for F coincides with the exact one.

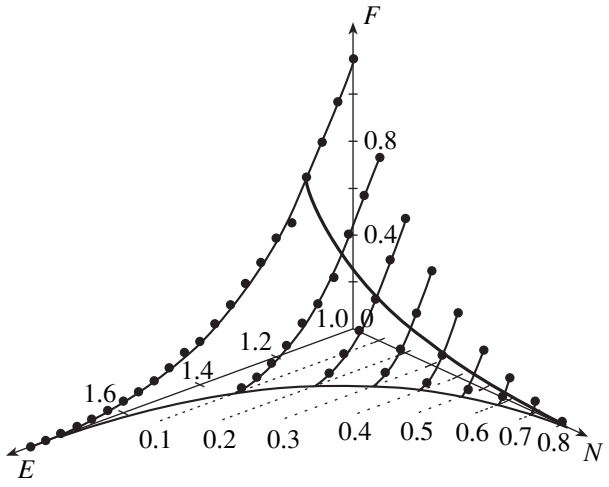


Fig. 7. The tunneling exponent $F(E, N)$ in the region $E > E_S = 1$. The lines show the semiclassical results, and the dots represent exact ones, obtained by solving the Schrödinger equation. The lines across the plot are the boundary of the region of classically allowed transitions $E_0(N)$ and the bifurcation line $E_1(N)$.

3.3. Classically Allowed Transitions

We now show that our regularization procedure allows one to obtain a subset of classical overbarrier solutions existing at sufficiently high energies. This subset is interesting because it extends to the boundary of the region of classically allowed transitions, $E = E_0(N)$. In principle, finding this boundary is purely a problem of classical mechanics; indeed, in mechanics of two degrees of freedom, this boundary can be found numerically by solving the Cauchy problem for given E and N and all possible oscillator phases (see Section 2.3). However, if the number of degrees of freedom is much larger, this classical problem becomes quite complicated, because a high-dimensional space of Cauchy data has to be spanned. As an example, a stochastic Monte Carlo technique was developed in [35] to deal with this problem in the field theory context. The approach below is an alternative to the Cauchy methods.

We first recall that all classical overbarrier solutions with given energy and excitation number satisfy the T/θ boundary value problem with $T = 0$, $\theta = 0$. We cannot reach the "allowed" region of the EN plane without regularization, because we have to cross the line $E_0(N)$ corresponding to the excited sphaleron configurations in the final state. But the excited sphalerons no longer exist among the solutions of the regularized boundary value problem at any finite value of ϵ. This suggests that the regularization makes it possible to enter the region of classically allowed transitions and, after taking an appropriate limit, obtain classical solutions with finite values of E and N.

By definition, the classically allowed transitions have $F = 0$. We therefore expect that, in the "allowed" region of initial data, the regularized problem has the property that

$$F_\epsilon(E, N) = \epsilon f(E, N) + O(\epsilon^2).$$

In view of inverse Legendre formulas (12) and (13), the values of T and θ must be of order ϵ,

$$T = \epsilon \tau(E, N), \quad \theta = \epsilon \vartheta(E, N),$$

where the quantities τ and ϑ are related to the initial energy and excitation number (see Eqs. (12), (13)) as

$$\tau = -\lim_{\epsilon \to 0} \frac{\partial}{\partial E} \frac{F_\epsilon}{\epsilon} = -\frac{1}{2} \frac{\partial}{\partial E} T_{\mathrm{int}}(E, N), \tag{18}$$

$$\vartheta = -\lim_{\epsilon \to 0} \frac{\partial}{\partial N} \frac{F_\epsilon}{\epsilon} = -\frac{1}{2} \frac{\partial}{\partial N} T_{\mathrm{int}}(E, N), \tag{19}$$

where we have used Eq. (17). We thus expect that the region of classically allowed transitions can be invaded by taking a fairly sophisticated limit $\epsilon \longrightarrow 0$ with $\tau \equiv T/\epsilon = \text{const}$, $\vartheta \equiv \theta/\epsilon = \text{const}$. For the allowed transitions, the parameters τ and ϑ are analogous to T and θ.

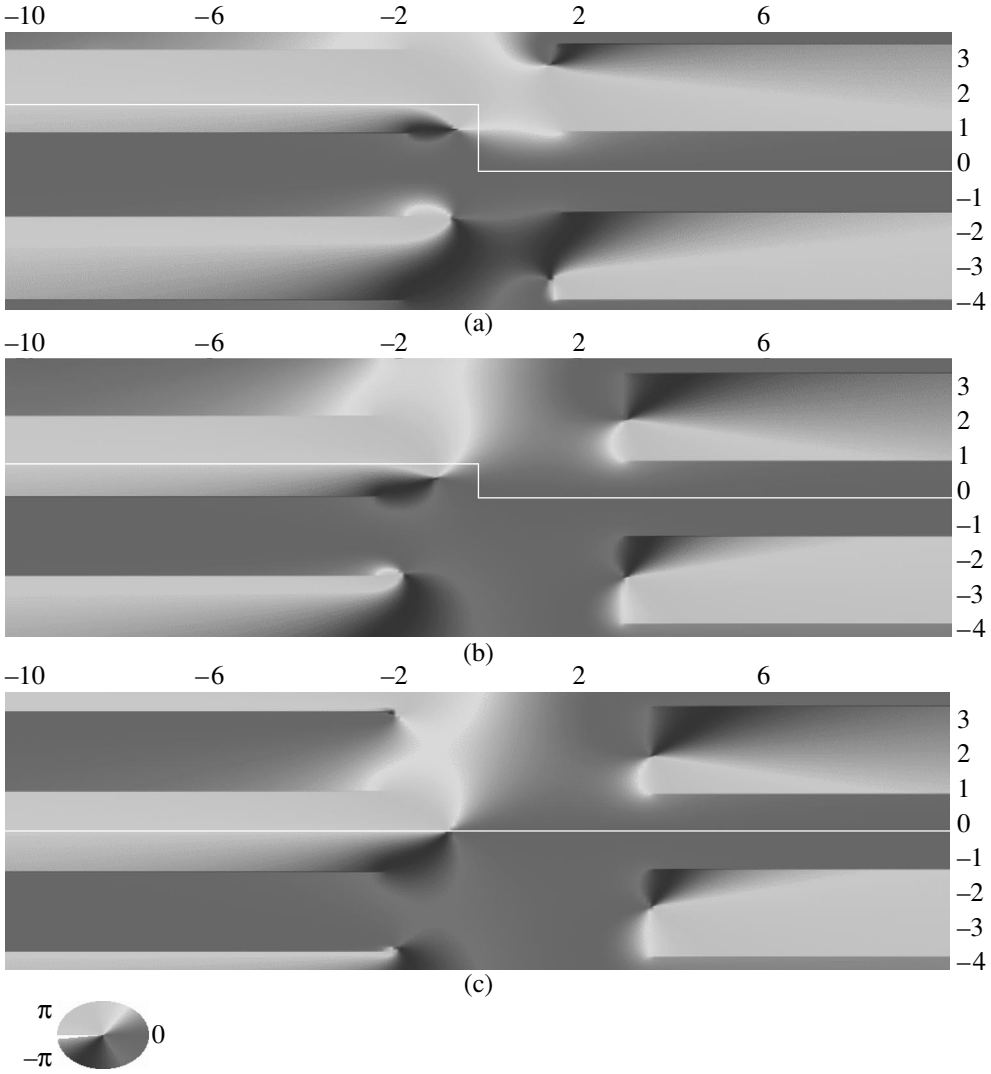


Fig. 8. The phase of the tunneling coordinate in the complex time plane at three points of the curve $\tau = 380$, $\vartheta = 130$. Figures 8a, 8b, and 8c correspond to $\epsilon = \epsilon_a = 0.01$, $\epsilon = \epsilon_b = 0.0048$, and $\epsilon = \epsilon_c = 0$, respectively. The asymptotics for $X \longrightarrow -\infty$ and $X \longrightarrow +\infty$ correspond to $\arg X = \pi$ and 0. The contour in the time plane is plotted with the white line.

Solving the regularized T/θ boundary value problem allows one to construct a single solution for given E and N. On the other hand, for $\epsilon = 0$, there are more classical overbarrier solutions: they form a continuous family labeled by the initial oscillator phase. Thus, taking the limit $\epsilon \longrightarrow 0$ gives a subset of overbarrier solutions, which should therefore obey some additional constraint. It is almost obvious that this constraint is that the interaction time T_{int} (Eq. (15)) is minimal. This is shown in Appendix B.

The subset of classical overbarrier solutions obtained in the $\epsilon \longrightarrow 0$ limit of the regularized T/θ procedure extends to the boundary of the region of classically allowed transitions. We now consider what happens as this boundary is approached from the "classically allowed" side. At the boundary $E_0(N)$, the unregularized solutions tend to excited sphalerons, and the interaction time T_{int} is, therefore, infinite. This is consistent with (18) and (19) only if τ and ϑ become infinite at the boundary. Hence, to obtain a point of the boundary, we take the further limit,

$$(E_0(N), N) = \lim_{\substack{\tau/\vartheta = \text{const} \\ \tau \to +\infty}} (E(\tau, \vartheta), N(\tau, \vartheta)).$$

Different values of τ/ϑ correspond to different points of the line $E_0(N)$. We thus find the boundary of the region of classically allowed transitions without initial-state simulation.

We have checked this procedure numerically. The limit $\epsilon \longrightarrow 0$ indeed exists—the values of E and N tend to the point of the EN plane that corresponds to the classically allowed transition. The phase of the tunneling

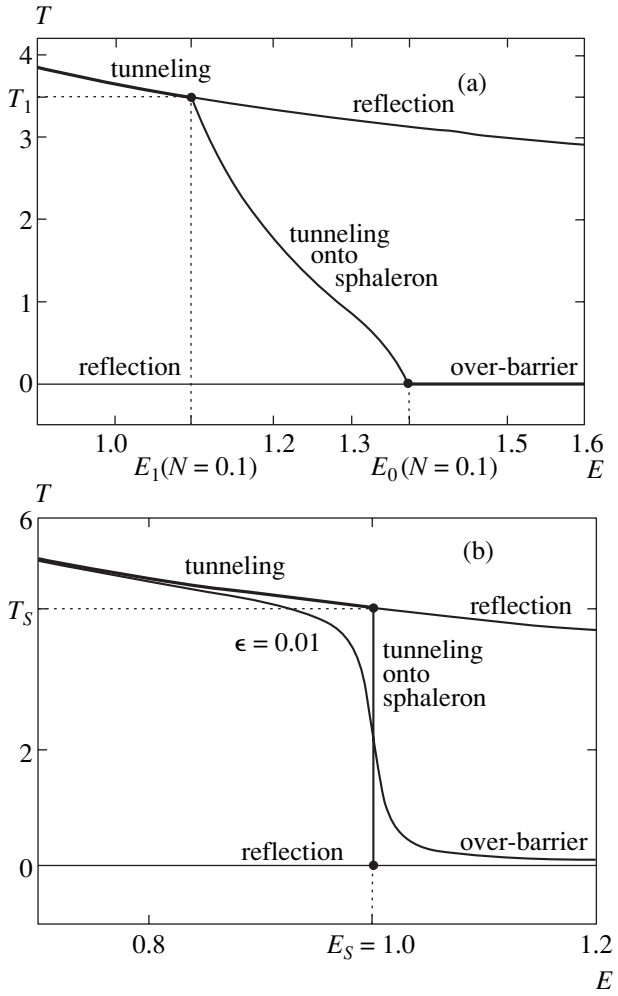


Fig. 9. Dependence of the parameter $T = -\partial F/\partial E$ on the energy for (a) the two-dimensional model with fixed $N = 0.1$ and (b) the one-dimensional model (see Appendix C). Different lines correspond to different branches of classical solutions of the T/θ boundary value problem. The branches labeled "reflection" end up on the wrong side of the barrier. Figure 9b also contains a line with nonzero ϵ.

coordinate $X(t)$ in the complex time plane is shown in Fig. 8 for the three points (Figs. 8a–8c) of the curve $\tau \equiv T/\epsilon = 380$, $\vartheta \equiv \theta/\epsilon = 130$. Point (a) lies deep inside the tunneling region, $E_a < E_1(N_a)$; point (c) corresponds to the overbarrier solution with $T = 0$, $\theta = 0$, $\epsilon = 0$; and point (b) is in the middle of the curve. The branch points of the solution, the cuts, and the contour are clearly seen on these graphs.[6]

It is worth noting that the left branch points move down as T and θ approach zero. Solutions close enough to the boundary $E_0(N)$ have the left branch point in the lower complex half-plane (see Fig. 8). Therefore, the corresponding contour can be continuously deformed to the real time axis. These solutions still satisfy the reality conditions asymptotically (see Fig. 6) but show nontrivial complex behavior at any finite time.

The regularized T/θ procedure makes it possible to approach the boundary of the region of classically allowed transitions from both sides. The points at this boundary are obtained by taking the limits $T \longrightarrow 0$, $T/\theta = \text{const}$ of the tunneling solutions and $\tau \longrightarrow +\infty$, $\tau/\vartheta = \text{const}$ of the classically allowed ones. Because $\tau^* \equiv \tau/\vartheta = T/\theta$ by construction, the lines $\tau^* = \text{const}$ are continuous at the boundary $E_0(N)$, although they may have discontinuity of the derivatives. The variable τ^* can be used to parameterize the curve $E_0(N)$.

4. CONCLUSIONS

We conclude that classical solutions describing transmissions of a bound system through a potential barrier with different values of energy and the initial oscillator excitation number form three branches. These branches merge at bifurcation lines $E_0(N)$ and $E_1(N)$. Solutions from different branches describe physically different transition processes. Namely, solutions at low energies $E < E_1(N)$ describe conventional potential-like tunneling. At $E > E_0(N)$, they correspond to unsuppressed overbarrier transitions. At intermediate energies, $E_1(N) < E < E_0(N)$, physically relevant solutions describe transitions on top of the barrier. This branch structure is shown in Fig. 9a, where the period $T = -\partial F/\partial E$ obtained numerically for solutions from the different branches is plotted as a function of energy for $N = 0.1$.

We note that the qualitative structure of branches in a model with internal degrees of freedom is similar to the structure of branches in one-dimensional quantum mechanics (see Appendix C). The latter is shown in Fig. 9b. The features of solutions in both cases are similar, although the solutions ending up on top of the barrier are degenerate in energy in the one-dimensional case and, hence, are not physically interesting.

In this paper, we introduced a regularization technique that allows smoothly connecting solutions from different branches. Its advantage is that it automatically chooses the physically relevant branch. This technique is particularly convenient in numerical studies: we have seen that it makes it possible to cover the whole interesting region of the parameter space. We applied this technique to baryon-number violating processes in electroweak theory [16].

ACKNOWLEDGMENTS

The authors are indebted to V. Rubakov and C. Rebbi for numerous valuable discussions and criticism; A. Kuznetsov, W. Miller, and S. Sibiryakov for helpful discussions; and S. Dubovsky, D. Gorbunov, A. Penin, and P. Tinyakov for stimulating interest. We wish to

[6] The phase of the tunneling coordinate changes by π around the branch point. The points where the phase of the tunneling coordinate changes by 2π correspond to zeros of $X(t)$.

thank the Boston University Center for Computational Science and Office of Information Technology for allocation of supercomputer time. This research was supported by the Russian Foundation for Basic Research (grant no. 02-02-17398), grant of the President of the Russian Federation no. NS-2184.2003.2, US Civilian Research and Development Foundation for Independent States of FSU (CRDF) award no. RP1-2364-MO-02, and under the DOE (grant no. US DE-FG02-91ER40676). F.B. is supported by the Swiss Science Foundation (grant no. 7SUPJ062239).

APPENDIX A

T/θ Boundary Value Problem

The semiclassical method for calculating the probability of tunneling from a state with a few parameters fixed was developed in [13–15, 32] in the context of field theory models and in [3–5, 11, 12] in quantum mechanics. Here, we outline the method adapted to our model of two degrees of freedom.

1. Path integral representation of the transition probability. We begin with the path integral representation for the probability of tunneling from the asymptotic region $X \longrightarrow -\infty$ through a potential barrier. Let the incoming state $|E, N\rangle$ have fixed energy and oscillator excitation number, and have support only for $X \ll 0$, well outside the range of the potential barrier. The inclusive tunneling probability for states of this type is given by

$$\mathcal{T}(E, N) = \lim_{t_f - t_i \to \infty} \left\{ \int_0^{+\infty} dX_f \int_{-\infty}^{+\infty} dy_f \times \left| \langle X_f, y_f | \exp(-i\hat{H}(t_f - t_i)) | E, N \rangle \right|^2 \right\}, \tag{A.1}$$

where $\hat{H}$ is the Hamiltonian operator. This probability can be reexpressed in terms of the transition amplitudes

$$\mathcal{A}_{fi} = \langle X_f, y_f | \exp(-i\hat{H}(t_f - t_i)) | X_i, y_i \rangle \tag{A.2}$$

and the initial-state matrix elements

$$\mathcal{B}_{ii'} = \langle X_i, y_i | E, N \rangle \langle E, N | X_i', y_i' \rangle \tag{A.3}$$

as

$$\mathcal{T}(E, N) = \lim_{t_f - t_i \to \infty} \left\{ \int_0^{+\infty} dX_f \int_{-\infty}^{0} dX_i dX_i' \times \int_{-\infty}^{+\infty} dy_i dy_i' dy_f \mathcal{A}_{fi} \mathcal{A}_{i'f}^* \mathcal{B}_{ii'} \right\}. \tag{A.4}$$

The transition amplitude and its complex conjugate have the familiar path integral representation

$$\mathcal{A}_{fi} = \int [d\mathbf{x}] \Big|_{\substack{\mathbf{x}(t_i) = \mathbf{x}_i \\ \mathbf{x}(t_f) = \mathbf{x}_f}} \exp(iS[\mathbf{x}]),$$
$$\mathcal{A}_{i'f}^* = \int [d\mathbf{x}'] \Big|_{\substack{\mathbf{x}'(t_i) = \mathbf{x}_i' \\ \mathbf{x}'(t_f) = \mathbf{x}_f}} \exp(-iS[\mathbf{x}']), \tag{A.5}$$

where $\mathbf{x} = (X, y)$ and S is the action of the model. To obtain a similar representation for the initial-state matrix elements, we rewrite $\mathcal{B}_{ii'}$ as

$$\mathcal{B}_{ii'} = \langle X_i, y_i | \hat{P}_E \hat{P}_N | X_i', y_i' \rangle, \tag{A.6}$$

where $\hat{P}_N$ and $\hat{P}_E$ denote the projectors onto the respective states with the oscillator excitation number N and the total energy E. It is convenient to use the coherent state formalism for the y oscillator and choose the momentum basis for the X coordinate. In this representation, the kernel of the projector operator $\hat{P}_E \hat{P}_N$ becomes

$$\langle q, b | \hat{P}_E \hat{P}_E | p, a \rangle = \frac{1}{(2\pi)^2} \int d\xi d\eta \times \exp\left(-iE\xi - iN\eta + \frac{i}{2} p^2 \xi + \exp(i\omega\xi + i\eta) \bar{b} a \right) \delta(q - p),$$

where $|p, a\rangle$ is the eigenstate with the respective eigenvalues p and a of the center-of-mass momentum $\hat{p}_X$ and the y-oscillator annihilation operator $\hat{a}$. It is straightforward to express this matrix element in the coordinate representation using the formulas

$$\langle y | a \rangle = \sqrt[4]{\frac{\omega}{\pi}} \exp\left(-\frac{1}{2} a^2 + \sqrt{2\omega} a y - \frac{1}{2} \omega y^2 \right),$$
$$\langle X | p \rangle = \frac{1}{\sqrt{2\pi}} \exp(ipX).$$

Evaluating the Gaussian integrals over a, b, p, and q, we obtain

$$\mathcal{B}_{ii'} = \int d\xi d\eta \exp\left\{ -iE\xi - iN\eta - \frac{i}{2} \frac{(X_i - X_i')^2}{\xi} + \frac{\omega}{1 - \exp(-2i\omega\xi - 2i\eta)} \times \left[\frac{y_i^2 + y_i'^2}{2} (1 + \exp(-2i\omega\xi - 2i\eta)) - 2 y_i y_i' \exp(-i\omega\xi - i\eta) \right] \right\}, \tag{A.7}$$

where we omit the preexponential factor depending on η and ξ. For the subsequent formulation of the bound-

ary value problem, it is convenient to introduce the notation

$$T = -i\xi, \quad \theta = -i\eta.$$

Then, combining integral representations (A.7) and (A.5) and rescaling the coordinates, energy, and excitation number as $\mathbf{x} \longrightarrow \mathbf{x}/\sqrt{\lambda}$, $E \longrightarrow E/\lambda$, $N \longrightarrow N/\lambda$, we finally obtain

$$\mathcal{T}(E, N) = \lim_{t_f - t_i \to \infty} \left\{ \int_{-i\infty}^{+i\infty} dT d\theta \int [d\mathbf{x} d\mathbf{x}'] \times \exp\left\{ -\frac{1}{\lambda} F[\mathbf{x}, \mathbf{x}'; T, \theta] \right\} \right\}, \tag{A.8}$$

where

$$F[\mathbf{x}, \mathbf{x}'; T, \theta] = -iS[X, y] + iS[X', y'] - ET - N\theta + B_i(\mathbf{x}_i, \mathbf{x}_i'; T, \theta). \tag{A.9}$$

Here, the nontrivial initial term B_i is

$$B_i = \frac{(X_i - X_i')^2}{2T} - \frac{\omega}{1 - \exp(2\omega T + 2\theta)} \times \left[\frac{1}{2}(y_i^2 + y_i'^2)(1 + \exp(2\omega T + 2\theta)) - 2y_i y_i' \exp(\omega T + \theta) \right]. \tag{A.10}$$

In (A.8), $\mathbf{x}$ and $\mathbf{x}'$ are independent integration variables, while $\mathbf{x}_f' \equiv \mathbf{x}_f$ (see Eq. (A.5)).

2. The boundary value problem. For small λ, path integral (A.8) is dominated by a stationary point of the functional F. Therefore, to calculate the tunneling probability exponent, we extremize this functional with respect to all the integration variables $X(t)$, $y(t)$, $X'(t)$, $y'(t)$, T, and θ. We note that, because of the limit $t_f - t_i \longrightarrow +\infty$, variation with respect to the initial and final values of the coordinates leads to boundary conditions imposed at asymptotic $t \longrightarrow \pm\infty$ rather than at finite times t_i, t_f. We also note that the stationary points may be complex.

Variation of functional (A.9) with respect to the coordinates at intermediate times gives second-order equations of motion, in general complexified,

$$\frac{\delta S}{\delta X(t)} = \frac{\delta S}{\delta y(t)} = \frac{\delta S'}{\delta X'(t)} = \frac{\delta S'}{\delta y'(t)} = 0. \tag{A.11a}$$

The boundary conditions at the final time $t_f \longrightarrow +\infty$ are obtained by extremization of F with respect to $X_f \equiv X_f'$ and $y_f \equiv y_f'$. These are

$$\dot{X}_f = \dot{X}_f', \quad \dot{y}_f = \dot{y}_f'. \tag{A.11b}$$

It is convenient to write the conditions at the initial time (obtained by varying X_i, y_i, X_i', and y_i') in terms of the asymptotic quantities. At the initial time instant $t_i \longrightarrow -\infty$, the system moves in the region $X \longrightarrow -\infty$, well outside the range of the potential barrier. Equations (A.11a) in this region describe free motion of decoupled oscillators, and the general solution takes the form

$$X(t) = X_i + p_i(t - t_i),$$

$$y(t) = \frac{1}{\sqrt{2\omega}}[a\exp(-i\omega(t - t_i)) + \bar{a}\exp(i\omega(t - t_i))],$$

and similarly for $X'(t)$ and $y'(t)$. For the moment, a and $\bar{a}$ are independent variables. In terms of the asymptotic variables X_i, p_i, a, $\bar{a}$, the initial boundary conditions become

$$p_i = p_i' = -\frac{X_i - X_i'}{iT},$$
$$a' + \bar{a}' = a\exp(\omega T + \theta) + \bar{a}\exp(-\omega T - \theta), \tag{A.11c}$$
$$a + \bar{a} = a'\exp(-\omega T - \theta) + \bar{a}'\exp(\omega T + \theta).$$

Variation with respect to the Lagrange multipliers T and θ gives a relation between the values of E, N, and the initial asymptotic variables (where we use boundary conditions (A.11c)):

$$E = \frac{p_i^2}{2} + \omega N, \qquad N = a\bar{a}. \tag{A.11d}$$

Equations (A.11a)–(A.11d) constitute the complete set of saddle-point equations for the functional F.

The variables X' and y' originate from the conjugate amplitude $\mathcal{A}_{i'f}^*$ (see Eq. (A.5)), which suggests that they are complex conjugate to X and y. Indeed, the ansatz $X'(t) = X^*(t)$, $y'(t) = y^*(t)$ is compatible with boundary value problem (A.11). The Lagrange multipliers T and θ are then real, and problem (A.11) may be conveniently formulated on the contour $ABCD$ in the complex time plane (see Fig. 2).

We now have only two independent complex variables $X(t)$ and $y(t)$, which have to satisfy the classical equations of motion in the interior of the contour,

$$\frac{\delta S}{\delta X(t)} = \frac{\delta S}{\delta y(t)} = 0. \tag{A.12a}$$

The final boundary conditions (see Eq. (A.11b)) become the reality conditions for the variables $X(t)$ and $y(t)$ at the asymptotic part D of the contour,

$$\mathrm{Im}X_f = 0, \quad \mathrm{Im}y_f = 0, \quad \mathrm{Im}\dot{X}_f = 0, \quad \mathrm{Im}\dot{y}_f = 0, \quad t \longrightarrow +\infty. \tag{A.12b}$$

Seemingly complicated initial conditions (A.11c) simplify when written in terms of the time coordinate $t' =$

$t + iT/2$ running along part AB of the contour. We again write the asymptotic from of a solution, but now along the initial part AB of the contour,

$$X = X_0 + p_0(t' - t_i),$$

$$y = \frac{1}{\sqrt{2\omega}}[u\exp(-i\omega(t' - t_i)) + v\exp(i\omega(t' - t_i))].$$

In terms of X_0, y_0, u, and v, boundary conditions (A.11c) become

$$\mathrm{Im}X_0 = 0, \quad \mathrm{Im}p_0 = 0, \qquad v = u^* e^{\theta}. \tag{A.12c}$$

Finally, we write Eqs. (A.11d) in terms of the asymptotic variables along the initial part of the contour,

$$E = \frac{p_0^2}{2} + \omega N, \qquad N = \omega u v. \tag{A.13}$$

These equations determine the Lagrange multipliers T and θ in terms of E and N. Alternatively, we can solve problem (A.12) for given values of T and θ and find the values of E and N from Eqs. (A.13), which is more convenient computationally.

Given a solution to problem (A.12), the exponent F is the value of functional (A.9) at this saddle point. We thus obtain expression (8) for the tunneling exponent. The exponent F is now expressed in terms of S_0 in Eq. (9), the action of the system integrated by parts. The nontrivial boundary term B_i (Eq. (A.10)) is canceled by the boundary term coming from integration by parts. We note that we did not use constraints (A.13) to obtain formula (8), and we therefore still have to extremize (8) with respect to T and θ (see discussion in Section 2.2).

Classical problem (A.12) is conveniently called the T/θ boundary value problem. Equations (A.12b) and (A.12c) imply eight real boundary conditions for two complex second-order differential equations (A.12a). However, one of these real conditions is redundant: Eq. (A.12b) implies that the (conserved) energy is real, and, therefore, the condition $\mathrm{Im}p_0 \longrightarrow 0$ is automatically satisfied (we note that the oscillator energy $E_{\mathrm{osc}} = \omega u v = \omega e^{\theta} u u^*$ is real). On the other hand, system (A.12) is invariant under time translations along the real axis. This invariance is fixed, e.g., by requiring that $\mathrm{Re}X$ take a prescribed value at a prescribed large negative time t_0' (we note that other ways may be used instead; in particular, for $E < E_1(N)$, it is convenient to impose the constraint $\mathrm{Re}\dot{X}(t = 0) = 0$. Together with the latter requirement, we have exactly eight real boundary conditions for the system of two complexified (i.e., four real) second-order equations.

APPENDIX B

A Property of Solutions to the T/θ Problem in the Case of Overbarrier Transitions

For given E and N, there is only one overbarrier classical solution, which is obtained in the limit $\epsilon \longrightarrow 0$ of the regularized T/θ procedure. To see what singles out this solution, we analyze the regularized functional

$$F_\epsilon[q] = F[q] + 2\epsilon T_{\mathrm{int}}[q], \tag{B.1}$$

where q denotes the variables $\mathbf{x}(t)$, $\mathbf{x}'(t)$ and T, θ together. The unregularized functional F has a valley of extrema $q^e(\varphi)$ corresponding to different values of the initial oscillator phase φ. Clearly, at small ϵ, the extremum of F_ϵ is close to a point in this valley with the phase extremizing $T_{\mathrm{int}}[q^e(\varphi)]$,

$$\frac{d}{d\varphi} T_{\mathrm{int}}[q^e(\varphi)] = 0. \tag{B.2}$$

Hence, the solution q_ϵ^e of the regularized T/θ boundary value problem tends to the overbarrier classical solution, with T_{int} extremized with respect to the initial oscillator phase.

Because $U_{\mathrm{int}}(\mathbf{x}) > 0$, T_{int} is a positive quantity with at least one minimum. In a normal situation, there is only one saddle point of F_ϵ, and, hence, solving the T/θ boundary value problem gives the classical solution with the time of interaction minimized.

APPENDIX C

Classically Allowed Transitions: A One-Dimensional Example

The difficulties with bifurcations of classical solutions emerge in quite a general class of quantum-mechanical models. To illustrate this statement, we consider one-dimensional quantum mechanics, where the result is given by the well-known WKB formula. We show that the origin of the above difficulties can also be seen in one-dimensional model. Implementation of the regularization technique is explicit in the one-dimensional case. This makes it easy to see how our technique allows us to smoothly join the classical solutions relevant to the tunneling and allowed transitions.

In quantum mechanics of one degree of freedom, only one variable $X(t)$ is present, which describes the motion of a particle of mass $m = 1$ through a potential barrier $U(X)$. The motion is free in the asymptotic regions $X \longrightarrow \pm\infty$. The semiclassical calculation of the tunneling exponent is performed by solving the classical equation of motion

$$\frac{\delta S}{\delta X(t)} = 0$$

on contour *ABCD* in the complex time plane, with the condition that the solution is real in the asymptotic past (region *A*) and asymptotic future (region *D*). The relevant solutions tend to $X \longrightarrow -\infty$ and $X \longrightarrow +\infty$ in regions *A* and *D*, respectively. The auxiliary parameter *T* is related to the energy of the incoming state by the requirement that the energy of the classical solution is *E*. The exponent for the transition probability is

$$F = 2\mathrm{Im}S - ET. \tag{C.1}$$

We note that these boundary conditions resemble the ones on the tunneling coordinate *X* in the two-dimensional system.

In quantum mechanics of one degree of freedom, contour *ABCD* may be chosen such that points *B* and *C* are the turning points of the solution. Then, the solution is also real at part *BC* of the contour. Indeed, a real solution at part *BC* of the contour oscillates in the upside-down potential, *T*/2 is equal to the half-period of oscillations, and points *B* and *C* are the two different turning points, $\dot{X} = 0$. Continuation of this solution from point *C* to the positive real times in accordance with the equation of motion corresponds to real-time motion, with zero initial velocity, towards $X \longrightarrow +\infty$; the coordinate $X(t)$ stays real on part *CD* of the contour. Likewise, the continuation back in time from point *B* leads to a real solution in part *AB* of the contour. The reality conditions are, thus, satisfied at *A* and *D*. The only contribution to *F* comes from the Euclidean part of the contour, and it can be checked that expression (C.1) reduces to

$$F(E) = 2\int_{X_B}^{X_C} \sqrt{2(U(X)-E)}\,dX, \tag{C.2}$$

which is the standard WKB result.

The solutions appropriate for the classically forbidden and classically allowed transitions apparently belong to different branches. As the energy approaches the height of the barrier U_0 from below, the amplitude of the oscillations in the upside-down potential decreases, while the period *T* tends to a finite value determined by the curvature of the potential at its maximum. On the other hand, the solutions for $E > U_0$ always run along the real time axis, and, hence, the parameter *T* is always zero. Therefore, the relevant solutions do not merge at $E = U_0$, and $T(E)$ has a discontinuity at $E = U_0$. The regularization technique of Section 3.1 removes this discontinuity and allows smooth transitions through the point $E = U_0$. The only difference with quantum mechanics of multiple degrees of freedom is that, in the latter case, bifurcation points exist not only at the boundary of the region of classically allowed transitions but also well inside the region of classically forbidden transitions (but still at $E > E_S$; see the Introduction and Section 2.3).

To illustrate the situation, we consider an exactly solvable model with

$$U(X) = \frac{1}{\cosh^2 X}.$$

We implement our regularization technique by formally changing the potential

$$U(X) \longrightarrow e^{-i\epsilon}U(X), \tag{C.3}$$

which leads to the corresponding change of the classical equations of motion. Here, ϵ is a real regularization parameter, the smallest parameter in the model. At the end of the calculations, we take the limit $\epsilon \longrightarrow 0$.

We do not change the boundary conditions in our regularized classical problem, i.e., we still require $X(t)$ to be real in the asymptotic future on the real time axis and $X(t')$ to be real as $t' \longrightarrow -\infty$ on part *A* of contour *ABCD*. Then, the conserved energy is real. The sphaleron solution $X(t) = 0$ now has a complex energy (because the potential is complex). Hence, the solutions of our classical boundary value problem necessarily avoid the sphaleron, and we can expect that the solutions behave smoothly in energy.

The general solution to the regularized problem is

$$\sqrt{\frac{E}{e^{-i\epsilon}-E}}\sinh X = -\cosh(\sqrt{2E}(t-t_0)),$$

where t_0 is the integration constant. The value of $\mathrm{Im}\,t_0$ is fixed by the requirement that $\mathrm{Im}X = 0$ at positive time $t \longrightarrow +\infty$,

$$\mathrm{Im}t_0 = \frac{T}{2} - \frac{1}{2\sqrt{2E}}\arg[e^{-i\epsilon}-E].$$

The residual parameter $\mathrm{Re}\,t_0$ represents the real-time translational invariance present in the problem. The condition that the coordinate *X* is real on the initial part *AB* of the contour gives the relation between *T* and *E*,

$$\frac{T}{2} = \frac{1}{\sqrt{2E}}\{\pi + \arg(e^{-i\epsilon}-E)\}. \tag{C.4}$$

For $\epsilon = 0$ and $E < 1$, the original unregularized result $T/2 = \pi/\sqrt{2E}$ is reproduced.

We now analyze what happens in the regularized case in the vicinity of the would-be special value of energy, $E = E_S \equiv 1$. It is clear from Eq. (C.4) that *T* is

now a smooth function of E. Away from $E = 1$, Eq. (C.4) can be written as

$$\frac{T}{2} = \begin{cases} \dfrac{\pi}{\sqrt{2E}}, & \text{forbidden region}, \quad 1 - E \gg \epsilon, \\ \dfrac{\epsilon}{\sqrt{2E}(E-1)}, & \text{allowed region}, \quad E - 1 \gg \epsilon. \end{cases} \tag{C.5}$$

Deep enough in the region of forbidden transitions, where $1 - E \gg \epsilon$, the argument in Eq. (C.4) is nearly zero and we return to the original tunneling solution. When E crosses the region of size of order ϵ around $E = 1$, the argument rapidly changes from $O(\epsilon)$ to $-\pi$, and, hence, $T/2$ changes from $\pi/\sqrt{2}$ to nearly zero. Thus, at $E > 1$, we obtain a solution that is very close to the classical overbarrier transition, and the contour is also very close to the real axis. This is shown in Fig. 9. We conclude that, at small but finite ϵ, the classically allowed and classically forbidden transitions merge smoothly.

For $E < 1$, the limit $\epsilon \longrightarrow 0$ is straightforward. For $E > 1$, a somewhat more careful analysis of the limit $\epsilon \longrightarrow 0$ is needed. It follows from Eq. (C.5) that the limit $\epsilon \longrightarrow 0$ with a constant finite $T < \pi/\sqrt{2}$ leads to solutions with $E = 1$. Classical overbarrier solutions of the original problem with $E > E_S \equiv 1$ are obtained in the limit $\epsilon \longrightarrow 0$ if T also tends to zero while $\tau = T/\epsilon$ is kept finite. Different energies correspond to different values of τ. This is what one expects: classical overbarrier transitions are described by the solutions on the contour with $T \equiv 0$.

REFERENCES

1. Z. Huang, T. Feuchtwang, P. Cutler, and E. Kazes, Phys. Rev. A **41**, 32 (1990).
2. S. Takada and H. Nakamura, J. Chem. Phys. **100**, 98 (1994).
3. W. Miller, J. Chem. Phys. **53**, 3578 (1970).
4. W. Miller and T. George, J. Chem. Phys. **56**, 5668 (1972).
5. T. George and W. Miller, J. Chem. Phys. **57**, 2458 (1972).
6. W. H. Miller, Adv. Chem. Phys. **25**, 69 (1974).
7. M. Wilkinson, Physica D (Amsterdam) **21**, 341 (1986).
8. M. Wilkinson and J. Hannay, Physica D (Amsterdam) **27**, 201 (1987).
9. S. Takada, P. Walker, and W. Wilkinson, Phys. Rev. A **52**, 3546 (1995).
10. S. Takada, J. Chem. Phys. **104**, 3742 (1996).
11. G. F. Bonini, A. G. Cohen, C. Rebbi, and V. A. Rubakov, quant-ph/9901062.
12. G. F. Bonini, A. G. Cohen, C. Rebbi, and V. A. Rubakov, Phys. Rev. D **60**, 076004 (1999).
13. V. A. Rubakov, D. T. Son, and P. G. Tinyakov, Phys. Lett. B **287**, 342 (1992).
14. A. N. Kuznetsov and P. G. Tinyakov, Phys. Rev. D **56**, 1156 (1997).
15. F. Bezrukov, C. Rebbi, V. Rubakov, and P. G. Tinyakov, hep-ph/0110109.
16. F. Bezrukov, D. Levkov, C. Rebbi, *et al.*, Phys. Rev. D **68**, 036005 (2003).
17. A. M. Perelomov, V. S. Popov, and M. V. Terent'ev, Zh. Éksp. Teor. Fiz. **51**, 309 (1966) [Sov. Phys. JETP **24**, 207 (1966)].
18. V. S. Popov, V. Kuznetsov, and A. M. Perelomov, Zh. Éksp. Teor. Fiz. **53**, 331 (1967) [Sov. Phys. JETP **26**, 222 (1967)].
19. N. Makri and W. Miller, J. Chem. Phys. **89**, 2170 (1988).
20. K. Thompson and N. Makri, J. Chem. Phys. **110**, 1343 (1999).
21. K. Kay, J. Chem. Phys. **107**, 2313 (1997).
22. N. Maitra and E. Heller, Phys. Rev. Lett. **78**, 3035 (1997).
23. W. Miller, J. Phys. Chem. A **105**, 2942 (2001).
24. A. N. Kuznetsov and P. G. Tinyakov, Mod. Phys. Lett. A **11**, 479 (1996).
25. M. Davis and E. Heller, J. Chem. Phys. **75**, 246 (1981).
26. E. Heller and M. Davis, J. Chem. Phys. **85**, 309 (1981).
27. E. Heller, J. Phys. Chem. **99**, 2625 (1994).
28. T. Banks, G. Farrar, M. Dine, *et al.*, Nucl. Phys. B **347**, 581 (1990).
29. V. I. Zakharov, Phys. Rev. Lett. **67**, 3650 (1991).
30. G. Veneziano, Mod. Phys. Lett. A **7**, 1661 (1992).
31. V. A. Rubakov, hep-ph/9511236.
32. V. A. Rubakov and P. G. Tinyakov, Phys. Lett. B **279**, 165 (2002).
33. F. R. Klinkhamer and N. S. Manton, Phys. Rev. D **30**, 2212 (1984).
34. S. Y. Khlebnikov, V. A. Rubakov, and P. G. Tinyakov, Nucl. Phys. B **367**, 334 (1991).
35. C. Rebbi and R. Singleton, hep-ph/9502370.