Dynamics of the formation of an event horizon

A. A. Shatskiĭ*)

M. V. Lomonosov Moscow State University, 119899 Moscow, Russia

A. Yu. Andreev

P. N. Lebedev Physics Institute, Russian Academy of Sciences, 117924 Moscow, Russia

(Submitted 17 November 1998)
Zh. Éksp. Teor. Fiz. **116**, 353–368 (August 1999)

The radial motion of matter in a gravitational field with a symmetry center in a comoving reference frame is investigated for a realistic equation of state of matter. The dynamics of the formation of an event horizon is investigated. © *1999 American Institute of Physics.* [S1063-7761(99)00108-0]

1. INTRODUCTION

The formation of a black-hole event horizon has attracted a great deal of attention on the part of physicists for a long time. An enormous amount of material has been written on this subject (see, for example, Refs. 1–4); nevertheless, the treatment of this problem within the general theory of relativity has created more questions than known solutions.

One of the main questions concerning this problem is still the reciprocal influence of accreting matter on a black hole. The motion of test particles in the field of a black hole has so far been considered for the most part, but they, as we know, do not exhibit a reciprocal influence, which can be enormous when a falling particle achieves the speed of light as it crosses the event horizon.

1. In this paper we consider the special, but physically real case of spherically symmetric accretion on a central body without allowance for rotation. The following notation is adopted: the speed of light c and the gravitational constant G are set equal to unity. In these units the gravitational radius for a given mass M is $r_g=2M$, i.e., the radius of the event horizon in free space for the same mass concentrated at the center.

Let us devise a likely model for the evolution of the system. We assume that our system is a cooling massive star having a radius R_0 and a gravitational radius $R_{G_0}=2M(R_0)$, where $R_{G_0}<R_0$. The matter comprising this body is initially at rest (''dust'' with the equation of state $P=\alpha\varepsilon$, where P is the pressure in the matter, ε is the energy density, and α is a constant). In the next moment the matter begins to fall freely.[1)] If it is assumed that the gravitational fields are not excessively strong and that the dust density[2)] is fairly small in the initial moment, a force field with a finite energy is needed to retain it in the initial moment. After this field is removed, the dust leaves the system and ceases to interact with it after a time of the order of the size of the system, i.e., after a time much shorter than the time during which the dust manages to partially settle and the gravitational fields increase dramatically. Thus, this model is physically consistent.

What subsequently happens to the system? The dust begins to fall toward the center of the body, increasing its mean density and the gravitational radius $r_g(r)$ for the mass $M(r)$ at a certain radius r. If we would ignore the reciprocal influence of the pressure of the moving matter on the dynamics of the system and on its gravitational field, then after all of the matter has unavoidably fallen and the inequality

$$r_g(r)=2M(r)\geqslant r \tag{1}$$

holds at one of the points r of the system, an event horizon would form at that point according to Schwarzchild's solution for a gravitational field in a vacuum, i.e., the velocity of the falling matter relative to the $r=\text{const}$ surfaces would reach the speed of light (see below). Is this what actually happens? The attainment of the speed of light by the matter causes a change in the sign of the interval and is therefore an invariant event, which does not depend on the choice of the reference frame.

An attempt to solve this problem in a reference frame which is stationary at infinity leads at once to a contradiction. In fact, the analytically exact, nonstationary model, in principle, cannot be studied. If, however, a simplification is made and it is assumed that the system is quasistationary at a certain moment in time in the range from one radius to a certain radius known to be large, but still far smaller than the dimensions of the system, then it can be stated, at the very least, that the components $g_{tt}(r)$ and $g_{rr}(t)$ of the metric have singularities (zeros and poles) in this reference frame.[3)] When the parameters of the system are chosen so that there would be a region in space where the inequality (1) is sure to be satisfied, it becomes clear that the metric does not have singularities, regardless of whether the inequality (1) is satisfied.

This can be shown by assuming that if a singularity appears at a certain point r_0, the component of the metric near it can be represented in the form

$$g_{ii}(r)\approx\text{const}(r-r_0)^{y_i},$$

where y_i is a certain number. When such a metric is substituted into the equations, it is found that they do not have a solution for any $y_i\neq 0$.

 © 1999 American Institute of Physics

This apparently indicates that the singularities and thus the horizon of the rapidly moving matter are eliminated (the right-hand side of the Einstein equations, which is equal to zero in a vacuum, becomes singular in the presence of ultrarelativistic falling matter when the radial component of the three-velocity tends to unity and the radial and temporal components of the four-velocity tend to infinity; this is also the reason for the elimination of the singularities of the metric).

However, in reality all this stems from the inapplicability of the quasistationary approximation in the case of strong gravitational fields. It is inapplicable because the passage of time in the system is highly nonuniform due to the nonuniformity of the component $g_{tt}(r)$ of the metric. This causes the picture, which appears to be stationary far from the center, to become highly nonstationary to an observer approaching the symmetry center of the system.

Nevertheless, this does not remove the question posed: do an horizon and a black hole appear in the real nonstationary case?

2. An answer to the question posed can be found by selecting a comoving reference frame. The problem was solved in this frame in Ref. 1 (Sec. 103) in the special case of $\alpha=0$ (see below). Matter is at rest in the reference frame chosen, and its motion can be evaluated only from the variation of the ''circumferential'' or photometric distances r, which are related to the center of the system and are defined as the circumferences of the respective circles around the center: $2\pi r$. When the radius r is defined as such, it is convenient to represent the metric in the form

$$ds^2=e^{\nu}dt^2-e^{\lambda}dR^2-r^2(d\theta^2+\sin^2\theta\,d\varphi^2). \tag{2}$$

Here R is the coordinate of a dust particle in the comoving reference frame or its index, and e^{ν}, e^{λ}, and r are functions of R and t. It is noteworthy that at zero pressure, i.e., when $\alpha=0$, we have $\nu=0$, i.e., the reference frame is simultaneously synchronous.

To solve the problem posed we write out the Einstein equation in the comoving reference frame:

$$r'^2e^{-\lambda}(1+r\nu'/r')-e^{-\nu}(2r\ddot{r}+\dot{r}^2-r\dot{r}\dot{\nu})=1+8\pi\alpha r^2\varepsilon, \tag{3a}$$

$$2\dot{\mu}'+\dot{\mu}\mu'-\dot{\lambda}\mu'-\nu'\dot{\mu}=0, \tag{3b}$$

$$\left(\lambda+2\mu+\frac{2}{1+\alpha}\ln\varepsilon\right)=0, \tag{3c}$$

$$\left(\nu+\frac{2\alpha}{1+\alpha}\ln\varepsilon\right)'=0. \tag{3d}$$

Here $\mu=2\ln r$, a prime denotes differentiation with respect to R, and a dot denotes differentiation with respect to t. Equations (3) were derived in Ref. 1 [Eqs. (2), (5), and (6) of problem 5 in Sec. 100].

It follows from (3d) that

$$\nu=-\frac{2\alpha}{1+\alpha}\ln\varepsilon+f^*(t)$$

and that by transforming the time t in the interval element (2) the function $f^*(t)$ can be set equal to $[2\alpha/(1+\alpha)]\ln\varepsilon_*$, where ε_* is a constant with the dimensions of energy density, which expresses the measurement scale of ε. Then

$$\nu=-\frac{2\alpha}{1+\alpha}\ln\frac{\varepsilon}{\varepsilon_*}. \tag{4}$$

We next assign the indices R to the dust particles so that $r=R$ in the initial moment. Under such initial conditions $r'(R,t)$ corresponds to $(n_0/n)^{1/3}$, where $n(R,t)$ is the concentration of dust particles, and n_0 is its value at the initial moment.

Let us now ascertain the conditions which must be imposed on the initial distribution of the dust. The most important among them is that the inequality (1) need not hold within the matter at the initial moment. It means that there is no horizon in all space in the initial moment. It thus imposes an upper limit on the initial density of the dust and on the initial dimensions of the system. More specifically, if the initial density distribution of the dust is set equal to $\varepsilon_0(R)$, then, according to (1), the maximum radius of the body $R_{\max}$ is uniquely specified by the expression

$$R_{\max}=2\int_0^{R_{\max}}4\pi\varepsilon_0(R)R^2\,dR. \tag{5}$$

Then, it follows from (3c) and (4) that

$$\frac{\partial}{\partial t}[\alpha(\lambda+2\mu)-\nu]=0$$

or

$$\nu=\alpha[\lambda+2\mu+f^*(R)], \tag{6}$$

where $f^*(R)$ is an arbitrary function that depends on the initial conditions.

3. Let us now find the initial values for all the variables in our problem. We have already assigned these values for r and ε. From (4) it follows that

$$\nu_0=-\frac{2\alpha}{1+\alpha}\ln\frac{\varepsilon_0}{\varepsilon_*}. \tag{7}$$

To find the initial value of λ we take advantage of the fact that the problem has already been solved for $\alpha=0$, and we can therefore utilize the familiar expression for $\lambda_0|_{\alpha=0}$ from Ref. 1 (Sec. 103.6):

$$\lambda_0(R)=-\ln[1-S(R)], \tag{8}$$

where for $\alpha=0$ we have

$$S(R)=2M(R)/R, \tag{9}$$

and $M(R)$ is the mass within the radius R at the initial moment.

The expression for $S(R)$ for an arbitrary value of α is the same. It can be obtained from Eq. (4) in problem 5 of Sec. 100 in Ref. 1, where the Einstein equations in matter in the comoving reference frame were found for a system with a symmetry center. We write out this equation:

$$-e^{-\lambda}\left[\mu''+\frac{3}{4}\mu'^2-\frac{\mu'\lambda'}{2}\right]$$

$$+\frac{1}{r^2}+\frac{1}{2e^{\nu}}[\dot{\lambda}\dot{\mu}+(\dot{\mu})^2/2]=8\pi\varepsilon. \qquad (10)$$

Expressing μ in terms of r ($\mu=\ln r^2$) and combining similar terms, we can bring this expression into the form

$$8\pi r'\varepsilon r^2=-[r(r'^2e^{-\lambda}-1)]'+\frac{r'}{e^{\nu}}[\dot{\lambda}r\dot{r}+(\dot{r})^2]. \qquad (11)$$

Taking into account the expression (8), as well as the fact that, according to the expression (100.23) in Ref. 1, the equality

$$2M(r)=\int_0^r 8\pi\varepsilon(\tilde{r},t)\tilde{r}^2\,d\tilde{r}|_{t=\text{const}}$$

holds for the initial moment in time, when $\dot{r}=0$ and $r'=1$, we obtain the expression (9) for $S(R)$ after preliminarily integrating (11) over R from 0 to R.

Substituting the expression (8) into (6) and using (7), we find

$$f^*(R)=-\frac{2}{1+\alpha}\ln\frac{\varepsilon_0}{\varepsilon_*}+\ln[1-S(R)]-\ln R^4. \qquad (12)$$

4. Now, plugging (6) into (3b) and dividing everything by $\dot{\mu}\mu'$, we obtain the expression

$$\frac{1}{\dot{r}}(2\ln\mu'+\mu-\lambda)^{\cdot}=\frac{\nu'}{r'}=\alpha\frac{[\lambda+2\mu+f^*(R)]'}{r'}. \qquad (13)$$

Since $e^{-\nu}(2r\ddot{r}+\dot{r}^2-r\dot{r}\dot{\nu})=(e^{-\nu}r\dot{r}^2)^{\cdot}/\dot{r}$ and introducing the notation

$$U(R,t)=(\dot{r})^2,\quad Q(R,T)=r'^2e^{-\lambda}, \qquad (14)$$

we see that Eq. (3a) can be written as an equation for U:

$$\frac{\dot{U}}{\dot{r}}+aU=\sigma, \qquad (15)$$

where

$$a(R,t)=\frac{1}{r}\left(1-\frac{r\dot{\nu}}{\dot{r}}\right),$$

$$\sigma(R,t)=\frac{1}{r}\left[Q\left(1+\frac{r\nu'}{r'}\right)-1-8\pi\alpha r^2\varepsilon\right]e^{\nu}.$$

This equation has a solution which satisfies the initial conditions:

$$U(R,t)=\frac{1}{\gamma^*(R,t)}\int_0^t\gamma^*(R,\tilde{t})\sigma(R,\tilde{t})\dot{\tilde{r}}\,d\tilde{t},$$

$$\gamma^*(R,t)=\exp\left[\int_0^t a(R,\tilde{t})\dot{\tilde{r}}\,d\tilde{t}\right]. \qquad (16)$$

Finding U, we can obtain an expression for the square of the velocity of the matter relative to the $r=$ const surfaces from the form of the metric (2) (see Appendix 1):

$$V^2(R,t)=Ue^{-\nu}e^{\lambda}/r'^2. \qquad (17)$$

The expression for γ^* can easily be found:

$$\gamma^*(R,t)=C(R)re^{-\nu},$$

where the multiplier $C(R)$ for $\gamma^*(R,t)$, which does not depend on t, can be taken out of the integral sign in (16) and canceled; therefore, it can be set equal to unity. Then

$$\gamma^*\sigma=r'^2e^{-\lambda}\left(1+\frac{r\nu'}{r'}\right)-1-8\pi\alpha r^2\varepsilon.$$

Alternatively, since Eq. (13) can now be rewritten as

$$\frac{(2\ln\mu'+\mu-\lambda)^{\cdot}}{\dot{r}}=\frac{(\ln Q)^{\cdot}}{\dot{r}}=\frac{\nu'}{r'}, \qquad (18)$$

we obtain

$$\gamma^*\sigma=\frac{(r(Q-1))^{\cdot}}{\dot{r}}-8\pi\alpha r^2\varepsilon. \qquad (19)$$

Equation (16) can now be rewritten in the form

$$U=\frac{e^{\nu}}{r}[r(Q-1)-R(Q_0-1)+2\alpha m(R,t)], \qquad (20)$$

where we have introduced the notation

$$m(R,t)=\int_t^0 4\pi\tilde{\varepsilon}\tilde{r}^2\dot{\tilde{r}}\,d\tilde{t}=\int_r^R 4\pi\tilde{\varepsilon}\tilde{r}^2\,d\tilde{r}|_{R=\text{cosnt}}. \qquad (21)$$

5. For $\alpha=0$, using (6), (9), (18), and (20), we can easily obtain an analytically exact expression for U and V:

$$U_{\alpha=0}=S(R)\left(\frac{R}{r}-1\right)=\frac{2M(R)}{r}-S(R). \qquad (22)$$

Substituting this expression into (17), we obtain the following expression for the velocity:

$$V^2_{\alpha=0}=\frac{2M(R)/r-S(R)}{1-S(R)}. \qquad (23)$$

Hence $V_{\alpha=0}=1$ when $r=r_g=2M(R)$. This coincides with the results in Sec. 100 of Ref. 1, where the problem has already been solved for this case.

6. Finally, let us consider the location of the horizon[4)] in the presence of a nonzero pressure. For this purpose we plug the expressions (20) and (8) for $e^{-\lambda_0}$ into formula (17). After some relatively simple transformations, we ultimately obtain

$$r=\frac{2M(R)+2\alpha m(R,t)}{1-Q(1-V^2)}. \qquad (24)$$

As will be shown in Appendix 1, the horizon appears at the point and at the time where the velocity of the falling matter relative to the $r=$ const surfaces reaches unity, i.e., where $V=1$. In addition, the speed of light relative to the falling matter at this site is also, as always, equal to unity.

Hence, according to (24) and (9), the horizon radius r_{hor} is given by the formula

$$r_{\text{hor}}=2M(R)+2\alpha m(R,t_{\text{hor}}). \qquad (25)$$

Thus, the horizon is displaced to a larger radius in comparison to the value in a vacuum $r_{0_{\rm hor}}=2M(R)$ by $2\alpha m(R,t_{\rm hor})$. In this case the quantity $m(R,t)$ has the meaning of the mass which would accumulate if we would join layers of dust with the initial radius R and the thickness $d\tilde{r}(\tilde{t})$ to one another up to the radius $r(R,t)$ at the moment when this $d\tilde{r}$ layer passes through the joining point.

7. Regarding the possible values of α, we note that $\alpha=0$ corresponds to dustlike matter without interactions between the particles. The results obtained for them are the same [see (23)] as the results for test particles in a central field of mass M (see Sec. 101 in Ref. 1). However, of course, such an equation of state of matter cannot correspond to reality near the horizon. It is reasonable to assume that the ultrarelativistic equation of state of matter, in which $\alpha=1/3$, holds near the horizon. Therefore, the location of the horizon should probably be sought with just such a value of α.

8. When $\alpha\neq 0$, it would appear that the falling matter should be slowed under the action of the pressure gradient, and the horizon should therefore form later, i.e., be displaced toward smaller values of r, but, as we have just shown, it is displaced toward larger values of r by $2\alpha m(R,t_{\rm hor})$. What is the reason for this contradiction? It can be seen from the initial equations (3) that the reason should be sought in Eq. (3a). For this purpose we explore Eqs. (3a) and (4) in the initial moment for the case $\alpha\ll 1$. At that moment $\dot{r}=0$ and $r'=1$; therefore, we write

$$[1-S(R)]\left(1-2ar\frac{\varepsilon'}{\varepsilon}\right)-2r\ddot{r}\left(\frac{\varepsilon}{\varepsilon_*}\right)^{2\alpha}\approx 1+8\pi\alpha r^2\varepsilon.$$

Since

$$(\varepsilon/\varepsilon_*)^{2\alpha}\approx 1+2\alpha\ln\frac{\varepsilon}{\varepsilon_*},\quad S(R)=\frac{2M(R)}{R},\quad r=R,$$

after performing some relatively simple transformations, we obtain the following expression in the linear approximation with respect to α:

$$\ddot{r}=-\frac{GM(R)}{r^2}\left[1-2\alpha\ln\frac{\varepsilon}{\varepsilon_*}\right]-\frac{\nabla P}{\rho}\left[1-\frac{2GM(r)}{rc^2}\right]-4\pi\alpha Gr\rho, \tag{26}$$

where $\rho(R,r)=\varepsilon(R,r)/c^2$ is the density of the matter. Here, for clarity we use the ordinary (Gaussian) system of units with $G\neq 1$ and $c\neq 1$. It can be seen from (26) that the first term corresponds to the ordinary Newtonian force of gravity, and the second term corresponds to the interaction force between the particles, i.e, the pressure gradient (just this force is the cause of the slowing of the fall of the matter in the first stage). The remaining terms do not appear in the equation of motion in the Newtonian approximation (the corrections in square brackets are also ignored in that case), but, as we have already seen, the last term begins to dominate over the second term at high energies; therefore, a shift of the horizon toward larger radii appears. Thus, the contradiction has been resolved. Physically this corresponds to the ''gravity of pressure'' in the general theory of relativity, which surpasses the gradient terms at high energies.

9. The analysis performed allows us to draw the following conclusions.

First, a shift of the horizon toward a larger radius in comparison to the Schwarzchild radius due to the ''gravity of pressure'' has been discovered. We stress that this effect is purely dynamic and is not observed in the static case (after all the matter has fallen).

Second, according to the results in Appendix 2, the evolution of the entire system at a constant value of α is completely specified by the energy density distribution profile in the initial moment, i.e., for example, by the normalized density distribution of the matter and by the value of the parameter S at an arbitrary point on this distribution.

If the evolution of only one spherical layer of matter with the index R must be described, it is completely specified by three dimensionless parameters in the initial moment in that layer and, in this sense, does not depend on the initial distribution of the matter in the system below and above that layer. However, this in no way signifies the independence of the spherical layers in general, since just these three parameters, as will be seen from Appendix 2, govern the interaction of the layers. Consequently, integration of the system leads to a complete family of self-similar solutions.

Third, according to Appendix 2, a local extremum appears on the $V(R)|_{t=\rm const}$ curve for a specific choice of initial parameters, and when $V=1$, it leads to the formation of a second apparent horizon in the system (an analog of the second horizon in the Reissner–Nordström and Kerr–Newman solutions for an electrically charged rotating static black hole; for an interpretation of these solutions, see, for example, Refs. 5 and 6).

APPENDIX 1

We have heretofore used the term horizon to refer to a trapping surface, or an apparent horizon, as it is called in the literature.

Let us ascertain the difference between an event horizon and an apparent horizon in greater detail in an example. We assume that we already have a stationary black hole of mass M and that there is an apparent horizon at $r=2M$. Now we assume that another chunk of matter with a mass δM falls into our black hole. After it falls, the radius of the apparent horizon increases to $2(M+\delta M)$. Thus, if an observer is placed between these radii before the additional chunk of matter falls, he would then be outside the black hole, but after the chunk of matter falls he would be inside it. The concept of an event horizon is global and is determined by the entire course of evolution of the black hole or, stated differently, by all the mass which falls into it at any time.

The existence of an apparent horizon, which specifies a black hole locally, is sufficient for the existence of a black hole. As follows from our arguments, in the spherically symmetric case the two horizons ultimately coincide and form a static black hole described by Schwarzchild's solution. Therefore, we shall henceforth use the term horizon to refer to the apparent horizon.

Let us prove that the horizon in a system with spherical symmetry forms at the moment a falling particle with a non-

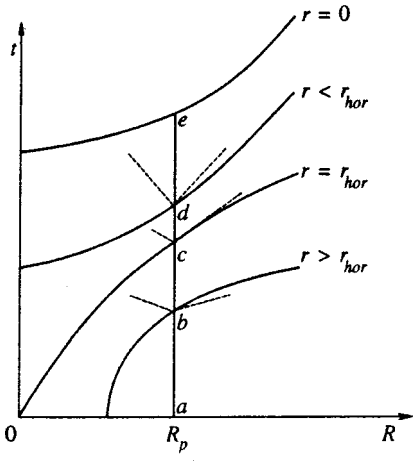

FIG. 1.

zero rest mass attains the speed of light relative to the $r=\text{const}$ surfaces at the same point. For this purpose we write the law of motion for the particle in the form

$$r(R,T)=R-\int_0^t \sqrt{U(R,\tilde{t})}\,d\tilde{t}. \tag{27}$$

We now assume that we are located on a dust particle with the index R_∞ and that we are tracking a dust particle with the index R_p, which sends us a light beam passing through the radii $r_p(R_p,t)$, from the large radius $r_\infty(t)$. The criterion for determining that the dust particle has not yet reached the horizon is the fact that we still see light from it, i.e., as the light propagates it still crosses the radii $r>r_p$. Therefore, the criterion for determining that the dust particle has reached the horizon is an event in which the light propagating from R_p can no longer cross the radii $r>r_p$. Let us express this criterion mathematically.

In Fig. 1 the vertical straight line $abcde$ denotes the world line of an R_p dust particle in the coordinates R and t of the comoving reference frame from the moment of rest (a) to the center of the system (e) at $r=0$. In this case of solid curves passing through points e, d, c, and b denote, respectively, lines of constant values of $r(R,t)$ for $r=0$, $r<r_{\text{hor}}$, $r=r_{\text{hor}}$, and $r>r_{\text{hor}}$. The dashed lines emerging from these points denote the cones within which light emitted by the R_p dust particle can propagate (light cannot propagate outside these cones). Therefore, according to the criterion indicated above, the horizon forms at the point where the cone is tangent to the $r=\text{const}$ line. In the figure this line is designated as $r=r_{\text{hor}}$, and it passes through point c. For clarity, Fig. 1 shows that the light cone intersects lines with $r>r_p$ at point b; therefore, there is still no horizon at that point. This figure also shows that at point d the light cone is located entirely above the $r=\text{const}$ curve passing through point d. Consequently, this light cone intersects only lines with $r<r_p$, and therefore point d is already located below the horizon.

Let us examine the expression (27) on one of the $r=\text{const}$ curves and take its complete differential on that curve:

$$0=dR-\sqrt{U}\,dt-\frac{1}{2}\int_0^t \frac{U'(R,\tilde{t})}{\sqrt{U(R,\tilde{t})}}dR\,d\tilde{t},$$

or

$$\sqrt{U}\,\frac{dt}{dR}\bigg|_{r=\text{const}}=1-\frac{1}{2}\int_0^t \frac{U'(R,\tilde{t})}{\sqrt{U(R,\tilde{t})}}d\tilde{t}. \tag{28}$$

Next, differentiating (27) with respect to R, we obtain the following expression for r':

$$r'=1-\frac{1}{2}\int_0^t \frac{U'}{\sqrt{U}}d\tilde{t}.$$

Using this expression, we obtain the following equation from (28):

$$\frac{dt}{dR}\bigg|_{r=\text{const}}=\frac{r'}{\sqrt{U}}. \tag{29}$$

Thus, we have found an expression for the slope of an $r=\text{const}$ curve relative to the R axis.

To find the slope of a light cone, by definition, for light we have $ds^2=0$. It thus follows from (2) that

$$\frac{dt}{dR}\bigg|_{\text{light}}=\sqrt{e^{\lambda-\nu}}. \tag{30}$$

According to the foregoing statements, the criterion for the absence of a horizon is the condition

$$\frac{dt}{dR}\bigg|_{\text{light}}<\frac{dt}{dR}\bigg|_{r=\text{const}}. \tag{31}$$

Substituting in it the expressions (29) and (30) and using (17), we obtain this criterion in the form

$$|V|=\frac{\sqrt{Ue^{\lambda-\nu}}}{r'}<1. \tag{32}$$

Here, according to (29), the rate of motion of the matter relative to the $r=\text{const}$ lines has the form

$$|V|=\frac{dl}{d\tau}\bigg|_{r=\text{const}}=\sqrt{e^{\lambda-\nu}}\,\frac{dt}{dR}\bigg|_{r=\text{const}}.$$

Thus, the assertion that a horizon forms at the moment the matter attains the velocity $V=1$ relative to the $r=\text{const}$ surfaces has been proved. The horizon surface separates regions in which r is space-similar and time-similar.

APPENDIX 2

To solve the equations describing collapse, we first bring them into dimensionless forms. For this purpose it is convenient to introduce the notation

$$x=r/R,\quad \gamma=\frac{\rho_0(R)}{\langle\rho\rangle}=\frac{8\pi\varepsilon_0(R)R^2}{3S(R)}.$$

In this Appendix we find the ranges of permissible values of γ and S, investigate the character of collapse at these

values of the parameters, and obtain numerical solutions for V^2. We must first of all know the form of the function $r'(x)$. Differentiating (27), we obtain[5)]

$$r'(x)=1+\frac{1}{2}R\int_1^x[\ln U(R,\tilde{x})]'\,d\tilde{x}. \tag{33}$$

Unfortunately, an analytically exact expression for r' can be found only in the case $\alpha=0$; the character of collapse can be assessed exactly only at that value of α. However, the main features of that character, as will be seen below, remain the same as in the case $\alpha\neq 0$. Therefore, let us first investigate the case $\alpha=0$.

Thus, we should find $r'(R,x)$. According to the expression (22) for U, we obtain

$$\ln[U(R,x)]=\ln[S(R)]+\ln\left(\frac{1}{x}-1\right).$$

Introducing the notation $y=r'-x$ and taking into account that $x'=y/R$, we have

$$(\ln U)'=\frac{S'}{S}-\frac{y}{Rx(1-x)}.$$

The substitution of this expression into (33) gives

$$y(x)+x-1=\frac{1}{2}\int_1^x\left[\frac{RS'}{S}-\frac{y(\tilde{x})}{\tilde{x}(1-\tilde{x})}\right]d\tilde{x}. \tag{34}$$

Differentiating (34) with respect to x, we obtain

$$\frac{\partial y(x)}{\partial x}+a^*(x)y(x)=\sigma^*(R), \tag{35}$$

where we have introduced the notation

$$a^*(x)=\frac{1}{2x(1-x)},\quad \sigma^*(R)=\frac{RS'}{2S}-1=-\frac{3}{2}+\frac{3}{2}\gamma.$$

As can be seen, Eq. (35) coincides in form with Eq. (15), and the initial conditions, $y|_{t=0}=0$, are the same; therefore, the method used to solve it is similar. The solution has the form

$$y(x)=r'-x=\sigma^*\left[\sqrt{\frac{1-x}{x}}\arctan\sqrt{\frac{1-x}{x}}-(1-x)\right]. \tag{36}$$

Let us find the domain of r'. First, the condition for compression of the matter has the form $r'\leqslant 1$. Second, the condition that dust layers with different R do not intersect[6)] has the form $r'>0$. Thus,

$$0<r'\leqslant 1. \tag{37}$$

We assume that the $V^2(R)$ curve for $t=t_m=\text{const}$ has a local extremum, and we presume (to fix ideas) that it is a maximum. Then the horizon appears specifically at the local maximum, i.e., the point $R=R_{\text{extr}}$. We now find the condition for a maximum. First, at that point we should have $V^2(R_{\text{extr}},x)=V^2_{\text{extr}}$. Second, since it is the first point at which the velocity of the matter attains the value V_{extr} and since the rate of collapse increases with time, in the vicinity of this point we should have $V^2<V^2_{\text{extr}}$, or

$$\begin{aligned}&\frac{\partial V^2(R,t_m)}{\partial R}>0,\quad R<R_{\text{extr}},\\ &\frac{\partial V^2(R,t_m)}{\partial R}<0,\quad R>R_{\text{extr}}.\end{aligned} \tag{38}$$

If it turns out that (38) holds with opposite inequality signs, there will be a local minimum on the $V^2(R)$ curve at the point R_{extr} at the moment the velocity V_{extr} is attained at that point; i.e., the matter will attain the velocity V_{extr} last at that point.

The condition for an extremum is written in the form

$$\frac{\partial V^2(R,x)}{\partial R}=0,$$

where, according to (23),

$$V^2(R,x)|_{\alpha=0}=\frac{1-1/x}{1-a/S(R)}.$$

Differentiating this expression with respect to R, we obtain

$$\left.\frac{\partial V^2}{\partial R}\right|_{t=t_m}=\frac{S/R}{1-S}\left[-y+\frac{1-3\gamma}{1-S}\right], \tag{39}$$

where it has been taken into account that $x'=y/R$ and $S'/S=(3\gamma-1)/R$. Then, with allowance for the fact that $0<x\leqslant 1$, $0<\gamma\leqslant 1$, $-1<y\leqslant 0$, and $0<S<1$, the condition (38) can be rewritten in the form

$$\begin{aligned}&-y>\frac{3\gamma-1}{1-S},\quad R<R_{\text{extr}},\\ &-y<\frac{3\gamma-1}{1-S},\quad R>R_{\text{extr}}.\end{aligned} \tag{40}$$

If we introduce the notation $z=\sqrt{(1-x)/x}$ and take into account that, according to formula (23), $z=V\sqrt{1/S-1}$, from (40) we obtain

$$\begin{aligned}&\frac{3}{2}(1-\gamma)\left[z\arctan(z)-\frac{z^2}{1+z^2}\right]-\frac{3\gamma-1}{1-S}>0,\quad R<R_{\text{extr}},\\ &\frac{3}{2}(1-\gamma)\left[z\arctan(z)-\frac{z^2}{1+z^2}\right]-\frac{3\gamma-1}{1-S}<0,\quad R>R_{\text{extr}},\end{aligned} \tag{41}$$

or for the extremum point we can write

$$(1-\gamma)\left[z\arctan(z)-\frac{z^2}{1+z^2}+\frac{2}{1-S}\right]-\frac{4/3}{1-S}=0.$$

This formula can be used to construct the plot of $\gamma(S,V_{\text{extr}})$ separating positive and negative values of the derivative $(V^2)'$ and to determine the character of the extremum. The corresponding curves for various values of V_{extr} are shown in Fig. 2. The regions where $V'>0$ are located above and to the right of them, and the regions where $V'<0$ are located below and to the left of them.

It is seen from Fig. 2 that there can be (for a definite choice of the distribution profile of the matter in the initial

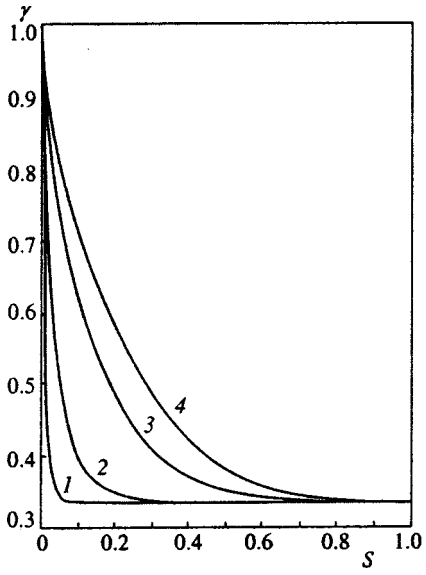

FIG. 2. Plots of $\gamma(S)$ for the existence of a velocity maximum for $t=\text{const}$, $\alpha=0$, and $V_{\text{extr}}=0.1$ (curve *1*), 0.3 (*2*), 0.7 (*3*), and 1.0 (*4*).

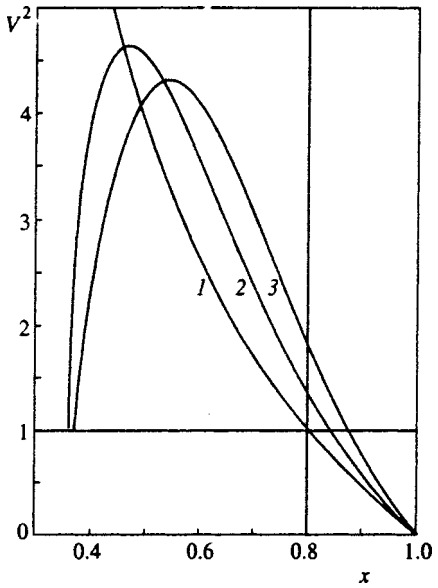

FIG. 3. Plots of $V^2(x)$ for the initial conditions $S=0.8$, $\gamma=0.6$, $\eta=0.1$, and $\alpha=0$ (curve *1*), 0.15 (*2*), and 0.33 (*3*). The vertical straight line corresponds to the position of the apparent horizon for $\alpha=0$.

moment and of the parameter S at a certain point R_*) two values of R, at which $V=1$ at a certain moment in time, and, therefore, the appearance of a second horizon in the system is possible.

The appearance of a second horizon is no news in the physics of black holes (see, for example, the Reissner–Nordström or Kerr–Newman solution in Ref. 7).

The results obtained in this Appendix apply to the case of the absence of pressure, although the case $\alpha=1/3$ is of experimental interest. Therefore, we used formulas (4), (6), (8), (9), (12), (17), (18), (20), and (33) to introduce new dimensionless variables ($\hat{\nu}=\nu-\nu_0$, $\hat{\lambda}=\lambda-\lambda_0$, $\hat{U}=Ue^{-\nu_0}$, and $\hat{\varepsilon}=\varepsilon/\varepsilon_0$) and equations for them. The initial conditions for them take the form

$$\hat{\nu}_0=\hat{\lambda}_0=\hat{U}_0=0, \quad r'_0=\hat{\varepsilon}_0=1.$$

Designating the new coordinates as $x=r/R$ and $\xi=R/R_*$ ($R_*=\text{const}$) and introducing the parameters[7)]

$$h=\xi\partial_\xi\nu_0, \quad h_{\text{cr}}=S\frac{1+3\alpha\gamma}{1-S}, \quad \eta=h/h_{\text{cr}},$$

we obtain equations for the new variables in the form

$$e^{\hat{\nu}/\alpha}=x^4e^{\hat{\lambda}}, \quad \hat{\varepsilon}=(e^{-\hat{\nu}/\alpha})^{(1+\alpha)/2},$$

$$\ln(r'^2e^{-\hat{\lambda}})=\int_1^x\left[\xi\frac{\partial_\xi\hat{\nu}}{r'}+\frac{h}{r'}\right]d\tilde{x}, \tag{42}$$

$$\hat{U}=e^{\hat{\nu}}\left[r'^2e^{-\hat{\lambda}}(1-S)-1+\frac{S}{x}-\alpha\frac{3\gamma S}{x}\int_1^x\hat{\varepsilon}\tilde{x}^2\,d\tilde{x}\right],$$

$$r'=1+\frac{1}{2}\int_1^x\xi\partial_\xi(\ln\hat{U})\,d\tilde{x}-\frac{h}{2}(1-x).$$

In the new variables the velocity is

$$V^2=\frac{\hat{U}e^{\hat{\lambda}-\hat{\nu}}}{(r')^2(1-S)}.$$

Hence $x_{\text{hor}}=S+2\alpha\tilde{m}$, where

$$\tilde{m}=\frac{m}{R}=\varepsilon_0R^2\int_x^1 4\pi\hat{\varepsilon}\tilde{x}^2\,d\tilde{x}, \quad \varepsilon_0R^2=\frac{3\gamma S}{8\pi}. \tag{43}$$

We note that this formula and formula (36) can be used to find the corrections δr_{hor} in (25) for the displacement of the horizon in the linear approximation with respect to α since, according to Eq. (103.11) from Ref. 1 for $\alpha=0$, we have

$$8\pi\varepsilon r^2=\frac{2M'}{r'}=\frac{8\pi\varepsilon_0R^2}{r'},$$

or

$$\hat{\varepsilon}x^2=1/r'.$$

This expression can be substituted into (43) and a quadrature expression can be obtained for the sought-for correction.

In addition, we numerically integrated the equations for the case $\alpha\neq0$ using a difference scheme, and the results for various values of α are presented in Fig. 3. According to (25), the plots of $V^2(x)$ are displaced upward and to the right, as they should be, as α is increased from $\alpha=0$ to $\alpha=1/3$.

The numerical calculations confirm that the analytical results of this Appendix remain valid for the real equation of state of matter: $P=\alpha\varepsilon$. Figure 4 shows plots of $\gamma(S)$, $\xi(S)$, and $\eta(S)$ for the special case of a Gaussian density distribution: $\varepsilon_0(\xi)/\varepsilon_0(0)=\exp(-3\xi^2)$. Comparing this figure with Fig. 2, we see that the $\gamma(S)$ curve in Fig. 4 crosses the $\gamma(S)$ curves in Fig. 2 in the downward direction roughly at the point $S\approx0.92$, if we proceed from $\xi=0$ to $\xi=1$. As can be

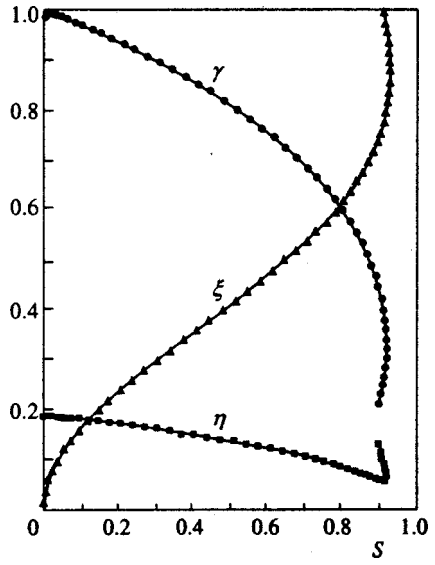

FIG. 4. Plots of $\gamma(S)$, $\eta(S)$, and $\xi(S)$ for $\varepsilon(\xi)/\varepsilon(0)=\exp(-3\xi^2)$.

seen in Fig. 4, the point $S\approx 0.92$ corresponds to $\xi\approx 0.85$ and $\eta\approx 0.1$; therefore, since the region where $V'>0$ is located above and to the right of the curves in Fig. 2 and the region where $V'<0$ is located below and to the left of these curves, the point $\xi\approx 0.85$ should be a local maximum on the $V(\xi)$ curve for a constant value of t.

This analytical result is confirmed by a numerical calculation of $V(\xi)|_{t=\text{const}}$ curves, whose results are shown in Fig. 5 with the predicted maxima.

To conclude this Appendix we would like to say a few words regarding the initial characteristics and distribution of the matter.

When the equations of the model were brought into dimensionless form, it was found that the solution for a spherical layer of matter with the index R is completely specified by three dimensionless parameters in the initial moment in that layer: $0<S<1$, $0<\gamma<1$, and $0<\eta<1$. This corresponds to assigning the initial conditions for the gravitational potential and two parameters which determine the distribution of the matter and the pressure gradient near the point under consideration. Thus, upon integration we at once find a whole family of self-similar solutions[8)], which can be characterized by these three parameters alone and which contains the dependence on the other layers of matter above and below the radius R considered.

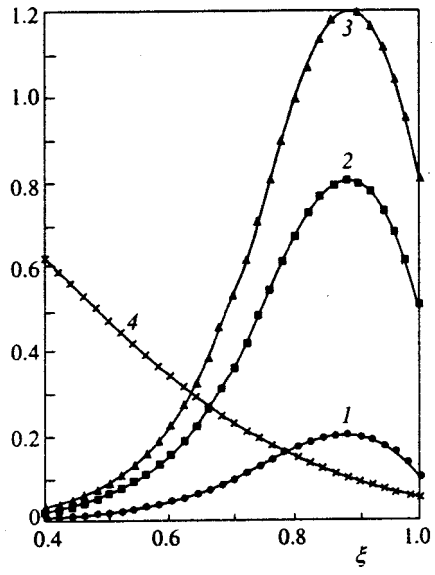

FIG. 5. Plots of $V^2(\xi)$ for $t=\text{const}$ and various initial values of the velocity V^2 (curves *1–3*), as well as a plot of $\varepsilon(\xi)/\varepsilon(0)=\exp(-3\xi^2)$ (curve *4*).

We thank N. S. Kardashev, V. L. Ginzburg, B. V. Komberg, V. N. Lukash, and Yu. M. Bruk, as well as all the participants in the seminars of the Division of Theoretical Physics and the Astrocosmic Center of the P. N. Lebedev Physics Institute of the Russian Academy of Sciences for fruitful discussions of this work and for their important comments.

With our sincerest gratitude we share the memory of D. A. Kirzhnits, with whom we formulated the ideas and initial approaches used in this work.

This work was supported by the Russian Fund for Fundamental Research (Grant No. 96-15-96616).

*)E-mail: aas@srdlan.npi.msu.su

1)Only the gravitational forces and the forces of interaction between the dust particles, i.e., the pressure, act on the matter.

2)In case under discussion the word ''dust'' does not, in any way, imply the absence of interactions between the particles, except in the case $\alpha=0$.

3)When there is an horizon, the components of the metric in the reference frame which is stationary at infinity, like the Schwarzchild metric, must acquire singularities as the surface of that horizon is approached.

4)Here we are dealing with the so-called apparent horizon, or, as it is also called, trapping surface, whose definition has been given in Refs. 5 and 6. As for the event horizon, as will be seen below, after all the matter falls, it coincides with the apparent horizon (for further details, see Appendix 1).

5)Here and below $d\tilde{x}=(\dot{r}\,dt/R)|_{R=\text{const}}$ and $\partial/\partial x=(R/\dot{r})(\partial/\partial t)|_{R=\text{const}}$.

6)We consider just such a case, since intersection of the layers would signify the appearance of an infinite density already at a radius $r\neq 0$, which is forbidden by cosmic censorship (see, for example, Ref. 6).

7)The significance of h_{cr} is that, according to (3a), $\ddot{r}=0$ when $h=h_{\text{cr}}$ in the initial moment, i.e., the system is in an unstable equilibrium. Here and below $\partial_\xi=\partial/\partial\xi$.

8)For further information on this subject, see, for example, Ref. 3.

[1] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, 4th ed., Pergamon, New York (1975) [Russ. original of newer edition, Nauka, Moscow (1988)].
[2] M. A. Podurets, Dokl. Akad. Nauk SSSR **154**, 300 (1964) [Sov. Phys. Dokl. **9**, 1 (1964)].
[3] M. W. Choptuik, Phys. Rev. D **44**, 3124 (1991); E-prints archive gr-qc/9607034.
[4] P. S. Wesson, J. Math. Phys. **19**, 2283 (1978).
[5] R. Penrose, *Structure of Space-Time*, W. A. Benjamin, New York (1968) [Russ. transl., Bibfizmat, Mogilev (1972)].
[6] I. D. Novikov and V. P. Frolov, *Physics of Black Holes*, Kluwer Academic, Dordrecht–Boston (1989) [Russ. original, Nauka, Moscow (1986)].
[7] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, W. H. Freeman, San Francisco (1973) [Russ. transl., Mir, Moscow (1977)].

Translated by P. Shelnitz
Edited by S. J. Amoretty

High orders of perturbation theory. Are renormalons significant?

I. M. Suslov*)

P. L. Kapitza Institute of Physical Problems, Russian Academy of Sciences, 117334 Moscow, Russia
(Submitted 25 February 1999)
Zh. Éksp. Teor. Fiz. **116**, 369–389 (August 1999)

According to Lipatov [Sov. Phys. JETP **45**, 216 (1977)], the high orders of perturbation theory are determined by saddle-point configurations, i.e., instantons, which correspond to functional integrals. According to another opinion, the contributions of individual large diagrams, i.e., renormalons, which, according to t'Hooft [*The Whys of Subnuclear Physics: Proceedings of the 1977 International School of Subnuclear Physics (Erice, Trapani, Sicily, 1977)*, A. Zichichi (Ed.), Plenum Press, New York (1979)], are not contained in the Lipatov contribution, are also significant. The history of the conception of renormalons is presented, and the arguments in favor of and against their existence are discussed. The analytic properties of the Borel transforms of functional integrals, Green's functions, vertex parts, and scaling functions are investigated in the case of φ^4 theory. Their analyticity in a complex plane with a cut from the first instanton singularity to infinity (the Le Guillou–Zinn-Justin hypothesis [Phys. Rev. Lett. **39**, 95 (1977); Phys. Rev. B **21**, 3976 (1980)] is proved. It rules out the existence of the renormalon singularities pointed out by t'Hooft and demonstrates the nonconstructiveness of the conception of renormalons as a whole. The results can be interpreted as an indication of the internal consistency of φ^4 theory. © *1999 American Institute of Physics.*
[S1063-7761(99)00208-5]

1. INTRODUCTION

Many problems in theoretical physics can be reduced to a calculation of functional integrals of the type

$$I=\int D\varphi \exp(-S_0\{\varphi\}-gS_{\text{int}}\{\varphi\}), \tag{1}$$

whose expansion in the coupling constant g gives an ordinary perturbation scheme. In 1977 Lipatov[1] proposed a method for calculating the long-range expansion coefficients of the integrals (1) on the basis of the following simple idea. If the function $F(g)$ is expanded in a series

$$F(g)=\sum_{N=0}^{\infty} F_N g^N,$$

the Nth expansion coefficient can be calculated from the formula

$$F_N=\oint_C \frac{dg}{2\pi i}\frac{F(g)}{g^{N+1}}, \tag{2}$$

where the contour C encompasses the point $g=0$ in the complex plane. Taking the integral (1) as $F(g)$, we obtain

$$I_{N-1}=\frac{1}{2\pi i}\int dg \int D\varphi \exp(-S_0\{\varphi\} - gS_{\text{int}}\{\varphi\}-N\ln g), \tag{3}$$

and the appearance of an exponential function with a large exponent at large N raises hope that the saddle-point method can be employed. Lipatov's idea is to seek the saddle point in (3) with respect to g and φ simultaneously: it exists for all the cases of interest and is realized on a function localized in space, which is called an instanton. Moreover, the conditions for applicability of the saddle-point method are satisfied for large values of N.

The Lipatov technique, which was originally applied to scalar φ^4 theories,

$$S_0\{\varphi\}+gS_{\text{int}}\{\varphi\}=\int d^dx\left\{\frac{1}{2}(\nabla\varphi)^2+\frac{1}{2}m^2\varphi^2+\frac{1}{4}g\varphi^4\right\}, \tag{4}$$

was subsequently generalized to vector fields,[2] scalar electrodynamics,[3,4] Yang–Mills fields,[5] fermion fields,[6] etc. (see the collection of articles in Ref. 7). The ultimate goal was to apply it to theories of practical interest, viz., quantum electrodynamics[8,9] and quantum chromodynamics (QCD).[10,11] As was pointed out already in Lipatov's first paper,[1] knowledge of the first few coefficients and their asymptotes permits approximate reconstruction of the Gell-Mann–Low function, opening up a direct route to the solutions of the problem of confinement and electrodynamics at short distances.

However, a conception which raised some doubts regarding the Lipatov technique originated in 1977. It was initiated in a paper by Lautrup,[12] which contained the following curious remark. The typical result of calculations based on the Lipatov technique has a functional form,

$$I_N=ca^N\Gamma(N+b)\approx ca^N N^{b-1}N!, \tag{5}$$

and the natural interpretation of it is that there is a factorially large number of diagrams of the same order $(ag)^N$. However, in general, such an interpretation is incorrect, since there are examples of individual Nth-order diagrams having a value $\sim N!$. The latter are diagrams (Fig. 1) which contain

 © 1999 American Institute of Physics

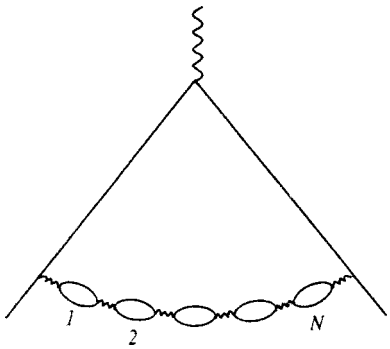

FIG. 1. Example of a diagram for quantum electrodynamics, which makes a contribution $\sim N!$ to the Nth order of perturbation theory.[12]

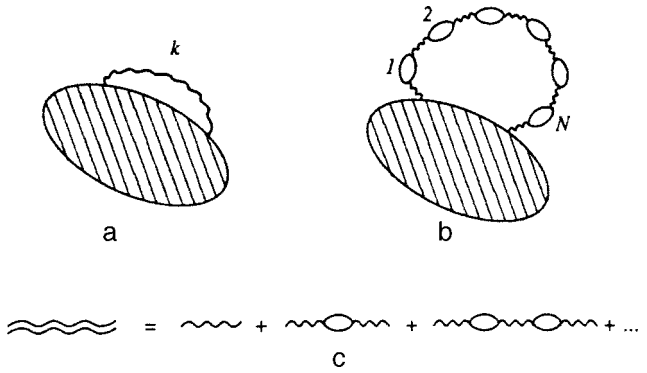

FIG. 2. More general conception of renormalons.

long chains of ''bubbles'' (bubble diagrams). Such factorial contributions of individual diagrams are called renormalons, since they appear only in renormalizable theories.[1)] Lautrup's example (Fig. 1) was related to quantum electrodynamics, but similar diagrams exist in QCD and four-dimensional φ^4 theory.

Strictly speaking, nothing followed from Lautrup's remark: the Lipatov technique is based on a formal calculation of the functional integral (3) and does not rely in any way on a statistical analysis of diagrams. It is natural to expect that the renormalon contributions have already been taken into account in the instanton result (5). In fact, no far-reaching claims were made in Ref. 12 or in the relevant publications appearing shortly thereafter.[13,14]

However, the tone of the publications advocating this doctrine subsequently changed dramatically. The reason was t'Hooft's lecture,[15] which was delivered at a seminar in the same year, 1977. The term renormalon was used in it for the first time, and it was asserted that renormalons are not contained in the instanton contribution (5). The authors of the subsequent publications[16–30] considered t'Hooft's opinion to be self-evident and did not trouble themselves with argumentation.

A convenient language for discussion, viz., the analytic properties of Borel transforms, was introduced in t'Hooft's lecture. The Borel transformation

$$F(g)=\sum_{N=0}^{\infty} F_N g^N=\sum_{N=0}^{\infty}\frac{F_N}{N!}\int_0^{\infty} dx\, x^N e^{-x} g^N = \int_0^{\infty} dx\, e^{-x}\sum_{N=0}^{\infty}\frac{F_N}{N!}(gx)^N,$$

which factorially takes into account the convergence of series, is widely used in the theory of divergent series.[30] It is convenient to rewrite it in the form

$$F(g)=\int_0^{\infty} dx\, e^{-x} B(gx), \tag{6}$$

$$B(z)=\sum_{N=0}^{\infty}\frac{F_N}{N!} z^N \tag{7}$$

by introducing the Borel transform $B(z)$ of the function $F(g)$. The Borel transform for a function with the expansion coefficients (5)

$$B_I(z)=\sum_N ca^N N^{b-1} z^N \sim (1-az)^{-b}, \quad za\to 1 \tag{8}$$

has a singularity at the point $z=1/a$.

t'Hooft arrived at this result in a different way, without reference to the Lipatov technique. Rewriting the integral (1) and the definition of the Borel transform (6) in the form

$$I=\int D\varphi \exp(-S\{\varphi\}/g), \tag{9}$$

$$F(g)=\int_0^{\infty} dx\, e^{-x/g} B(x), \tag{10}$$

which can be accomplished by means of the replacements $\varphi\to\varphi/\sqrt{g}$ and $x\to x/g$ in (4) and (6)[2)], yields the Borel transform of the integral (9):

$$B_I(z)=\int D\varphi\,\delta(z-S\{\varphi\})=\oint_{z=S\{\varphi\}}\frac{d\sigma}{|\nabla S\{\varphi\}|}, \tag{11}$$

where the latter integration is carried out over the hypersurface $z=S\{\varphi\}$. If an instanton $\varphi_c(x)$, i.e., a classical solution with a finite action, exists for the integral (9), then $\delta S\{\varphi_c\}=0$ and the partial derivatives $\partial S/\partial\varphi_i$ with respect to all the variables φ_i comprising $D\varphi$ vanish; therefore, $\nabla S\{\varphi_c\}=0$ and the Borel transform (11) has a singularity at the point

$$z=S\{\varphi_c\}, \tag{12}$$

which coincides with $1/a$. In addition, there are singularities at the points $mS\{\varphi_c\}$, which correspond to solutions containing m infinitely distant instantons. If it is assumed that the singularity (12) is closest to the origin of coordinates, the result (5) of the Lipatov technique is reproduced. However, t'Hooft allowed the existence of singularities differing from those of the instanton type: in this case the asymptotic behavior of the expansion coefficients can be specified by the singularity which is closest to the origin of coordinates.

t'Hooft regarded renormalons as a possible mechanism for the appearance of the new singularities. Let us use an arbitrary diagram for quantum electrodynamics and single out the line of a virtual photon with the index k (or an interaction line in φ^4 theory) in it (Fig. 2a): it corresponds to

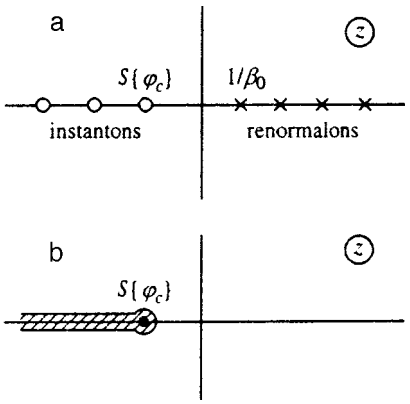

FIG. 3. a) Picture of singularities for φ^4 theory given by t'Hooft.[15] b) Analyticity region according to the results of our work.

integration over a region of large momenta of the type

$$\int_{k>k_0} d^4k\, k^{-2m},$$

where m is an integer. If we assume that all the renormalizations have been performed, the integral converges and $m \geqslant 3$. Inserting a chain of N ''bubbles'' into the photon line (Fig. 2b), we obtain the integral[3)]

$$\int_{k>k_0} d^4k\, k^{-2m}(\beta_0 \ln k^2)^N \sim \beta_0^N \int_0^\infty dt\, t^N e^{-(m-2)t} \sim \left(\frac{\beta_0}{m-2}\right)^N N!. \tag{13}$$

Borel summation of a sequence of such diagrams gives singularities at the points

$$z = \frac{m-2}{\beta_0}, \quad m = 3,4,5,\ldots. \tag{14}$$

The constant β_0 is the first nonvanishing expansion coefficient of the Gell-Mann–Low function (Sec. 2), and with consideration of the sign relationships ($S\{\varphi_c\}<0$, $\beta_0>0$) t'Hooft obtained the picture of singularities for φ^4 theory shown in Fig. 3a.

It is not difficult to see that t'Hooft's arguments regarding renormalons leave some cardinal questions unanswered:

Why should significance be attached to individual sequences of diagrams, which make up an infinitesimal fraction of their total number when $N \to \infty$?

How do we know that the renormalons have not already been taken into account in the instanton contribution (5)?

However, the general formulation of the question of the possibility of contributions of a noninstanton nature to the asymptotics of the expansion coefficients has meaning: it uncovers a hole in the mathematical foundation of the Lipatov technique. In fact, let the function $f(x)$ have a short maximum at the point x_0 and a slow tail at large values of x (Fig. 4), so that the contributions to the integral $\int f(x)\,dx$ from the vicinity of the maximum and from the tail region are comparable. An investigation of the integral for a saddle point discloses a maximum at x_0 and (provided it is sufficiently sharp) the formal applicability of the saddle-point method; however, a calculation of the integral in the saddle-point approximation will be erroneous, since the contribution of the tail will be lost. If such tails are present in the integral (3), the Lipatov technique can be incorrect.[4)]

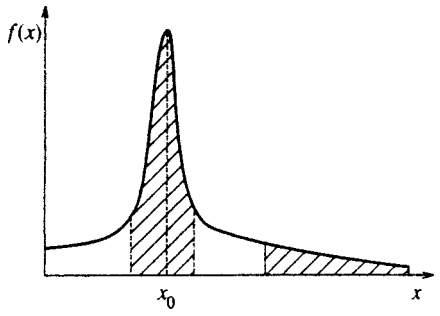

FIG. 4. Example of a function for which the saddle-point method is applicable, but gives an incorrect result.

The essential lack of nonsaddle-point methods for calculating functional integrals makes it impossible to directly investigate the contribution of possible tails. But are there constructive arguments pointing to their existence? In principle, such arguments exist, but they have a fairly intuitive and ambiguous character and do not hold up to criticism when they are closely examined (Sec. 2). As a result, the conception of renormalons has been in a dialectic equilibrium, i.e., it has not been proved or disproved. This uncertainty has caused the interest in high orders of perturbation theory to drop sharply and Lipatov's program[1] to remain uncompleted. For example, a preliminary result for quantum electrodynamics was obtained in 1978 (Ref. 9), but the parameters b and c in the asymptotic relation $I_N = ca^N\Gamma(N/2+b)$ have not yet been calculated. Moreover, the first result for QCD appeared in 1991 (Ref. 10) and was recently revised,[11] but it is still unsatisfactory (Sec. 4), although the foundation for such calculations was completely ready in 1980.[5,6] Finally, the attempts to reconstruct the Gell-Mann–Low function have been confined to φ^4 theory.[33–35]

A reawakening of interest in asymptotic estimates has recently been observed, but it has been confined almost exclusively to the renormalon doctrine.[21–30] In particular, it is generally accepted (see Zakharov's review article[21]) that renormalons determine the perturbation asymptotics in QCD. However, the work within the renormalon approach has already raised some doubts: the summation of larger sequences of diagrams leads to dramatic renormalization of the renormalon contribution and renders the common coefficient in front of them totally indefinite;[30] in fact, it is impossible to state that it does not vanish. On the other hand, the use of the Lipatov technique has provided significant progress in the theory of disordered systems[36] and in the theory of turbulence.[37]

This paper presents a detailed discussion of the existing arguments in favor of renormalons, which are shown to be unsound (Sec. 2). The analytic properties of Borel transforms are investigated in the example of φ^4 theory (Sec. 3), and their analyticity in a complex plane with a cut from the first instanton singularity to infinity is demonstrated (Fig. 3b). It rules out the existence of the renormalon singularities indi-

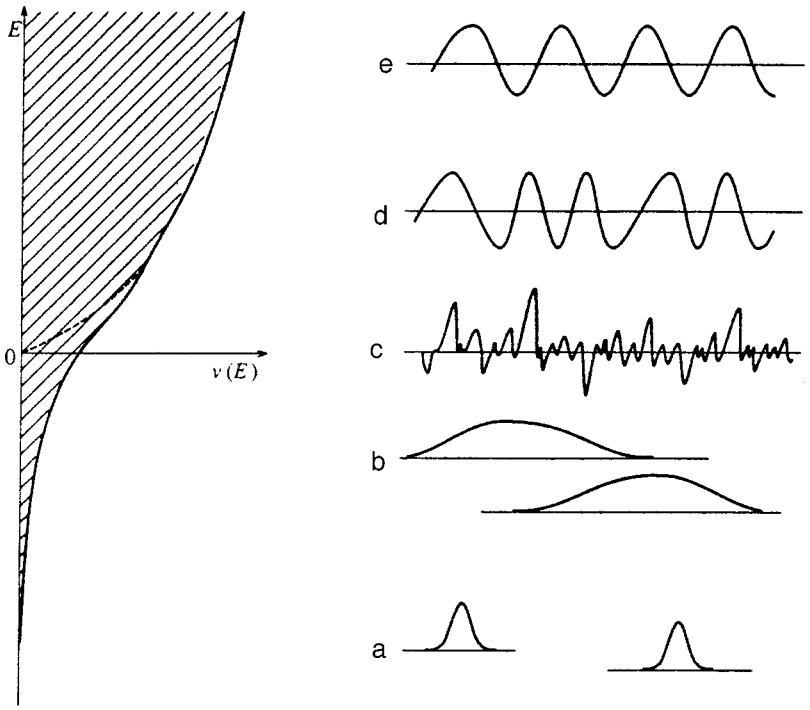

FIG. 5. Left-hand figure — density of states $\nu(E)$ in the presence (solid curve) and absence (dashed curve) of a random potential, right-hand figure — schematic representations of the eigenfunctions of Eq. (15).

cated by t'Hooft (Fig. 3a) and demonstrates the baselessness of the conception of renormalons as a whole.

An hypothesis that Borel transforms have the analytic properties indicated above was advanced by Le Guillou and Zinn-Justin[38] and underlies one of the most effective methods for summing perturbation series, which is known as the ''Borel conform.'' The results obtained below provide the mathematical foundation of this method.

2. PROS AND CONS

Let us discuss the arguments in the literature that point to the existence of noninstanton contributions in the integral (3).

1. There have been numerous semi-intuitive assertions which reduce to the notion that *instantons do not exhaust all of physics*. This thesis is correct as long as the way in which it is understood is correct, but in the present case it is not relevant.

Historically, instantons first appeared when the saddle-point approximation was employed in the original integral (1). It was substantiated only in a narrow region of parameters, and thus instantons did not, in fact, exhaust all of physics. In the Lipatov technique the situation changed, because the saddle-point approximation is used not in the integral (1), but in the expression (3) for the expansion coefficients. Since only large values of N are considered, only a limited role is assigned to instantons at the outset; however, the saddle-point method is now always applicable[5)], and there is a basis to assume that everything is determined by instantons.

Let us illustrate the foregoing statements in the example of the Schrödinger equation with a random potential $V(x)$:

$$[\hat{p}^2/2m + V(x)]\Psi(x) = E\Psi(x). \tag{15}$$

At large negative values of E its eigenfunctions (Fig. 5) are localized on the infrequent fluctuations of the random potential (a and b), at large positive values of E they are similar to plane waves (d and e), and in the vicinity of the bare spectrum edge at $E=0$ they have a highly broken, fractal character (c). The problem of investigating Eq. (15) can be reformulated in the language of an effective field theory, viz., φ^4 theory with the ''incorrect'' sign for g (Refs. 36 and 39). In this case the typical wave functions of localized states are described by instantons. The changes in the situation observed as E increases can be described in the following manner in terms of instantons: At first, instantons have a small radius and a sparse distribution, i.e., they form an ideal gas (a); then the radius of the instantons increases, and their density rises, i.e., they begin to interact with one another (b); they then condense (c), and an instanton crystal forms (d and e). Only the case in Fig. 5a corresponds to applicability of the saddle-point method in an integral of the type (1), and thus the standard instanton approximation is very poor.

Let us examine this situation from the standpoint of perturbation theory with respect to the random potential $V(x)$. An ideal instanton crystal (e) corresponds to a plane wave, i.e., the zeroth order of perturbation theory. In a nonideal crystal (d) the higher orders have a larger role; in the vicinity of the bare spectrum edge (c) all the diagrams are of the same order of magnitude, so that the high and low orders of perturbation theory are equally significant. In the region of localized states (a and b) the dominant role shifts to the high orders: These states are not manifested in any finite order of perturbation theory, and discarding the low-order contributions does not influence their properties in any way. We see that Lipatov's conception (high orders are determined by instantons) fits excellently into the existing physical picture.

Thus, the status of instantons in the integral (1) differs greatly from that in the integral (3). In our opinion, this

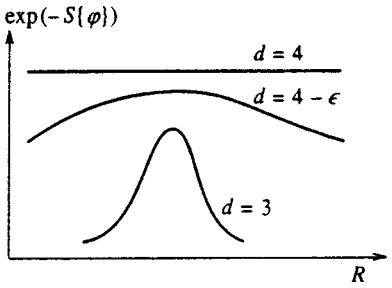

FIG. 6. Dependence of the integrand (1) on the instanton radius in d-dimensional φ^4 theory.

accounts for the position taken by t'Hooft, since he wrote a classical paper on instantons[40] specifically in the original integral (1).

2. Relationship to the logarithmic situation.[15] Renormalons exist only in renormalizable theories, but not in superrenormalizable theories. If a theory is superrenormalizable, an upper bound of the type $a^N g^N$ can be obtained for the contribution of an individual diagram, and the appearance of the multiplier $N!$ in the asymptotic relation (5) can be attributed only to a factorially large number of diagrams. Renormalons and, thus, a new mechanism for the appearance of factorial contributions appear in renormalizable theories. It can be expected that this mechanism is associated with the formation of the tails in the integral (3) and is disregarded in the Lipatov technique.

In this argument everything except the last conclusion is correct. We can illustrate this in the case of φ^4 theory, which is renormalizable for $d=4$ and superrenormalizable for $d<4$. Among the large set of integrations concealed in the symbol $D\varphi$ in the integral (3), we can single out one for which the limiting transition $d\to 4$ is associated with qualitative changes: it is the integration over the instanton radius R (Fig. 6). For d significantly smaller than 4 (for example, $d=3$), the integrand $\exp(-S\{\varphi\})$ has a sharp maximum as a function of R and allows saddle-point integration; when $d=4-\epsilon$, the maximum becomes gently sloping, and when $d=4$, the instanton action $S\{\varphi_c\}$ does not depend on R at all. In the latter case the integral diverges, leading to the logarithmic situation. We see (see the curve for $d=4-\epsilon$ in Fig. 6) that the ''activation'' of renormalon contributions is, in fact, related to the appearance of slow tails in the integral (3), but these tails are taken into account in the Lipatov technique.[36]

A certain enigma can arise. If the Lipatov technique is a saddle-point method, then how can it cover the definitely nonsaddle-point situation for $d=4-\epsilon$? The fact is that the saddle point in a functional integral essentially never reduces to a simple maximum achieved at a single point: The maximum is degenerate in a certain space of finite dimensionality. Accordingly, a finite number of integrations should be performed exactly, rather than in the saddle-point approximation. However, if the integration is performed exactly over a certain variable (for example, R), it is of no significance whether the degeneracy is exact ($d=4$) or approximate ($d=4-\epsilon$). Nevertheless, in the latter case technical difficulties arise, and the corresponding methods (constrained instantons[41,42]) have until recently been poorly developed.[36]

It is thus clear that even in cases where slow tails actually appear in the integral (3), the Lipatov procedure is sufficiently flexible and contains broad possibilities for dealing with them.

3. The limit $n\to\infty$. There is an opinion that the existence of renormalon contributions can easily be proved by treating the n-component φ^4 theory in the limit $n\to\infty$ (and the analogous models in QCD and electrodynamics):[16] The multiplier n corresponds to a closed loop, and renormalon graphs containing the maximum possible number of loops are isolated with respect to the large parameter n. Although diagrams of the same order, but with a smaller number of loops, can make comparable contributions at large values of N due to the combination multipliers, they have a slower dependence on n; therefore, the renormalons cannot be canceled identically.

This argument is valid in any finite order with respect to $1/n$. However, a detailed investigation of the structure of the $1/n$ expansion[18,19] reveals the presence of numerous cancellations, and although the situation cannot be totally elucidated, the question is not resolved on the level of simple arguments of the type indicated above.

It is not difficult to identify the crux of the problem here. As an example, let us consider the self-energy $\Sigma(p,m)$ of φ^4 theory; it is clear from a diagrammatic analysis for $m=0$ and values of the momentum p close to the truncation value Λ that the $(N+1)$th expansion coefficient for $\Sigma(p,0)-\Sigma(0,0)$ has the form of a polynomial in n

$$p^2\{A_N(N)n^N+A_{N-1}(N)n^{N-1}+\ \ldots\ +A_1(N)n+A_0(N)\}, \tag{16}$$

in which the coefficient $A_N(N)$ is specified by renormalon graphs:

$$A_N(N)=\text{const}\cdot\left(-\frac{1}{16\pi^2}\right)^N N!. \tag{17}$$

If it is assumed that the renormalon contribution is contained in the instanton contribution, the expression (16) should transform into the Lipatov asymptotic relation at large N [see Eq. (130) in Ref. 43 for $M=1$ and $p\approx\Lambda$]:

$$p^2\alpha\beta^n N^{(n+6)/2}\left(-\frac{3}{8\pi^2}\right)^N N!\left[\Gamma\left(\frac{n+2}{2}\right)\right]^{-1}$$
$$\times\int_0^\infty dy\, y^{(n+5)/3}K_1(y)^2, \tag{18}$$

where $\alpha,\beta\sim 1$, and $K_1(x)$ is a modified Bessel function of the second kind. It is easily seen that an equality between (16) and (18) is impossible when $n\to\infty$. This is a manifestation of the ''noncancelability'' of renormalons.

However, the usual condition for applicability of the Lipatov technique, $N\gg 1$ at large n is generally replaced by a more rigid condition, for example, $N\gg n$, and n then has a bound of the type

$$n\lesssim n_0(N), \tag{19}$$

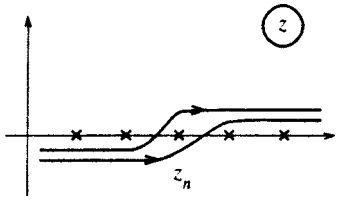

FIG. 7. Ambiguity of the choice of the integration path in the Borel integral (6) in the presence of renormalon singularities.

which precludes going to the limit $n\to\infty$. If it is taken into account that the Lipatov asymptotic relation has limited accuracy ($\sim 1/N$ in relative units), then the correct formulation of the question is as follows. Can we construct an interpolation polynomial of type (16) with the high-order coefficient (17) which would approximate the function (18) within an assigned accuracy in the interval $0\leqslant n\leqslant n_0(N)$, where $n_0(N)\to\infty$ as $N\to\infty$? The answer to this question is positive (see the Appendix); therefore, the assumption that the renormalon contribution is contained in the instanton contribution does not lead to contradictions.

4. *Relationship to a Landau pole*.[16,19] It is not difficult to see (Fig. 2c) that the summation of a sequence of renormalon diagrams corresponds to ''dressing'' the interaction. The relationship between the renormalized charge g and the bare charge g_0 is then given by the familiar expression[44–46]

$$g_0=\frac{g}{1-\beta_0 g\ln(\Lambda^2/m^2)}, \tag{20}$$

which contains a pole at the point

$$\Lambda_c^2=m^2e^{1/\beta_0 g}. \tag{21}$$

Under the literal interpretation of this pole in the spirit of the early paper by Landau and Pomeranchuk,[47] a graphic physical meaning can be assigned to renormalon singularities (Refs. 16 and 19).[6)]

The dependence of the perturbation series on the truncation parameter Λ has the structure

$$c_{-1}\Lambda^2+c_0\ln\Lambda^2+c_1\Lambda^{-2}+c_2\Lambda^{-4} + \ldots + c_n\Lambda^{-2n}+\ldots. \tag{22}$$

The first two terms are eliminated by a renormalization procedure, and the remaining terms, in principle, remain, but vanish in the limit $\Lambda\to\infty$. Because of the pole in (20), values of Λ greater than Λ_c are unattainable in principle, and unremovable uncertainties of the type

$$\Lambda_c^{-2n}=m^2e^{-n/\beta_0 g} \tag{23}$$

appear in the theory. Similar uncertainties are generated by renormalon singularities, whose existence on the positive semiaxis leads to ambiguity in the choice of the integration path in the Borel integral (6). The path can be drawn to the right or the left of the nth singularity (Fig. 7), producing an uncertainty in reconstructing the function from its Borel transform:

$$\delta F(g)\sim\oint_{z\approx z_n}dz\,e^{-z/g}B(z)\sim e^{-z_n/g}, \tag{24}$$

which, with allowance for the relation $z_n=n/\beta_0$, coincides with (23).

Of course, the literal interpretation of the Landau pole seems archaic, but after some modification of the arguments presented, a real meaning can be assigned to it. As we know,[46] the dependence of the charge g on the distance scale Λ^{-1} is given by the equation

$$\frac{dg}{d\ln\Lambda^2}=\beta(g)=\beta_0g^2+\beta_1g^3+\ldots, \tag{25}$$

whose solution has a character that depends on the behavior of the Gell-Mann–Low function $\beta(g)$. The pole in (20) is eliminated if $\beta(g)$ changes sign or behaves as g^α with $\alpha<1$ at large g. If, on the other hand, $\beta(g)$ is positive and increases as g^α with $\alpha>1$ when $g\to\infty$, the pole is preserved, and the theory is internally inconsistent because it is impossible to determine $g(\Lambda)$ for all Λ (Ref. 46). In the latter case the position of the pole is given by the equation

$$\ln\frac{\Lambda_c^2}{m^2}=\int_g^\infty\frac{dx}{\beta(x)}, \tag{26}$$

which, for small values of g, leads to the result

$$\Lambda_c^2=\text{const}\cdot m^2e^{1/\beta_0 g}, \tag{27}$$

which is distinguished from (21) only by an insignificant constant multiplier. Thus, the existence of renormalon singularities seems fairly convincing for internally inconsistent theories. Conversely, there is no basis for them in ''good'' theories.[7)]

Since the behavior of the function $\beta(g)$ at $g\gtrsim 1$ is unknown, the presence or absence of renormalon singularities is a matter of belief. We stress, however, the following circumstance. Factorial contributions of individual diagrams exist in all field theories in which the expansion of the β function (25) begins from the quadratic term: then the interaction is described on the k^{-1} scale by a formula of the type (20) with the replacement of Λ by k, whose expansion gives $(\beta_0\ln k^2)^N$ in the Nth order [see (13)].[8)] Resolution of the question of the internal inconsistency of a theory requires knowledge of all the coefficients in the expansion (25). Therefore, it would be incorrect to assume that the formal existence of renormalons is an indication of the internal inconsistency of a theory.

3. ANALYTIC PROPERTIES OF THE BOREL TRANSFORMS OF φ^4 THEORY

3.1. Expansion of the class of Borel transformations

For the ensuing treatment it is convenient to expand the class of Borel transformations, setting

$$F(g)=\int_0^\infty dx\,e^{-x}x^{b_0-1}B(gx),$$

$$B(g)=\sum_{N=0}^\infty\frac{F_N}{\Gamma(N+b_0)}g^N \tag{28}$$

with the arbitrary parameter $b_0>0$, instead of using (6) and (7). If $B(g)$ and $\tilde{B}(g)$ are the Borel transforms corresponding to the parameters b_0 and b_1 (to fix ideas, we set $b_1>b_0$), it is not difficult to derive the conversion formula:

$$\tilde{B}(g)=\frac{1}{\Gamma(b_1-b_0)}\int_0^\infty dx\,\frac{x^{b_1-b_0-1}}{(1+x)^{b_1}}B\left(\frac{g}{1+x}\right). \tag{29}$$

We determine the analyticity region of $B(g)$ by constructing the so-called Mittag–Leffler star,[30] i.e., by drawing cuts for all the singular points to infinity along rays drawn through these points from the origin of coordinates. If g lies in the analyticity region of $B(g)$, the integration path in (29) does not pass through its singularities and $\tilde{B}(g)$ is also analytic. If g_c is a singular point of $B(g)$, the path of the integral (29) for $g=g_c$ unavoidably passes through g_c, generating a singularity in $\tilde{B}(g)$. For the interesting case of power-law singularities we have the correspondence rules

$$B(g)=A\Gamma(-\beta)\left(\frac{g_c-g}{g_c}\right)\to\tilde{B}(g)=A\Gamma(-\beta-b_1+b_0)\times\left(\frac{g_c-g}{g_c}\right)^{\beta+b_1-b_0} \tag{30}$$

for noninteger $\beta+b_1-b_0$ and

$$B(g)=A\Gamma(-\beta)\left(\frac{g_c-g}{g_c}\right)^{\beta}\to\tilde{B}(g)A\frac{(-1)^{n+1}}{n!}\left(\frac{g_c-g}{g_c}\right)^{n}\ln\left(\frac{g_c-g}{g_c}\right), \tag{31}$$

if $\beta+b_1-b_0=n$ is an integer.

We see that the analyticity region for all the Borel transforms is identical and that it is sufficient to establish it for any fixed b_0. The choice $b_0=1/2$ is convenient for investigating functional integrals, since a simple result is obtained in that case for the Borel transform of an exponential function:

$$F(g)=e^{-g}\to B(g)=\frac{\cos(2\sqrt{g})}{\sqrt{\pi}}=\frac{1}{2\sqrt{\pi}}\{\exp(2i\sqrt{g}+\text{c.c.})\}, \tag{32}$$

which preserves its exponential form. This permits writing an explicit expression for the Borel transform of the functional integral (1):

$$B_I(g)=\frac{1}{2\sqrt{\pi}}\int D\varphi\,\exp(-S_0\{\varphi\})\times[\exp(2i\sqrt{gS_{\text{int}}\{\varphi\}})+\text{c.c.}]. \tag{33}$$

The integrand is a regular function, and the analyticity region of $B_I(g)$ is determined by the condition for convergence of the integral.

3.2. Analyticity outside the negative semiaxis

For simplicity, let us consider scalar φ^4 theory. Generalization to the n-component case is trivial and reduces to only a complication of the notation. We assume that $m^2>0$, bearing in mind the subsequent analytic continuation to arbitrary complex m^2.

The integral (33) for φ^4 theory is defined well for positive values of g, since its convergence is determined by an exponential function of $-S_0\{\varphi\}$ and is obvious after the Fourier transformation of $\varphi(x)$:

$$S_0\{\varphi\}=\frac{1}{2}\int d^dx\,\{(\nabla\varphi)^2+m^2\varphi^2\}=\frac{1}{2}\sum_k\,(k^2+m^2)|\varphi_k|^2. \tag{34}$$

For the analytic continuation to complex g we turn the integration path in (33), setting

$$g=\tilde{g}e^{i\Psi},\quad \varphi=\tilde{\varphi}e^{-i\Psi/4}, \tag{35}$$

where $\tilde{g}$ and $\tilde{\varphi}$ are real, and $\tilde{g}>0$. Then the integral in (33) takes the form

$$\int D\tilde{\varphi}\,\exp(-S_0\{\tilde{\varphi}\}e^{-i\Psi/2})[\exp(2i\sqrt{\tilde{g}S_{\text{int}}\{\tilde{\varphi}\}})+\text{c.c.}] \tag{36}$$

and converges for $-\pi<\Psi<\pi$. Thus, the Borel transform is analytic outside the negative semiaxis.

3.3. Analyticity within a circle

We utilize the formal technique used in Refs. 3 and 41 and introduce the function

$$(R\{\varphi\})=\frac{S_0\{\varphi\}^2}{4S_{\text{int}}\{\varphi\}}. \tag{37}$$

We can then rewrite (33) in the form

$$B_I(g)=\frac{1}{2\sqrt{\pi}}\int D\varphi\times\exp\left(-\left[1-i\left(\frac{g}{R\{\varphi\}}\right)^{1/2}\right]S_0\{\varphi\}\right)+\text{c.c.}, \tag{38}$$

and after the replacement of $R\{\varphi\}$ by the constant R_0, it is analytic within the circle $|g|<R_0$. Let us now have

$$R\{\varphi\}\geqslant R_0 \tag{39}$$

for all φ, i.e., R_0 is the exact lower bound of $R\{\varphi\}$.

Setting $g=-|g|e^{i\gamma}$ $(-\pi\leqslant\gamma\leqslant\pi)$, we have the inequality

$$|B_I(g)|\leqslant\frac{1}{2\sqrt{\pi}}\int D\varphi\left\{\exp\left(-\left[1-\left|\frac{g}{R_0}\right|^{1/2}\times\cos\frac{\gamma}{2}\right]S_0\{\varphi\}\right)+\exp(-S_0\{\varphi\})\right\}, \tag{40}$$

which ensures convergence of the integral in (33) and, consequently, its analyticity within the circle $|g|<R_0$.

To find R_0, we consider the variational problem of minimizing $R\{\varphi\}$. It yields the equation

$$-\Delta\varphi(x)+m^2\varphi(x)-C\varphi^3(x)=0. \tag{41}$$

where

$$C=\frac{S_0\{\varphi\}}{2S_{\rm int}\{\varphi\}},$$

which, after the replacement $\varphi(x)\to\varphi(x)/\sqrt{C}$, transforms into the standard equation of an instanton of φ^4 theory. Using it, we can easily show that $R_0=S\{\varphi_c\}$, which establishes the required analyticity region (Fig. 3b). Questions concerning the absence of instantons in massive four-dimensional theory[47] are discussed in Sec. 3.5.

Apart from the integral (1), some other functional integrals containing products of the type $\varphi(x_1)\varphi(x_2)\ldots\varphi(x_M)$ in the preexponential factor are of interest. The presence of such products does not influence the convergence, and all the proofs performed remain unchanged.

3.4. Invariance relative to algebraic operations

As t'Hooft pointed out,[15] the singularities of Borel transforms are not shifted when algebraic operations are performed on the original functions. This can easily be proved for a modified definition of the Borel transform (10), which differs from (6) and (28), since

$$F(g)=F_0+F_1g+F_2g^2+F_3g^3+\ldots,$$

$$B(z)=F_0\delta(z)+\frac{F_1}{0!}+\frac{F_2}{1!}z+\frac{F_3}{2!}z^2+\ldots \tag{42}$$

and $B(z)$ contains a δ-function singularity at zero. The transformation of (10) by means of the replacement $g\to1/z$ reduces to a Laplace transformation and allows inversion. It can be used to express the Borel transform of the product $F_3(g)=F_1(g)F_2(g)$ in terms of the known Borel transforms of the factors:

$$B_3(z)=\int_0^z dz'\,B_1(z')B_2(z-z'). \tag{43}$$

It can easily be seen that the δ-function singularity in $B_3(z)$ corresponds to the definition (42) and that the singular points for finite z coincide with the singular points of $B_1(z)$ and $B_2(z)$ (see the analogous reasoning in Sec. 3.1). In particular, the Borel transform g^n is the function $z^{n-1}/\Gamma(n)$, which is analytic for integer values of n, and multiplication of the function by g^n does not alter its analytic properties in the Borel plane.

If $F_2(z)=1/F_1(z)$, then

$$\delta(z)=\int_0^z dz'\,B_1(z')B_2(z-z') \tag{44}$$

and the δ-function singularity on the left-hand side cancels out with the δ-function singularities in $B_1(z)$ and $B_2(z)$. At finite values of z the right-hand side contains singularities corresponding to singular points of $B_1(z)$ and $B_2(z)$, which are absent on the left-hand side and, therefore, compensate one another. This is possible only if $B_2(z)$ has singularities at the same points as $B_1(z)$.

The proof of the analogous statements for linear operations, viz., summation, differentiation, integration, etc., is trivial.

The standard definition of the Borel transform (6) is obtained from (10) and (42) when $F_0=0$ after the replacement $F(g)\to gF(g)$. In this case the δ-function singularities disappear, and the remaining singularities are preserved at the same points due to the insignificance of the multiplier g. The definition (6) corresponds to the definition (28) with $b_0=1$, and, by virtue of Sec. 3.1, the analysis performed can be extended to arbitrary b_0.

Since all the quantities entering into the theory, viz. the Green's functions, vertex parts, etc., can be expressed in terms of functional integrals with identical analytic properties using algebraic operations (see the end of Sec. 3.3), their singular points in the Borel plane are the same as for the integral (1).

3.5. Renormalization procedure

The absence of ultraviolet divergences was implicitly assumed above. In φ^4 theory this is correct for $d<2$. For $2\leqslant d\leqslant4$ a continual theory without divergences can be devised by introducing counterterms into the Lagrangian.[46,50] In the simple case where only renormalization of the mass is required ($2\leqslant d<4$) the corresponding term in (4) is rewritten in the form

$$m_0^2\varphi^2=(m^2+\Delta m^2)\varphi^2=(m^2+Ag+Bg^2+Cg^3+\ldots)\varphi^2, \tag{45}$$

where the coefficients $A,B,C,\ldots$ are chosen so as to cancel the divergences. When counterterms are present, the analytic properties of integrals of the type (1) become more complicated, since the coupling constant appears not only in the combination $g\varphi^4$, but also in the form of $g\varphi^2$, $g^2\varphi^2$, etc. One of the types of renormalon-related activity involved specifically the introduction of additional terms into the Lagrangian and tracing the renormalon singularities that appear.[16,18,19] A question arises in this case in regard to the cancellation of singularities by selecting the coefficients in front of the additional terms so as to achieve an absence of divergences, for which an unequivocal answer could not be obtained.

A simpler route is to explicitly introduce regularization and to use renormalization-group equations. Here we have in mind the so-called truncation scheme:[51] the vertex parts are calculated as functions of the bare charge g_0 and the truncation parameter Λ using perturbation theory, then scaling functions which depend only on g_0 are obtained, and, finally, renormalized vertices, which depend on the renormalized charge g, are constructed.[50] In this case the explicit introduction of counterterms is not required, but all the details associated with their presence are taken into account, since the fundamental possibility of eliminating the divergences is essentially used to write the renormalization-group equations.

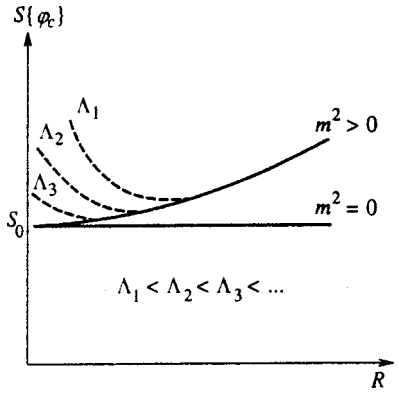

FIG. 8. Dependence of the action $S\{\varphi\}$ on the instanton radius R in four-dimensional φ^4 theory in the absence of regularization (solid curves) and for a finite truncation parameter Λ (dashed lines).

The simplest regularization method consists of substituting $\varphi\epsilon(\hat{p})\varphi$ for the term $(\nabla\varphi)^2$ in (4), which is brought into the form $-\varphi\Delta\varphi=\varphi\hat{p}^2\varphi$, where $\hat{p}$ is the momentum operator. If

$$\epsilon(p)=\epsilon(-p), \quad \epsilon(p)\geqslant 0, \tag{46}$$

then both the entire structure of the instanton calculations[49] and the proofs presented above are preserved. The only change occurs in the equation of the instanton (41), which is brought into the form

$$\epsilon(\hat{p})\varphi(x)+m^2\varphi(x)-\varphi^3(x)=0. \tag{47}$$

When the regularization

$$\epsilon(p)=p^2+p^6/\Lambda^4 \tag{48}$$

is employed, the dependence of the action $S\{\varphi\}$ on the instanton radius R in four-dimensional φ^4 theory has the form shown in Fig. 8.[9)] If $\Lambda=\infty$, there is degeneracy with respect to the instanton radius in the massless theory,[1] while there are no instantons in the massive theory[49] because of the monotonic dependence of $S\{\varphi\}$ on R. At finite values of Λ a minimum appears on the plot of $S\{\varphi\}$ versus R at $m^2>0$ (the dashed curves in Fig. 8), and instantons appear in the massive theory. Their action $S\{\varphi_c\}$ determines the positions of the singularities in the Borel plane. For $\Lambda\to\infty$ and arbitrary $m^2>0$ the value of $S\{\varphi_c\}$ tends to the instanton action of the massless theory, and the positions of the singularities do not depend on m.[10)]

The renormalization-group equations (in the Callan–Symanzik form) are valid for the vertices $\Gamma^{L,N}$ with N free ends and L two-line links:[50]

$$\left[\frac{\partial}{\partial\ln\Lambda^2}+\beta(g_0)\frac{\partial}{\partial g_0}+\left(L-\frac{N}{2}\right)\eta(g_0)-L\,\eta_2(g_0)\right]\Gamma^{L,N}(g_0,\Lambda)=0. \tag{49}$$

Writing out three such equations with different L and N, we can express the scaling functions $\beta(g_0)$, $\eta(g_0)$, and $\eta_2(g_0)$ in terms of the vertices $\Gamma^{L,N}(g_0,\Lambda)$ using algebraic operations and differentiation operations, which do not shift the positions of the singularities in the Borel plane. In the limit $\Lambda\to\infty$, where Eq. (49) is valid, the dependence of the scaling functions on Λ disappears,[50] and their singularities in the Borel plane correspond to the massless theory.

The Gell-Mann–Low function $\beta(g_0)$ defines the relationship between the renormalized charge g and the bare charge g_0. Let the functions F_0 and F_1 be such that $F_0(g_0)\equiv F_1(g)$. The relationship between the corresponding Borel transforms B_0 and B_1 [in the sense of the definition (10)] can easily be found for an infinitesimal charge transformation, $g_0=g+2\beta(g)\delta\Lambda/\Lambda$ [see (25)]:

$$B_1(z)=B_0(z)+\frac{2\delta\Lambda}{\Lambda}\times\int_0^z dy\,[B_0(y)+yB_0'(y)B_\beta(z-y)], \tag{50}$$

where $B_\beta(z)$ is the Borel transform of the function $\beta(g)/g$. Equation (50) is analogous to Eq. (43); therefore, the analytic properties do not change as a result of the transformation of the charges.

The vertices $\Gamma^{L,N}$ diverge as $\Lambda\to\infty$, but they become finite after separation of the divergent Z factors from them and the transition from the bare to the renormalized charge. Since the Z factors are, in turn, expressed in terms of the vertices $\Gamma^{L,N}$ (Ref. 50), the renormalized vertices have the required analytic properties.

The dependence of the scaling functions on the renormalization scheme is given by the following conversion formulas:[51]

$$\tilde{\beta}(q(g))=\beta(g)\frac{dq(g)}{dg},$$

$$\tilde{\eta}(q(g))=\eta(g)-\beta(g)\frac{d\ln p(g)}{dg},$$

$$\tilde{\eta}_2(q(g))=\eta_2(g)-\beta(g)\frac{d\ln p_2(g)}{dg}. \tag{51}$$

The forms of the converted functions $q(g)$, $p(g)$, and $p_2(g)$ for standard renormalization schemes (truncation, subtraction, etc.) in them are expressed in terms of the vertices $\Gamma^{L,N}$, so that the analytic properties of the scaling functions are identical in all the schemes. In general, the analytic properties of the converted functions require additional investigation.

4. CONCLUSIONS

The results in Sec. 3 rule out the existence of renormalon singularities in φ^4 theory. If the arguments in Sec. 2 regarding the relationship to a Landau pole are considered convincing, φ^4 theory cannot be internally inconsistent. The same conclusion can be drawn on the basis of solid-state applications: mathematically, a reasonable model of a disordered system reduces exactly to φ^4 theory,[36,39] and the internal inconsistency of φ^4 theory would signify the impossibility, in principle, of obtaining a mathematical description of this model. Therefore, a revision of the results in Refs. 34 and 35, in which indications of the internal inconsistency of φ^4

theory were obtained on the basis of an approximate reconstruction of the Gell-Mann–Low function, is urgently needed.

The results of Sec. 3 refer only to φ^4 theory and cannot be extended directly to other field theories; however, along with the qualitative arguments in Sec. 2, they demonstrate the futility of the conception of renormalons as a whole. Therefore, it would be of interest to generalize the method of proof used in Sec. 3 to other cases.

In quantum chromodynamics (QCD) the renormalon doctrine presently prevails.[20–30] However, the specific details of QCD in this context have never been stressed. For example, t'Hooft,[15] speaking about QCD, gives explanations within φ^4 theory, and the term ''naive non-Abelianization,'' which essentially means the disregard of the particular details of QCD, appeared in some more recent publications.[26,30] On the other hand, in QCD there is a special reason for the belief in renormalons, which has a purely phenomenological character. It can be established by treating experimental data that the contribution of the higher orders has a momentum dependence $\propto 1/q^4$ (Ref. 22). This dependence can easily be obtained from renormalon graphs, but, as is generally assumed, it cannot be obtained within the instanton method. The latter is based on the results in Refs. 10 and 11, according to which the instanton contribution is proportional to $1/q^{18}$. However, it can easily be seen that a contribution $\sim 1/q^4$ appeared in Refs. 10 and 11, but contained divergences which the authors found difficult to eliminate;[11)] therefore, the corresponding term was ''transported'' to the renormalon sector with the reasoning that ''this term contributes to the renormalon singularity, rather than to the instanton singularity'' (Ref. 10, p. 287). If there are no renormalon singularities, this contribution can be simply discarded; at present, there is therefore no valid Lipatov asymptotic relation for QCD.

This work was stimulated by lengthy discussions with P. G. Sil'vestrov, whom we thank for opposing the renormalon doctrine, for the critical remarks, and for general assistance in becoming familiar with the situation. We also thank B. L. Ioffe, L. N. Lipatov, and the participants in the seminars at the Institute of Physical Problems, the P. N. Lebedev Physics Institute, the Institute of Theoretical and Experimental Physics, and the St. Petersburg Nuclear Physics Institute for interest in this work and for useful discussions.

This work was carried out with financial support from the INTAS (Grant 96-0580) and the Russian Fund for Fundamental Research (Project 96-02-19527).

APPENDIX

Construction of interpolation polynomials

The polynomial of degree N which coincides with the function $f(x)$ at the points $x_0, x_1, x_2, \ldots, x_N$, is defined by the Lagrange formula:[53]

$$P_N(x)=\sum_{k=0}^{N}\frac{f(x_k)}{\psi'(x_k)}\frac{\psi(x)}{(x-x_k)}, \tag{A1}$$

$$\psi(x)=(x-x_0)(x-x_1)(x-x_2)\ldots(x-x_N), \tag{A2}$$

and the interpolation error is given by the expression

$$R_N(x)=f(x)-P_N(x)=\frac{f^{(N+1)}(\xi)}{(N+1)!}\psi(x), \tag{A3}$$

where ξ belongs to the interval (x_0, x_N).

The function (18), which is of interest to us, behaves as q^n at $n\leqslant N$ with slowly varying q, so that $\ln q\sim\ln N$. Disregarding these slow variations and omitting the common multiplier in (16) and (18), we have

$$f(x)=q^x, \quad A_N\sim\frac{1}{6^N N^3}. \tag{A4}$$

Since $|\psi(x)|\leqslant\Delta^{N+1}$ in the interval $0\leqslant x\leqslant\Delta$, we obtain

$$|R_N(x)|\leqslant\frac{(\ln q)^{N+1}q^{\Delta}}{(N+1)!}\Delta^{N+1}, \tag{A5}$$

and the interpolation error is small for

$$\Delta\leqslant N/\ln N. \tag{A6}$$

To investigate the dependence of the coefficient A_N on the positions of the points x_k, we set $\psi(x)=\mathrm{Re}\,\psi(x+i0)$, and calculating $\ln\psi(x+i0)$ using the Euler–MacLaurin formula, for $\psi(x)$ we obtain the expression

$$\psi(x)=F(x)\sin G(x). \tag{A7}$$

In particular, for the power-law arrangement of points

$$x_k=(k/N)^{\alpha}\Delta, \quad k=0,1,\ldots,N, \tag{A8}$$

at $\alpha\gg 1$ we have

$$F(x)=(-1)^N\sqrt{x(\Delta-x)}\exp\{N[\alpha(x/\Delta)^{1/\alpha}+\ln\Delta-\alpha]\},$$
$$G(x)=\pi N(x/\Delta)^{1/\alpha}. \tag{A9}$$

For a high-order coefficient of the polynomial (A1) we obtain

$$A_N=\sum_{k=0}^{N}\frac{f(x_k)}{\psi'(x_k)}\sim\exp\{(\alpha-\ln\Delta)N\} \tag{A10}$$

(the sum is determined by the term with $k=1$), and for $\alpha\sim\ln N$ the coefficient A_N can be factorially small or factorially large, depending on the relationship between α and $\ln\Delta$, so that the required value of (A4) falls in the range of variation. Thus, the required polynomial (16) exists in the interval $0\leqslant n\leqslant n_0$, where $n_0\sim N/\ln N$.

*)E-mail: suslov@kapitza.ras.ru

1)In a broader sense, a renormalizable theory is one in which the divergences are eliminated by renormalizing a finite number of parameters. According to more precise terminology, such theories are subdivided into superrenormalizable (renormalizable ''with a surplus'') and renormalizable in the narrow sense (marginally renormalizable); the latter, which gave their name to renormalons, are characterized by the logarithmic situation, which is needed for the appearance of factorial contributions (see below).

2)t'Hooft omitted the multipliers of the form g^n, since integrals of the type (1) usually appear in the form of a ratio and such multipliers cancel out.

3)In quantum electrodynamics and QCD a polarization loop gives the multiplier $k^2\ln k^2$, and a photon (gluon) propagator gives $1/k^2$; in four-dimensional φ^4 theory a closed loop corresponds to $\ln k^2$, and an interaction line corresponds to a constant. In all cases a chain of N ''bubbles'' corresponds to $(\ln k^2)^N$.

[4)]The validity of the saddle-point method can be substantiated for convergent finite-multiplicity integrals of functions of the type $\exp[\lambda F(x)]$ in the limit $\lambda\to\infty$ (Ref. 31). The integral (3) can be brought into the form indicated above, but, in general, it contains both ordinary ultraviolet divergences and divergences associated with the transition to an infinite number of integrations. The ratio between the two integrals of type (1) must be finite (after the appropriate renormalizations), but each of them taken individually need not be finite.

[5)]Of course, instantons exist only in a part of the region of parameters, but this is not a restriction in the Lipatov technique: the values of a, b, and c in the asymptotic relation (5) are calculated exactly, and they allow analytic continuation as functions of the physical parameters.

[6)]It is assumed below that $\beta_0>0$. For asymptotically free theories, in which $\beta_0<0$, similar arguments are valid in regard to so-called infrared renormalons. The latter are obtained from integrals of the type (13) with $m=1,0,-1,-2,\ldots$ in the region of small momenta.

[7)]In particular, the result (27) is valid when the expansion (25) is truncated at a finite number of terms provided the resultant polynomial is positive. On this basis it is easy to draw the erroneous conclusion that the high-order terms of the expansion of the β function are insignificant. Parisi's arguments[16,17] regarding the momentum dependence of Borel transforms demonstrate just this point. In fact, the character of the solution of Parisi's equations[17] depends significantly on the behavior of $\beta(g)$ at infinity. In particular, they are easily solved for the model function $\beta(g)=\beta_0 g^2/(1+\lambda g)$ with $\lambda\gg 1$ and lead to a result which differs qualitatively from the one-loop analog.

[8)]The concrete form of the renormalon diagrams can differ somewhat in different theories. For example, in φ^4 theory the significant diagrams do not reduce to chains of ''bubbles'' (Fig. 2c), but form a so-called parquet.[48]

[9)]This dependence can easily be obtained by characterizing an instanton by two parameters, viz., its radius and amplitude, and performing a variational estimation of the action. In the theory of disordered systems this corresponds to the optimal-fluctuation method.[35]

[10)]For the dimensionalities $2\leqslant d<4$ the influence of Λ on the properties of the instanton is insignificant, and the role of the renormalizations reduces to the fact that the renormalized mass appears in the equation of the instanton.[36] The dependence of $S\{\varphi_c\}$ on m is preserved in this case.

[11)]Such divergences also appear in φ^4 theory, and a procedure for eliminating them is known.[36]

[1] L. N. Lipatov, Zh. Éksp. Teor. Fiz. **72**, 411 (1977) [Sov. Phys. JETP **45**, 216 (1977)].
[2] E. Brezin, J. C. Le Guillou, and J. Zinn-Justin, Phys. Rev. D **15**, 1544 (1977).
[3] C. Itzykson, G. Parisi, and J. B. Zuber, Phys. Rev. Lett. **38**, 306 (1977).
[4] A. P. Bukhvostov and L. N. Lipatov, Zh. Éksp. Teor. Fiz. **73**, 1658 (1977) [Sov. Phys. JETP **46**, 871 (1977)].
[5] E. B. Bogomolny and V. A. Fateyev, Phys. Lett. B **71**, 93 (1977); L. N. Lipatov, A. P. Bukhvostov, and E. I. Malkov, Phys. Rev. D **19**, 2974 (1979).
[6] G. Parisi, Phys. Lett. B **66**, 382 (1977).
[7] *Large Order Behavior of Perturbation Theory*, J. C. Le Guillou and J. Zinn-Justin (Eds.), Amsterdam (1990).
[8] C. Itzykson, G. Parisi, and J. B. Zuber, Phys. Rev. D **16**, 996 (1977); R. Balian, C. Itzykson, G. Parisi, and J. B. Zuber, Phys. Rev. D **17**, 1041 (1978).
[9] E. B. Bogomolny and V. A. Fateyev, Phys. Lett. B **76**, 210 (1978).
[10] I. I. Balitsky, Phys. Lett. B **273**, 282 (1991).
[11] S. V. Faleev and P. G. Silvestrov, Nucl. Phys. B **463**, 489 (1996).
[12] B. Lautrup, Phys. Lett. B **69**, 109 (1977).
[13] S. Chadha and P. Olesen, Phys. Lett. B **72**, 87 (1977).
[14] P. Olesen, Phys. Lett. B **73**, 327 (1977).
[15] G. t'Hooft, in *The Whys of Subnuclear Physics: Proceedings of the 1977 International School of Subnuclear Physics (Erice, Trapani, Sicily, 1977)*, A. Zichichi (Ed.), Plenum Press, New York (1979).
[16] G. Parisi, Phys. Lett. B **76**, 65 (1978); Nucl. Phys. B **150**, 163 (1979).
[17] G. Parisi, Phys. Rep. **49**, 215 (1979).
[18] F. David, Nucl. Phys. B **209**, 433 (1982); **234**, 237 (1984); **263**, 637 (1986).
[19] M. C. Bergere and F. David, Phys. Lett. B **135**, 412 (1984).
[20] A. H. Mueller, Nucl. Phys. B **250**, 327 (1985).
[21] G. B. West, Phys. Rev. Lett. **67**, 1388 (1991).
[22] V. I. Zakharov, Nucl. Phys. B **385**, 452 (1992).
[23] L. S. Brown and L. J. Yaffe, Phys. Rev. D **45**, R398 (1992); L. S. Brown, L. J. Yaffe, and C. Zhai, Phys. Rev. D **46**, 4712 (1992).
[24] G. Grunberg, Phys. Lett. B **304**, 183 (1993).
[25] A. H. Mueller, Phys. Lett. B **308**, 355 (1993).
[26] M. Beneke *et al.*, Phys. Lett. B **307**, 154 (1993); **348**, 513 (1995); Nucl. Phys. B **452**, 563 (1995); **472**, 529 (1996); Phys. Rev. D **52**, 3929 (1995).
[27] D. J. Broadhurst, Z. Phys. C **58**, 339 (1993).
[28] A. I. Vainstein and V. I. Zakharov, Phys. Rev. Lett. **73**, 1207 (1994); Phys. Rev. D **54**, 4039 (1996).
[29] C. N. Lovett-Turner and C. V. Maxwell, Nucl. Phys. B **432**, 147 (1994).
[30] S. V. Faleev and P. G. Silvestrov, Nucl. Phys. B **507**, 379 (1997).
[31] G. H. Hardy, *Divergent Series*, Clarendon Press, Oxford (1949) [Russ. transl., IL, Moscow (1951)].
[32] Yu. V. Sidorov, M. V. Fedoryuk, and M. I. Shabunin, *Lectures on the Theory of Functions of Complex Variables* [in Russian], Nauka, Moscow (1976).
[33] V. S. Popov, V. L. Eletskiĭ, and A. V. Turbiner, Zh. Éksp. Teor. Fiz. **74**, 445 (1978) [Sov. Phys. JETP **47**, 232 (1978)].
[34] D. I. Kazakov, O. V. Tarasov, and D. V. Shirkov, Teor. Mat. Fiz. **38**, 15 (1979).
[35] Yu. A. Kubyshin, Teor. Mat. Fiz. **58**, 137 (1984).
[36] I. M. Suslov, Usp. Fiz. Nauk **168**, 503 (1998) [Phys. Usp. **41**, 441 (1998)].
[37] G. Falkovich, I. Kolokolov, V. Lebedev, and A. Migdal, Phys. Rev. E **54**, 4896 (1996).
[38] J. C. Le Guillou and J. Zinn-Justin, Phys. Rev. Lett. **39**, 95 (1977); Phys. Rev. B **21**, 3976 (1980).
[39] M. V. Sadovskiĭ, Usp. Fiz. Nauk **133**, 223 (1981) [Sov. Phys. Usp. **24**, 96 (1981)].
[40] G. t'Hooft, Phys. Rev. D **14**, 3433 (1976).
[41] Y. Frishman, Phys. Rev. D **19**, 540 (1979).
[42] I. Affleck, Nucl. Phys. B **191**, 429 (1981).
[43] I. M. Suslov, Zh. Éksp. Teor. Fiz. **111**, 220 (1997) [JETP **84**, 120 (1997)].
[44] l. D. Landau, A. A. Abrikosov, and I. M. Khalatnikov, Dokl. Akad. Nauk SSSR **95**, 497, 773, 1177 (1954).
[45] V. B. Berestetskiĭ, E. M. Lifshitz, and L. P. Pitaevskiĭ, *Quantum Electrodynamics*, Pergamon Press, Oxford (1982) [Russ. original, Nauka, Moscow (1980)].
[46] N. N. Bogoliubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields*, 2nd Am. ed., Wiley, New York (1980) [Russ. original, Nauka, Moscow (1976)].
[47] l. D. Landau and I. Ya. Pomeranchuk, Dokl. Akad. Nauk SSSR **102**, 489 (1955).
[48] I. T. Dyatlov, V. V. Sudakov, and K. A. Ter-Martirosian, Zh. Éksp. Teor. Fiz. **32**, 77 (1957) [Sov. Phys. JETP **5**, 631 (1957)].
[49] V. G. Makhankov, Phys. Lett. A **61**, 431 (1977).
[50] E. Brezin, J. C. Le Guillou, and J. Zinn-Justin, in *Phase Transitions and Critical Phenomena, Vol. 6*, C. Domb and M. S. Green (Eds.), Academic Press, New York (1976).
[51] A. A. Blalimirov and D. V. Shirkov, Usp. Fiz. Nauk **129**, 407 (1979) [Sov. Phys. Usp. **22**, 860 (1979)].
[52] I. M. Suslov, Zh. Éksp. Teor. Fiz. **106**, 560 (1994) [JETP **79**, 307 (1994)].
[53] A. O. Gel'fond, *The Calculus of Finite Differences* [in Russian], Nauka, Moscow (1967).

Translated by P. Shelnitz
Edited by S. J. Amoretty

Bremsstrahlung spectrum for α decay and quantum tunneling

E. V. Tkalya*)

D. V. Skobel'tsyn Scientific-Research Institute of Nuclear Physics, 119899 Moscow, Russia
(Submitted 9 March 1999)
Zh. Éksp. Teor. Fiz. **116**, 390–409 (August 1999)

Spectra of the electromagnetic radiation arising during α decay of atomic nuclei as a consequence of the motion of the α particle through a Coulomb potential barrier and in the Coulomb field of the daughter nucleus are calculated via a quantum-mechanical approach. The contributions of the $E1$ and $E2$ multipoles are calculated. Model problems of emission during motion of a charged particle through a spherically symmetric, rectangular potential barrier and a ''cut-off'' Coulomb barrier are treated. Numerical calculations are performed for ^{210,214}Po and ^{226}Ra nuclei. Emission spectra are derived for an α particle propagating along classical trajectories in these potentials. © *1999 American Institute of Physics.* [S1063-7761(99)00308-X]

1. INTRODUCTION

Tunneling through a potential barrier is one of many purely quantum phenomena which have no analog in classical physics. Experimentally, the process of tunneling is observed by recording particles that have already passed through the barrier. The possibility exists in principle, however, of investigating the motion of a particle inside the potential barrier. This is so because as it passes through a region of space with a nonzero potential gradient a charged particle emits photons. For these there is no barrier. By recording the emission spectrum it is possible to assess the properties of the potential barrier. This additional diagnostic channel can be useful when studying the most diverse physical phenomena in which tunneling of charged particles occurs (e.g., in a tunneling microscope).

One well-known tunneling process is α decay of atomic nuclei. It is often accompanied by γ emission with a discrete spectrum corresponding to transitions from excited levels of the daughter nuclei populated during α decay. Besides this, there is also emission with a continuous spectrum, which has come to be called bremsstrahlung.[1–5] The nature of this emission and its connection with tunneling through a potential barrier are the subject of the present paper.

The experimental data on the bremsstrahlung spectra during α decay of ^{210,214}Po, ^{226}Ra, and ^{244}Cm nuclei have been published by two different research groups, a Russian/Italian[1] and a Japanese one.[2,3] These groups worked with different nuclides. Taking this fact into account, it is nevertheless concluded that the results of the two sets of experiments do not agree.[2]

An effort to explain the bremsstrahlung spectrum for α decay of ^{210}Po was made in Ref. 2 based on the quasiclassical approximation developed in Ref. 4 for a tunneling charge. However, a quantum-mechanical calculation[5] of the electrical-dipole radiation with the $E1$-transition operator taken in simplified form did not completely reproduce the shape of the experimental spectrum in Ref. 2 or the quasiclassical curve obtained there.

In the present work, within the framework of the quasistationary approximation in quantum mechanics we consider a model problem of emission during motion through a spherically symmetric, rectangular potential barrier (it is solved analytically and gives an upper bound on the spectrum) and problems of emission during motion through an ordinary Coulomb barrier and a Coulomb barrier cut off at some point R_2 (see Fig. 1). For these cases we also consider the problem of emission in the classical approach. Spectra of the electrical dipole and electric-quadrupole γ radiation of an α particle are calculated for three specific nuclei.

The system of units $\hbar = c = 1$ is adopted in this work.

2. EMISSION PROBABILITY

Emission ''under the barrier'' is a special case of emission from a region forbidden for classical motion. Processes of this kind are well known. An example is emission of light during atomic transitions. The electron wave functions of bound states in an attractive Coulomb field have nonzero amplitudes, including in the regions inaccessible to classical motion. Despite the rapid decay of the wave function in the indicated regions, the contribution of the latter to the probability of photon emission is completely real and easily estimated. In such a calculation the energies and wave functions of the atomic states are found with the help of the time-independent Schrödinger equation, and the emission probability—by perturbation theory. Alpha decay can also be described within the framework of the single-particle model with time-independent wave functions.[6] Correspondingly, an attempt can be made to treat the emission of a bremsstrahlung photon in analogy with atomic transitions within the framework of perturbation theory.[5] Obviously, in such an approach it is impossible to determine where the photon was emitted—under the barrier or outside of it. It is possible to speak of the contribution of a region to the emission probability only in the sense of the total integral.

We will calculate in the center-of-mass system of the α particle and the daughter nucleus. For brevity we will speak

 © 1999 American Institute of Physics

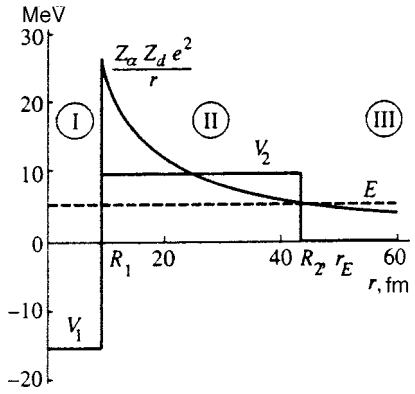

FIG. 1. Coulomb potential barrier for ^{210}Po, spherically symmetric potential step, and Coulomb barrier cut off at the point $r=R_2=r_E$.

of the wave functions and coordinates of the α particle, but by that we mean the wave functions and coordinates of the relative motion of the α particle and daughter nucleus.

The spectrum of the emitted γ particles in the transition of an α particle from the initial state ψ_i to the final state ψ_f is described by the formula (see, e.g., Ref. 7)

$$\frac{dW_\gamma}{d\omega}=2\pi|\langle\psi_f|\hat{H}_{\rm int}|\psi_i\rangle|^2\rho_f, \tag{1}$$

where ρ_f is the density of final states of the α particle and γ quantum, and $\hat{H}_{\rm int}$ is the Hamiltonian of the interaction of the α-particle current with the electromagnetic field.

If ω is the energy of the emitted bremsstrahlung photons, K' is the momentum of the α particle at infinity in the final state, and m is its mass, then

$$\rho_f=\frac{mK'}{(2\pi)^3}\,d\Omega_\alpha\,\frac{\omega^2}{(2\pi)^3}\,d\Omega_\gamma.$$

The interaction of the α-particle current $\mathbf{j}_{fi}$ with a photon having wave function $\mathbf{A}_\lambda$ (a plane wave with polarization λ and momentum $\mathbf{q}$ is written as

$$\langle\psi_f|\hat{H}_{\rm int}|\psi_i\rangle=\int d^3r\mathbf{j}_{fi}(\mathbf{r})\mathbf{A}_\lambda(\mathbf{r}). \tag{2}$$

The standard form of the transition current can be written

$$\mathbf{j}_{fi}(\mathbf{r})=\frac{Z_{\rm eff}e}{2mi}(\psi_f^*(\mathbf{r})\nabla\psi_i(\mathbf{r})-(\nabla\psi_f^*(\mathbf{r}))\psi_i(\mathbf{r})), \tag{3}$$

where e is the charge of the proton and the effective charge $Z_{\rm eff}$ in the center-of-mass system depends on the multipole character of the transition. It can be expressed approximately in terms of the atomic number A and charge Z of the decaying nucleus. For an electrical-dipole transition

$$Z_{\rm eff}^{E1}\approx[2(A-4)-4(Z-2)]/A,$$

and for an electrical quadrupole transition

$$Z_{\rm eff}^{E2}\approx Z_\alpha=2$$

(see Ref. 7).

We expand the wave function in electric multipoles A_{Lm}^E and magnetic multipoles A_{Lm}^M:[7]

$$\mathbf{A}_\lambda(\mathbf{r})=\sqrt{\frac{2\pi}{\omega}}\lambda\sum_{L,m}\sqrt{2\pi(2L+1)}i^L D_{m,\lambda}^L[\varphi_{\mathbf{q}},\theta_{\mathbf{q}},0)$$
$$\times(i\lambda\mathbf{A}_{Lm}^E(\mathbf{r})+\mathbf{A}_{Lm}^M(\mathbf{r}))]. \tag{4}$$

Here $D_{m,\lambda}^L$ are the Wigner D functions. The electromagnetic vector potentials are chosen in the Coulomb gauge

$$\mathbf{A}_{Lm}^E(\mathbf{r};\omega)=\sqrt{\frac{L+1}{2L+1}}\,j_{L-1}(\omega r)\mathbf{Y}_{LL-1;m}(\mathbf{n})$$
$$-\sqrt{\frac{L}{2L+1}}\,j_{L+1}(\omega r)\mathbf{Y}_{LL+1;m}(\mathbf{n}),$$
$$\mathbf{A}_{Lm}^M(\mathbf{r};\omega)=j_L(\omega r)\mathbf{Y}_{LL;m}(\mathbf{n}), \tag{5}$$

where $j_L(x)$ is the spherical Bessel function, $\mathbf{Y}_{LJm}(\mathbf{n})$ are the vector spherical harmonics

$$\mathbf{Y}_{LJm}(\mathbf{n})=\sum_{m_1,m_2}(Lm_1 1m_2|Jm)Y_{Lm_1}(\mathbf{n})\xi_{m_2},$$

$(Lm_1 1m_2|Jm)$ are the Clebsch–Gordan coefficients, $Y_{Lm_1}(\mathbf{n})$ are the spherical functions, and $\xi_{m_2}(m_2=\pm1,0)$ are the components of the spin vector of the photon in the cyclical basis.

Since only emission of electrical photons is possible in the problem under consideration,[8] in what follows we will work with the potential $\mathbf{A}_{Lm}^E$. Also, besides the electric-dipole radiation, we also obtain formulas for the spectrum of the electric-quadrupole radiation, for the following reasons. First, the effective charge in the $E1$ transition is much less than the effective charge in the $E2$ transition. Therefore, in particular for fission of heavy nuclei, when the daughter nuclei (fission fragments) have similar charge-to-mass ratios, the electric-dipole radiation can be suppressed and the $E2$ radiation can make a substantial contribution to the spectrum. (The formulas obtained in the present work also apply to this case.) Second, the region in which the wave functions are nonzero and consequently contribute to the radiation is much larger than a nucleus. Third, the energies of the α particles for decay of nuclei with short lifetimes are very large. For such nuclei the region of photon energies substantially greater than 1 MeV becomes accessible in the spectral measurements. At these energies not only can a retardation effect be manifested, but also a contribution from the $E2$ radiation.

As for the wave functions of the α particles, we construct them for the three potentials depicted in Fig. 1, that is, for a spherically symmetric potential step, a Coulomb potential barrier, and a Coulomb potential cut off at some point R_2.

3. RECTANGULAR POTENTIAL STEP

The most suitable objects of study in connection with bremsstrahlung during α decay are even–even nuclei. When they decay a single-line transition from the ground state 0^+ of the mother nucleus to the ground state 0^+ of the daughter nucleus is often observed. In this case, the α particle is emitted by the mother nucleus only as an S wave.

Let us first consider a model problem in which the role of the Coulomb potential barrier is played by a spherically symmetric rectangular-step potential, as shown in Fig. 1. Noting that the angular and radial variables separate in a spherically symmetric field, we represent the wave function of the initial state as

$$\psi_i(\mathbf{r})=\phi_i(r)Y_{00}(\mathbf{n_r}).$$

The radial part of the wave function $\phi_i(r)$ is the solution of the Schrödinger equation

$$\rho^2\frac{d^2\phi_L(\rho)}{d\rho^2}+2\rho\frac{d\phi_L(\rho)}{d\rho}+[\rho^2-L(L+1)]\phi_L(\rho)=0 \tag{6}$$

for $L=0$. The variable ρ, referring to Fig. 1, is equal to $\rho=kr$ in region I, where $k=\sqrt{2m(E-V_1)}$; $\rho=\kappa r$ in region II, where $\kappa=\sqrt{2m(V_2-E)}$; and $\rho=Kr$ in region III, where $K=\sqrt{2mE}$. The potentials V_1 and V_2 are respectively the well depth and barrier height, and E is the energy of the α particle.

The solution of Eq. (6) for $L=0$ that is regular at the origin and satisfies the condition at infinity that it be a divergent spherical wave is obviously

$$\phi_i(r)=\begin{cases} a_0\dfrac{\sin(kr)}{r}, & 0\leqslant r<R_1,\\ b_0^{(1)}\dfrac{e^{-\kappa(r-R_1)}}{r}+b_0^{(2)}\dfrac{e^{\kappa(r-R_1)}}{r}, & R_1\leqslant r<R_2,\\ c_0\dfrac{e^{iK(r-R_2)}}{r}, & R_2\leqslant r.\end{cases} \tag{7}$$

We determine the coefficient c_0 from the normalization condition on the current of the transmitted wave, i.e., when the particle current

$$\mathbf{j}(\mathbf{r})=\frac{1}{2mi}[\psi_i^*(\mathbf{r})\nabla\psi_i(\mathbf{r})-(\nabla\psi_i^*(\mathbf{r}))\psi_i(\mathbf{r})]$$

through the surface of the sphere of larger radius is equal to 1:

$$\oint \mathbf{j}ds=4\pi r^2 j_r=1.$$

Noting that the radial component of the particle current for the wave function (7) in region III is $|c_0|^2 v_\alpha 4\pi r^2$, where $v_\alpha=K/m$ is the velocity of the α particles, we obtain for the normalization constant $c_0=1/\sqrt{v_\alpha}$.

The matching conditions for the wave functions and their derivatives at the points R_1 and R_2 give the values of the coefficients a_0, $b_0^{(1)}$, and $b_0^{(2)}$ and the energy E as functions of the magnitudes of the potentials and the widths of regions I and II.

The wave function of the final state should have the asymptotic form ''plane wave+converging spherical wave.'' [10] Therefore we begin its construction in region III. Employing the well-known expansion of a plane wave

$$e^{i\mathbf{K'r}}=4\pi\sum_{L=0}^{\infty}\sum_{m=-L}^{L} i^L Y_{Lm}^*(\mathbf{n_{K'}})Y_{Lm}(\mathbf{n_r})j_L(K'r),$$

we can write the unknown wave function in the form

$$4\pi\sum_{L=0}^{\infty}\sum_{m=-L}^{L} i^L Y_{Lm}^*(\mathbf{n_{K'}})Y_{Lm}(\mathbf{n_r})[j_L(K'r)+C_L h_L^{(2)}(K'r)].$$

where C_L is a constant and $h_L^{(2)}$ is the Hankel function of the second kind,[11] defined by the general relation $h^{(\frac{1}{2})}(x)=j_L(x)\pm in_L(x)$, where n_L is the spherical Neuman function.

The radial part of the wave function of the final state in regions I and II is written analogously as the wave function of the initial state. As a result, we obtain the wave function of the final state satisfying the conditions at the origin and at infinity:

$$\psi_f(\mathbf{r})=4\pi\sum_{L=0}^{\infty}\sum_{m=-L}^{L} i^L Y_{Lm}^*(\mathbf{n_{K'}})Y_{Lm}(\mathbf{n_r})\phi_{f_L}(r), \tag{8}$$

$$\phi_{f_L}(r)=\begin{cases} A_L j_L(k'r), & 0\leqslant r<R_1,\\ [B_L^{(1)}h_L^{(1)}(\kappa'r)+B_L^{(2)}h_L^{(2)}(\kappa'r)], & R_1\leqslant r<R_2,\\ [j_L(K'r)+C_L h_L^{(2)}(K'r)], & R_2\leqslant r.\end{cases} \tag{9}$$

If the energy of the α particle in the final state is equal to E', then the momenta in Eqs. (8) and (9) are

$$k'=\sqrt{2m(E'-V_1)},\quad \kappa'=\sqrt{2m(V_2-E')},$$

$$K'=\sqrt{2mE'}.$$

The four matching equations for the wave functions and their derivatives at the points $r=R_1$ and $r=R_2$ for the P and D waves are solved analytically. Thus, the function of the final state of the α particle is defined over all space.

Let us turn to the emission spectrum. The key element in its calculation is evaluation of the matrix element from Eq. (2) or, taking into account the expansion (4), evaluation of the matrix element

$$\int d^3r\,\mathbf{j}_{fi}(\mathbf{r})\mathbf{A}_{Lm}^E(\mathbf{r}). \tag{10}$$

Substituting the transition current (3) in explicit form in expression (10) simplifies it. Integrating the second of the terms making up the transition current (3) by parts while taking into account the Coulomb gauge of the electric monopole

$$\nabla\mathbf{A}_{Lm}^E=0$$

and the fact that both wave functions vanish at the origin while the initial wave function vanishes at infinity, we find that

$$\int d^3r\,\mathbf{j}_{fi}(\mathbf{r})\mathbf{A}_{Lm}^E(\mathbf{r})=\frac{Z_{\text{eff}}e}{mi}\int d^3r\,\mathbf{A}_{Lm}^E(\mathbf{r})\psi_f^*(\mathbf{r})\nabla\psi_i(\mathbf{r}). \tag{11}$$

From here on we can proceed in several different ways. Let us consider some of them.

3.1. Spectrum in "velocity form"

The first approach consists in direct calculation of $\nabla\psi_i(\mathbf{r})$ using the gradient formula (see, e.g., Ref. 7), according to which

$$\nabla\psi_i(\mathbf{r})=\nabla\phi_i(r)Y_{00}(\mathbf{n_r})=-\frac{d\phi_i(r)}{dr}\,\mathbf{Y}_{01;0}(\mathbf{n_r}). \tag{12}$$

We substitute this result into Eq. (11). Next, using (4) and (5) and integrating over all the angular variables, we find the following expression from formula (1) for the spectrum of the emitted $E1$ photons:

$$\frac{dW_\gamma^{(E1)}}{d\omega}=\frac{8(Z_{\rm eff}^{E1}e)^2}{3\pi}\frac{K'\omega}{m}\left(|R_0(\omega)|^2+\frac{1}{5}(R_0(\omega)R_2^*(\omega)+R_0^*(\omega)R_2(\omega))+\frac{3}{25}|R_2(\omega)|^2\right). \tag{13}$$

The emission spectrum (13) is the spectrum of unpolarized photons, for which all polarizations of the emitted quantum have been summed over. The radial integrals in expression (13) are defined as

$$R_{0,2}(\omega)=\int_0^\infty dr\,r^2\phi_{f_1}^*(r)j_{0,2}(\omega r)\frac{d\phi_i(r)}{dr}. \tag{14}$$

For most physical problems the approximation $\omega r\ll 1$ is valid, which allows us to neglect the matrix element $R_2(\omega)$ in comparison with $R_0(\omega)$ and replace $j_0(\omega r)$ by 1. In this case, relations (13) and (14) simplify substantially:

$$\frac{dW_\gamma^{(E1)}}{d\omega}=\frac{8(Z_{\rm eff}^{E1}e)^2}{3\pi}\frac{K'\omega}{m}|R_\nabla(\omega)|^2, \tag{15}$$

$$R_\nabla(\omega)=\int_0^\infty dr\,r^2\phi_{f_1}^*(r)\frac{d\phi_i(r)}{dr} \tag{16}$$

(the radial matrix element R_∇, as before, depends on the energy of the emitted photon since ω enters into the wave function of the final state (9) by way of the momenta k', κ', and K').

Formulas (13)–(16) give the $E1$ bremsstrahlung spectrum in the so-called "velocity form."[12] [Following Ref. 12, we have emphasized this fact by the symbol ∇ in the notation for the matrix element R_∇ in Eq. (16)].

3.2. Spectrum in "length form"

The form of the spectrum which in the theory of bremsstrahlung is sometimes called the "length form"[12] (we denote the corresponding matrix element below as R_r) follows from Eq. (11) with the help of the relation, well known in quantum mechanics,[9]

$$\langle\psi_f|\nabla|\psi_i\rangle=m\omega\langle\psi_f|\mathbf{r}|\psi_i\rangle.$$

Substituting it into Eq. (11) and carrying out manipulations similar to those described above, we obtain instead of formulas (15) and (16) for the spectrum and radial matrix element the following approximate expressions:

$$\frac{dW_\gamma^{(E1)}}{d\omega}=\frac{8(Z_{\rm eff}^{E1}e)^2}{3\pi}K'm\omega^3|R_r(\omega)|^2, \tag{17}$$

$$R_r(\omega)=\int_0^\infty dr\,r^2\phi_{f_1}^*(r)r\phi_i(r). \tag{18}$$

For such a rapidly decreasing potential as a rectangular potential step, we can calculate using any of the above formulas. The only point in question here is the greater or lesser degree of complexity in the calculation of the integrals in the radial matrix elements. However, for a Coulomb potential it is necessary to use another method.

3.3. Calculation with a potential gradient

The most universal way of calculating the matrix element (11) was employed in Ref. 5. Taking into account the definition $\hat{\mathbf{p}}=-i\nabla$ and commutativity of the momentum operator $\hat{\mathbf{p}}$ with the Hamiltonian of free motion $\hat H_0$, from the equation of motion for the components of the momentum vector in the Heisenberg representation

$$\hat p_j=-i[\hat p_j,\hat H]$$

(the full Hamiltonian is $\hat H=\hat H_0+\hat V$, where $\hat V$ is the interaction) we at once obtain

$$\psi_f^*(\mathbf{r})\nabla\psi_i(\mathbf{r})=\frac{1}{\omega}\psi_f^*(\mathbf{r})(\nabla V(\mathbf{r}))\psi_i(\mathbf{r}). \tag{19}$$

Substituting the latter result into Eq. (11), we can easily calculate the desired transition matrix element.

Making use of the spherical symmetry of the potential, we represent it in the form

$$V(\mathbf{r})=V(r)\sqrt{4\pi}Y_{00}(\mathbf{n_r}).$$

From the gradient formula we obtain

$$\nabla V(\mathbf{r})=-\sqrt{4\pi}\,\frac{dV(r)}{dr}\mathbf{Y}_{01;0}(\mathbf{n_r}).$$

On the other hand, the wave function of the initial state is an S wave: $\psi_i(\mathbf{r})=\phi_i(r)Y_{00}(\mathbf{n_r})=1/\sqrt{4\pi}\phi_i(r)$. Employing this fact, we can rewrite Eq. (19) as

$$\psi_f^*(\mathbf{r})(\nabla V(\mathbf{r}))\psi_i(\mathbf{r})=\psi_f^*(\mathbf{r})\left(-\frac{dV(r)}{dr}\right)\phi_i(r)\mathbf{Y}_{01;0}(\mathbf{n_r}). \tag{20}$$

Comparison of formulas (20) and (12) shows that they are quite similar. If in formula (12) we make the substitution $d\phi_i(r)/dr\to\phi_i(r)dV(r)/dr$, then repeating the derivation of formulas (13) and (14) taking the factor $1/\omega$ in expression (19) into account, we obtain for the emission spectrum

$$\frac{dW_\gamma^{(E1)}}{d\omega}=\frac{8(Z_{\rm eff}^{E1}e)^2}{3\pi}\frac{K'}{m}\frac{1}{\omega}\left(|\tilde R_0(\omega)|^2+\frac{1}{5}(\tilde R_0(\omega)\tilde R_2^*(\omega)+\tilde R_0^*(\omega)\tilde R_2(\omega))+\frac{3}{25}|\tilde R_2(\omega)|^2\right), \tag{21}$$

where

$$\tilde R_{0,2}(\omega)=\int_0^\infty dr\,r^2\phi_{f_1}^*(r)j_{0,2}(\omega r)\frac{dV(r)}{dr}\phi_i(r). \tag{22}$$

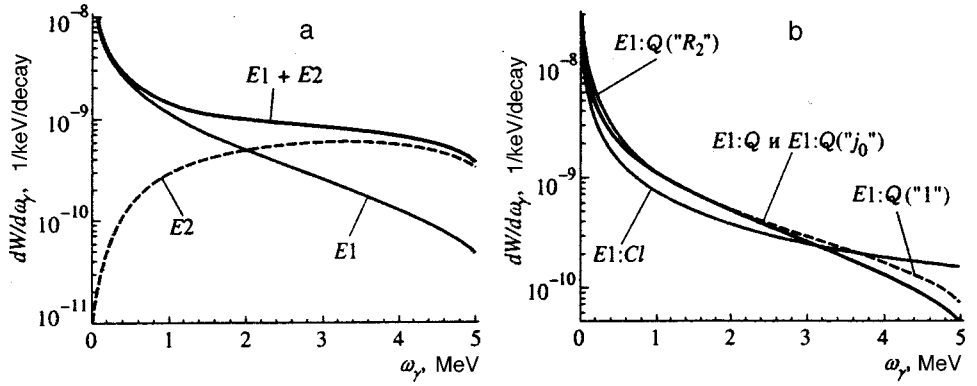

FIG. 2. Emission spectra for a spherically symmetric potential step imitating α decay of ^{210}Po. a—Quantum $E1$ and $E2$ spectra; b—$E1$ spectra: Cl—classical [formula (45)], Q—Quantum [exact formula (21)], $Q(\text{``}j_0\text{''})$—approximation when only the radial matrix element $\tilde{R}_0(\omega)$ is left in formula (21), $Q(\text{``}1\text{''})$—approximation used in Ref. 5, $Q(\text{``}R_2\text{''})$—quantum-mechanical spectrum from formula (21) with matrix elements (23) in which the contribution of the first term (i.e., the matrix elements at the point R_1) has been neglected.

Formulas (21) and (22) allow us to calculate the emission spectrum systematically. If the radial part of the potential has the form of a spherically symmetric step, i.e.,

$$V(r)=V_1\theta(R_1-r)+V_2\theta(r-R_1)\theta(R_2-r),$$

where $\theta(x)$ is the Heaviside step function, then

$$\frac{dV(r)}{dr}=(-V_1+V_2)\delta(r-R_1)-V_2\delta(r-R_2).$$

Therefore, the integrals (22) can be evaluated at once with the help of delta functions, and the matrix elements simplify:

$$\tilde{R}_{0,2}(\omega)=(-V_1+V_2)R_1^2\phi^*_{f_1}(R_1)j_{0,2}(\omega R_1)\phi_i(R_1) - V_2R_2^2\phi^*_{f_1}(R_2)j_{0,2}(\omega R_2)\phi_i(R_2). \quad (23)$$

The electric-dipole emission spectra, calculated according to formula (21), are shown in Fig. 2, where both matrix elements $\tilde{R}_0(\omega)$ and $\tilde{R}_2(\omega)$ as given by formula (23) have been used or the simplified expression has been used, where the contribution of the terms containing $\tilde{R}_2(\omega)$ has been neglected in formula (21). The barrier parameters were chosen in order to model α decay of ^{210}Po (i.e., reproduce the lifetime of the nucleus approximately and the energy of the emitted α particle exactly; see Sec. 5 for more details. It can be seen, in particular, that it is not necessary to take the second term [proportional to $j_2(\omega r)$] in the expression for the electric $E1$ multipole (5) into account in this case: the graphs of the spectra with the exact operator and the approximate $E1$ operator coincide. As for the approximation developed in Ref. 5, it works well at relatively low energies and somewhat more poorly at higher energies. [The scheme used in Ref. 5 is equivalent to discarding the terms with $\tilde{R}_2(\omega)$ in the formula for the emission probability (21) while simultaneously using the approximation $j_0(\omega r)\to 1$ in the matrix element $\tilde{R}_0(\omega)$ in formula (22)]. Note should also be made of the considerable contribution to the radial integral $\tilde{R}_0(\omega)$ in formulas (22) and (23) and, accordingly, to the emission probability at low energies of the left-hand boundary of the potential step, specifically, the point $r=R_1$ (see Fig. 2).

Let us turn now to an estimate of the electric-quadrupole emission. The calculation is analogous to that of the $E1$ emission spectrum. The wave function (5) of the photon must be taken with angular momentum $L=2$. As a result, the radial wave function of the finite state (9) enters into the matrix elements with $L=2$. The final expression for the $E2$ spectrum has the form

$$\frac{dW_\gamma^{(E2)}}{d\omega}=\frac{8(Z_{\text{eff}}^{E2}e)^2}{3\pi}\frac{K'}{m}\frac{1}{\omega}\frac{9}{5}(|\tilde{R}_1(\omega)|^2), \quad (24)$$

where

$$\tilde{R}_1(\omega)=\int_0^\infty dr\, r^2\phi^*_{f_2}(r)j_1(\omega r)\frac{dV(r)}{dr}\phi_i(r). \quad (25)$$

We limit the discussion here to the first term in the expression for the electric multipole with $L=2$ in formula (5). In fact, this is equivalent to using the Siegert theorem (see, e.g., Ref. 7) in the case when the emission probability is calculated in ''length form'' in terms of the electric-quadrupole moment of the transition of the system. The contribution of the second term, as will become clear below in the instance of the $E1$ spectrum, is substantially less.

The $E2$ emission spectrum for a rectangular potential step is plotted in Fig. 2. As expected, the $E2$ emission at relatively large energies of the γ photons becomes first substantial and then dominant in the spectrum.

3.4. ''Length form'' and the Siegert theorem

Let us consider the approach based on the Siegert theorem in more detail. Toward this end, we make use of a well-known expression for the longitudinal potential[7]

$$\mathbf{A}^Y_{Lm}(\mathbf{r};\omega)=\sqrt{\frac{L}{2L+1}}\,j_{L-1}(\omega r)\mathbf{Y}_{LL-1;m}(\mathbf{n}) + \sqrt{\frac{L+1}{2L+1}}\,j_{L+1}(\omega r)\mathbf{Y}_{LL+1;m}(\mathbf{n}). \quad (26)$$

Comparison of expression (26) and formula (5) for the electric potential $\mathbf{A}^E_{Lm}$ leads to the approximate identity

$$\mathbf{A}^{E}_{Lm}(\mathbf{r};\omega)\approx\sqrt{\frac{L+1}{L}}\mathbf{A}^{Y}_{Lm}(\mathbf{r},\omega), \tag{27}$$

which is valid provided

$$\frac{j_{L+1}(\omega r)}{j_{L-1}(\omega r)}\ll 1. \tag{28}$$

Condition (28) is satisfied over a much wider range of energies than the condition $\omega r\ll 1$ used in the derivation of formulas (15)–(18). In fact, even for $x\equiv\omega r=1$ the relations

$$j_2(x)/j_0(x)<1/10, \quad j_3(x)/j_1(x)<1/30$$

are valid (see Ref. 11). Therefore, the formulas obtained with the help of relation (27) work somewhat better than those obtained for $\omega r\ll 1$.

The longitudinal potential A^{Y}_{Lm} in the form (26) is derived (see Ref. 7) with the help of the gradient formula from the expression

$$\mathbf{A}^{Y}_{Lm}(\mathbf{r};\omega)=\frac{1}{\omega}\nabla j_L(\omega r)Y_{Lm}(\mathbf{n}).$$

Using the property of $\mathbf{A}^{Y}_{Lm}$, and also relation (27), we obtain the following approximate expression for the electric multiple:

$$\mathbf{A}^{E}_{Lm}(\mathbf{r};\omega)\approx\frac{1}{\omega}\sqrt{\frac{L+1}{L}}\nabla j_L(\omega r)Y_{Lm}(\mathbf{n}). \tag{29}$$

To calculate further, we need the equation of continuity for the transition current (3). It is derived in the standard way. The Schrödinger equation for the wave function of the initial state is multiplied on the left by the complex-conjugate of the wave function of the final state, and the Schrödinger equation for the complex-conjugate of the wave function of the final state is multiplied on the right by the wave function of the initial state. The one equation is then subtracted from the other, and we obtain the desired equation of continuity

$$\nabla\mathbf{j}_{fi}(\mathbf{r})=iZ_{\rm eff}e\,\omega\,\psi^{*}_{f}(\mathbf{r})\psi_i(\mathbf{r}). \tag{30}$$

Integrating by parts in Eq. (10) taking relations (29) and (30) into account, we find the following relation for the matrix element of the interaction of the transition current with the field:

$$\int d^3r\,\mathbf{j}_{fi}(\mathbf{r})\mathbf{A}^{E}_{Lm}(\mathbf{r};\omega)\approx$$

$$-iZ_{\rm eff}e\sqrt{\frac{L+1}{L}}\int d^3r\,\psi^{*}_{f}(\mathbf{r})j_L(\omega r)Y_{Lm}(\mathbf{n})\psi_i(\mathbf{r}).$$

The emission spectrum is now calculated the same way it was done above. Finally, we have

$$\frac{dW^{(EL)}_{\gamma}}{d\omega}=\frac{4(2L+1)(L+1)}{L\pi}(Z^{E1}_{\rm eff}e)^2K'm\omega|R_{jL}(\omega)|^2,$$

where

$$R_{jL}(\omega)=\int_0^{\infty}dr\,r^2\phi^{*}_{f_L}(r)j_L(\omega r)\phi_i(r).$$

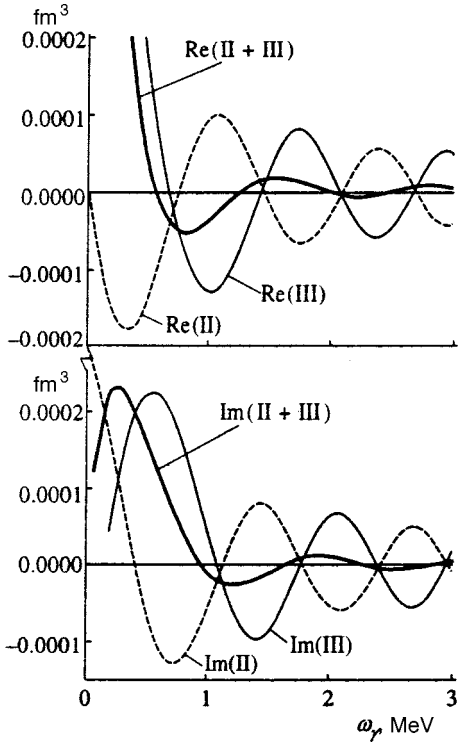

FIG. 3. a—real part, b—imaginary part of the radial integral of the dipole moment of the transition (31) for a spherically symmetric potential step in regions II (under the barrier) and III (the region of free motion).

From these formulas in the limit $\omega r\ll 1$ and using the expansion

$$j_L(\omega r)\approx(\omega r)^L/(2L+1)!!$$

we can easily obtain formulas in ''length form'' for emission of any multipolity, including relations (17) and (18) obtained above.

Depending on the purpose of the problem, the shape of the potential, the form of the wave function, etc., it is possible to use one of the formulas obtained in the present section to calculate bremsstrahlung spectra.

Let us consider, for example, the region of formation of the $E1$ bremsstrahlung spectrum in ''length form,'' proceeding from formula (18). If we are talking here of the transition between two states of the discrete spectrum, then the expression

$$Z^{E1}_{\rm eff}e\int_0^{\infty}dr\,r^2\phi^{*}_{f_1}(r)r\phi_i(r) \tag{31}$$

would have the meaning of the dipole moment of the transition of the system. In our case, the dimensions of expression (31) are fm^3. Nevertheless, taking this correspondence into account, for simplicity we will speak here of the dipole moment of the transition. Figure 3 shows how the real and imaginary parts of the radial integral in expression (31) behave, with the contributions from regions II and III shown separately. It can be seen that on the whole, region III dominates. However, the contributions from both regions are similar in magnitude but have different signs. Thus, the di-

pole moments of the transition in regions II and III cancel out for the most part in the given problem geometry.

4. COULOMB POTENTIAL BARRIER

Let us now consider the emission spectrum of an α particle as it traverses a Coulomb potential barrier. The potential acting on the α particle is written as

$$V(r)=V_1\theta(R_1-r)+\frac{Z_dZ_\alpha e^2}{r}\theta(r-R_1), \tag{32}$$

where Z_d is the charge of the daughter nucleus. The wave functions of the α particle in region I are the same as in the previous section; specifically, they are spherical Bessel functions $j_L(kr)$. In the region where the repulsive Coulomb potential acts (for $r\geqslant R_1$) the radial wave functions of the α particle are found by solving the time-independent Schrödinger equation

$$\frac{d^2\phi_L(r)}{dr^2}+\frac{2}{r}\frac{d\phi_L(r)}{dr}+\left(2mE-2m\frac{Z_dZ_\alpha e^2}{r}-\frac{L(L+1)}{r^2}\right)\phi_L(r)=0,$$

which as a result of the substitutions

$$\phi_L(r)=\frac{u_L(r)}{r},\quad K=\sqrt{2mE},\quad \rho=Kr,\quad \eta=\frac{Z_dZ_\alpha e^2}{K/m}$$

reduces to the Coulomb wave equation[11]

$$\frac{d^2u_L(\rho)}{d\rho^2}+\left(1-\frac{2\eta}{\rho}-\frac{L(L+1)}{\rho^2}\right)u_L(\rho)=0. \tag{33}$$

The solutions of Eq. (33) are the regular $F_L(\eta,\rho)$ and irregular $G_L(\eta,\rho)$ Coulomb wave functions.[11] These functions have the asymptotic limits

$$F_L(\eta,\rho)\overset{\rho\to\infty}{=}\sin(\rho-\eta\ln 2\rho-L\pi/2+\sigma_L),$$

$$G_L(\eta,\rho)\overset{\rho\to\infty}{=}\cos(\rho-\eta\ln 2\rho-L\pi/2+\sigma_L),$$

where

$$\sigma_L=\arg\Gamma(L+1+i\eta)$$

(see Ref. 11). Therefore, in the initial state the combination of functions

$$\sqrt{\frac{1}{K/m}}\frac{G_0(\eta,\rho)+iF_0(\eta,\rho)}{r}.$$

will have the asymptotic limit of a diverging spherical wave normalized to unit particle current. Using this expression, we write the radial wave function of the initial state in the form

$$\phi_i^C(r)=\begin{cases} a_0\dfrac{\sin(kr)}{r}, & 0\leqslant r<R_1,\\ \sqrt{\dfrac{1}{K/m}}\dfrac{G_0(\eta,\rho)+iF_0(\eta,\rho)}{r}, & R_i\leqslant r.\end{cases} \tag{34}$$

The coefficient a_0 is determined from the matching condition at the point R_1. The equation for the energy of the quasi-stationary state has the form

$$\cot(kR_1)=\frac{K}{k}\frac{G_0'(KR_1)+iF_0'(KR_1)}{G_0(KR_1)+iF_0(KR_1)}.$$

We construct the wave functions of the final state in analogy with functions (8) and (9). It can be easily seen that when the Coulomb interaction is switched off, i.e, when $\eta=0$ and correspondingly $\sigma_L=0$ holds, the function

$$4\pi\sum_{L=0}^{\infty}\sum_{m=-L}^{L}i^Le^{i\sigma_L}Y_{Lm}^*(\mathbf{n}_{\mathbf{K}'})Y_{Lm}(\mathbf{n}_{\mathbf{r}})\frac{F_L(\eta,K'r)}{K'r}$$

goes over to the expansion of a plane wave in spherical Bessel functions. The combination of functions

$$\frac{G_L(\eta,K'r)-iF_L(\eta,K'r)}{K'r}.$$

has the asymptotic limit of a ''converging spherical wave.'' Thus, we can write the wave functions of the final state as

$$\psi_f^C(\mathbf{r})=4\pi\sum_{L=0}^{\infty}\sum_{m=-L}^{L}i^Le^{i\sigma_L}Y_{Lm}^*(\mathbf{n}_{\mathbf{K}'})Y_{Lm}(\mathbf{n}_{\mathbf{r}})\phi_{fL}^C(r),$$

$$\phi_{fL}^C(r)=\begin{cases} A_Lj_L(k'r), & 0\leqslant r<R_1,\\ \dfrac{F_L(\eta,K'r)}{K'r}+C_L\dfrac{G_L(\eta,K'r)-iF_L(\eta,K'r)}{K'r}, & R_1\leqslant r.\end{cases} \tag{35}$$

Using the two matching equations for the wave functions and their derivatives at the point R_1 it is easy to find to find the coefficients A_L and C_L and thereby determine the wave functions of the final state over all space.

The emission spectrum is calculated according to formulas (21) and (22) with functions (34) and (35). Substituting the derivative of the potential (32)

$$\frac{dV(r)}{dr}=\left(-V_1+\frac{Z_dZ_\alpha e^2}{r}\right)\delta(r-R_1)-\frac{Z_dZ_\alpha e^2}{r^2}\theta(r-R_1),$$

into the formulas for the radial matrix elements leads to the following expressions for the matrix elements, which should be substituted in formula (21) in place of formulas (22):

$$\tilde{R}_{0,2}(\omega)=(-V_1R_1+Z_\alpha Z_de^2)R_1\phi_{f_1}^{C*}(R_1)j_{0,2}(\omega R_1)\phi_i^C(R_1)-Z_\alpha Z_de^2\int_{R_1}^{\infty}dr\,\phi_{f_1}^{C*}(r)j_{0,2}(\omega r)\phi_i^C(r). \tag{36}$$

The integral in expression (36) with wave functions (34) and (35) converges and can be calculated numerically.

5. TRUNCATED COULOMB BARRIER

To verify that the shape of the barrier has an affect on the bremsstrahlung spectrum, let us consider a potential of the form

$$V(r)=V_1\theta(R_1-r)+\frac{Z_dZ_\alpha e^2}{r}\theta(r-R_1)\theta(R_2-r), \quad (37)$$

which is a Coulomb potential barrier cut off at the point $r=R_2$. Taking the results of the two preceding sections into account, the wave functions of the α particle in the indicated potential have the following form:

the wave function of the initial state $\tilde{\psi_i^C}(\mathbf{r})=\tilde{\phi_i^C}(r)Y_{00}(\mathbf{n_r})$, where

$$\tilde{\phi_i^C}(r)=\begin{cases} a_0\dfrac{\sin(kr)}{r}, & 0\leqslant r<R_1, \\ b_0^{(1)}\dfrac{F_0(\eta,Kr)}{r}+b_0^{(2)}\dfrac{G_0(\eta,Kr)}{r}, & R_1\leqslant r<R_2, \\ \sqrt{\dfrac{1}{K/m}}\dfrac{e^{iK(r-R_2)}}{r}, & R_2\leqslant r; \end{cases} \quad (38)$$

the wave function of the final state

$$\tilde{\psi_f^C}(\mathbf{r})=4\pi\sum_{L=0}^{\infty}\sum_{m=-L}^{L} i^L Y^*_{Lm}(\mathbf{n_{K'}})Y_{Lm}(\mathbf{n_r})\tilde{\phi_{fL}^C}(r),$$

$$\tilde{\phi_{fL}^C}(r)=\begin{cases} A_L j_L(k,r), & 0\leqslant r<R_1, \\ B_L^{(1)}\dfrac{F_L(\eta,K'r)}{K'r}+B_L^{(2)}\dfrac{G_L(\eta,K'r)}{K'r}, & R_1\leqslant r<R_2, \\ (j_L(K'r)+C_L h_L^{(2)}(K'r)), & R_2\leqslant r. \end{cases} \quad (39)$$

Substituting the derivative of the potential (37)

$$\frac{dV(r)}{dr}=\left(-V_1+\frac{Z_dZ_\alpha e^2}{r}\right)\delta(r-R_1)-\frac{Z_dZ_\alpha e^e}{r^2}\times\theta(r-R_1)\theta(R_2-r)-\frac{Z_dZ_\alpha e^2}{r}\delta(r-R_2)$$

along with wave functions (38) and (39) into formula (22), we obtain the following formula for the matrix elements:

$$\begin{aligned}\tilde{R}_{0,2}(\omega)=&(-V_1R_1+Z_dZ_\alpha e^2)R_1\\ &\times\tilde{\phi_{f_1}^{C*}}(R_1)j_{0,2}(\omega R_1)\tilde{\phi_i^C}(R_1)\\ &-Z_dZ_\alpha e^2\int_{R_1}^{R_2}dr\,\tilde{\phi_{f_1}^{C*}}(r)j_{0,2}(\omega r)\tilde{\phi_i^C}(r)\\ &-Z_dZ_\alpha e^2R_2\tilde{\phi_{f_1}^{C*}}(R_2)j_{0,2}(\omega R_2)\tilde{\phi_i^C}(R_2). \quad (40)\end{aligned}$$

The emission spectrum calculated according to formula (21) with the matrix element (40) is plotted in Fig. 4. The Coulomb barrier, potential step, and cut-off Coulomb barrier, for which the spectra depicted here were calculated, model α decay of ^{214}Po. The Coulomb barrier was cut off at the point $R_2=r_E$. This is the right-hand boundary of the spherically symmetric potential step. Its height V_2 was determined by

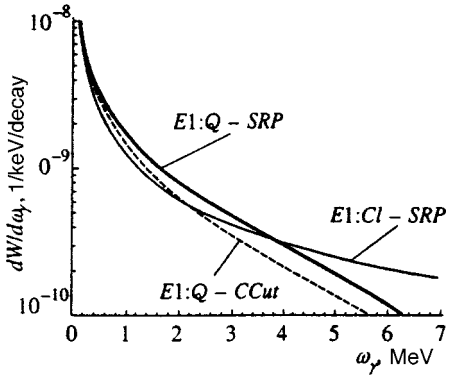

FIG. 4. Experimental data for ^{214}Po and ^{226}Ra (corrected data of Ref. 1) and ^{210}Po (from Refs. 2 and 3). Graphs $E1$: $Q-C$ and $E2$: $Q-C$—quantum-mechanical calculation of the $E1$ and $E2$ emission spectra for tunneling through a Coulomb barrier; the graphs $E1$: $Cl-C$—classical calculation of $E1$ emission for motion in the Coulomb field of a nucleus from the point of closest approach r_E to infinity.

equating the quasiclassical transmissivity coefficients of the Coulomb barrier and the potential step, i.e., in fact from the equality of the arguments of the exponentials

$$\int_{R_1}^{r_E}dr\sqrt{2m(V_2-E)}=\int_{R_1}^{r_E}dr\sqrt{2m\left(\frac{Z_dZ_\alpha e^2}{r}-E\right)}.$$

For V_2 this gives

$$V_2=E\left(1+\left(\frac{\arccos\sqrt{R_1/r_E}-\sqrt{R_1/r_E}\sqrt{1-R_1/r_E}}{1-R_1/r_E}\right)^2\right).$$

It can be seen from Fig. 4 that the quantum-mechanical emission spectra for the spherically symmetric potential step and for the cut-off Coulomb barrier for energies $\omega\leqslant 1$ MeV are very similar, i.e., at low energies the emission spectrum for these potentials is formed mainly in region III. This statement is in agreement with the behavior of the radial matrix elements in Fig. 3. This is completely natural for $\omega\leqslant 1$ MeV. The region $r\sim 1/\omega\sim 40$ fm (where both of the considered potential barriers already terminate) will work most effectively only for extremely high-energy γ photons with $\omega\approx 1/r\approx 5-7$ MeV.

6. EMISSION DURING MOTION ALONG A CLASSICAL TRAJECTORY

Let us consider the emission spectrum of an α particle moving in a Coulomb potential along a classical trajectory from the point of closest approach r_E, where the initial kinetic energy of the α particle is equal to zero, to infinity, where its kinetic energy is equal to E.

The electric-dipole bremsstrahlung spectrum is described by the well-known expression[8]

$$\frac{dW_\gamma^{(E1)}}{d\omega}=\frac{2(Z_{\rm eff}^{E1}e)^2}{3\pi}\frac{1}{\omega}|\mathbf{w}_\omega|^2, \quad (41)$$

in which $\mathbf{w}_\omega$ is the Fourier transform of the acceleration $\mathbf{w}(t)$:

$$\mathbf{w}_\omega=\int_0^\infty dt\mathbf{w}(t)e^{i\omega t}.$$

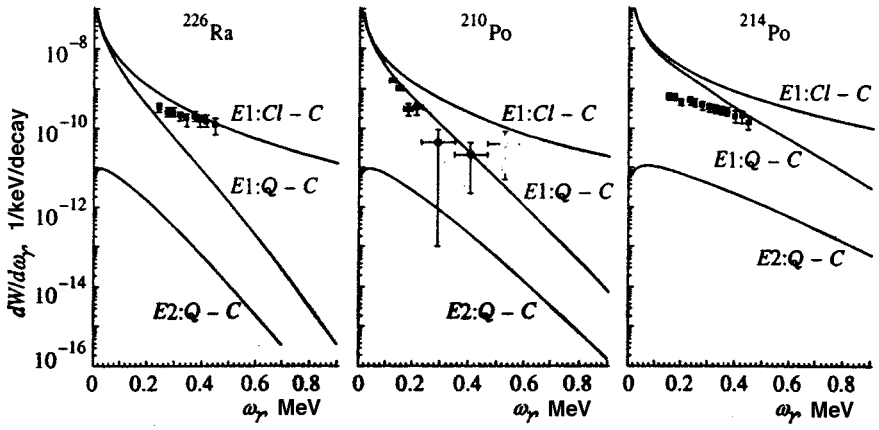

FIG. 5. Experimental data for ^{214}Po, ^{226}Ra (corrected values taken from Ref. 1) and ^{210}Po (from Refs. 2 and 3). $E1$: Q–C traces and $E2$: Q–C traces— quantum calculation of the $E1$ and $E2$ emission spectra for tunneling through the Coulomb barrier; $E1$:Cl–C traces— classical calculation of the $E1$ emission for motion in the Coulomb field of the nucleus from the point of minimum approach r_E to infinity.

The acceleration is determined from the equation of motion

$$w(t)=\frac{Z_\alpha Z_d e^2}{m r(t)^2} \tag{42}$$

with initial conditions

$$r(t=0)=r_E, \quad v(t=0)=0.$$

Integrating Eq. (42), we find the equation of the trajectory

$$\sqrt{\frac{r(t)}{r_E}}\sqrt{\frac{r(t)}{r_E}}-1+\ln\left(\sqrt{\frac{r(t)}{r_E}}+\sqrt{\frac{r(t)}{r_E}-1}\right)=\frac{v_\alpha t}{r_E}, \tag{43}$$

where $v_\alpha=\sqrt{2E/m}$. This dependence $r(t)$ is used to calculate the Fourier transform of the acceleration from Eq. (42) and, correspondingly, the emission spectrum.

In the case of a potential step it is also possible to find the ''classical'' electric-dipole emission spectrum. Toward this end, we draw a sloping line from the point at which the α particle ''exits'' from the rectangular potential barrier (this is the intersection point of the energy line E and the outer vertical wall of the barrier, located in Fig. 1 at $r=R_2$) to the point at which it intersects the r axis, $r=R_3$, where $R_3>R_2$. (That is to say, the α particle does not ''fall'' straight down upon appearing at the outer wall of the rectangular barrier but rather ''rolls down'' along a sloped plane.) The equations of motion in this case have the form

$$m\mathbf{w}(t)=E/\Delta R, \quad R_2\leqslant r\leqslant R_3,$$

$$m\mathbf{w}(t)=0, \quad R_3<r, \tag{44}$$

where $\Delta R=R_3-R_2$, and the initial conditions are analogous to those considered above: $r(t=0)=R_2$, $v(t=0)=0$. Integrating Eq. (44), we find the equation of the trajectory

$$r(t)=\frac{E}{2m\Delta R}t^2+R_2, \quad 0\leqslant t\leqslant T,$$

$$r(t)=v_\alpha t+R_3, \quad T<t,$$

where $T=2\Delta R/v_\alpha$ is the time it takes the particle to move from R_2 to R_3.

The Fourier transform of the acceleration

$$\mathbf{w}_\omega=\int_0^T dt\,\frac{E}{m\Delta R}e^{i\omega t}=v_\alpha e^{i\omega\Delta R/v_\alpha}\frac{\sin(\omega\Delta R/v_\alpha)}{\omega\Delta R/v_\alpha}$$

in the limit $\Delta R\to 0$ is v_α. Therefore, for the classical $E1$ emission spectrum of a particle ''falling'' in the field of the vertical potential wall we obtain

$$\frac{dW_\gamma^{(E1)}}{d\omega}=\frac{2(Z_{\rm eff}^{E1}e)^2}{3\pi}\frac{1}{\omega}v_\alpha^2. \tag{45}$$

Spectra corresponding to formulas (41) for various nuclei are plotted in Fig. 5. As for the classical emission spectrum for the vertical potential wall (45), it is shown in Figs. 2 and 4. Note that in contrast to the usual Coulomb case, for a potential step the quantum-mechanical $E1$ spectrum lies above the classical spectrum in the low-energy region.

7. RESULTS OF CALCULATIONS FOR ^{210,214}Po AND ^{226}Ra

We present a sequence of calculations of α-decay emission spectra of specific nuclei for the problem with a Coulomb barrier. We choose the size of the well for α decay of a nucleus with atomic number A according to the formula

$$R_1=1.2[(A-4)^{1/3}+4^{1/3}] \ \text{fm}$$

(the bremsstrahlung spectrum varies only slightly as R_1 varies within reasonable limits). Next, proceeding from the experimental value of the energy of the α particle in the center-of-mass system E, we calculate the values of the Coulomb functions and their derivatives in the initial state at the point $r=R_1$. From the equation for the energy of the quasi-stationary state for fixed E we find a set of values for the potential V_1. The magnitude of the latter is the main factor determining the lifetime of the quasi-stationary state. The bremsstrahlung spectrum, as was noted in Ref. 5, is insensitive to V_1. However, this is valid only in the absence of bound states with $L=1,2$ in a potential well of depth V_1. Levels with $L=1$ can lie quite low. In addition, the $E1$ resonance in the bremsstrahlung spectrum at these levels is narrow. Therefore, to calculate the $E1$ spectrum out to photon energies $\omega\approx 1$ MeV requires several values of the potential V_1. The $E2$ emission spectrum is another matter. $E2$ resonances are very broad. A level with $L=2$ is present and is found quite close to the $L=0$ level for all values of V_1

except for the one for which the well has the smallest depth. For the nuclei considered here this value of the potential V_1 is positive. The majority of the photon spectra presented in the figures were calculated for just such potential wells with smallest depth. In this case, clearly, the situation is possible in which the energy of the final state is less than V_1 and the momentum k' becomes imaginary. In this case it is more convenient to work with a combination of Hankel functions in region I, using the fact that $j_L(x)=h_L^{(1)}(x)+h_L^{(2)}(x)$. In the remaining cases the calculation is analogous to the case $V_1<0$.

At present, the most reliable experimental data available are for the ^{210}Po nucleus.[2,3] Figure 5 displays, in addition to the indicated experimental values, curves calculated in the quantum-mechanical approach according to formulas (21) and (36), and also (24) and (25), of the $E1$ and $E2$ emission spectra for passage of an α particle through a Coulomb barrier. These include curves corresponding to $E1$ emission (41) arising upon the motion along a classical trajectory in the Coulomb field of a nucleus from the point of closest approach, where the kinetic energy is equal to zero, to infinity, where the kinetic energy is equal to E (43). Our quantum-mechanical result for the electric-dipole emission spectrum in a Coulomb potential is in excellent agreement with the calculation of Ref. 5. At present, there is probably no basis for speaking of a disparity between the experimental data[2,3] and the theoretical calculations obtained here and in Ref. 5.

The experimental data for the ^{210}Po and ^{226}Ra nuclei are from Ref. 1. These results were later subjected to the criticism of the authors of Refs. 2 and 3. The misunderstanding is partly due to the fact that Ref. 1 presents data on the emission spectrum of γ radiation at an angle of 90° to the direction of motion of the α particles (this is noted in the text of the article). The data were not averaged over angle. However, this fact is in no way reflected in the graphs. Corrected values for ^{210}Po and ^{226}Ra, i.e., values averaged over the angles of emission of the γ quantum, were provided for the present publication by N. V. Eremin,[1)] who performed all of the measurements for Ref. 1 in his own time. Now the experimental points for ^{226}Ra lie (see Fig. 5) systematically above the quantum-mechanical $E1$-emission probability curve for tunneling through the Coulomb barrier and are very close to the line corresponding to classical $E1$ emission in a Coulomb field. However, the experimental values for ^{214}Po relative to the analogous calculated curves behave completely differently. They lie systematically below the quantum-mechanical $E1$-emission spectrum for the Coulomb potential. Such fundamentally different behavior of the experimental data seems quite strange. Still more problematic is the fact that the experimental values for such different α-emitters as ^{210}Po and ^{226}Ra (the energies of the α particles are, respectively, 7.687 MeV and 4.785 MeV in the laboratory system) essentially coincide. From Fig. 5 it may be surmised that in addition to the bremsstrahlung in the experiment with ^{226}Ra discussed here, bremsstrahlung arising in the scattering of α particles on neighboring nuclei was also measured. However, the data on ^{214}Po speak otherwise. It follows from them that the emission process in the form considered here simply does not exist. Instead, some process is being measured whose probability is practically independent of the energy of the α particles. It is clear that this cannot be the bremsstrahlung investigated in the present work if only because for ^{214}Po the spectrum in the region $\omega_\gamma \approx 100-200$ keV should already be close to the classical limit. The question is probably still open, and it is necessary to carry out new experiments with increased measurement accuracy and rejection of processes involving scattering on neighboring nuclei.

In conclusion let use address briefly another question touched on, in particular, in Ref. 3. This is the problem of the so-called ''destructive interference of amplitudes in the under-barrier and outside regions.'' According to Ref. 3, the fact that the quantum curve of the $E1$ spectrum for a Coulomb field lies below the classical curve (Fig. 5) can be explained by destructive interference of amplitudes in the indicated regions of space. Note that the discussion in Ref. 3 is about one and the same amplitude whose magnitude it is proposed to estimate separately in the under-barrier and outside regions. According to the model of the emission process used in the present work, such interference of regions must be understood as contributions of integrals over different regions of space in the calculation of the transition matrix elements to the total radial integral. In this sense the example with a spherically symmetric potential step is instructive. (This barrier is convenient because in contrast to the Coulomb case the barrier boundary is found at the sample point $r=R_2$ for the initial and final states, and there is no ''transitional'' region.) For a potential step the quantum curve lies above the classical curve up to energies of several MeV [Fig. 2b]. It would seem that a ''constructive interference of regions'' should take place. Let us consider the spectrum in ''length form,'' formed by the dipole moment of the transition (31). Indeed, in Fig. 3 there are energies of the emitted photons at which the transition dipole moments have the same sign and add. However, for most of the energies the signs of the dipole moments in the under-barrier and outside regions are different, and the contributions to the total radial integral largely cancel, i.e., ''destructive interference'' occurs, if we follow the logic of Ref. 3. Let us turn now to an analysis of formulas for the matrix elements of the potential gradient (23). Here it is generally unclear how to treat the radial matrix elements from the viewpoint of interference of regions. In contrast to r, the operator dV/dr is localized at the point $r=R_2$, i.e., precisely at the boundary of two regions. The situation thus depends on how the spectrum is calculated, on the operator whose matrix element we are calculating. There is nothing similar in the case of different diagrams or amplitudes. Therefore, if we are speaking of ''interference of regions,'' we must first specify the operator about whose spatial distribution we are talking, and second bear in mind the explicit limitations of the concept under discussion.

The author is grateful to A. M. Dykhne, N. V. Eremin, and N. P. Yudin for useful discussions of a number of questions touching on this paper.

The present work was partially supported by the Russian Fund for Fundamental Research (Grants No. 98-02-16070a

and No. 98-02-16529a) and a grant in support of Leading Scientific Schools (No. 96-15-96481).

*)E-mail: tkalya@ibrae.ac.ru

1)Corrected values of the spectra for ^{210}Po and ^{226}Ra will be published by the authors of Ref. 1 in Phys. Rev. C in 1999.

[1] A. D. Arrigo, N. V. Eremin, G. Fazio, *et al.*, Phys. Lett. B **332**, 25 (1994).
[2] J. Kasagi, H. Yamazaki, N. Kasajima, *et al.*, Phys. Rev. Lett. **79**, 371 (1997).
[3] J. Kasagi, H. Yamazaki, N. Kasajima, *et al.*, J. Phys. G **23**, 1451 (1997).
[4] M. I. Dyakonov and I. V. Gornyi, Phys. Rev. Lett. **76**, 3542 (1996).
[5] T. Papenbrock and G. F. Bertch, Phys. Rev. Lett. **80**, 4141 (1998).
[6] M. A. Preston, *Physics of the Nucleus* (Addison-Wesley, Reading, Mass., 1962).
[7] J. M. Eisenberg and W. Greiner, *Nuclear Theory: Mechanisms of Excitation of the Nucleus. Electromagnetic and Weak Interactions*, 2nd ed., (North-Holland, Amsterdam, 1976).
[8] L. A. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, 4th ed. (Pergamon Press, Oxford, 1975).
[9] L. A. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-Relativistic Theory*, 3rd ed. (Pergamon Press, Oxford, 1977).
[10] A. I. Akhiezer and V. B. Berestetskii, *Quantum Electrodynamics* (Wiley, New York, 1965).
[11] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1965).
[12] M. Ya. Amus'ya, *Bremsstrahlung Radiation* [in Russian] (Energoatomizdat, Moscow, 1990)

Translated by Paul F. Schippnick

Many-electron tunneling in atoms

B. A. Zon*)

Voronezh State University, 394693 Voronezh, Russia
(Submitted 8 December 1998)
Zh. Éksp. Teor. Fiz. **116**, 410–417 (August 1999)

A formula for describing the N-electron ionization of atoms by a dc field and laser radiation in the tunneling regime is derived theoretically, and numerical examples for noble-gas atoms are presented. © *1999 American Institute of Physics.* [S1063-7761(99)00408-4]

1. INTRODUCTION

The first to observe multi-electron ionization of alkali-earth atoms by laser light were Suran and Zapesochnyĭ[1] (a review of their paper and of earlier work can be found in Ref. 2). At present such studies constitute one of the main avenues of research in the physics of the interaction of strong laser light and atoms.[3]

To interpret the experimental data, a number of theoretical models have been proposed, some dealing with the direct effect of laser light on atomic electrons,[4–8] others considering the formation of highly stripped ions due to the inelastic scattering of previously emitted ions by the parent ion.[9–11] These models make it possible to explain many features of the phenomenon.[12–16] Nevertheless, the difficulties inherent in the theoretical description of the formation of highly stripped ions in a laser field not associated with inelastic scattering[4–8] make it impossible to properly apply these mechanisms in the interpretation of the experimental data.

At the same time, relatively simple formulas provide a satisfactory description of the formation of singly charged ions in a laser field.[17–19] Eichmann *et al.*[20] developed an empirical generalization of these formulas (known as the formulas of the ADK (Ammosov–Delone–Kraĭnov) theory) to describe the formation of highly stripped ions. Hence it would be reasonable to generalize the existing theory of tunneling in atoms to the case of nonsequential multiple ionization of atoms. The aim of the present work is to solve this problem.

An obvious analog of the phenomenon being discussed is the Josephson effect in solid-state theory. Zakhar'ev[21] has proposed a number of ideas concerning the difference between single-particle and multiparticle tunneling effects. A comparison of these ideas with the results of the present work shows that for tunneling in atoms the difference is not as trivial as described in Ref. 21.

2. ASYMPTOTIC BEHAVIOR OF THE MULTI-ELECTRON WAVE FUNCTION

Let us recall some facts that will make the understanding of the main concept of the proposed model easier. As well is known, to describe optical transitions in complex atoms, Bates and Damgaard[22] modified the Slater model (see Ref. 23), basically retaining the nodeless nature of the Slater orbitals. Unlike the Slater model, the effective nuclear charge ceases to be a fitting parameter for the valence electrons of an atom, since it coincides with the residual ion charge. The effective principal quantum number, however, is uniquely determined by the electron binding energy. All this means that one must consider the asymptotic region of electron motion, where the atomic potential is of the Coulomb type. The high accuracy of oscillator-strength calculations achieved by the Bates–Damgaard method (see Sobelman's book cited in Ref. 24) and the clear physical justification allows us to use this methods for calculating other atomic characteristics determined by large electron–nucleus distances.

Another characteristic determined by large electron–nucleus distances, where the electron–external-field interaction becomes comparable to the attraction energy of the residual atom, is the tunneling probability. Thus, the Bates–Damgaard method can be used to describe the tunneling effect. Such a procedure has been recently developed in Ref. 25 for tunneling calculations in Rydberg molecules, and some conditions for the applicability of the method were also estimated.

Suppose that N equivalent (i.e., belonging to the same atomic shell) electrons are removed from an atom via tunneling. The asymptotic behavior of the radial part of the N-electron wave function in the Bates–Damgaard approximation is determined by the product of properly symmetrized one-electron asymptotic wave functions:

$$\psi_{\nu lm}(\mathbf{r}) = C_{\nu l} b^{-3/2} \left(\frac{r}{b}\right)^{\nu-1} \exp\left(-\frac{r}{b}\right) Y_{lm}\left(\frac{\mathbf{r}}{r}\right),$$

$$C_{\nu l} = (2\pi\nu)^{-1/2} \left(\frac{2}{\nu}\right)^{\nu} L(\varepsilon),$$

$$L(\varepsilon) = \left(\frac{1-\varepsilon}{1+\varepsilon}\right)^{(l+1/2)/2} (1-\varepsilon^2)^{-\nu/2}, \tag{1}$$

where Z is the residual-ion charge, $a = \hbar^2/\mu e^2$ is the Bohr radius, μ and e are the electron mass and charge (the absolute value of the charge), $b = \alpha\nu/Z$, and $\varepsilon = (l+1/2)/\nu$. The constant $C_{\nu l}$ in (1) is determined in the quasiclassical approximation without the assumption that $l \ll \nu$ (this assumption was adopted in Ref. 19). In view of this, the function $L(\varepsilon)$ arises, with $L(\varepsilon) \to 1$ as $\varepsilon \to 0$. In this limit the expression for the constant $C_{\nu l}$ in (1) becomes formula (11) of Ref. 19 with an inaccuracy corrected (the number $e = 2.718\cdots$ should be omitted).

 © 1999 American Institute of Physics

The expression for $C_{\nu l}$ in (1) was obtained under the assumption that $\varepsilon<1$. For $\varepsilon>1$, the quasiclassical approximation breaks down, with the result that numerical methods must be used to calculate $C_{\nu l}$ (see, e.g., Ref. 26).

The principal quantum number ν is determined by the electron binding energy. If E_1/e, E_2/e, etc., are the first, second, etc., ionization potentials of the atom, the principal quantum number of the jth electron detached from the atom is $\nu_j=(2aE_j/Z^2e^2)^{-1/2}$. But if the electrons are equivalent, i.e., belong to the same shell, and are detached from the atom simultaneously, for all these electrons

$$\nu=\left(\frac{2aE_N}{NZ^2e^2}\right)^{-1/2}, \tag{2}$$

where $E_N=\Sigma_{j=1}^{N}E_j$ is the binding energy of the N electrons. Note that in this model the asymptotic behavior of the bound-electron wave function [Eq. (1)] depends on how many electrons get detached from the atom. This partially allows for multi-electron effects in the initial state of the atom. Formula (2) is similar to the one derived by Eichmann *et al.*[20] empirically.

Now we consider N-electron ionization as the detachment from the atom of an N-electron cloud, a quasiparticle of sorts that has a mass $N\mu$ and carries a charge $-Ne$. Here in the region in which ionization takes place we consider the distances between the electrons in the cloud to be much smaller than the distance between the atomic core and the center of mass of the cloud. Denoting the distance between the ith and jth electrons by $\mathbf{x}_{ij}$ and the radius vector of the position of the cloud center of mass by $\mathbf{R}$, we can write

$$x_{ij}\ll R. \tag{3}$$

Since the interaction between atom and laser field is considered in the dipole approximation, the effect of the field on N individual electrons is equivalent to the effect of the field on a quasiparticle with a charge $-Ne$ located at the point $\mathbf{R}$. As for the interaction between the quasiparticle and the Coulomb field of the atomic core, the error introduced by ignoring this field is of order $(x_{ij}/R)^2$, which is small in view of the adopted condition (3).

The problem that we must solve to describe mathematically the model is similar to the one that appears, say, in nuclear α-decay theory, i.e., we must construct the quasiparticle wave function $\Psi^{(N)}_{\{\nu lm\}}(\mathbf{R},\{\mathbf{x}_i\})$ at large distances from the residual system using the one-particle wave functions of the system in the initial state (symbols in braces represent sets of quantum numbers or the coordinates of individual particles). To solve this problem we examine the asymptotic behavior of the function $\Psi^{(N)}_{\{\nu lm\}}$ as $R\to\infty$, which is represented by the product of the asymptotic one-electron wave functions (1). Clearly, the radial dependence of (1) introduces the factor $\exp\{-NR/b\}(R/b)^{N(\nu-1)}$ into the asymptotic function $\Psi^{(N)}_{\{\nu lm\}}$. To obtain the angular dependence we must specify the way in which the variables $\mathbf{R}$ and $\{\mathbf{x}_i\}$ are introduced. Since in a linearly polarized field the problem is axisymmetric, the projections of the orbital angular momenta of the noninteracting electrons on the direction of polarization are conserved. Hence it is convenient to leave the azimuthal angles φ_i the same as in the initial spherical system of coordinates with the atomic nucleus as its origin. The change of variables will affect only the absolute values $\{r_i\}$ and the polar angles $\{\theta_i\}$. The behavior as $\theta\to0$ of the Legendre polynomials in the spherical functions in (1) is given by the formula

$$P_l^{|m|}(\cos\theta_i)\sim(-1)^{|m|}\frac{(l+|m|)!}{(l-|m|)!}\frac{\sin^{|m|}\theta_i}{2^{|m|}|m|!}$$
$$=(-1)^{|m|}\frac{(l+|m|)!}{(l-|m|)!}\frac{(r_i^2-r_{iz}^2)^{|m|/2}}{2^{|m|}|m|!r_i^{|m|}}.$$

Replacing r_i with R and r_{iz} with R_z and introducing the parabolic coordinates $\xi=R+R_z$ and $\eta=R-R_z$ for the cloud center of mass, we can write the asymptotic N-electron wave function for $\xi\gg\eta$:

$$\Psi^{(N)}_{\{\nu lm\}}(\mathbf{R},\{\mathbf{x}_i\})=B\phi(\xi,\eta)\chi(\{r_i,\theta_i\})\times\prod_{j=1}^{N}\frac{\exp\{im_j\varphi_j\}}{\sqrt{2\pi}},$$

$$B=a^{-3/2}C^N_{\nu l}\left(\frac{Z}{\nu}\right)^{3N/2}\left(l+\frac{1}{2}\right)^{N/2}\times\prod_{j=1}^{N}\frac{(-1)^{|m_j|}}{|m_j|!}\left[\frac{(l+|m_j|)!}{(l-|m_j|)!}\right]^{1/2},$$

$$\phi(\xi,\eta)\sim\exp\left[-\frac{N(\xi+\eta)}{2b}\right]\left(\frac{\xi}{2b}\right)^{N(\nu-1)}\left(\frac{\eta}{\xi}\right)^{M/2}, \tag{4}$$

where $M=\Sigma_{j=1}^{N}|m_j|$ and χ is the wave function (normalized to unity) of the electron's motion within the cloud. Note that among the $2N$ variables $\{r_i,\theta_i\}$ there are only $2(N-1)$ independent variables. The function $\phi(\xi,\eta)$ describes the motion of the cloud's center of mass.

3. TUNNELING PROBABILITY

Further calculations of the tunneling probability are carried out by the standard procedure,[17,27] with allowance for the fact that the mass of the electron cloud is $N\mu$ and the charge is $-Ne$. Inserting the function $\phi(\xi,\eta)$ in the Schrödinger equation describing the motion along the parabolic coordinate ξ as $\xi\to\infty$,

$$\frac{d}{d\xi}\left(\xi\frac{d\phi}{d\xi}\right)+\left(\beta-\frac{E_N N\mu}{2\hbar^2}\xi\right)\phi=0,$$

we arrive at a formula for the constant associated with the separation of the variables:

$$\beta=\frac{N}{b}\left[N(\nu-1)-\frac{M-1}{2}\right]. \tag{5}$$

Here we ignore the centrifugal potential since it decays rapidly as $\xi\to\infty$.

Assuming that the external field $F(t)$ varies slowly, we adopt the quasiclassical approximation for the wave function $\phi_F(\xi,\eta)$ describing the motion of the electron cloud's center of mass in the field. In the subbarrier region,

$$\phi_F(\xi,\eta)=\kappa\left(\frac{\xi|p(\xi)|}{\hbar}\right)^{-1/2}\exp\left\{\frac{1}{\hbar}\int_{\xi_1}^{\xi}|p(\xi)|d\xi\right\}, \quad (6)$$

$$p(\xi)=\hbar\left(-\frac{E_N N\mu}{2\hbar^2}+\frac{\beta}{\xi}+\frac{1}{4\xi^2}+\frac{N^2 e\mu}{4\hbar^2}F\xi\right)^{1/2}, \quad (7)$$

where ξ_1 is the larger root of the equation $p(\xi)=0$. By comparing the expression (6) with the function $\phi(\xi,\eta)$ in (4) at a point ξ_0 inside the interval

$$\frac{2\hbar^2\beta}{E_N N\mu}\simeq b\ll\xi_0\ll\frac{2E_N}{NeF}=\frac{eZ}{b\nu F} \quad (8)$$

we find κ:

$$\kappa(\eta;\xi_0)\simeq\left(\frac{N\xi_0}{2b}\right)^{1/2}\exp\left\{-\frac{1}{\hbar}\int_{\xi_0}^{\xi_1}|p(\xi)|d\xi\right\}\phi(\xi_0,\eta). \quad (9)$$

The condition for the existence of the interval (8) yields a restriction on the external field,

$$F\ll F_a\equiv\frac{eZ}{b^2\nu}=\frac{e}{a^2}\left(\frac{Z}{\nu}\right)^3, \quad (10)$$

which differs from the condition that arises in the description of the one-electron tunneling effect only in the way in which ν is defined. Note that if ν is much larger than unity, which is the case, say, for Rydberg states, the inequality (10) is replaced by a stronger one,

$$F<\frac{Z^3 e}{16\nu^4 a^2}, \quad (11)$$

which follows from the condition for the existence of a potential barrier.[28]

Equations (6) and (9) define the function $\phi_F(\xi,\eta)$ outside the barrier. With allowance for (8), we have an expression for the square of the absolute value of this function:[27]

$$|\phi_F(\xi,\eta)|^2=\frac{\hbar N\xi_0}{2b\xi p(\xi)}\left(\frac{\xi_0}{2b}\right)^{2N(\nu-1)}\left(\frac{\eta}{\xi_0}\right)^M\times\exp\left\{-\frac{N\eta}{b}-\frac{16\hbar^2}{3N^2\mu eF}\left(\frac{E_N N\mu}{2\hbar^2}\right)^{3/2}-\beta\left(\frac{2\hbar^2}{E_N N\mu}\right)^{1/2}\ln\frac{NeF\xi_0}{8E_N}\right\}. \quad (12)$$

Using (2) and (5), we can easily see that the dependence on the arbitrary parameter ξ_0 has in fact dropped out of (12):

$$|\phi_F(\xi,\eta)|^2=\frac{\hbar N(\eta/b)^M}{2^M\xi p(\xi)}\left(\frac{2F_a}{F}\right)^{2N(\nu-1)-M+1}\times\exp\left(-\frac{N\eta}{b}-\frac{2NF_a}{3F}\right). \quad (13)$$

The ionization probability is determined by the flux of the probability density (13) through a plane perpendicular to the z axis:[27]

$$W_{\nu l}^{(N)}(F)\sim 2\pi\int_0^\infty v_z|\phi_F(\xi,\eta)|^2\rho\,d\rho,$$

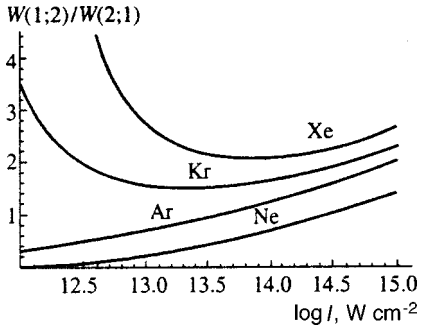

FIG. 1. Probability ratio of triply-charged noble-gas ion formation by two different channels (see the text).

$$v_z=\frac{2p(\xi)}{N\mu},\quad \rho=\sqrt{\xi\eta},\quad d\rho\simeq\sqrt{\frac{\xi}{\eta}}d\eta.$$

Substituting (4) and (13) in this formula yields

$$W_{\nu l}^{(N)}(F)=\frac{\pi\hbar}{a^2\mu}\frac{M!(l+1/2)^N C_{\nu l}^{2N}}{2^{M-2}N^{M+1}}\left(\frac{Z}{\nu}\right)^{3N-1}\times\prod_{j=1}^{N}\frac{(l+|m_j|)!}{(|m_j|!)^2(l-|m_j|)!}\left(\frac{2F_a}{F}\right)^{2N(\nu-1)-M+1}\times\exp\left\{-\frac{2NF_a}{3F}\right\}. \quad (14)$$

This formula determines the N-electron tunneling probability in a dc field to within a factor accounting for the overlap of the wave functions of the electrons remaining in the atom and the wave functions of the same electrons in the initial state. Obviously, this factor cannot exceed unity (a more accurate evaluation can be done only numerically). Note that the factor N in the exponent of the exponential function in (14) in no way exhausts the dependence of the exponent on N, as Zakhar'ev[21] believed. In view of Eqs. (2) and (10), the dependence on N is much more complicated and is determined by the spectrum of the particular atom. Figure 1 depicts a numerical example illustrating this statement.

We will assume that

$$F(t)=F_0\cos\omega t, \quad (15)$$

where ω is the frequency of the laser field. As is well known, tunneling in a laser field is possible when the Keldysh[29] parameter $\gamma=(\sqrt{2\mu E_1}/eF)\omega$ is small (here E_1 is the binding energy per electron). Applying the technique developed by Keldysh[29] to a particle of mass $N\mu$ and charge $-Ne$, we can easily see that N-electron tunneling is possible when the parameter

$$\gamma_N=\frac{\sqrt{2\mu E_N/N}}{eF}\omega \quad (16)$$

is small. Since the energy of detachment of each subsequent electron increases with the number of the electron, N-electron tunneling requires field values that are lower than those required by an N-electron tunneling cascade.

We substitute (15) in (14) and average the result over the time interval $t\in[-\pi/2\omega,\pi/2\omega]$ (see Ref. 18).[1)] In view of

the inequality (10), the resulting integral can be evaluated by the saddle-point method. If the condition (11) is met, the saddle point is $t=0$. The final formula is

$$W_{\nu l}^{(N)}(F_0)=\frac{\sqrt{3\pi}\hbar}{a^2\mu}\frac{M!(l+1/2)^N C_{\nu l}^{2N}}{2^{M-3/2}N^{M+3/2}}\left(\frac{Z}{\nu}\right)^{3N-1}$$
$$\times\prod_{j=1}^{N}\frac{(l+|m_j|)!}{(|m_j|!)^2(l-|m_j|)!}\left(\frac{2F_a}{F_0}\right)^{2N(\nu-1)-M+1/2}$$
$$\times\exp\left\{-\frac{2NF_a}{3F_0}\right\}. \qquad (17)$$

We note once more that the dependence on N of the exponent of the exponential function in (17) is not exhausted by the factor N written explicitly.

4. NUMERICAL EXAMPLES

Unfortunately, the above formulas cannot be related directly to an experiment, since, in addition to the possibility of direct N-fold ion formation there is the possibility of many cascade channels and other ionization mechanisms for which inelastic collisions of electrons and ions are responsible.[9–11] Comparing the theoretical results with experimental data requires solving certain kinetic equations, a problem that merits a separate investigation. Hence this section considers only a few illustrative examples.

Figure 1 depicts the probability ratio of triply-charged noble-gas ion formation by two different channels, $A\to A^+\to A^{3+}$ and $A\to A^{2+}\to A^{3+}$. The two probabilities are denoted by $W(1;2)$ and $W(2;1)$, respectively, and have the same dependence on the laser pulse length. Clearly, $W(1;2)/W(2;1)$ is not equal to unity, as follows from the Zakhar'ev results.[21]

The following result is also of interest: the probabilities of two-electron tunneling for neutral atoms may be higher than the probabilities of one-electron tunneling for the corresponding singly charged ions. For instance, for the Ar atom the two-electron tunneling probability exceeds the probability of a one-electron process for the Ar^+ ion at intensities $I>10^{14.88}\,W\,cm^{-2}$. The same result is true for Kr at $I>10^{14.76}\,W\,cm^{-2}$ and for Xe at $I>10^{14.34}\,W\,cm^{-2}$. At the same time, for the light noble gases He and Ne, the one-electron tunneling probabilities are two orders of magnitude higher that the probabilities of the two-electron process in the corresponding neutral atoms at $I\simeq 10^{15}\,W\,cm^{-2}$. These facts prove that there exists a wide range of experimental situations arising in multiphoton tunneling.

This work was stimulated by the work of Eichmann *et al.*[20] The author is grateful to Prof. W. Sandner for the interest in the work and to WE-Heraeus-Stiftung for the opportunity to participate in the seminar. Special thanks go to Prof. N. B. Delone and to the participants of his seminar at the Institute of General Physics of the Russian Academy of Sciences for fruitful discussions. This research was partially supported by the Russian Fund for Fundamental Research (Grant No. 97-02-18035).

*)E-mail: zon@niif.vsu.ru

1)The values $t\in[\pi/2\omega,3\pi/2\omega]$ yield $F(t)<0$, and the tunneling takes place in the direction opposite to the positive z semiaxis.

[1] V. V. Suran and I. P. Zapesochnyĭ, Pis'ma Zh. Tekh. Fiz. **1**, 973 (1975) [Sov. Tech. Phys. Lett. **1**, 420 (1975)].
[2] N. B. Delone, V. V. Suran, and V. A Zon, in *Multiphoton Ionization of Atoms*, S. L. Chin and P. Lambropoulos (Eds.), Academic Press, New York (1984), pp. 235–264.
[3] L. F. DiMauro and P. Agostini, Adv. At. Mol. Phys. **35**, 79 (1995).
[4] K. C. Kulander, Phys. Rev. A **36**, 2726 (1987); **38**, 778 (1988).
[5] C. A. Ullrich, U. J. Goosman, E. K. U. Gross, Phys. Rev. Lett. **74**, 872 (1995).
[6] P. Maragakis and P. Lambropoulos, Laser Phys. **7**, 679 (1997).
[7] F. H. M. Faisal, Phys. Lett. A **187**, 180 (1994); A. Becker and F. H. M. Faisal, Phys. Rev. A **50**, 3256 (1994); J. Phys. B **29**, L197 (1996); F. H. M. Faisal and A. Becker, Laser Phys. **7**, 684 (1996); F. H. M. Faisal, A. Becker, and J. Muth, Report at the *WE-Heraeus-Seminar ''Strong Field Phenomena''* (Berlin, July 6–10, 1998).
[8] P. A. Golovinski, Laser Phys. **7**, 655 (1997).
[9] K. Boyer and C. K. Rhodes, Phys. Rev. Lett. **54**, 1490 (1985).
[10] M. Yu. Kuchiev, JETP Lett. **45**, 404 (1987); J. Phys. B **28**, 5094 (1995); Phys. Lett. A **212**, 77 (1996).
[11] P. B. Corkum, Phys. Rev. Lett. **71**, 1994 (1993).
[12] B. Walker, B. Sheely, L. F. DiMauro, P. Agostini, K. J. Schafer, and K. C. Kulander, Phys. Rev. Lett. **73**, 1227 (1994).
[13] S. Augst, A. Talebpour, S. L. Chin, Y. Beaudoin, and M. Chaker, Phys. Rev. A **52**, R917 (1995).
[14] A. Talebpour, C. Y. Chien, and S. L. Chin, J. Phys. B **29**, 5725 (1996).
[15] A. Talebpour, C. Y. Chien, Y. Liang *et al.*, J. Phys. B **30**, 1721 (1997).
[16] S. Larochelle, A. Talebpour, and S. L. Chin, J. Phys. B **31**, 1201 (1998).
[17] B. M. Smirnov and M. I. Chibisov, Zh. Éksp. Teor. Fiz. **49**, 841 (1965) [Sov. Phys. JETP **22**, 585 (1966)].
[18] A. M. Perelomov, V. S. Popov, and M. V. Terent'ev, Zh. Éksp. Teor. Fiz. **50**, 1393 (1966) [Sov. Phys. JETP **23**, 924 (1966)].
[19] M. V. Ammosov, N. B. Delone, and V. P. Kraĭnov, Zh. Éksp. Teor. Fiz. **91**, 2008 (1986) [Sov. Phys. JETP **64**, 1191 (1986)].
[20] U. Eichmann, H. Maeda, and W. Sandner, Report at the *WE-Heraeus-Seminar ''Strong Field Phenomena''* (Berlin, July 6–10, 1998).
[21] B. N. Zakhar'ev, Izv. Akad. Nauk SSSR, Ser. Fiz. **47**, 859 (1983).
[22] D. Bates and A. Damgaard, Philos. Trans. R. Soc. London, Ser. A **242**, 101 (1949).
[23] P. Gombas, *Theorie und Lösungsmethoden des Mehrteilchenproblems der Wellenmechanik*, Basel (1950).
[24] I. I. Sobelman, *Atomic Spectra and Radiative Transitions*, Springer-Verlag, Berlin (1979).
[25] B. A. Zon, Zh. Éksp. Teor. Fiz. **112**, 115 (1997) [JETP **85**, 61 (1997)].
[26] A. A. Radtsig and B. M. Smirnov, *Reference Data on Atoms, Molecules, and Ions*, Springer-Verlag, Berlin (1985).
[27] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-relativistic Theory*, 3rd ed., Pergamon Press, Oxford (1977), ß 77, Problem 1.
[28] H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Systems*, Springer-Verlag, Berlin (1958).
[29] L. V. Keldysh, Zh. Eksp. Teor. Fiz. **47**, 1945 (1964) [Sov. Phys. JETP **20**, 1307 (1965)].

Translated by Eugene Yankovsky

On the theory of multiple scattering

V. I. Yurchenko*)

G. I. Budker Nuclear Physics Institute, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia

(Submitted 15 May 1998)
Zh. Éksp. Teor. Fiz. **116**, 418–435 (August 1999)

In this paper a kinetic equation is derived for the distribution function in the variable $q=2\sin(\vartheta/2)$ for the case of a scattering cross section of general form under the assumption that the region of multiple scattering (the diffusion region) is small. The limits of the kinetic equation are discussed, with no restrictions imposed on the scattering angles. It is found that the equation has a solution in the form of an integral. Finally, it is established that the solution is applicable over the entire range of angles, from 0 to 180°. © *1999 American Institute of Physics.* [S1063-7761(99)00508-9]

1. INTRODUCTION

The Moliere theory[1)] of multiple scattering[1,2] is generally accepted and used in modeling and calculating the passage of particles through matter (see, e.g., Refs. 3–6). The theory employs the small-angle approximation, $\sin\vartheta d\vartheta$ is replaced by $\vartheta d\vartheta$, and the scattering cross section has the form $\sigma(\delta)\propto\delta^{-4}$, where δ is the scattering angle. The distribution function found in Refs. 1 and 2 in the small-angle region is close to Gaussian, $f(\vartheta)\approx 2\exp(-\vartheta^2/\lambda^2)/\lambda^2$, i.e., describes the diffusion of particles in angle space due to multiple collisions with small-angle scattering. For the sake of brevity we will call the corresponding region of angles in a solution with a characteristic size λ the diffusion region. Moreover, in the solution we can specify a region known as the multiple-scattering region into which a particle lands if among the multiple sequence of collisions some were large-angle, and the limiting region of single scattering, where a particle experiences only one collision and the distribution function has the form $f(\vartheta)\propto\vartheta^{-4}$, i.e., follows the angular dependence of the cross section itself.

In this paper we will examine, for a cross section of general form and without resorting to the small-angle approximation, the kinetic equation for the distribution functions in the variable

$$q=2\sin\frac{\vartheta}{2}. \tag{1}$$

The variable q determines the momentum transfer, and the cross section can also be expressed in terms of a similar variable, $\chi=2\sin(\delta/2)$. Moreover, we have $\sin\vartheta d\vartheta=qdq$, which makes it possible to generalize the results obtained earlier in the small-angle approximation. Operating within this approach, we will obtain the solutions for the Rutherford and Mott cross sections. The accuracy of the results will be verified, in particular, by the Monte Carlo method over the entire range of angles.

2. THE KINETIC EQUATION AND ITS SOLUTION

We introduce the following notation: $\mathbf{i}$ is the vector of the direction of propagation of the initial particles (Fig. 1), $\sigma(\mathbf{n}\cdot\mathbf{n}')d\Omega/2\pi$ is the differential cross section of scattering particles with the initial direction $\mathbf{n}'$ into the solid angle $d\Omega$ along direction $\mathbf{n}$, and $f(\mathbf{n}\cdot\mathbf{i},t)d\Omega/2\pi$ is the number of particles in the solid angle $d\Omega$ after they have travelled the distance t. The starting kinetic equation is

$$\frac{\partial f(\mathbf{n}\cdot\mathbf{i},t)}{\partial t}=-Nf(\mathbf{n}\cdot\mathbf{i},t)\int\sigma(\mathbf{n}\cdot\mathbf{n}')\frac{d\Omega'}{2\pi}+N\int\sigma(\mathbf{n}\cdot\mathbf{n}')f(\mathbf{n}'\cdot\mathbf{i},t)\frac{d\Omega'}{2\pi}, \tag{2}$$

where N is the number of scattering centers per cubic centimeter and $d\Omega'$ corresponds to the direction $\mathbf{n}'$. Here all the quantities are expressed in terms of the cosines of the angles. It is important for our further discussion to note that $d\Omega'$ is the surface element of a unit sphere and that the integrals in Eq. (2) can be interpreted as integrals over this surface.

We introduce the vectors

$$\boldsymbol{\chi}=\mathbf{n}-\mathbf{n}',\quad \mathbf{q}=\mathbf{n}-\mathbf{i},\quad \mathbf{q}'=\mathbf{n}'-\mathbf{i}$$

and, in accordance with (1), replace the variables $\cos\delta$, $\cos\vartheta$, and $\cos\vartheta'$ by new variables (see Fig. 1),

$$\chi=2\sin\frac{\delta}{2}=|\mathbf{n}-\mathbf{n}'|,$$

$$q=2\sin\frac{\vartheta}{2}=|\mathbf{n}-\mathbf{i}|,$$

$$q'=2\sin\frac{\vartheta'}{2}=|\mathbf{n}'-\mathbf{i}|. \tag{3}$$

Here $f(\mathbf{n}\cdot\mathbf{i},t)\sin\vartheta d\vartheta=f(q,t)qdq$, the right-hand side of the equation contains an integral of $\sigma(\chi)[f(q')-f(q)]$, the quantities χ, q, and q' are the sides of a triangle, and $q'=|\mathbf{q}-\boldsymbol{\chi}|$. If we direct the z axis along $\mathbf{n}$, we have $d\Omega'=\chi d\chi d\varphi$, and the explicit dependence of q' on q, χ, and φ has the form

 © 1999 American Institute of Physics

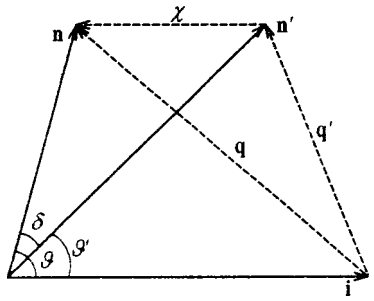

FIG. 1. Geometric representation of the variables χ, q, and q'.

$$q'^2=q^2-2(q_\perp\chi_\perp\cos\varphi+q_z\chi_z)+\chi^2,$$

$$\chi_\perp^2=\chi^2\left(1-\frac{\chi^2}{4}\right),\quad \chi_z=\frac{\chi^2}{2},$$

$$q_\perp^2=q^2\left(1-\frac{q^2}{4}\right),\quad q_z=\frac{q^2}{2}.$$

We derive a kinetic equation in which integration is carried out over a planar surface.

Two domains of integration in $\int\sigma(\chi)f(q')\chi d\chi d\varphi$ are important: the domain of small χ and the domain of small q', i.e., the neighborhoods of the maximum of the cross section and of the maximum of the distribution function. The distance between the centers of these two domains, points **n** and **i**, is q (Fig. 1). The first domain is of order $(\overline{\chi^2})^{1/2}$ in size, with $\overline{\chi^2}=(\sigma_t)^{-1}\int_0^2\chi^2\sigma(\chi)\chi d\chi$ (σt is the total cross section), and determines the diffusion of particles along the direction **n**. The second (diffusion) domain of the solution[2)] is

$$\lambda\ll 1, \tag{4}$$

in size (in this domain $f(q')\sim 1/\lambda^2\gg 1$) and describes the scattering of particles that travel at small angles to **n**; in particular, it describes large-angle scattering. We will assume that condition (4) is met. By examining specific examples we can show that if this condition is violated, energy losses by the particle travelling in matter become large and hence Eq. (2) breaks down.

Bearing all this in mind and taking $q\ll 1$, we can assume that, due to the smallness of χ, q, and q', the points represented by the vectors **i**, **n**, and **n**′ lie in plane tangent to the spherical surface at the point **n**. We introduce a cylindrical system of coordinates centered at the point **n**. The surface element is $d\Omega'=\chi d\chi d\varphi=d\boldsymbol{\chi}$, and

$$\frac{\partial f(q,t)}{\partial t}=N\int\sigma(\chi)[f(|\mathbf{q}-\boldsymbol{\chi}|,t)-f(q,t)]\frac{d\boldsymbol{\chi}}{2\pi}. \tag{5}$$

This is a kinetic equation in which the arguments are defined by (3) and the vector **q** lies in the integration plane, which we assume flat. Our approximation amounts to going over to a planar surface in Eq. (2) or, what is the same, to $q'^2=q^2-2q\chi\cos\varphi+\chi^2$. Here the small-angle approximation, which amounts to replacing χ and q with δ and ϑ, is not used and, as the further investigation will show, is superfluous. The range of the argument of $f(q)$ is assumed infinite, which is possible because we have $f(q)\ll 1$ for sufficiently large q if condition (4) is met.

By integrating (5) with respect to **q** we obtain the normalization condition, and by multiplying both sides of this equation by q^2 and integrating the result with respect to **q**, with $q^2d\mathbf{q}$ replaced by $(q'^2+2\mathbf{q}'\chi+\chi^2)d\mathbf{q}'$ in the first term on the right-hand side, we obtain $\overline{q^2}$:

$$\int_0^\infty f(q)qdq=1,\quad \overline{q^2}=\int_0^\infty f(q)q^3dq=Nt\sigma_t\overline{\chi^2}.$$

For Eq. (2) these integrals are (see Sec. 3)

$$\int_0^2 f(q)qdq=1,$$

$$\overline{q^2}=\int_0^2 f(q)q^3dq=2[1-\exp\{-Q_1\}],\quad Q_1=\frac{1}{2}Nt\sigma_t\overline{\chi^2}. \tag{6}$$

Now assume that the important domains of integration are far apart, $q>\lambda$, so that the transition to Eq. (5) is not so obvious. We estimate the error of the right-hand side when we go from Eq. (2) to Eq. (5). We assume that $f(q')$ is a smooth function in the first domain and $\sigma(\chi)$ in the second. To within terms of order χ^2, in the first domain we have

$$f(q')\approx f(q)+\chi\frac{\partial f}{\partial\chi}+\frac{1}{2}\chi^2\frac{\partial^2 f}{\partial\chi^2}.$$

After integrating over the solid angle Ω' ($d\Omega'=\chi d\chi d\varphi$), we obtain an expression for the contribution to the right-hand side of Eq. (2):

$$\frac{1}{4}N\sigma_t\overline{\chi^2}(\hat{L}+\hat{l})f(q),$$

$$\hat{L}=\frac{1}{q}\frac{\partial}{\partial q}q\frac{\partial}{\partial q},\quad \hat{l}=-\frac{1}{q}\frac{\partial}{\partial q}\frac{q^3}{4}\frac{\partial}{\partial q}.$$

We note in passing that in this new approximation (actually, the Fokker–Planck approximation; see, e.g., Ref. 7) the diffusion equation leads to the same integrals (6) as Eq. (2), but the separation into regions of multiple and single scattering in the solution, a physical result determined by the specific dependence of σ on χ, is lost.

As for the second domain, we can proceed in the same manner, i.e., expand $\sigma(\chi)=\sigma(|\mathbf{q}-\mathbf{q}'|)$ up to terms of order q'^2 and integrate over Ω' ($d\Omega'=q'dq'd\varphi$). The resulting contribution is $N\sigma(q)+\frac{1}{4}N\overline{q^2}(\hat{L}+\hat{l})\sigma(q)$.

Analogous approximations for Eq. (5) have no terms with $\hat{l}$ which yield the desired estimate. This justifies the use of Eq. (5) over the entire range of angles. First, if condition (4) is met, the particles appear in the region with $q\gg\lambda$ due to large-angle scattering. In this region the effect of diffusion (the terms with $\hat{L}$ and $\hat{l}$) can be ignored, and both equations for the number of particles in the interval dq over path t yield $f(q,t)qdq\approx Nt\sigma(q)qdq$, which corresponds to the probability of single scattering. Second, as q gets smaller, the difference between the right-hand sides decreases (due to the factor $q^2/4$ in the operator $\hat{l}$ in comparison to $\hat{L}$) and for

$q \sim \lambda$ becomes small. For instance, when $\sigma(q) \propto q^{-4}$ holds we have $\hat{l}\sigma(q) = -\frac{1}{8}q^2\hat{L}\sigma(q)$. If we allow for the fact that $f(q) \approx Nt\sigma(q)$ and $\overline{q^2} \approx Nt\sigma_t\overline{\chi^2} = 2Q_1$, the error of the right-hand side of Eq. (5) is estimated at $-NQ_1\hat{l}\sigma(q)$, which gives rise to the following error in the solution

$$\delta f(q) \sim -N\hat{l}\sigma(q)\int_0^t Q_1 dt = -\frac{1}{4}\overline{q^2}Nt\hat{l}\sigma(q).$$

Assuming that $\overline{q^2} \sim \lambda^2$, we find $(q > \lambda)$ that

$$\delta f(q) \sim \frac{\lambda^2}{4}\frac{1}{q}\frac{\partial}{\partial q}\frac{q^3}{4}\frac{\partial}{\partial q}f(q),$$

In the region where $f(q)$ decreases faster than $1/q^2$, δf is positive. In the case of the Rutherford cross section we have $f(q) \propto q^{-4}$ and $\delta f/f \sim \lambda^2/2$. The upper bound on λ is determined by the admissible error. For $\xi = \delta f/f \sim \lambda^2/2 < 5\%$ we obtain $\lambda_{\max} \sim 0.3$.

To estimate δf when $q < \lambda$, we use the normalization condition,

$$\int_0^{q_0} f(q)qdq = 1 - \int_{q_0}^{\infty} f(q)qdq \sim 1,$$

$$\int_0^{q_0} \delta f(q)qdq = -\int_{q_0}^{\infty} \delta f(q)qdq \sim \frac{\lambda^2}{4}\left[\frac{q^3}{4}\frac{\partial f}{\partial q}\right]_{q\sim q_0}.$$

In the last integral we have used the formula for δf, and $q_0 \sim \lambda$ is the value of q at which the error δf changes sign. If we assume that for q smaller than or of order λ the distribution function is close to Gaussian, $f(q) \approx 2\exp(-q^2/\lambda^2)/\lambda^2$, we find that the value of q at which $\delta f(q) = 0$ is $q_0 = \sqrt{2}\lambda$. This leads to a qualitative estimate for the diffusion region:

$$\frac{\int_0^{q_0}\delta f(q)qdq}{\int_0^{q_0} f(q)qdq} \sim \frac{\delta f}{f} \sim \frac{\lambda^2}{4}\left[\frac{q^3}{4}\frac{\partial f}{\partial q}\right]_{q=q_0} \sim -c\frac{\lambda^2}{2},$$

where $c = 2/e^2 \approx 0.3$ is a numerical factor.

As a result we conclude that for $q > \lambda$ the relative error $\xi(q)$ is of order $\lambda^2/2$ and is weakly dependent on q, and for $q < \lambda$ the relative error is of order $-0.3\lambda^2/2$. As for the limits of our approximation, we also note that for Eq. (2) $\overline{q^2} \approx 2Q_1 - Q_1^2$, i.e., Eq. (5) provides an overvalued $\overline{q^2}$ with a relative error $\sim \lambda^2/4$.

Thus, under quite reasonable restrictions, the equation is valid for all angles. With an infinite domain of definition this solution can be found by the Fourier–Bessel method.[2]. For $f(q,0) = \delta(1-\cos\vartheta) = \delta(q)/q$ we have

$$f(q,t) = \int_0^{\infty} \eta d\eta J_0(\eta q)\exp\{-Q(\eta)\},$$

$$Q(\eta) = Nt\int_0^{\infty} \sigma(\chi)\chi d\chi[1 - J_0(\eta\chi)]. \tag{7}$$

If condition (4) is met, we can retain, say, in the Rutherford cross section the dependence $\sigma(\chi) \propto \chi^{-4}$ for $\chi > 2$. This reduces the number of particles in the diffusion region because they scatter into the region $\chi > 2$, further reducing the normalization integral $I_{\text{norm}} = \int_0^2 f(q)qdq$, which equals unity if the upper limit of integration in infinite. These effects, however, are small and can be taken account by adjusting the number of particles in the diffusion region.

We note in conclusion that the small-angle approximation (the Moliere formula) can be obtained by replacing χ and q with δ and ϑ in (7). The general nature of the cross section is lost in this case.

3. THE RUTHERFORD CROSS SECTION

For the Rutherford cross section

$$\sigma_R(\chi) = \frac{2s^2\kappa(\chi)}{\chi^4},$$

where $s^2 = 4\pi e^4 z^2 Z(Z+1)(pv)^{-2}$, p is the momentum, v is the velocity of the scattered particle with charge z, and $\kappa(\chi)$ allows for the atomic form factor, which cuts off the cross section at small angles, we can use the Moliere theory since the formulas for $\sigma_R(\chi)$ and the solutions (7) coincide with their small-angle approximation. Below we will briefly discuss this theory in an interpretation that corresponds to the results of Sec. 2.

The initial parameters in this theory are the cutoff angle χ_a and $\chi_c^2 = Nts^2$. The cutoff angle[8] is defined by the relationship $\chi_a^2 = \chi_0^2(1.13 + 3.76(Zze^2/\hbar v)^2)$, where $\chi_0 = \hbar/pa = \hbar/(p \times 0.885a_0 Z^{-1/3})$, where a is the Fermi radius of the atom, and a_0 is the Bohr radius. The quantity χ_c^2 can be interpreted as the probability of a particle being scattered at an angle $\vartheta > 60°$ $(q > 1)$, since for large angles we have

$$\int_1^{\infty} f(q)qdq \approx Nt\int_1^{\infty} \sigma_R(q)qdq = \chi_c^2.$$

While the details of the dependence of σ_R on χ are not taken into account at small angles, the cutoff of the cross section is taken into account by the expression[1,2]

$$\int_0^k \kappa(\chi)\frac{d\chi}{\chi} = \ln\frac{k}{\chi_a} - \frac{1}{2},$$

with $\kappa(k) = 1$ for $k \gg \chi_a$. Such an approach is possible if

$$\lambda \gg \chi_a. \tag{8}$$

Partitioning the integral for Q into two integrals,

$$\int_0^k \cdots + \int_k^{\infty} \cdots,$$

where $\chi_a \ll k \ll \delta q$, and assuming that the interval δq of the argument in the solution is small $(\delta q \ll \lambda)$, we can limit ourselves to values $\eta < 1/\delta q$ in the Fourier expansion, i.e., cut off the upper limit in (7) at $1/\delta q \ll 1/\chi_a$. Then we have $k\eta \ll 1$ and, putting $1 - J_0(\eta\chi) \approx \eta^2\chi^2/4$ in the first integral and $\kappa(\chi) = 1$, we arrive at the approximate expression[2]

$$Q^R(\eta) = \frac{1}{4}\chi_c^2\eta^2\left[b - \ln\left(\frac{1}{4}\chi_c^2\eta^2\right)\right],$$

$$b = \ln\frac{\chi_c^2}{\chi_a'^2}, \quad \ln\chi_a' = \ln\chi_a - \frac{1}{2} + C,$$

$$C = 0.577, \quad \chi_a'^2 = 1.167\chi_a^2, \tag{9}$$

which reduces to

$$Q^R(\eta)=\frac{u^2}{4}\left(1-\frac{1}{B}\ln\frac{u^2}{4}\right),\quad u^2=B\chi_c^2\eta^2,$$

for $B-\ln B=b$.

Ignoring the second term in the parentheses, we obtain the Gaussian distribution of the particles in the diffusion region,

$$f(q,t)=2\exp(-q^2/\lambda^2)/\lambda^2$$

with the mean-square value

$$\lambda=\chi_c\sqrt{B}$$

of the quantity q. From the relation $B=\ln(\lambda^2/\chi_a'^2)$, condition (8), and the equation for B it immediately follows that[3)]

$$B\gg 1,\quad \chi_a\ll\chi_c\ll\lambda\ll 1,$$

where the chain of inequalities is augmented by condition (4). Since $\sigma_t\approx s^2/\chi_a^2$, for the average number of collisions we have $Nt\sigma_t\approx\chi_c^2/\chi_a^2\gg 1$.

A remark concerning λ is in order. The equation for B can be replaced by the equation $\lambda^2=\chi_c^2\ln(\lambda^2/\chi_a'^2)$, or [with the integral divided in the same way as in the derivation of formula (9)]

$$\lambda^2=Nt\int_0^{\alpha\lambda}\chi^2\sigma_R(\chi)\chi d\chi,\quad \alpha=\exp\{1-C\}=1.53,$$

which differs from the formula for $\overline{q^2}$ with cross section cutoff (or cutoff of the upper limit in the integral) at $\chi=\alpha\lambda$. We see that as the length t of the path of a particle and hence λ increase, more and more collisions with ever increasing collision angles become involved in the diffusion process, and the range of angles contributing to diffusion increases in proportion to λ. To a certain extent this explains the meaning of the equation for B and shows that the existing recommendations that the upper limit should be set roughly to $\hbar/pr_N$ (this value is determined by the nuclear form factor,[9] with r_N the radius of the nucleus) are correct for $\overline{q^2}$, i.e., for estimates. The use of the Moliere formulas is preferable for the size of the diffusion region and in practical cases where $\alpha\lambda\ll\hbar/pr_N$ holds.

Expanding $\exp\{-Q^R(\eta)\}$ and the solution (7) in power series in $1/B$ yields[1,2]

$$f_R(q)\approx\frac{2\exp\{-X^2\}+B^{-1}f^{(1)}(X)+B^{-2}f^{(2)}(X)+\cdots}{\lambda^2},\tag{10}$$

where $X=q/\lambda$, and

$$f^{(n)}(X)=\frac{1}{n!}\int_0^\infty u\,du J_0(Xu)\exp\left\{-\frac{u^2}{4}\right\}\left[\frac{u^2}{4}\ln\frac{u^2}{4}\right]^n$$

are universal functions (i.e., functions of X only). The formulas and tables that are commonly used in calculating the functions $f^{(1)}$ and $f^{(2)}$ can be found in Ref. 2. In the limit $q\gg\lambda$ the term with

$$f^{(1)}(X)=2\exp\{-Z\}(Z-1)[\overline{\mathrm{Ei}}(Z)-\ln Z]-2[1-2\exp\{-Z\}]$$

$(Z=X^2)$ provides the main contribution, $f^{(1)}(X)\approx 2/X^4$, and

$$f_R(q)\approx\frac{2\chi_c^2}{q^4}.\tag{11}$$

Allowing for the fact that $\int_0^\infty\xi d\xi f^{(n)}(\xi)=0$ holds for $n\geqslant 1$ (see Ref. 1), we arrive at an expression for the normalization integral:

$$I_{\mathrm{norm}}=1-\exp\left\{-\frac{4}{\lambda^2}\right\}-B^{-1}F^{(1)}\left(\frac{2}{\lambda}\right)-B^{-2}F^{(2)}(2/\lambda),$$

where $F^{(n)}(X)=\int_X^\infty\xi d\xi f^{(n)}(\xi)$, and, for $X\gg 1$, we have

$$F^{(1)}\approx X^{-2},\quad F^{(2)}\approx 2X^{-4}\ln X^2.$$

The difference between I_{norm} and unity is small (of order $\lambda^2/4B$) and can be removed by introducing a correction factor into the diffusion part of the solution.

Note that here Eqs. (9) and (10) were derived only with condition (8) taken into account; condition (4) ensured the validity of solution (7). If condition (8) is not met, the number of collisions is small, the fraction of the nonscattered particles in the solution, equal to $\exp(-Nt\sigma_t)$, is large, and the concepts of a multiple-scattering region and its size λ lose all meaning. Here the quantity $(\overline{\chi^2})^{1/2}$ becomes the characteristic small size. The solution (7) remains valid since in the small-angle region $(q\ll 1)$ the transition to a flat domain of integration is always possible, while for $q\gg(\overline{\chi^2})^{1/2}$ Eqs. (2) and (5) yield close results.

An alternative approach consists in solving Eq. (2) exactly in the form of a series:

$$f(\cos\vartheta,t)=\sum_{l=0}^\infty\frac{2l+1}{2}P_l(\cos\vartheta)\exp\{-Q_l(t)\},$$

$$Q_l(t)=\int_0^\pi Nt\sigma(\cos\delta)\sin\delta d\delta[1-P_l(\cos\delta)],\tag{12}$$

where P_l are Legendre polynomials, and $f(\cos\vartheta,0)=\delta(1-\cos\vartheta)$. [This yields to the above formulas (6)]. For the Rutherford cross section we have[2,10]

$$Q_l\approx\frac{1}{2}\chi_c^2\eta_l^2\left(\ln\frac{2}{\chi_a}+\frac{1}{2}-S_l\right),$$

$$S_l=\frac{1}{2}\int_{-1}^1\frac{1-P_l(x)}{1-x}dx=\sum_{k=1}^l\frac{1}{k},\tag{13}$$

where $\eta_l=\sqrt{l(l+1)}$.

Let us reduce the series (12) to integral form. (For the small-angle approximation this problem has been examined by Bethe.[2]). Allowing for the fact that the representation of $P_l(\cos\vartheta)$ in the form of a hypergeometric function yields a series in powers of q^2 and that for small angles $P_l(\cos\vartheta)\approx 1-\eta_l^2q^2/4$, we replace $P_l(\cos\vartheta)$ by the function $J_0(\eta_lq)$ in which the expansion of $J_0(\eta_lq)$ for small angles is the same. The resulting series can be transformed as follows:

$$\sum_{l=0}^\infty\frac{2l+1}{2}g(\eta_l)=\frac{1}{2}\sum_{l=0}^\infty\frac{1}{2}(g(\eta_{l+1})+g(\eta_l))\Delta_l,$$

where $\Delta_l = x_{l+1} - x_l$, and $x_l = \eta_l^2$. Each term in the sum on the right-hand side is an approximation of the integral of $g(x)$ in the interval (x_l, x_{l+1}) by the trapezoid rule, i.e., the series reduces to the integral

$$\frac{1}{2}\int_0^\infty g(x)dx = \int_0^\infty g(\eta)\,\eta d\eta.$$

Thus, if we replace $\cos\vartheta$ with q, $\cos\delta$ with χ, $\sqrt{l(l+1)}$ with η, and $P_l(\cos\vartheta)$ with $J_0(\eta q)$, we arrive at a general relationship between the series (12) and the integral (7). For the Rutherford cross section we have (since $S_l \approx \ln\eta_l + C + \frac{1}{6}\eta_l^{-2} + \cdots$),

$$Q_l \approx \frac{1}{2}\chi_c^2\eta_l^2\left(\ln\frac{2}{\chi_a} + \frac{1}{2} - C - \ln\eta_l\right),$$

i.e., we arrive at formula (9). This approach will be used in Sec. 4 to study the case of the Mott cross section (spin-$\frac{1}{2}$ particles).

4. THE MOTT CROSS SECTION

For the cross section

$$\sigma_M(\chi) = \sigma_R(\chi)\left(1 - \frac{1}{4}\beta^2\chi^2\right), \quad \beta = \frac{v}{c},$$

we represent $Q(\eta)$ in (7) as $Q^M(\eta) = Q^R(\eta) + \delta Q(\eta)$, where

$$\delta Q(\eta) = Nt\int_0^\infty \delta\sigma(\chi)\chi d\chi[1 - J_0(\eta\chi)],$$

$$\delta\sigma(\chi) = -\frac{1}{2}s^2\beta^2\chi^{-2}.$$

The integral diverges if $\delta\sigma(\chi)$ is not cut off at $\chi = 2$. No such problem arises for a solution in the form of a series, and instead of (13) we obtain

$$Q_l \approx \frac{1}{2}\chi_c^2\eta_l^2\left(\ln\frac{2}{\chi_a} + \frac{1}{2} - S_l\right) + \delta Q_l,$$

$$\delta Q_l = -\frac{1}{2}\chi_c^2\beta^2 S_l, \tag{14}$$

i.e., for the Mott cross section we have the series (12) with (14). Putting $S_l \approx \ln\eta_l + C$, we reduce the series to an integral (see Sec. 3). Note that the resulting value of

$$\delta Q(\eta) = -\frac{1}{2}\chi_c^2\beta^2(\ln\eta + C)$$

can be found from the relation

$$\delta Q(\eta) = Nt\lim_{\varepsilon\to 0}\left(\int_\varepsilon^2 \delta\sigma(\chi)\chi d\chi - \int_\varepsilon^\infty \delta\sigma(\chi)J_0(\eta\chi)\chi d\chi\right),$$

i.e., with only the first integral cut off. This approximation is sufficiently accurate for the case we are examining here.

After u is replaced by $\lambda\eta$ and X by q/λ, we get

$$f_M(q) = \frac{1}{\lambda^2}\int_0^\infty u du J_0(Xu)\exp\left\{-\frac{u^2}{4}\right\} \times \exp\left\{\frac{u^2}{4B}\ln\frac{u^2}{4} - \delta Q(u)\right\},$$

$$\delta Q(u) = -\frac{1}{2}\chi_c^2\beta^2\left(a + \frac{1}{2}\ln\frac{1}{4}u^2\right), \quad a = \ln\frac{2}{\lambda} + C.$$

Here we must bear in mind that in this case the correction to the Rutherford cross section is much smaller than the latter for $\chi \leqslant \lambda$ and is comparable to the latter when $\chi \sim 1$ holds, i.e., is important in the large-angle region. Hence it is enough to find only the correction to the function $f^{(1)}$ in (10), which provides the main contribution at large angles, while the variations of λ and hence of the parameter B caused by changes in the cross section can be ignored.

We expand the second exponential function in the integrand to within second-order terms in $1/B$ and keep only the first-order term in $\delta Q(u)$, ignoring the terms with products of $\delta Q(u)$ and powers of $1/B$. The result is

$$f_M(q) = f_R(q) + \frac{1}{2B}\beta^2\psi(\lambda, X), \tag{15}$$

$$\psi(\lambda, X) = \exp\{-X^2\}[2a + \ln X^2 - \overline{\mathrm{Ei}}(X^2)]. \tag{16}$$

Formula (16) follows from the fact that

$$\psi(X) = \int_0^\infty u du J_0(Xu)\exp\left\{-\frac{u^2}{4}\right\}\left(a + \frac{1}{2}\ln\frac{u^2}{4}\right).$$

The first term in the square brackets yields $2a\exp(-X^2)$. We find the derivative of the remaining integral $I(X)$ with respect to X. Allowing for the fact that

$$\frac{dJ_0(Xu)}{dX} = \frac{u}{X}\frac{dJ_0(Xu)}{du}$$

and integrating by parts, we arrive at the equation

$$\frac{d(X^2 I(X))}{d(X^2)} = \frac{f^{(1)}(X)}{2} - \exp\{-X^2\}.$$

Solving it, we obtain (16).

Note that in the diffusion region the contribution of the second term in (15), $\sim B^{-1}\times\ln(1/\lambda)$, is small, as expected. For instance, the term with $f^{(1)}$ yields a contribution of order $B^{-1}\lambda^{-2}$. The main contribution comes from the single-scattering region. In this region, in the first order in χ_c^2 we obtain

$$f_M(q) \approx 2\chi_c^2\left(1 - \frac{1}{4}\beta^2 q^2\right)\frac{1}{q^4}. \tag{17}$$

5. THE MONTE CARLO METHOD

The accuracy of the solutions (10) and (15) with (16) was verified by summing of the corresponding series [Eqs. (12)–(14)] and that of the Rutherford cross section, also by solving Eq. (2) by the Monte Carlo method with the cross section in the form[2,8] $\sigma_R(\chi) = 2s^2(\chi^2 + \chi_a^2)^{-2}$. The total cross section is $\sigma_t = (s/\chi_a)^2/(1 + \chi_a^2/4)$.

In the Monte Carlo method, to obtain the distribution function $f(q)$ with the required statistical accuracy one must accumulate a fairly large number of events in which the random angle ϑ is determined after a particle has travelled a path of length t. The direct method consists in Monte Carlo calculations of the number of collisions n on the path t by the Poisson distribution with a mean $\bar{n}=Nt\sigma_t$ and a Monte Carlo calculation in each collision of the scattering angle δ combined with a calculation of a new value of ϑ until all collisions are exhausted.

To calculate δ, the random value F of the integral of the distribution function for the quantity $\chi=2\sin(\delta/2)$ is generated:

$$F(\chi)=2\chi_a^2\left(1+\frac{\chi_a^2}{\chi_g^2}\right)\int_\chi^{\chi_g}\frac{\chi d\chi}{(\chi^2+\chi_a^2)^2},$$

$$\chi_g=2, \quad 0\leqslant F\leqslant 1. \tag{18}$$

In view of the fact that $F=F(\chi)=(1-\chi^2/\chi_g^2)/(1-\chi^2/\chi_a^2)^{-1}$, we have

$$\chi=2\sin\frac{\delta}{2}=\chi_a\sqrt{\frac{1-F}{F+\chi_a^2/\chi_g^2}}.$$

The new value of the angle ϑ can be found, say, from the expression

$$q=2\sin\frac{\vartheta}{2}=\sqrt{r_1^2+r_2^2-2r_1r_2\cos\nu},$$

where ν is the random azimuthal angle and r_1 and r_2 are determined by the value of ϑ before the collision:

$$r_1=2\cos\frac{\delta}{2}\sin\frac{\vartheta}{2}, \quad r_2=2\sin\frac{\delta}{2}\cos\frac{\vartheta}{2}.$$

Such an approach is effective when we need to find $f(q)$ in the diffusion region. However, in the single-scattering region, where the probability of finding a particle is low, it is difficult to achieve the required accuracy. To determine $f(q)$ in the entire region we can proceed as follows. We divide the domain χ into intervals (χ_{i-1},χ_i), $i=1,\ldots,N$, $\chi_0=0$, $\chi_N=2$, and divide events into mutually exclusive classes with two selection rules: the ith class has no collisions with $\chi>\chi_i$ and at least one collision with the value of χ in the ith interval. The events for which only the first rule holds belong to the first class. Here the distribution function is

$$f(q)=\sum_{i=1}^{N} W_i f_i(q),$$

where $f_i(q)$ is the distribution function for the class with number i, and W_i is the probability of events of the given class taking place. The formulas for W_i and the distributions by the number of collisions in classes are given in the Appendix. This division into classes makes it possible to obtain histograms for $f(q)$ with the occupancy of the channels in the number of events being close to uniform. As the values of the quantities

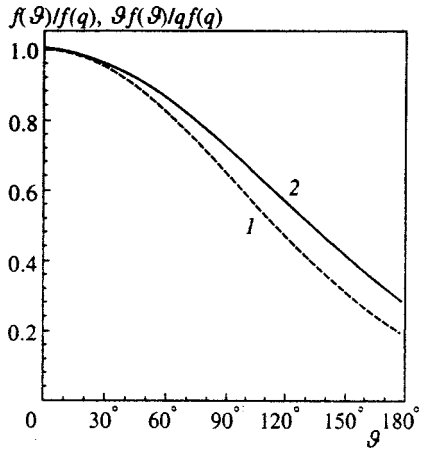

FIG. 2. Dependence of the ratios $f(\vartheta)/f(q)$ (curve *1*) and $\vartheta f(\vartheta)/qf(q)$ (curve *2*) on ϑ for the Rutherford cross section. The curves were calculated for cases 1, 2, and 3 of Table I and for these cases are indistinguishable in the figure.

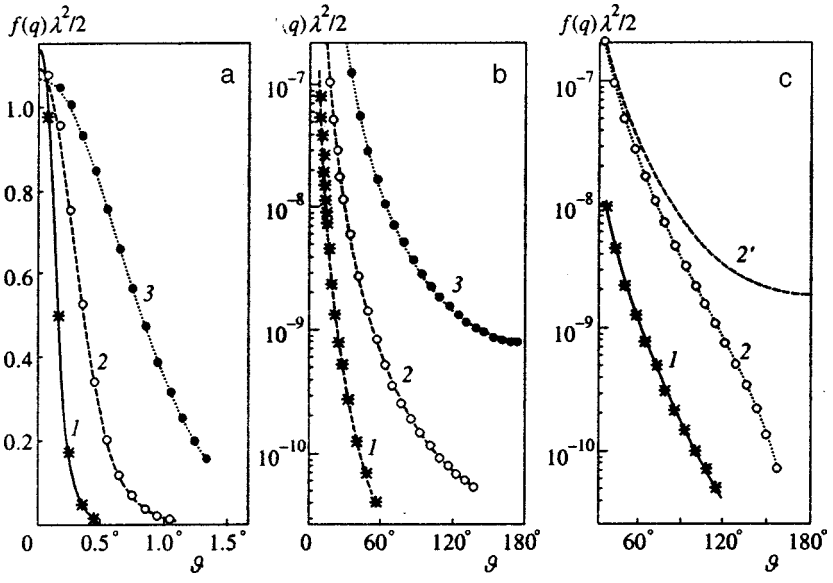

FIG. 3. The ϑ-dependence of $f(q)\lambda^2/2$: (a) and (b), the results of calculations of $f(q)$ by formula (10) for the cases 1, 2, and 3 of Table I (curves *1*, *2*, and *3*) and by the Monte Carlo method (dots); (c), the results of calculations of $f(q)$ by formulas (15) and (16) for the cases 1 and 2 of Table I (curves *1* and *2*) and by the series (12) with (14) (dots); the curve *2′* was obtained for $\psi(\lambda,X)=0$.

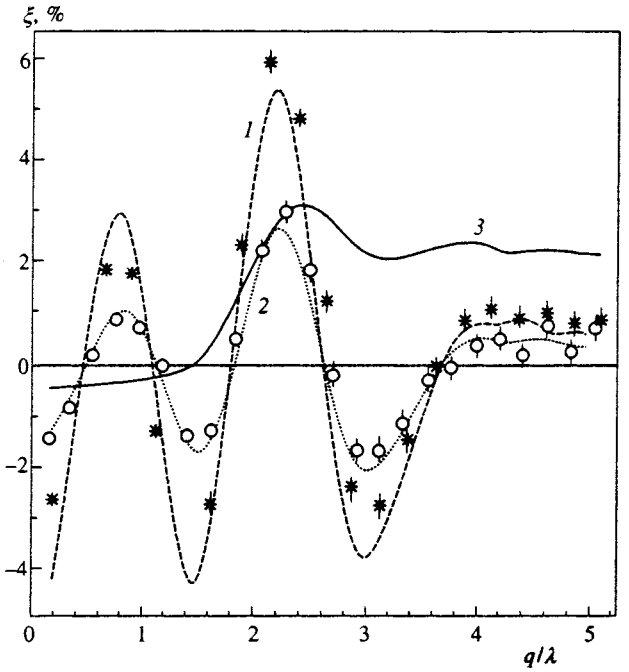

FIG. 4. The dependence of ξ on q/λ: curves *1*, *2*, and *3* represent the deviation of $f(q)$ calculated by (10) from the value calculated by the series (12) with (13) for the cases 1, 2, and 7 of Table I, and the dots indicate the deviations from the Monte Carlo results for the cases 1 and 2.

$$h=\sum_{i=1}^{N}\frac{W_i m_i}{M_i},\quad S^2=\sum_{i=1}^{N} m_i\left(\frac{W_i}{M_i}\right)^2,$$

in the channels build up (here M_i is the number of selected events of the ith class, m_i is the number of events of the ith class, and S^2 is the variance of h), the fact that another event of the ith class finds itself in a channel adds W_i/M_i to h and adds W_i^2/M_i^2 to S^2, i.e., accumulating two histograms is sufficient. The quantity m_i/M_i corresponds to $f_i(q)$. The normalization of h and S to the value of the channel yields $f(q)$ and the statistical error.

6. DISCUSSION

First we note that Eqs. (11) and (17) yield correct probabilities of infrequent collisions, $f(q)qdq \approx Nt\sigma(q)qdq$, i.e., the above solutions can be used in the entire range of angles. In Fig. 2 we compare the solution (10) with the small-angle approximation for the Rutherford cross section. The difference become appreciable when the angles are large, i.e., when, in accordance with (11), we have $f(\vartheta)/f(q)\approx(q/\vartheta)^4$ and $\vartheta f(\vartheta)/qf(q)\approx(q/\vartheta)^3$ and they reach their corresponding limits $(2/\pi)^4$ and $(2/\pi)^3$ at $\vartheta=180°$. Such comparison is meaningless for the Mott cross section, since in the small-angle representation the cross section becomes negative for $\vartheta>120°/\beta^2$.

A comparison of the solutions with more exact results is done in Figs. 3 and 4 and Tables I and II. Figures 3a and 3b clearly indicate agreement between the solution (10) and the solution obtained by the Monte Carlo method over the entire range of angles, and Fig. 3c indicates agreement between the solution (15), (16) and the series (12), (14) and also the difference between the solution (15), (16) and the solution (10) in the large-angle region.

Table I lists the data on pions scattered by gold ($Z=79$ and $A=196.97$; ρ is the density); $\xi_{\max}$, $\xi(0)$, and $\xi(2)$ are the maximum and boundary values of the relative deviation $\xi(q)=[f(q)-f_{GS}(q)]/f_{GS}(q)$ in the interval $0\leqslant q\leqslant 2$. Here f is the solution (10), and f_{GS} is the Goudsmith–Saunderson[10] solution given by Eq. (12) and (13). The nature of the ξ vs. q dependence for large values of B (curve *3* in Fig. 4) at $q\sim\lambda$ intersects the axis $\xi=0$ and the value $\xi(2)$ agree with the estimates of Sec. 2. The value $\xi_{\max}$ lands in the multiple region $q/\lambda\sim 2$. As λ decreases, $\xi(0)$ and $\xi_{\max}$ first decrease and then increase. This fact and the behavior of curves *1* and *2* in Fig. 4 can be explained by the insufficient number of terms in the expansion (10) for small values of B. In this case, in the multiple-scattering region, the error introduced by the solution (10) is estimated at $\xi(0)\sim f^{(3)}(0)/2B^3$, i.e., is of the order of the first discarded term, $f^{(3)}=5.94$. In the multiple region $q/\lambda\sim 2$ the error is larger and, since here $f^{(0)}$ is small and $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$ are of the same order, $\xi_{\max}\sim 1/B^2$. Assuming the in the single region $\xi\sim\lambda^2/2$, we find that $\xi(0)>\lambda^2/2$ for $\lambda^2<5.94/B^3$, or $\chi_c^2<5.94/B^4$, i.e., in almost all problems encountered in practical cases the solution in the single region will be more exact.

Table I shows that the approximate value of the upper bound on the size of the diffusion region is $\lambda_{\max}\sim 0.3$. Note that at $\lambda=0.3$ the normalization integral is equal to 0.998 and that the mean-square value of q in the interval $0\leqslant q\leqslant 2$ deviates from the value obtained by formula (6) by less than 1%. As λ diminishes, the accuracy of these quantities increases.

The approximations (9) and (13) have their limits, since they were derived under the assumption that $\eta, l\ll 1/\chi_a$. The

TABLE I. The Rutherford scattering cross sections (for pions) at $E_{\text{kin}}=50$ MeV and $\chi_a=2.77\times10^{-4}$.

No.	ρt, g/cm^2	λ	B	$\bar{n}$	$\xi(0)$, %	$\xi_{\max}$, %	$\xi(2)$, %
1	0.00375	3.51×10^{-3}	4.93	32.7	−5.38	5.52	
2	0.015	8.14×10^{-3}	6.61	131	−1.63	2.68	0.01
3	0.060	1.81×10^{-2}	8.21	523	−0.76	1.59	0.03
4	0.24	3.96×10^{-2}	9.77	2.09×10^{3}	−0.43	1.11	0.12
5	0.4	5.26×10^{-2}	10.34	3.49×10^{3}	−0.35	1.06	0.25
6	2	0.127	12.11	1.74×10^{4}	−0.35	1.74	1.01
7	4	0.185	12.86	3.49×10^{4}	−0.47	3.17	2.06
8	10	0.304	13.85	8.71×10^{4}	−0.93	8.63	5.24
9	40	0.640	15.34	3.49×10^{5}	−3.57	50.7	38.6

TABLE II. The Mott scattering cross sections (for electrons) at $E_{\text{kin}}=15$ MeV and $\chi_a=6.89\times10^{-4}$.

No.	ρt, g/cm^2	λ	B	$\bar{n}$	$\xi(0)$, %	$\xi_{\max}$, %	$\xi(1)$, %
1	0.00375	9.10×10^{-3}	5.01	34.8	−4.99	5.25	–
2	0.015	2.10×10^{-2}	6.68	139	−1.57	2.61	0.06
3	0.030	3.15×10^{-2}	7.49	278	−1.05	2.00	0.12
4	0.06	4.68×10^{-2}	8.28	556	−0.75	1.61	0.23
5	0.12	6.93×10^{-2}	9.06	1.11×10^{3}	−0.58	1.43	0.44
6	0.24	0.102	9.84	2.23×10^{3}	−0.50	1.51	0.90
7	0.4	0.135	10.41	3.71×10^{3}	−0.49	1.86	1.57
8	1	0.224	11.42	9.28×10^{3}	−0.66	4.04	4.04
9	2	0.328	12.17	1.86×10^{4}	−1.07	8.66	8.66
10	4	0.477	12.93	3.71×10^{4}	−1.99	16.8	16.8

series (12) must be cut off at $l<\eta_m\sim1\chi_a$, where η_m is the value of η at which $Q(\eta)$ assumes its maximum value Q_m; otherwise it diverges. If we employ the approximation (9) and an infinite upper limit, the integrals (7) are also divergent. The value of $\exp(-Q_m)$ must be small. For instance, in the diffusion region the term in the series (12) corresponding to η_m must be small in comparison to $f(0)\approx2/\lambda^2$. For the cases *1*, *2*, and *3* of Table I, $\log[(2\eta_m+1)\exp(-Q(\eta_m))/2f(0)]$ takes the values −6.1, −19, and −72, respectively. For $q<4\lambda$ the agreement between (12) with (13) and the results of Monte Carlo calculations for case 1 is still satisfactory (Fig. 4). This leads to the approximate lower limit $\lambda_{\min}/\chi_a\sim10$, but the series does not ensure sufficient accuracy in the single region, where the value of $f(q)$ is small (the upper bound on $\xi(2)$ obtained by the Monte Carlo method for this case is less than 1%).

The data on scattering in aluminum for the Mott cross sections are listed in Table II, where ξ is the relative deviation of (15) with (16) from the series (12) with (14) in the interval $0\leqslant q\leqslant1$, and yield the same results. (A comparison at $q=2$ is almost impossible in view of the very small value of $f(2)$ and the approximate nature of the solutions.)

Note that in the above examples, for $\lambda\sim\lambda_{\max}$ the amount of energy lost by a particle in matter is large, i.e., when the particle path in matter is properly limited the condition $\lambda<\lambda_{\max}$ is met.

7. CONCLUSION

Our research has shown that the starting point in deriving the theory was the idea that the size λ of the diffusion region is small, and no need for the small-angle approximation arises. For values of λ interesting from the practical viewpoint, the kinetic equation derived for a cross section of a general form with no restrictions imposed on the scattering angles can be applied in the entire range of scattering angles, and the results obtained in the small-angle approximation can be generalized by simply replacing ϑ with $q=\sin(\vartheta/2)$.

Over a broad range of values of λ interesting from the practical viewpoint, the solutions obtained for the single region are much more accurate than those obtained for the multiple and multiple-scattering regions, i.e., the series (10) for the last two regions must be refined. (Here we do not discuss the separate problem of the accuracy of χ_a, which is equally important for all the solutions. We only note that in the single region the solution is basically determined by the parameter χ_c^2.)

The work contains a general result, a kinetic equation with a solution in the form of an integral, which is a generalization of the results obtained earlier in the small-angle approximation (the Rutherford cross section), and a new result, the distribution function for spin-$\frac{1}{2}$ particles (the Mott cross section). A basis for allowing for the nuclear form factor and the absorption of particles in a nucleus has been developed. The results can be employed when allowing for multiple scattering in step-by-step passage of particles through the detector (computer simulation), when the particle path is interrupted and in each small displacement all interaction processes are taken into account.

APPENDIX

By P_i we denote the probability of χ having its value in the interval $(0,\chi_i)$, and by m_i, n_i, and $n-m_i-n_i$ the number of collisions in an event with the value of χ in the intervals $(0,\chi_{i-1})$, (χ_{i-1},χ_i), and $(\chi_i,2)$, correspondingly, with n the total number of collisions. The probability of such an event occurring is determined by a polynomial distribution and is equal to

$$\frac{n!}{m_i!n_i!(n-m_i-n_i)!}P_{i-1}^{m_i}p_i^{n_i}(1-P_i)^{n-m_i-n_i},$$

where $p_i=P_i-P_{i-1}$.

The probability of an event of the ith class occurring with the numbers of collisions n and n_i is (in this formula we put $n-m_i-n_i=0$, exclude the case with $n_i=0$, and allow for the fact that the value of n in the events obeys the Poisson distribution with the mean value $\bar{n}$)

$$\omega_i(n,n_i)=\exp\{-\bar{n}\}\frac{\bar{n}^n}{n!}\frac{n!}{(n-n_i)!n_i!}\times P_{i-1}^{n-n_i}(p_i^{n_i}-\delta_{0n_i}),\quad i\neq1,$$

where δ_{0n_i} is the Kronecker delta. Reasoning in the same manner, we arrive at a formula for the probability of an event of the first class occurring with the number of collisions n:

$$\omega_1(n)=\exp\{-\bar{n}\}\frac{1}{n!}(\bar{n}P_1)^n.$$

The total probability W_i of finding an event in the ith class can be obtained by summing the above probabilities over n and n_i. When summing over n_i, we perform the transformation $P_{i-1}^{n-n_i}p_i^{n_i}=P_i^n a^{n-n_i}b^{n_i}$ (here $a+b=1$) and, allowing for the fact that the probabilities in the binomial distribution sum to unity, obtain

$$\sum_{n_i}\omega_i(n,n_i)=\exp\{-\overline{n}\}\frac{\overline{n}^n}{n!}(P_i^n-P_{i-1}^n),\quad i\neq 1.$$

Summing over n and allowing for the fact that the probabilities in the Poisson distribution sum to unity, we get

$$W_1=\exp\{-\overline{n}(1-P_1)\},$$

$$W_i=\exp\{-\overline{n}(1-P_i)\}-\exp\{-\overline{n}(1-P_{i-1})\},\quad i\neq 1.$$

Since $1-P_i=F_i=F(\chi_i)$, where F_{χ_i} is defined in (18) at $\chi_g=2$, we finally arrive at the formula

$$W_1=\exp\{-\overline{n}F_1\},$$

$$W_i=\exp\{-\overline{n}F_i\}-\exp\{-\overline{n}F_{i-1}\},\quad i=2,\ldots,N.$$

Note that $F_N=0$, and hence $\Sigma W_i=1$, i.e., the division into classes encompasses all the events.

As shown by the formula for $\omega_1(n)$, the number of collisions in the events of the first class normalized to unity (divided by W_1) is distributed according to the Poisson law with the mean $\overline{n}(1-F_i)$. In this class, the values of χ in collisions are selected from the interval $(0,\chi_i)$, in accordance with formula (18) in which we put $\chi_g=\chi_i$. For the other classes we define the number of collisions as $n=m_i+n_i$. Passing in the formula for $\omega_i(n,n_i)$ to the variables m_i and n_i, we find that the probability of the values of m_i and n_i occurring in the ith class is proportional to the product of two factors:

$$\sim\frac{a^{m_i}}{m_i!}\left(\frac{b^{n_i}}{n_i!}-\delta_{0n_i}\right)\sim\exp\{-a\}\frac{a^{m_i}}{m_i!}\left(\exp\{-b\}\frac{b^{n_i}}{n_i!}-\exp\{-b\}\delta_{0n_i}\right),\quad a=\overline{n}P_{i-1},\quad b=\overline{n}p_i.$$

Thus, m_i obeys the Poisson distribution with the mean $\overline{n}(1-F_{i-1})$ except for the case $n_i=0$. After m_i and n_i are generated randomly the n collisions are sampled, and among these n_i collisions are sampled uniformly and at random with the value of χ occurring in the interval (χ_{i-1},χ). In the remaining m_i collisions the values of χ occur in the interval $(0,\chi_{i-1})$.

*)E-mail: yurchenko@vxinpz.inp.nsk.su

1)A brief exposition of this theory is given in Sec. 3.

2)In this paper we will assume that most collisions are small-angle. This causes diffusive broadening of the initial beam of particles in an angular region $\sim\lambda$. Note that we assume that $\chi^2\approx\delta^2$ and $q^2\approx\overline{\vartheta^2}$.

3)For $B>1$ the value of B can be found by the method of successive approximations: $B^{(n)}=b+\ln B^{(n-1)}$. For instance, for $B^{(0)}=1$ we have $B^{(3)}=b+\ln(b+\ln b)\approx b+(1+1/b)\ln b$. A possible refinement $\tilde{B}=b+1/2b^2+(1+0.958/b)\ln b$ yields, for $b>2.6$ ($B>4$), a value of B with an error smaller than 1.4×10^{-2}%, while the error introduced by the formula $B=b+\ln\tilde{B}$ is smaller than 2×10^{-3}%.

[1] G. Moliere, Z. Naturforsch. A **3a**, 78 (1948).
[2] H. A. Bethe, Phys. Rev. **89**, 256 (1953).
[3] A. D. Bukin, V. N. Ivanchenko, M. Yu. Lel'chuk *et al.*, Preprint 84-33 of the Nuclear Physics Institute, Siberian Branch of the USSR Academy of Sciences, Novosibirsk (1984).
[4] A. D. Bukin, N. A. Grozina, M. S. Dubrovin *et al.*, Preprint 94-20 of the Nuclear Physics Institute, Siberian Branch of the Russian Academy of Sciences, Novosibirsk (1994).
[5] R. Brun, F. Bruyant, N. Maire *et al.*, GEANT3, CERN preprint DD/EE/84-1, Geneva (1987).
[6] N. F. Shul'ga and S. P. Fomin, Zh. Éksp. Teor. Fiz. **113**, 58 (1998) [JETP **86**, 32 (1998)].
[7] V. G. Levich, Yu. A. Vdovin, and V. A. Myamlin, *Theoretical Physics: An Advanced Text*, Vol. 2: *Statistical Physics and Electromagnetic Processes in Matter*, Wiley, New York (1971).
[8] G. Z. Moliere, Z. Naturforsch. A **2a**, 133 (1947).
[9] B. Rossi and K. Greisen, Rev. Mod. Phys. **13**, 240 (1941).
[10] S. A. Goudsmith and J. L. Saunderson, Phys. Rev. **57**, 24 (1940); **58**, 36 (1940).

Translated by Eugene Yankovsky

Polarization effects in the interaction between light and multilevel quantum systems

K. A. Nasyrov and A. M. Shalagin*)

Institute of Automation and Electrometry, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia

(Submitted 6 January 1999)
Zh. Éksp. Teor. Fiz. **116**, 436–457 (August 1999)

We develop a theory that describes the interaction of radiation and multilevel particles on the basis of the classical description of the orientation of angular momentum. For examples of application of this theory we take the problems of stimulated Raman scattering and of parametric generation of light in four-wave mixing. We find that the efficiency of the radiation processes largely depends on the polarization of the pump waves and on the types of optical transitions (ΔJ) and to a much lesser extent on the value of the angular momentum J. We also explain some polarization phenomena observed in experiments on four-wave mixing.
© *1999 American Institute of Physics.* [S1063-7761(99)00608-3]

1. INTRODUCTION

Lately there has been an upsurge of interest in the problem of the interaction of radiation with multilevel systems. This, in particular, is true of radiation frequency conversion in resonant processes and of inversionless amplification in multilevel systems. The theoretical analysis of such problems, which is very complicated in itself, becomes even more complicated when polarization phenomena accompanying such processes come into the picture. The thing is that in studying such phenomena we cannot limit ourselves to the model of nondegenerate states of quantum systems (atoms and molecules) and we are forced to allow for degeneracy of the energy states in the orientation of angular momentum. In the case of strong fields, such problems sometimes allow an exact solution when the angular momentum J is small ($J=0,1$), but at large values of J the difficulty of finding such a solution increases immeasurably.

Fortunately, in the limit $J \gg 1$ we go over to the classical way of describing the orientation of angular momentum, which simplifies the above problems to such an extent that the representations of the model of nondegenerate states again come to the fore. More than that, in this approximation the well-known results of the model of nondegenerate states can be slightly modified and then fully used for solving problems in which the degeneracy of atomic or molecular states is important. The present paper is devoted to the substantiation of the above assertion and details examples of the successful use of the proposed method in specific problems.

The plan of the paper is as follows. In Sec. 2, to develop the results of Ref. 1, we display a method of passing to the classical description of the orientation of angular momentum in the kinetic equations for the density matrix. This passage is done via a transition in the form of the well-known Wigner transformation for translational motion. As a result, we derive equations for the elements of the density matrix in which the dependence on the magnetic quantum numbers M is transformed into a dependence on the angles specifying the orientation of angular momentum (as parameters). In other words, the equations become exactly the same as in the model of nondegenerate states. In the next order of the approximation in the parameter $1/J$, the equations acquire terms describing the dynamics of the orientation of angular momentum driven by an external perturbation.

In Secs. 3 and 4 the proposed method is used to analyze polarization effects in resonant radiative processes.

Section 3 examines the problem of amplifying a combination wave in a three-level system. We analyze the dependence of the amplification coefficient for the combination wave on the wave polarization and on the polarization of the pump wave and also on the types of atomic and molecular transitions involved in the pumping and generation of the combination wave.

Section 4 studies the problem of resonant four-wave parametric interaction. We succeed in explaining certain experimentally detected features of the polarization of the parametrically generated wave as a function of the polarizations of the pump waves. More than that, we find that the efficiency of generation depends significantly on the choice of polarization conditions. We give specific recommendations for such a choice.

2. INTERACTION BETWEEN A QUANTUM SYSTEM AND AN EXTERNAL PERTURBATION WITH THE ORIENTATION OF THE ANGULAR MOMENTUM DESCRIBED IN CLASSICAL TERMS

To pass from the quantum mechanical description of the orientation of angular momentum to the classical description, we follow the results of our earlier work.[1–3]

We take a quantum system subjected to an external perturbation $\hat{V}$. Its behavior is described by the well-known mathematical apparatus of the density matrix $\hat{\rho}$, which makes it possible to allow for various types of relaxation processes (see, e.g., Refs. 4–6). The method by which we propose to pass to the classical description of rotational states can be used only within this apparatus. The equation for the density matrix that can be used for a broad spectrum

 © 1999 American Institute of Physics

of problems has the form (in the interaction representation)

$$\frac{d\hat{\rho}}{dt}+\hat{\Gamma}(\hat{\rho})=\frac{i}{\hbar}[\hat{\rho},\hat{V}], \tag{1}$$

where relaxation processes are accounted for by the term $\hat{\Gamma}(\hat{\rho})$ (this term may be very complicated and can include integral operations), while the dynamic part is represented by the commutator $[\hat{\rho},\hat{V}]$.

In the so-called JM-representation, the elements of the density matrix $\hat{\rho}$ are $\rho(aJM|a'J'M')$. Here J is the value of the angular momentum, M is the value of the projection of the angular momentum on the quantization axis, and a represents the set the other quantum numbers, which characterize the eigenstate of the quantum system with the unperturbed Hamiltonian.

To make the formulas less cumbersome, we introduce the compact notation

$$\rho_{nn'}(M|M')=\rho(aJM|a'J'M'),$$

$$V_{nn'}(M|M')=V(aJM|a'J'M'),$$

by incorporating J into the set of quantum numbers n. In the present paper we focus on dynamic processes [the right-hand side of Eq. (1)], while in relation to relaxation process we adopt the simplest model of relaxation constants (see, e.g., Refs. 4 and 6), so that Eq. (1) written for the components of the density matrix becomes

$$\frac{d\rho_{nn'}(M|M')}{dt}+\Gamma_{nn'}[\rho_{nn'}(M|M')-\rho^0_{nn'}(M|M')]$$

$$=\frac{i}{\hbar}\sum_{n_1,M_1}[\rho_{nn_1}(M|M_1)V_{n_1n'}(M_1|M')$$

$$-V_{nn_1}(M|M_1)\rho_{n_1n'}(M_1|M')], \tag{2}$$

where the $\rho^0_{nn'}(M|M')$ are the values of the density matrix elements in the absence of the perturbation $\hat{V}$, and the relaxation constants $\Gamma_{nn'}$ are independent of the quantum numbers M and M'.

Obviously,

$$\rho_{nn'}(M|M')=\rho^*_{n'n}(M'|M),$$

$$V_{nn'}(M|M')=V^*_{n'n}(M'|M),$$

in view of the hermiticity of $\hat{\rho}$ and $\hat{V}$.

Now we assume, as a basis for passing to the classical description of the orientation of angular momentum, that the values of the quantum numbers J in Eq. (2) are large ($J,J_1,J'\gg 1$). In the classical picture, angular momentum has a well-defined direction or, in other words, its projection on an arbitrarily chosen direction is well-defined. From the viewpoint of the elements of the density matrix, $\rho_{nn'}(M|M')$, which describe the coherence between the states with angular momentum projections M and M', this means that for an arbitrarily chosen quantization axis the value of $\rho_{nn'}(M|M')$ is essentially nonzero only if there is a small difference between M and M', at least in the limit $|M-M'|\ll J$. More precisely, the effective interval $|M-M'|$ characterizing the quantum indeterminacy in the angular momentum projection must not exceed the limits specified by the characteristic angle scale in the given problem. Here it is natural to interpret the average $\overline{M}=(M+M')/2$ as the classical value of the projection of angular momentum on the chosen quantization axis.

When this requirement is met, it is convenient to introduce the new variables

$$\rho_{nn'}(\overline{M},m)=\rho_{nn'}(M|M'),$$

$$\overline{M}=(M+M)'/2, \quad m=M-M',$$

and similarly for the matrix element of the interaction and the matrix elements of other operators. In view of what has been said earlier, under classical conditions the value of $\rho_{nn'}(\overline{M},m)$ must be a rapidly decreasing function of m: the interval $|m|<m_{\text{eff}}\ll J$ where $\rho_{nn'}(\overline{M},m)$ is essentially nonzero, is the measure of indeterminacy in the magnitude of the angular momentum projection in conditions close to classical. Note that these requirements are met automatically if applied to the interaction matrix element. Indeed, when we have dipole interaction, $V_{nn'}(\overline{M},m)$ is finite only for $m=0,\pm 1\ll J$, and in the event of an interaction of arbitrary multipole order κ we have $|m|\leqslant\kappa$, so that for $J\gg\kappa$ the value of $V_{nn'}(\overline{M},m)$ is a ''sharp'' function of m, while the dependence on $\overline{M}$ is always smooth in the interval $-J<\overline{M}<J$, according to the explicit form of the coefficients of vector addition in terms of which $V_{nn'}(\overline{M},m)$ can be expressed.

These facts suggest that it may be useful to introduce a new representation for the density matrix and for $\hat{V}$ via the transformation

$$\rho_{nn'}(\overline{M},\phi)=\frac{1}{2\pi}\sum_m \exp\{im\phi\}\rho_{nn'}(\overline{M},m),$$

$$V_{nn'}(\overline{M},\phi)=\sum_m \exp\{im\phi\}V_{nn'}(\overline{M},m). \tag{3}$$

The factor $1/2\pi$ in the transformation for $\rho_{nn'}$ is introduced so that the following normalization condition holds:

$$\sum_M\int_0^{2\pi}\rho_{nn}(\overline{M},\phi)d\phi=\rho_{nn}, \tag{4}$$

where ρ_{nn} is the population of the state n. For other operators the transition to the new representation is carried out in the same way as for $\hat{V}$. The inverse of (3) (for $\rho_{nn'}$) is

$$\rho_{nn'}(\overline{M},m)=\int_0^{2\pi}\exp\{-im\phi\}\rho_{nn'}(\overline{M},\phi)d\phi, \tag{5}$$

and similarly for $V_{nn'}$.

Using (3) and (5), we can establish that the quantum mechanical average of a physical quantity A is calculated according to the rule

$$\langle A\rangle=\sum_{n,n'}\sum_{M,M'}A_{nn'}(M|M')\rho_{n'n}(M'|M)$$

$$=\sum_{n,n'}\sum_M\int A_{nn'}(\overline{M},\phi)\rho_{n'n}(\overline{M},\phi)d\phi. \tag{6}$$

Note that as for the variables referring to angular momentum, the averaging in (6) is according to a rule commonly used for classical distribution functions. This is to be expected since the transformation (3) is in direct analogy with the transition to the well-known Wigner representation for translational motion, and for this representation we have exactly the same type of averaging. On the other hand, the corresponding averaging rule indicates that the choice of the representation for the density matrix is an important stage on the way to the classical description.

Applying a transformation of type (3) to Eq. (2), we arrive at an intermediate equation:

$$\left(\frac{d}{dt}+\Gamma_{nn'}\right)\rho_{nn'}(\bar{M},\phi)-\Gamma_{nn'}\rho^0_{nn'}(\bar{M},\phi)$$
$$=\frac{i}{2\pi\hbar}\sum_{n_1,m_1,m_2}\exp\{i(m_1+m_2)\phi\}$$
$$\times\left[\rho_{nn_1}\left(\bar{M}+\frac{m_2}{2},m_1\right)V_{n_1n'}\left(\bar{M}-\frac{m_1}{2},m_2\right)\right.$$
$$\left.-V_{nn_1}\left(\bar{M}+\frac{m_2}{2},m_1\right)\rho_{n_1n'}\left(\bar{M}-\frac{m_1}{2},m_2\right)\right], \tag{7}$$

where $m_1=M-M_1$ and $m_2=M_1-M'$.

Next we discuss a situation that is close to the classical. This means that the angular momenta are large and that their projections on a specified direction have a small indeterminacy (in the case of Eq. (7), about the value $\bar{M}$). As mentioned earlier, under such conditions $\rho_{nn'}(\bar{M},m)$ is a sharp function of the second argument and a smooth function of the first argument (the matrix element $V_{nn'}(\bar{M},m)$ automatically has these properties). This fact makes it possible, first, to consider the first arguments as continuous and, second, to use the expansion in the small corrections $m_1/2$ and $m_2/2$ in these arguments, e.g.,

$$\rho_{nn_1}\left(\bar{M}+\frac{m_2}{2},m_1\right)=\rho_{nn_1}(\bar{M},m_1)$$
$$+\frac{m_2}{2}\frac{\partial}{\partial\bar{M}}\rho_{nn_1}(\bar{M},m_1)+\cdots,$$

$$V_{n_1n'}\left(\bar{M}-\frac{m_1}{2},m_2\right)=V_{n_1n'}(\bar{M},m_2)$$
$$-\frac{m_1}{2}\frac{\partial}{\partial\bar{M}}V_{n_1n'}(\bar{M},m_2)+\cdots. \tag{8}$$

We keep only the first two terms in the expansion, which means that Eq. (7) becomes

$$\left[\frac{d}{dt}+\Gamma_{nn'}\right]\rho_{nn'}(\bar{M},\phi)-\Gamma_{nn'}\rho^0_{nn'}(\bar{M},\phi)$$
$$=\frac{i}{\hbar}\sum_{n_1}[\rho_{nn_1}(\bar{M},\phi)V_{n_1n'}(\bar{M},\phi)$$
$$-V_{nn_1}(\bar{M},\phi)\rho_{n_1n'}(\bar{M},\phi)]$$
$$+\frac{1}{2}\sum_{n_1}[\{\rho_{nn_1}(\bar{M},\phi),V_{n_1n'}(\bar{M},\phi)\}$$
$$-\{V_{nn_1}(\bar{M},\phi),\rho_{n_1n'}(\bar{M},\phi)\}]. \tag{9}$$

Here $\{\cdots\}$ is what is called the Poisson bracket, which operates according to the rule

$$\{A,B\}=\frac{1}{\hbar}\left[\frac{\partial A}{\partial\bar{M}}\frac{\partial B}{\partial\phi}-\frac{\partial A}{\partial\phi}\frac{\partial B}{\partial\bar{M}}\right].$$

The first term on the right-hand side of Eq. (9) is the principal term, and the second is of the next order of smallness in the parameter $1/J\ll 1$, in powers of which the expansions in (8) are done. However, the principal term may vanish if the interaction operator causes transitions only between the sublevels of a single energy state ($V_{nn_1}(\bar{M},\phi)=\delta_{nn_1}V(\bar{M},\phi)$), so that only the second term is left on the right-hand side of Eq. (9) and the equation becomes

$$\frac{d\rho}{dt}=\frac{\partial V}{\partial\phi}\frac{\partial\rho}{\partial\mathscr{M}}-\frac{\partial V}{\partial\mathscr{M}}\frac{\partial\rho}{\partial\phi},\quad \mathscr{M}=\hbar\bar{M}.$$

(Here we have not written the label of the energy state explicitly, and the relaxation constants are assumed equal to zero.) This equation has the standard form of the classical Liouville equation and describes the classical precession of angular momentum in the field V. For this equation to have the classical form, we corrected the sign of the phase in the transformations (3) in comparison to that used in Ref. 1.

Note that the quantity $\hbar\bar{M}\equiv\mathscr{M}$ is the projection of the classical angular momentum on the z axis. Obviously, $\mathscr{M}$ and ϕ are canonically conjugate variables. This corroborates, among other things, the above interpretation of the angle ϕ in Eqs. (3).

Of course, in the general case described by Eqs. (9), the terms containing Poisson brackets are responsible for the precession of angular momentum, but this precession takes place in specific conditions, under which the external perturbation is capable of causing transitions between the energy level of the particle. This has a certain impact on the nature of such precession.

In the present paper we do not touch on the problem of analyzing the precession of angular momentum. We limit ourselves to the study of transitions between quantum states. Here in Eq. (9) we can drop the terms containing the classical Poisson bracket. The result is the equation

$$\left[\frac{d}{dt}+\Gamma_{nn'}\right]\rho_{nn'}(\bar{M},\phi)-\Gamma_{nn'}\rho^0_{nn'}(\bar{M},\phi)$$
$$=\frac{i}{\hbar}\sum_{n_1}\left[\rho_{nn_1}(\bar{M},\phi)V_{n_1n'}(\bar{M},\phi)\right.$$
$$\left.-V_{nn_1}(\bar{M},\phi)\rho_{n_1n'}(\bar{M},\phi)\right]. \quad (10)$$

Equation (10) differs from the equation of the model of nondegenerate states only in one respect: it contains the dependence on $\bar{M}$ and ϕ as parameters. This constitutes a radical simplification of the problem: if we know the solution to a problem in the model of nondegenerate states, allowing for degeneracy due to rotational motion amounts to replacing $V_{nn'}$ with $V_{nn'}(\bar{M},\phi)$ in the known solution and to averaging the desired characteristic according to rule (6).

In a classical situation the readily interpreted characteristic is the orientation in space of the angular momentum operator. This (classical) image is convenient when the external perturbation $\hat{V}$ involves a group of energy states whose angular momenta J_n have values close a certain value $\bar{J}$ within an interval $\Delta J \ll \bar{J}$. In the given system of coordinates, the orientation of angular momentum can be characterized by the azimuthal (ϕ) and polar (θ) angles. The azimuthal angle is present in the formulas, while the polar angle is related to the projection $\bar{M}$ of the angular momentum on the quantization axis and is naturally defined by the formula $\cos\theta=\bar{M}/\bar{J}$. The indeterminacy in defining $\cos\theta$, related to the spread of the values of J in the group of ''active'' levels, produces an error in specifying the direction of the angular momentum, $\sim\Delta J/\bar{J}$, which can be much smaller than the quantum indeterminacy in finding the direction of the angular momentum if $\bar{J}\gg 1$. The latter is $1/\sqrt{J}$ at best.

Thus, allowing for all the remarks we have made, we can introduce a new variable (instead of $\bar{M}$) in the elements of the density matrix and the external perturbation:

$$\rho_{nn'}(\theta,\phi)\equiv\bar{J}\rho_{nn'}(\bar{M},\phi), \quad V_{nn'}(\theta,\phi)\equiv V_{nn'}(\bar{M},\phi) \quad (11)$$

(we hope that retaining the same notation for V and ρ will not result in a misunderstanding). The normalization coefficient $\bar{J}$ was chosen so that in the new variables the average of a quantity A is calculated by the rule

$$\langle A\rangle=\sum_{n,n'}\int_0^\pi \sin\theta d\theta\int_0^{2\pi}d\phi A_{nn'}(\theta,\phi)\rho_{n'n}(\theta,\phi). \quad (12)$$

Here instead of (4) we have

$$\int_0^\pi \sin\theta d\theta\int_0^{2\pi}d\phi\rho_{nn}(\theta,\phi)=\rho_{nn}. \quad (13)$$

If we introduce (11) into Eq. (10), we get

$$\left[\frac{d}{dt}+\Gamma_{nn'}\right]\rho_{nn'}(\theta,\phi)-\Gamma_{nn'}\rho^0_{nn'}(\theta,\phi)$$
$$=\frac{i}{\hbar}\sum_{n_1}\left[\rho_{nn_1}(\theta,\phi)V_{n_1n'}(\theta,\phi)\right.$$
$$\left.-V_{nn_1}(\theta,\phi)\rho_{n_1n'}(\theta,\phi)\right]. \quad (14)$$

Here the orientation angles of the angular momentum acts as parameters.

The only question that remains is: what is the explicit form of the matrix elements $V_{nn'}(\theta,\phi)$? In each specific situation this problem can be solved fairly easily. Here we examine the case of the electric dipole interaction, where

$$V=-\boldsymbol{\mu}\cdot\mathscr{E}=-\sum_\sigma(-1)^\sigma\mathscr{E}_\sigma\mu_{-\sigma},$$

with $\mathscr{E}_\sigma$ and μ_σ the cyclic components of the electric-field and dipole-moment vectors. According to the Wigner–Eckart theorem, in the JM-representation the matrix elements for the cyclic components of the dipole-moment vector are[7]

$$(\mu_{nn'}(M|M'))_\sigma=\exp\{i\omega_{nn'}t\}(\tilde{\mu}_{nn'}(M|M'))_\sigma,$$

$$(\tilde{\mu}_{nn'}(M|M'))_\sigma=\frac{\mu_{nn'}}{\sqrt{2J+1}}C^{JM}_{J'M',1\sigma}.$$

Here $\omega_{nn'}$ is the transition frequency between states n and n', $C^{JM}_{J'M',1\sigma}$ is a vector addition coefficients, and $\mu_{nn'}$ is the reduced dipole-moment matrix element, which satisfies the property[8] $\mu_{nn'}=(-1)^{J-J'}\mu^*_{n'n}$.

Using the asymptotic behavior ($J,J'\gg 1$) of the vector addition coefficients,[7]

$$C^{JM}_{J'M',1\sigma}=\delta_{M-M',\sigma}D^1_{\sigma,J-J'}(0,\theta,0),$$

$$\cos\theta=\frac{\bar{M}}{\bar{J}}, \quad \bar{J}=\frac{J+J'}{2},$$

and applying the transformation (3) with allowance for (11), we obtain

$$(\mu_{nn'}(\theta,\phi))_\sigma=\exp\{i\omega_{nn'}t\}(\tilde{\mu}_{nn'}(\theta,\phi))_\sigma,$$

$$(\tilde{\mu}_{nn'}(\theta,\phi))_\sigma=\frac{\mu_{nn'}}{\sqrt{2\bar{J}}}D^{1*}_{\sigma,J-J'}(\phi,\theta,0), \quad (15)$$

where D is the Wigner rotation matrix:

$$D^1_{\sigma,\sigma'}(\phi,\theta,\gamma)=\exp\{-i\sigma\phi\}d^1_{\sigma,\sigma'}(\theta)\exp\{-i\sigma'\gamma\}.$$

The explicit form of this matrix and $d^1_{\sigma,\sigma'}$ can be found in Ref. 7.

Now let us assume that the external radiation field has a spectral component $\mathscr{E}_\omega$ with a frequency ω close to the frequency $\omega_{nn'}$ of the $n-n'$ transition, i.e.,

$$\mathscr{E}_\omega(t)=\mathbf{E}\exp\{-i\omega t\}+\mathbf{E}^*\exp\{i\omega t\}.$$

In this case what is known as the resonance approximation is valid, and for the matrix element of the dipole interaction

with the given spectral component of the electric field we obtain an expression in the $\theta\varphi$-representation:

$$V_{nn'}(\theta,\phi)=\exp\{-i\Omega_{nn'}t\}\tilde{V}_{nn'}(\theta,\phi),$$

$$\Omega_{nn'}=\omega-\omega_{nn'}, \quad \omega_{nn'}>0,$$

$$\tilde{V}_{nn'}(\theta,\phi)=-\sum_{\sigma}(-1)^{\sigma}E_{\sigma}(\tilde{\mu}_{nn'}(\theta,\phi))_{-\sigma}$$

$$=-\frac{\mu_{nn'}(-1)^{J'-J}}{\sqrt{2J}}\sum_{\sigma}E_{\sigma}D^{1}_{\sigma,J'-J}(\phi,\theta,0). \tag{16}$$

Here **E** is the complex-valued amplitude of the electric field of the selected wave.

In deriving the second relationship in (16) we used Eqs. (15) and the following property of D-matrices:[7] $D^{1*}_{-\sigma,-\sigma'}(\phi,\theta,\gamma)=(-1)^{\sigma'-\sigma}D^{1}_{\sigma,\sigma'}(\phi,\theta,\gamma)$.

The interaction matrix elements for the reversed transition can be calculated by using the hermiticity property:

$$V_{n'n}(\theta,\phi)=V^{*}_{nn'}(\theta,\phi). \tag{17}$$

If the external field contains spectral components that are in resonance with other transitions, the problem is solved in a similar manner.

In conclusion of this section we note that Eqs. (9) and (14) are a natural generalization of the equations obtained in Ref. 1 for the model of two-level particles in a resonant field. Earlier Ducloy,[9,10] using another method (based on the representation of coherent states), derived classical equations (i.e., classical with respect to the orientation of angular momentum) that contained only the elements of the density matrix of a two level atom that are diagonal in the energy levels. This particular case corresponds to conditions of incoherent excitation (a wide emission line). When these conditions are met, Eqs. (14) (and the corresponding equations in Ref. 1) give rise to equations that coincide with the Ducloy equations in Ref. 9 and 10.

Equations (9) and (14) can be generalized in a straightforward manner to the case where the translational motion of the particles is taken into account, with terms responsible for various type of collision incorporated into the new equations.

3. STIMULATED RAMAN SCATTERING

Below we discuss some examples of using Eqs. (14) to describe radiative processes in multilevel systems. We begin by examining the process of stimulated Raman scattering in a three-level system. The transition diagram is depicted in Fig. 1. A strong pump field in resonance with the $m-n$ transition creates the necessary conditions for amplifying the probe wave p in the $m-l$ transition. In its simplest variant, Raman scattering is a two-photon process. To avoid having to allow for one-photon process, we assume that in the absence of radiation the only level that is populated is n. Our problem is to describe the formation of polarization in the combination wave as depending on the pump wave polarization and the transition types.

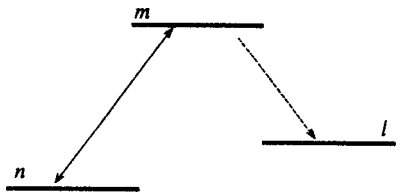

FIG. 1. The three-level scheme for Raman scattering. The pump wave is in resonance with the $n-m$ transition, and the combination wave is generated in the $m-l$ transition.

As is well known (see, e.g., Ref. 4), within the model of nondegenerate states, the expression for the polarization P^p of the medium at the frequency of the combination wave has the obvious structure

$$P^p=-iA\tilde{\mu}_{lm}\tilde{V}^{p}_{ml}|\tilde{V}_{nm}|^2\rho^{0}_{nn}. \tag{18}$$

Here A is a factor that incorporates the frequency detunings and the relaxation constants of the states involved in the process and may take into account the inhomogeneous broadening effect (this is unimportant for our further discussion since, as noted earlier, only the matrix elements that are written explicitly in (18) are to be modified), and ρ^{0}_{nn} is the population of the state n in the absence of radiation. Of course, Eq. (18) can also be easily derived from Eq. (14).

In accordance with the procedure developed in Sec. 2, generalizing Eq. (18) to the case where the atomic and molecular states are degenerate in the orientation of angular momentum in the limit $J\gg 1$ consists in introducing the dependence on the angles θ and ϕ into the quantities $\tilde{\mu}$, $\tilde{V}$, and ρ^{0}_{nn} and in integrating with respect to θ and ϕ:

$$\mathbf{P}^p=-iA\int\tilde{\mu}_{lm}(\theta,\phi)\tilde{V}^{p}_{ml}(\theta,\phi)|\tilde{V}_{nm}(\theta,\phi)|^2\rho^{0}_{nn}(\theta,\phi)$$
$$\times\sin\theta d\theta d\phi. \tag{19}$$

The explicit form of the functions $\tilde{\mu}_{lm}(\theta,\phi)$ and $\tilde{V}(\theta,\phi)$ is given in (15) and (16). We assume that in the absence of a pump wave the state n is not anisotropic, i.e., ρ^{0}_{nn} is independent of θ and ϕ. Then ρ^{0}_{nn} can be taken outside the integral sign, so that with allowance for the normalization condition (13) we get

$$\mathbf{P}^p=-i\frac{A}{4\pi}\rho^{0}_{nn}\int\tilde{\mu}_{lm}(\theta,\phi)\tilde{V}^{p}_{ml}(\theta,\phi)|\tilde{V}_{nm}(\theta,\phi)|^2$$
$$\times\sin\theta d\theta d\phi. \tag{20}$$

In steady-state conditions, the variation of the amplitude $\mathbf{E}^p$ of the probe wave as the wave propagates can be described by the equation (by virtue of the simplified Maxwell equations)

$$\left(\mathbf{n}\cdot\frac{\partial}{\partial\mathbf{r}}\right)\mathbf{E}^p=2\pi ik\mathbf{P}^p, \tag{21}$$

where $\mathbf{n}$ is the unit vector pointing along the direction of propagation of the probe wave. Here, obviously, the following condition must hold:

$$\mathbf{n}\cdot\mathbf{E}^p=0 \tag{22}$$

(the tranversality condition for a electromagnetic wave). If we substitute (20) in Eq. (21) and allow for (15) and (16), we arrive at an equation for the circular components of the electric field vector of the probe wave:

$$\left(\mathbf{n}\cdot\frac{\partial}{\partial\mathbf{r}}\right)E^p_\sigma=\frac{1}{2}\sum_{\sigma'}\alpha_{\sigma\sigma'}E^p_{\sigma'}, \tag{23}$$

where $\alpha_{\sigma\sigma'}$ is a matrix whose components are

$$\alpha_{\sigma\sigma'}=\frac{kA}{4}\frac{|\mu_{ml}|^2}{\hbar}\rho^0_{nn}\int D^{1*}_{\sigma\Delta_{lm}}(\phi,\theta,0)D^1_{\sigma'\Delta_{lm}}(\phi,\theta,0)$$
$$\times|\tilde{V}_{nm}(\theta,\phi)|^2\sin\theta d\theta d\phi, \tag{24}$$

with $\Delta_{lm}=J_l-J_m$.

The real part of this matrix determines the amplification of the combination wave, and for this reason we call it the amplification matrix for the combination wave. In the general case of an arbitrary polarization of the pump wave, the matrix $\alpha_{\sigma\sigma'}$ is not diagonal. However, using the properties of D-matrices and (24), we can see that the matrix α/A is Hermitian and hence the matrix α is a normal matrix, i.e., $\alpha\alpha^\dagger=\alpha^\dagger\alpha$. This implies[11] that there is a unitary transformation that diagonalizes this matrix. Accordingly, there must be ''normal'' combination waves, i.e., waves with polarizations that do not change in the course of propagation.

Suppose that a combination wave is propagating along the z axis, which we assume to be the polar axis; the direction of propagation of the pump way may be arbitrary. Then, in accordance with (22), the electric field vector of the combination wave has only two finite circular components ($\sigma=\pm1$), which means that for the amplification matrix $\alpha_{\sigma\sigma'}$ we can take a 2-by-2 matrix. We look for the solution of Eq. (23) in the form $E^p\propto\exp(\alpha z/2)$. For the eigenvalues $\alpha_\pm$ of the matrix α we have

$$\alpha_\pm=\frac{\alpha_{+1+1}+\alpha_{-1-1}}{2}$$
$$\pm\sqrt{\left(\frac{\alpha_{+1+1}-\alpha_{-1-1}}{2}\right)^2+\alpha_{+1-1}\alpha_{-1+1}}. \tag{25}$$

The circular components of each of two normal combination waves $\mathbf{E}^{p\pm}$ are related by formulas

$$(\alpha_+-\alpha_{+1+1})E^{p+}_{+1}=\alpha_{+1-1}E^{p+}_{-1}, \tag{26}$$

$$(\alpha_--\alpha_{-1-1})E^{p-}_{-1}=\alpha_{-1+1}E^{p-}_{+1}. \tag{27}$$

Obviously, in view of what was said above concerning the properties of the matrix α, the polarizations of the normal combination waves ($\mathbf{E}^{p+}$ and $\mathbf{E}^{p-}$) must be mutually orthogonal. As these two normal waves propagate in the medium, they are amplified differently (if the radicand in (25) is finite). If this is the case, one of the normal waves, more precisely $\mathbf{E}^{p+}$, finds itself in preferable conditions because for it the amplification coefficient is larger and hence this wave is generated spontaneously.

But if the radicand in (25) is zero, which is possible if the conditions

$$\alpha_{+1+1}=\alpha_{-1-1}, \tag{28}$$

$$\alpha_{+1-1}=0, \tag{29}$$

are met simultaneously, there is no preferable polarization in the combination wave in the amplification coefficient.

Using (24) and the explicit expressions for the Wigner D-matrices,[7] we can easily establish that the condition (28) is met if at least one of the transition of the problem is a Q-type transition ($\Delta_{nm}=0$ or $\Delta_{lm}=0$).

Now let us see when condition (29) is met. To this end, integrating in (24) with respect to ϕ and using the property[7] $d^j_{\sigma\sigma'}(\theta)=(-1)^{\sigma'-j}d^j_{-\sigma\sigma'}(\pi-\theta)$, we arrive at an expression for α_{+1-1}:

$$\alpha_{+1-1}=\frac{\pi kA}{4}\frac{|\mu_{lm}|^2|\mu_{nm}|^2}{\overline{J}\hbar^3}\rho^0_{nn}E_{+1}E^*_{-1}(-1)^{\Delta_{nl}}$$
$$\times\int d^1_{1\Delta_{lm}}(\theta)d^1_{1\Delta_{lm}}(\pi-\theta)d^1_{1\Delta_{nm}}(\theta)d^1_{1\Delta_{nm}}$$
$$\times(\pi-\theta)\sin\theta d\theta. \tag{30}$$

The integrand in (30) is always positive.

According to (30), the condition (29) is met when $E_{+1}=0$ or $E_{-1}=0$. Thus, not one of the normal combination waves has preference in the amplification coefficient if at least one transition is of the Q type and, simultaneously, one of the circular components of the pump wave is zero in the given system of coordinates. In all other cases one of the normal combination waves will be amplified to a greater extent.

Below we focus on the dependence of the polarization state of the normal combination wave $\mathbf{E}^{p+}$ with maximum amplification on the pump wave polarization and transition types.

We begin by examining in greater detail the case when one of the transitions is of the Q type. Then condition (28) is met, and from (25) and (26) we can obtain a relationship for the circular components of the vector $\mathbf{E}^{p+}$:

$$\frac{A}{|A|}|\alpha_{+1-1}|E^{p+}_{+1}=\alpha_{+1-1}E^{p+}_{-1}. \tag{31}$$

This implies $|E^{p+}_{+1}|=|E^{p+}_{-1}|$, which means that the polarization of the wave $\mathbf{E}^{p+}$ must be linear. How is this polarization oriented in relation of the pump wave polarization? We select the x and y axes so that the Cartesian components of the projection of the vector $\mathbf{E}$ of the pump wave on a plane perpendicular to the z axis are coupled by the formula

$$E_y=ibE_x, \quad b<1. \tag{32}$$

This choice of the x and y axes corresponds to a situation in which the x axis is directed along the major axis of the polarization ellipse formed by the projection of $\mathbf{E}$ onto the xy plane. Since the cyclic and Cartesian components of an arbitrary vector are related by the formulas

$$E_{+1}=-\frac{1}{\sqrt{2}}(E_x+iE_y), \quad E_{-1}=\frac{1}{\sqrt{2}}(E_x-iE_y),$$

$$E_x=\frac{1}{\sqrt{2}}(E_{-1}-E_{+1}),\quad E_y=\frac{i}{\sqrt{2}}(E_{+1}+iE_{-1}),\tag{33}$$

for the circular components of $\mathbf{E}$ we have

$$E_{+1}=-\frac{1-b}{\sqrt{2}}E_x,\quad E_{-1}=\frac{1+b}{\sqrt{2}}E_x.\tag{34}$$

Accordingly, Eq. (31) yields expressions for the Cartesian components of $\mathbf{E}^{p+}$:

$$E_x^{p+}=\frac{1}{\sqrt{2}}E_{-1}^{p+}\left(1-\frac{|A|}{A}\frac{\alpha_{+1-1}}{|\alpha_{+1-1}|}\right),$$

$$E_y^{p+}=\frac{i}{\sqrt{2}}E_{-1}^{p+}\left(1+\frac{|A|}{A}\frac{\alpha_{+1-1}}{|\alpha_{+1-1}|}\right).$$

Using Eq. (30) for α_{+1-1}, we get

$$E_x^{p+}=\frac{1}{\sqrt{2}}E_{-1}^{p+}[1+(-1)^{\Delta_{nl}}],$$

$$E_y^{p+}=\frac{i}{\sqrt{2}}E_{-1}^{p+}[1-(-1)^{\Delta_{nl}}].\tag{35}$$

Recall that we are dealing with a situation in which at least one transition is of the Q type. Here, if $\Delta_{nl}=0$ (both transitions are of the Q type), $E_y^{p+}=0$, i.e., the linearly polarized combination wave has the maximum amplification coefficient, and its field vector $\mathbf{E}^{p+}$ is directed along the major axis of the polarization ellipse of the pump wave. But if $|\Delta_{nl}=1|$, then $E_x^{p+}=0$, which means that $\mathbf{E}^{p+}$ is orthogonal to the major axis of the polarization ellipse of the pump wave.

Now we turn to the case where $|\Delta_{lm}|=|\Delta_{nm}|=1$. Formula (26), which links the different circular components of the combination wave $\mathbf{E}^{p+}$, becomes

$$[|E_{-1}|^2-|E_{+1}|^2+\sqrt{(|E_{-1}|^2-|E_{+1}|^2)^2+|\tfrac{2}{5}E_{+1}E_{-1}^*|^2}]E_{+1}^{p+}=\tfrac{2}{5}E_{+1}E_{-1}^*E_{-1}^{p+},\tag{36}$$

where $\Delta_{lm}=\Delta_{nm}=\pm1$. When $\Delta_{lm}=-\Delta_{nm}=\pm1$ holds, E_{-1} and E_{+1} in Eq. (36) must be interchanged.

We again select the orientation of the x and y axes so that the Cartesian components of the vector $\mathbf{E}$ of the pump wave are related via (32) but the circular components are still related via (34). In this system of equations, formula (36) yields an equation that connects the Cartesian components of $\mathbf{E}^{p+}$:

$$E_y^{p+}=ib'E_x^{p+},\tag{37}$$

where the ellipticity parameter b' for the combination wave is expressed in terms of b as follows:

$$b'=\frac{10b-1+b^2+\sqrt{1+98b^2+b^4}}{10b+1-b^2+\sqrt{1+98b^2+b^4}},\quad \text{at } \Delta_{lm}=\Delta_{nm}=\pm1,\tag{38}$$

TABLE I. Polarization of the combination wave with maximum amplification when the polarization of the pump wave is linear (indicated above the table).

↑

J_n-J_m \ J_1-J_m	−1	0	1
−1	↑	→	↑
0	→	↑	→
1	↑	→	↑

$$b'=\frac{-10b-1+b^2+\sqrt{1+98b^2+b^4}}{-10b+1-b^2+\sqrt{1+98b^2+b^4}},\quad \text{at } \Delta_{lm}=-\Delta_{nm}=\pm1.\tag{39}$$

Thus, we see that for the specified transitions the polarization of the normal combination wave with maximum amplification proves to be elliptic if the polarization of the pump wave is also elliptic.

We examine the limit $b\ll1$, i.e., the case in which the polarization of the pump wave is close to linear. Equations (38) and (39) then yield

$$b'=5b\quad \text{at } \Delta_{lm}=\Delta_{nm}=\pm1,$$

$$b'=-5b\quad \text{at } \Delta_{lm}=-\Delta_{nm}=\pm1.\tag{40}$$

In both cases the polarization of the combination wave is close to linear and is oriented in the same as the polarization of the pump wave, but the ellipticity parameter of the combination wave is five times larger than that of the pump wave. More than that, from (40) it follows that the electric vector of the combination wave rotates in the same direction as that of the pump wave if $\Delta_{lm}=\Delta_{nm}=\pm1$ and in the opposite direction if $\Delta_{lm}=-\Delta_{nm}=\pm1$.

In the other limit, $b\to1$, i.e., when the polarization of the pump wave is close to circular, Eqs. (38) and (39) yield

$$b'\to1,\quad \text{at } \Delta_{lm}=\Delta_{nm}=\pm1,$$

$$b'\to-1,\quad \text{at } \Delta_{lm}=-\Delta_{nm}=\pm1.$$

This means that the polarization of the normal combination wave with maximum amplification is also close to circular and that the electric vector of the combination wave rotates in the same direction as that of the pump wave if $\Delta_{lm}=\Delta_{nm}=\pm1$ and in the opposite direction if $\Delta_{lm}=-\Delta_{nm}=\pm1$.

In the particular case where the pump wave has simple polarizations (linear or circular), the normal modes of the combination wave propagating collinearly are waves of the same polarization or of orthogonal polarization. Tables I and II list the polarizations of the normal mode of the combination wave with maximum amplification according to the polarization of the pump wave and the type of transition. In other words, the data listed in Tables I and II indicate the polarization of the wave generated spontaneously.

Although our results were obtained with the orientation of angular momentum ($J\gg1$) described in classical terms, qualitatively they are valid for small values of J. For instance, it is known that for the three-level system

TABLE II. Polarization of the combination wave with maximum amplification when the polarization of the pump wave is circular (indicated above the table). The notation arb. means that the polarization of the combination wave can be arbitrary.

		↻	
J_n-J_m \ J_1-J_m	−1	0	1
−1	↻	arb.	↻
0	arb.	arb.	arb.
1	↻	arb.	↻

$1s_3-2p_7-1s_4$ in Ne, which was used by Rittner *et al.*[12] to design a Raman laser [the linearly polarized pump wave was in resonance with the $1s_3-2p_7$ transition ($\Delta_{nm}=1$) and the combination wave was generated in the $2p_7-1s_4$ transition ($\Delta_{lm}=0$)], the amplification coefficient for the combination wave whose polarization is orthogonal with respect to the polarization of the pump wave is larger than for the combination wave whose polarization coincides with that of the pump wave. This result agrees with our result obtained for the specified types of transition.

4. FOUR-WAVE PARAMETRIC MIXING

Let us examine the problem of parametric generation of a wave resulting from the action of three pump waves in a four-level quantum system. The possible patterns of a four-level system in which the wave is generated in the transition to the ground state are depicted in Fig. 2. Here ω_a, ω_b, and ω_c are the frequencies of the four pump waves and ω_d is the frequency of the generated wave.

Our problem amounts to establishing the dependence of the polarization of the parametric wave on the polarizations of the pump waves and types of transition and to finding the polarization conditions for the pump waves from the standpoint of the efficiency of generation of a parametric wave.

Within the model of nondegenerate states, we have the following expression (the approximation of low-intensity pump waves; see Ref. 13) for the polarization of the medium at the frequency ω_d of the generated wave, which can also be easily obtained from Eq. (14):

$$P^d=-iA\tilde{\mu}_{14}\tilde{V}^c_{43}\tilde{V}^b_{32}\tilde{V}^a_{21}, \tag{41}$$

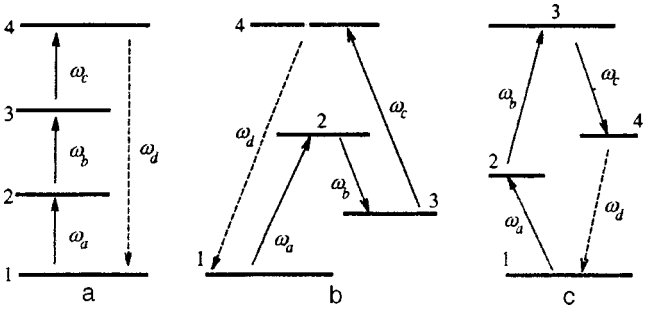

FIG. 2. Several four-level transition schemes used for parametric generation with four-wave mixing.

where $\tilde{\mu}_{14}$ is the matrix element of the dipole moment in the 1–4 transition, $\tilde{V}_{nm}$ is the matrix element of the potential of the interaction with the pump wave in resonance with the $n-m$ transition ($n,m=1,2,3,4$), and A is a proportionality factors, which incorporates the relaxation constants, the inhomogeneous broadening effect, the conditions for resonance between frequencies of the pump waves and the transitions in the quantum system, and the populations of the states. For the problem under discussion the specific form of the factor A is unimportant.

In accordance with the general approach formulated above, to generalize Eq. (41) to the case of states that are degenerate in the orientations of angular momentum, we must assign the dependence on the angles θ and ϕ to the dipole moment and the matrix elements of the interaction potential and then integrate over the various orientations of angular momentum:

$$\mathbf{P}^d=-iA\int \tilde{\mu}_{14}(\theta,\phi)\tilde{V}^c_{43}(\theta,\phi)\tilde{V}^b_{32}(\theta,\phi)\tilde{V}^a_{21}(\theta,\phi)\times\sin\theta d\theta d\phi. \tag{42}$$

The equation for the generation of the wave d has the form [cf. Eq. (21)]

$$\frac{\partial\mathbf{E}^d}{\partial z}=2\pi ik\mathbf{P}^d. \tag{43}$$

Using the expressions (15) and (16) and integrating in (42) with respect to the angle ϕ, we arrive at a relationship for the circular components of the polarization vector of the medium (the transition diagram in Fig. 2a with $\omega_{21},\omega_{32},\omega_{43}>0$):

$$P^d_\sigma=\sum_{\sigma_1\sigma_2\sigma_3}T_{\sigma\sigma_1\sigma_2\sigma_3}E^a_{\sigma_1}E^b_{\sigma_2}E^c_{\sigma_3}, \tag{44}$$

$$T_{\sigma\sigma_1\sigma_2\sigma_3}=iA(-1)^\sigma\mu_{14}\mu_{43}\mu_{32}\mu_{21}\frac{\pi}{2\bar{J}^2}\delta_{\sigma,\sigma_1+\sigma_2+\sigma_3}\times\int_0^\pi d^1_{\sigma\Delta_{14}}(\theta)d^1_{-\sigma_1\Delta_{21}}(\theta)d^1_{-\sigma_2\Delta_{32}}(\theta)d^1_{-\sigma_3\Delta_{43}}\times(\theta)\sin\theta d\theta,$$

where $\Delta_{nm}=J_n-J_m$ and $\Delta_{21}+\Delta_{32}+\Delta_{43}+\Delta_{14}=0$.

For the transition diagram in Fig. 2b ($\omega_{21}>0$, $\omega_{32}<0$, and $\omega_{43}>0$), the polarization of the medium has the form

$$P^d_\sigma=\sum_{\sigma_1\sigma_2\sigma_3}T_{\sigma\sigma_1\sigma_2\sigma_3}E^a_{\sigma_1}(-1)^{\sigma_2}(E^b_{-\sigma_2})^*E^c_{\sigma_3}, \tag{45}$$

while for the transition diagram in Fig. 2c ($\omega_{21}>0$, $\omega_{32}>0$, and $\omega_{43}<0$) we have

$$P^d_\sigma=\sum_{\sigma_1\sigma_2\sigma_3}T_{\sigma\sigma_1\sigma_2\sigma_3}E^a_{\sigma_1}E^b_{\sigma_2}(-1)^{\sigma_3}(E^c_{-\sigma_3})^* \tag{46}$$

with the same tensor $T_{\sigma\sigma_1\sigma_2\sigma_3}$ as in Eq. (44).

Equation (44) clearly shows that the tensor $T_{\sigma\sigma_1\sigma_2\sigma_3}$ depends significantly on the types of transitions considered

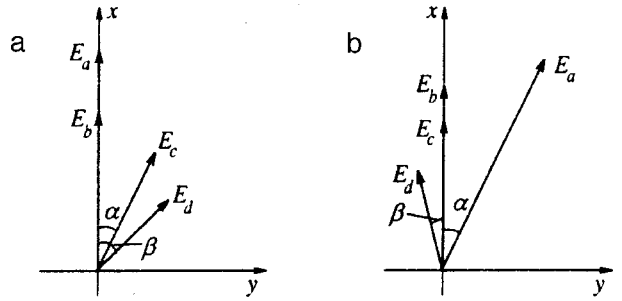

FIG. 3. Orientations of the electric field vectors of the pump waves and the parametrically generated wave: (a) the electric field vector of wave c is rotated through an angle α in the xy plane, and (b) the electric field vector of wave a is rotated through α. The angle β characterizes the orientation of the electric field vector of the generated wave.

here, and its final calculation is possibly only after these types have been specified. Note that this tensor has the following symmetry property:

$$T_{\sigma\sigma_1\sigma_2\sigma_3}=T_{-\sigma-\sigma_1-\sigma_2-\sigma_3}. \tag{47}$$

Because of the presence of the Kronecker delta in (44), many of its elements are zeros. To establish the full form of this tensor, it often suffices to calculate only three elements, T_{111-1}, T_{11-11}, and T_{1-111}, and then use the symmetry property to calculate the three remaining nonzero elements.

If pump waves with circular polarizations are used, the form of the tensor $T_{\sigma\sigma_1\sigma_2\sigma_3}$ imposes certain restrictions on the indices of the circular components of the polarizations of these waves. In particular, for the transition diagram of Fig. 2a, a parametric wave cannot be generated when all the pump waves have the same circular polarizations. The polarization of one of the pump waves must be orthogonal to the polarizations of the other two pump waves. On the other hand, for the transition diagram of Fig. 2b the generation of a parametric wave is allowed when all the pump waves have same circular polarizations and forbidden when the circular polarization of the pump waves in resonances with the 1–2 and 3–4 transitions are the same and the polarization of the wave in resonance with the 3–2 transition is orthogonal to those two.

As an example, we focus on the particular case involving the transitions $\Delta_{21}=-1$, $\Delta_{32}=1$, $\Delta_{43}=0$, and $\Delta_{14}=0$ and simple polarizations of the pump waves. Suppose that all three pump waves propagate in the same direction z and are linearly polarized, with the electric vectors of two of these waves oriented along the same axis (the x axis) and the electric vector of the third wave, e.g., the wave c, is rotated in relation to the x axis in the xy plane through an angle α (Fig. 3a). The electric field of wave c can be written

$$\mathbf{E}^c=\mathbf{n}_xE_x^c+\mathbf{n}_yE_y^c, \quad E_x^c=E^c\cos\alpha, \quad E_y^c=E^c\sin\alpha, \tag{48}$$

where $\mathbf{n}_x$ and $\mathbf{n}_y$ are the unit vectors along the x and y axes, and E_x^c and E_y^c are the components of the electric field of the wave c with the same phase. Below we are interested in the state of polarization of the parametric wave as a function of the angle α.

Before we go any further, we should note that if Eq. (44) is valid the superposition principle operates in each pump wave separately. This means that we can consider the contributions into the polarization $\mathbf{P}^d$ of the medium made degenerate by virtue of the fields E_x^c and E_y^c separately. We denote the polarization of the medium generated by the field component E_x^c by $\mathbf{P}^{d\|}$ and that generated by the field component E_y^c by $\mathbf{P}^{d\perp}$.

Since we are dealing with pump waves that are linearly polarized, it is convenient to calculate the components of the tensor T in Cartesian coordinates, using Eqs. (44) and the relationship (33) between Cartesian and circular components. Direct calculations under the specified conditions and with $\Delta_{21}=-1$, $\Delta_{32}=1$, $\Delta_{43}=0$, and $\Delta_{14}=0$ yield

$$\begin{aligned} &P_x^{d\|}=\tfrac{2}{15}BE_x^c, \quad P_y^{d\|}=0,\\ &P_x^{d\perp}=0, \quad P_y^{d\perp}=\tfrac{4}{15}BE_y^c, \end{aligned} \tag{49}$$

where B is a factor depending, in particular, on the amplitudes of the other pump waves.

Equation (43) implies that $\mathbf{E}^d\propto\mathbf{P}^d$, so that the pump waves are linearly polarized, and so is the generated wave. Equations (49) readily lead to a relationship between the angle β of the orientation of the polarization plane of the wave d and the angle α:

$$\tan\beta=\frac{P_y^{d\perp}}{P_x^{d\|}}=2\frac{E_y^c}{E_x^c}=2\tan\alpha. \tag{50}$$

In particular, when the angle α is small, the polarization plane of the parametrically generated wave proves to be rotated though an angle that is twice the given angle for wave c.

Now let us see what happens when, in contrast to the previous case, the polarization planes of waves b and c pass through the x axis and the polarization plane of wave a is rotated through the angle α in the xy plane (Fig. 2b), i.e., the field of wave a has the form

$$\mathbf{E}^a=\mathbf{n}_xE_x^a+\mathbf{n}_yE_y^a, \quad E_x^a=E^a\cos\alpha, \quad E_y^a=E^a\sin\alpha.$$

For this transition pattern the Cartesian components of the polarization vector of the medium at the frequency ω_d are

$$\begin{aligned} &P_x^{d\|}=\tfrac{2}{15}BE_x^a, \quad P_y^{d\|}=0,\\ &P_x^{d\perp}=0, \quad P_y^{d\perp}=-\tfrac{1}{15}BE_y^a. \end{aligned} \tag{51}$$

It is clear that the generated wave d has linear polarization rotated through the angle β with respect to the x axis in the xy plane. This angle is related to α as follows:

$$\tan\beta=\frac{P_y^d}{P_x^{d\|}}=-\frac{1}{2}\frac{E_y^a}{E_x^a}=-\frac{1}{2}\tan\alpha.$$

Thus, the polarization plane of wave d is found to be rotated in the opposite direction in comparison to wave a and through a smaller angle.

We have examined the case where all the pump waves are linearly polarized. A similar analysis can be carried out

when one of the waves, c or a, is elliptically polarized. For instance, let us examine the case where the electric field of wave c is $\mathbf{E}^c=\mathbf{n}_x E_x^c+i\mathbf{n}_y E_y^c$.

The polarization of such a way is elliptic and the ratio of the axes of the ellipse is $b^c=|E_y^c/E_x^c|$. Generalization to the case of the results (49) is obvious:

$$P_x^{d\parallel}=\tfrac{2}{15}BE_x^c, \quad P_y^{d\perp}=i\tfrac{4}{15}BE_y^c,$$

and in turn $E_x^d\propto P_x^{d\parallel}$ and $E_y^d\propto P_y^{d\perp}$. This implies that the field of wave d is related to the field of wave c by the formula $\mathbf{E}^d=Q(\mathbf{n}_x E_x^c+2i\mathbf{n}_y E_y^c)$, where Q is a proportionality factor. The ellipticity parameters of the generated wave is $b^d=|E_y^d/E_x^d|=2b^c$.

Thus, the presence of ellipticity in the polarization of the pump wave c generates ellipticity in the polarization of the parametrically generated wave, but with a larger ellipticity parameter.

If in the situation where the polarizations of waves b and c are linear and directed along the x axis and the polarization of wave a is elliptic we reason in the same way as in the above discussion, we can easily find that the generated wave is elliptically polarized with an ellipticity parameter related to the ellipticity parameter of wave a according to the formula $b^d=b^a/2$. We see that in this case the ellipticity parameter of wave d is smaller than that of wave a.

Note that the differences in the polarizations of the generated wave, which depend on the polarization of which wave, c or a, differs from linear, are fully determined by the differences in the transition types ($\Delta_{12}=1$, $\Delta_{34}=0$) and do not depend on the type of the four-level transition pattern.

The particular case of the transition types in the four-level system considered here corresponds to the conditions of the experiment conducted by Apolonskii *et al.*,[14] who studied the generation of radiation in four-wave mixing in vapors of sodium dimers (Na_2). The arrangement of levels and the transition patter in their experiment corresponded to a double Lambda scheme with the following values of angular momenta: $J_1=49$, $J_2=48$, $J_3=49$, and $J_4=49$. Such large values of J_i fully agree with the conditions for a classical description of the orientation of angular momentum. The results of Ref. 14 were obtained for coinciding linear polarizations and the three pump aves propagating in the same direction. Here the generated fourth wave had the same polarization. In a special measurement, Apolonskii *et al.*[14] carried out a rotation of the polarization plane of one pump wave (the wave c in the present notation) though a certain angle and found that the polarization plane of the generated wave rotates in the same direction but through a larger angle. Apolonskii *et al.*[14] gave no explanation of this effect. It can be explained quite convincingly on the basis of the description, developed in the present paper, of parametric generation with four-wave mixing. Figure 4 presents a quantitative comparison of the experimental data taken from Ref. 14 and the theoretical results obtained by formula (50) for the dependence of the angle of rotation of the polarization plane of the generated wave (angle β) on the angle α between the planes of polarization of wave c and of the other two pump waves. We see that the agreement is satisfactory, which

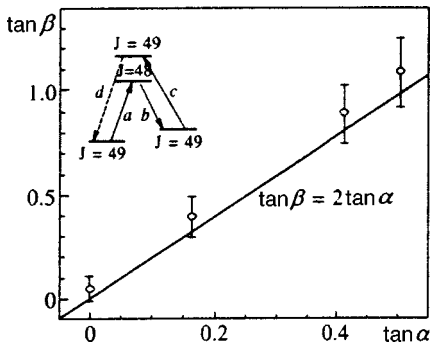

FIG. 4. Comparison of the theoretical curve and the experimental results of measurements of the dependence of the angle β specifying the orientation of the electric field vector of the generated wave on the corresponding angle α for wave c.

serves, in particular, as an additional argument in favor of the assertion that in the experiments of Apolonskii *et al.*[14] parametric generation was indeed achieved with four-wave mixing.

The key issue in the problem of converting the radiation frequency by using resonant radiative processes in gaseous media is the question of conversion efficiency. It is seen that the correct choice of the polarization conditions plays an important role here. For different transition types, different relationships between the polarizations of the pump waves are optimal. In this sense, four-level transition schemes can be divided into three classes. The first class consists of the transition schemes in which all transitions are of the Q type (all $\Delta=0$); the second class consists of transition schemes in which two transitions are of the Q type and the remaining transitions are of the P and R types; and the third class consists of transitions schemes in which there are no Q-type transitions (all $|\Delta|=1$).

A simple analysis based on (44) shows that if in a four-level scheme at least one transition has $\Delta=0$ (this means that there is at least one more transition with $\Delta=0$), it is advisable to use linear polarizations in all four pump waves. Here the gain in the intensity of the generated waves is 9/4 times that in the case of circular polarizations with the same pump-wave intensities. The use of the same linear polarization of the pump waves is the optimal choice for the first class of four-wave schemes. The generated wave will have the same polarization. The situation where the linear polarizations of two pump waves in resonance with the transitions with equal values of $|\Delta|$ coincide and the polarization of the remaining pump wave is orthogonal to those polarizations is the optimal choice for the second class of four-level schemes. In this case the polarization of the generated wave will coincide with the polarization of the latter pump wave. Note that in the experiment conducted by Apolonskii *et al.*[14] the situation where the polarizations of all the pump waves coincide is not the optimal choice. To optimize the process of generation of a parametric wave, the polarizations of the pump waves a and b should be linear and the same, while the polarization of wave c should be orthogonal to the first two. The generated wave will have the same polarization as that of wave c, and its intensity, according to (49), will be

four times higher than in the case where all three pump waves have coinciding linear polarizations.

The situation where all three waves are circularly polarized is the optimal choice for the third class of four-level schemes. Here the gain in the intensity of the generated waves will amount to the same factor of 9/4 in comparison to the case of linear polarizations of the pump waves. However, here we must be careful in selecting the correct relation between the polarizations of the waves. For instance, if we take the scheme in Fig. 2a, the circular polarizations of two pump waves in resonance with the transitions with equal values of $|\Delta|$ must coincide and the polarization of the remaining pump wave must be orthogonal to those polarizations. The generated wave will have a polarization coinciding with the polarization of the first two pump waves.

Strictly speaking, the conclusions drawn about the optimal polarization conditions are true only in the approximation of low-intensity pump waves. In condition where higher-order nonlinear effects become important, the conclusions may change, especially if we allow for possible relaxation processes in the system of magnetic sublevels (collisions that reorient the angular momentum). This general case merits a separate investigation.

5. CONCLUSION

In our study we have stressed the simplicity and productivity in solving problems associated with the interaction of quantum objects with an external perturbation when the orientation of angular momentum can be described by the use of classical tools. In particular, in dealing with the problem of the resonant interaction of radiation and a multilevel system we succeeded in formulating a procedure by which the well-known solutions of many problems obtained in the model of nondegenerate states can easily be generalized to the case of degeneracy in the orientation of angular momentum. Of course, the specific examples studied in this paper to first order in the intensities (Raman scattering in a three-level system) or in the amplitudes (four-wave mixing in a four-level system) of the pump fields can be analyzed in the common quantum mechanical approach, but this means one has to deal with complicated formulas involving the use of $3J$-, $6J$-, and $9J$-symbols, which implies knowing the appropriate mathematical tools that make it possible to operate with these mathematical objects (see, e.g., the paper of Lam and Abrams,[15] who examined the problem of the four-wave interaction with a two-level system in the JM-representation, and Refs. 16 and 17, where this problem was solved using the irreducible-tensor representation). Our approach yields results without resorting to complicated mathematical tools, since everything reduces to integrating simple expressions containing trigonometric functions. More than that, within this approach there will be only a slight increase in difficulties if we are forced to incorporate strong fields with arbitrary (elliptic) polarizations into the picture, while the ordinary quantum mechanical approach in solving such problems may not lead to an analytic solution.

We also found that to ensure maximum efficiency in radiation frequency conversion we must properly select the polarizations of the pump waves.

The authors are grateful to A. A. Apolonskiĭ for providing experimental data and for discussing the results. The work was made possible by a grant from the Russian Fund for Fundamental Research (Grant No. 98-02-17924).

*)E-mail: shalagin@iae.nsk.su

[1] K. A. Nasyrov and A. M. Shalagin, Zh. Éksp. Teor. Fiz. **81**, 1649 (1981) [Sov. Phys. JETP **54**, 877 (1981)].
[2] K. A. Nasyrov and A. M. Shalagin, Zh. Éksp. Teor. Fiz. **83**, 1685 (1982) [Sov. Phys. JETP **56**, 973 (1982)].
[3] K. A. Nasyrov and A. M. Shalagin, Kvant. Elektron. (Moscow) **9**, 1997 (1982) [Sov. J. Quantum Electron. **12**, 1298 (1982)].
[4] S. G. Rautian, G. I. Smirnov, and A. M. Shalagin, *Nonlinear Resonances in the Spectra of Atoms and Molecules* [in Russian], Nauka, Novosibirsk (1979); S. G. Rautian and A. M. Shalagin, *Kinetic Problems of Non-linear Spectroscopy*, North-Holland, Amsterdam (1991).
[5] *Nonlinear Spectroscopy*, N. Bloembergen (Ed.), Societa Italiana di Fisica (1977).
[6] V. S. Letokhov and V. P. Chebotaev, *Nonlinear Laser Spectroscopy*, Springer, Berlin (1977).
[7] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskiĭ, *Quantum Theory of Angular Momentum*, World Scientific, Singapore (1987).
[8] I. I. Sobelman, *Atomic Spectra and Radiative Transitions*, Springer, Berlin (1979).
[9] M. Ducloy, J. Phys. (Paris) **36**, 927 (1975).
[10] M. Ducloy, J. Phys. B **9**, 357 (1976).
[11] G. Korn and T. Korn, *Mathematical Handbook*, 2nd ed., McGraw-Hill, New York (1968).
[12] K. Rittner, A. Hope, T. Muller-Wirts, and B. Wellegehausen, IEEE J. Quantum Electron. **28**, 342 (1992).
[13] S. A. Akhmanov and R. V. Khokhlov, *Problems in Nonlinear Optics*, Gordon & Breach, New York (1973).
[14] A. Apolonskii, S. Balushev, U. Hinze, E. Tiemann, and B. Wellegehausen, Appl. Phys. B: Lasers Opt. **64**, 435 (1997).
[15] J. F. Lam and R. L. Abrams, Phys. Rev. A **26**, 1539 (1982).
[16] M. A. Yuratich and D. C. Hanna, J. Phys. B **9**, 729 (1976).
[17] G. G. Adonts and D. G. Akopyan, J. Phys. B **18**, 3407 (1985).

Translated by Eugene Yankovsky

Ultranarrow beams of electromagnetic radiation in media with a Kerr nonlinearity

V. E. Semenov, N. N. Rozanov,*) and N. V. Vysotina

Laser Physics Research Institute, All-Russian Scientific Center ''S. I. Vavilov State Optics Institute,'' 199034 St. Petersburg, Russia

(Submitted 19 January 1999)
Zh. Éksp. Teor. Fiz. **116**, 458–468 (August 1999)

The vector structure of a self-channeling electromagnetic field is determined by solving the complete system of Maxwell's equations in a transparent medium with a Kerr nonlinearity. Self-channelling with an asymmetric angular distribution of the field occurs at powers several times the critical self-focusing power. As the power is increased, a universal (self-similar) field structure develops in which only the scales change as the power is varied. Self-channelling with a channel width much smaller than the (linear) wavelength of the light, i.e., a ''needle of light'' with an extreme concentration of radiant power, is found to occur.
© *1999 American Institute of Physics.* [S1063-7761(99)00708-8]

1. INTRODUCTION

Extremely high concentrations of radiant energy can be achieved by reducing the duration of laser pulses (femto- and attosecond range) and by transverse compression of the beam (micron and submicron transverse dimensions). The latter process can occur in transparent media with a self-focusing type of nonlinearity in the refractive index.

The basic phenomenon of self-focusing of radiation was predicted by Askaryan in 1962[1] and was subsequently studied actively both experimentally and theoretically.[2–6] The main theoretical studies are based on a slowly varying amplitude (the quasioptical approximation), which assumes, in particular, that the beam is much wider than the wavelength of the light. In this approximation the beam is compressed to a point focus in a Kerr (with a cubic nonlinearity) medium when the power exceeds a critical value, so that the question of the final stage of self-focusing near a nonlinear focus is still open and can only be answered using a complete system of nonlinear Maxwell's equations. Several attempts have been made to analyze this considerably more complicated system for a real (two transverse dimensions) geometry, but only for a series of special cases.[6] Here an important restriction has been that beams with a strictly axisymmetric intensity distribution have been considered.[6] Even in the quasioptical approximation, beams of this sort carry a power considerably beyond the critical value, which usually indicates that they are unstable.[7] The instability of such beam against small perturbations has been demonstrated.[8] Another factor which limits the nonlinear focusing of radiation is the possible deviation of the nonlinearity of the medium from a Kerr nonlinearity or even the destruction of the medium. However, when the saturation intensity for the nonlinearity is high enough, this factor is less important than the ''nonparaxiality'' of ultranarrow beams (with widths comparable to or less than the wavelength of the light).

Thus, the question of the final stage of the self-focusing of radiation and possibility of forming ultranarrow beams (needles of light) in a medium with a nonlinearity close to a Kerr nonlinearity remains open. The task of this paper is to analyze this question when the assumption of strict axisymmetry of the radiant intensity distribution breaks down and its more complicated polarization structure is taken into account.

In Sec. 2 we describe the basic initial equations for the problem in the case of quasimonochromatic radiation which is examined throughout this article. Here the central point is the introduction of a vector potential similar to the Hertz vector with a definite gauge. As far as we know, this approach has not been applied before to nonlinear optics problems. Since the electric and magnetic field strengths are expressed in terms of the vector potential, this makes it possible to account for the vector character of the electromagnetic field in a simpler way. Another important circumstance for the following discussion is a rigorous treatment of the radiation power as the integral of the longitudinal component of the Poynting vector, averaged over the period of the optical oscillations.

In Sec. 3 the vector nature of the problem is further simplified by choosing a particular solution for which the vector potential contains only one nonzero cartesian component. This is justified by a numerical simulation illustrating the unimportance of mixing of the components of the vector potential. In the quasioptical limit, here we obtain the well known solution corresponding to linearly polarized radiation. It may be expected that other solutions will be characterized by a higher power (for the same propagation constant).

In Sec. 4, an approximate asymptotic solution of the nonlinear wave equation is introduced for the case of a beam power substantially exceeding the critical (threshold) self-focusing limit. The starting point is the assumption (confirmed by subsequent numerical calculations) that the light is predominantly linearly polarized and the field structure is approximate axisymmetric. Here a discrete set of stationary self-channelling regimes is determined; these differ in their topological charge (an azimuthal index characterizing the change in phase on going around the beam axis) and radial index (the number of zeroes in the radial function). Of these

 © 1999 American Institute of Physics

regimes, the fundamental regime, which corresponds to a nodeless radial function and the lowest power (for a given propagation constant), stands out and is all that is considered afterward. The complete vector structure of the field is constructed for it and found to be invariant (within changes in scale lengths) for arbitrarily high powers.

The next two sections are devoted to a numerical analysis of the problem. In Sec. 5 the iteration procedure for numerically solving the nonlinear Maxwell's equations is described. The results of the calculations are given in Sec. 6. The Conclusion contains a general discussion and the results.

2. BASIC EQUATIONS

We shall consider monochromatic radiation (at frequency ω) with electric fields $\tilde{\mathbf{E}}$ and magnetic fields $\tilde{\mathbf{H}}$ of the form

$$\tilde{\mathbf{E}}=\frac{1}{2}[\mathbf{E}(\mathbf{r})\exp(-i\omega t)+\text{c.c.}],$$

$$\tilde{\mathbf{H}}=\frac{1}{2}[\mathbf{H}(\mathbf{r})\exp(-i\omega t)+\text{c.c.}]. \tag{1}$$

In general, in a nonlinear medium fields also develop at frequency multiples (3ω, etc.). Their amplitudes, however, are negligible when (as assumed) the synchronism conditions are not satisfied. Then the nonlinear component of the electric induction can be written in the form[6,9]

$$\mathbf{D}_{nl}=A(\mathbf{E}\cdot\mathbf{E}^*)\mathbf{E}+\frac{B}{2}(\mathbf{E}\cdot\mathbf{E})\mathbf{E}^*$$

$$=\left(A+\frac{B}{2}\right)(\mathbf{E}\cdot\mathbf{E}^*)\mathbf{E}+\frac{B}{2}[(\mathbf{E}\cdot\mathbf{E})\mathbf{E}^*-(\mathbf{E}\cdot\mathbf{E}^*)\mathbf{E}]. \tag{2}$$

The relationship of the coefficients A and B in Eq. (2) depends on the mechanism of the nonlinearity. For a striction nonlinearity in gases and liquids we have $B=0$ and the term on the right hand side of Eq. (2) with the square brackets vanishes. As will shown later (Secs. 3–5), in the case to be examined below, where the polarization of the radiation is close to linear, this term is significantly smaller than the preceding one. Then the medium is isotropic and a scalar nonlinear permittivity can be introduced, with $\mathbf{D}=\varepsilon\mathbf{E}$, where

$$\varepsilon=\varepsilon_0+\varepsilon_2|\mathbf{E}|^2. \tag{3}$$

For self-focusing media, the nonlinearity coefficient satisfies $\varepsilon_2>0$. The magnetic permeability of the medium is assumed constant, with $\mu=1$.

We introduce the vector potential $\mathbf{\Pi}$ using the relation

$$\mathbf{H}=-\frac{i\omega}{c}\operatorname{curl}\mathbf{\Pi}. \tag{4}$$

The first Maxwell equation ($\operatorname{curl}\mathbf{E}-(i\omega/c)\mathbf{H}=0$) and the definition (4) imply

$$\mathbf{E}=-\nabla\phi+\frac{\omega^2}{c^2}\mathbf{\Pi}, \tag{5}$$

where ϕ is an arbitrary function (scalar potential). Substituting Eq. (4) in the second Maxwell equation ($\operatorname{curl}\mathbf{H}+(i\omega/c)\mathbf{D}=0$) and subjecting the scalar potential ϕ to the condition

$$\phi=-\frac{1}{\varepsilon}\operatorname{div}\mathbf{\Pi}, \tag{6}$$

we obtain using Eq. (5) an equation for the vector $\mathbf{\Pi}$,

$$\Delta\mathbf{\Pi}+\frac{\varepsilon\omega^2}{c^2}\mathbf{\Pi}-\nabla\ln\varepsilon\operatorname{div}\mathbf{\Pi}=0. \tag{7}$$

Based on Eqs. (5) and (6), the electric field strength $\mathbf{E}$ can be expressed in terms of the potential as follows:

$$\mathbf{E}=\frac{1}{\varepsilon}\operatorname{curl}\operatorname{curl}\mathbf{\Pi}. \tag{8}$$

We shall seek ''homogeneous beams'' with a field that falls off exponentially with distance from the beam axis and retain their form along the axis of propagation

$$\mathbf{\Pi}(x,y,z)=\mathbf{\Pi}(x,y)\exp(i\gamma z),$$

where γ is the real propagation constant. Here Eq. (7) for the components of the vector $\mathbf{\Pi}(x,y)$ is written in the form

$$\Delta_\perp\Pi_x-\gamma^2\Pi_x+\frac{\varepsilon\omega^2}{c^2}\Pi_x-\frac{\partial(\ln\varepsilon)}{\partial x}\times\left(\frac{\partial\Pi_x}{\partial x}+\frac{\partial\Pi_y}{\partial y}+i\gamma\Pi_z\right)=0,$$

$$\Delta_\perp\Pi_y-\gamma^2\Pi_y+\frac{\varepsilon\omega^2}{c^2}\Pi_y-\frac{\partial(\ln\varepsilon)}{\partial y}\times\left(\frac{\partial\Pi_x}{\partial x}+\frac{\partial\Pi_y}{\partial y}+i\gamma\Pi_z\right)=0, \tag{9}$$

$$\Delta_\perp\Pi_z-\gamma^2\Pi_z+\frac{\varepsilon\omega^2}{c^2}\Pi_z=0,$$

where $\Delta_\perp$ is the transverse Laplacian operator.

The power density S_z of the radiation is defined as the average of the z-component of the Poynting vector over the optical period,

$$S_z=\frac{c}{8\pi}\operatorname{Re}\mathbf{E}^*\times\mathbf{H}_z=\frac{c}{8\pi}\operatorname{Re}(E_x^*H_y-E_y^*H_x), \tag{10}$$

and the power P as the integral of the power density over the transverse cross section S,

$$P=\frac{c}{8\pi}\operatorname{Re}\left(\int_S(E_x^*H_y-E_y^*H_x)dS\right). \tag{11}$$

It is convenient to use the normalized variables

$$x'=kx,\quad y'=ky,\quad z'=kz,\quad \gamma'=\gamma/k,\quad k=\sqrt{\varepsilon_0}\,\omega/c,$$

$$\mathbf{E}'=\mathbf{E}/E_n,\quad \mathbf{H}'=\mathbf{H}/\sqrt{\varepsilon_0}E_n,\quad \mathbf{\Pi}'=\mathbf{\Pi}\omega^2/c^2E_n, \tag{12}$$

$$\varepsilon'=1+\varepsilon_{nl},\quad \varepsilon_{nl}=\frac{\varepsilon_2}{\varepsilon_0}|E_n|^2|\mathbf{E}'|^2.$$

Here we choose the normalization coefficient E_n so that the nonlinearity parameter is $C_{nl}=(\varepsilon_2/\varepsilon_0)|E_n|^2=1$. Then the final solutions have the form

$$\Delta'_\perp \Pi'_x - \gamma'^2 \Pi'_x + \varepsilon' \Pi'_x - \frac{\partial(\ln \varepsilon')}{\partial x'} \times \left(\frac{\partial \Pi'_x}{\partial x'} + \frac{\partial \Pi'_y}{\partial y'} + i\gamma' \Pi'_z \right) = 0,$$

$$\Delta'_\perp \Pi'_y - \gamma'^2 \Pi'_y + \varepsilon' \Pi'_y - \frac{\partial(\ln \varepsilon')}{\partial y'} \times \left(\frac{\partial \Pi'_x}{\partial x'} + \frac{\partial \Pi'_y}{\partial y'} + i\gamma' \Pi'_z \right) = 0, \tag{13}$$

$$\Delta'_\perp \Pi'_z - \gamma'^2 \Pi'_z + \varepsilon' \Pi'_z = 0,$$

with

$$\mathbf{E}' = \frac{1}{\varepsilon'} \mathrm{curl}' \, \mathrm{curl}' \{ \mathbf{\Pi}' \exp(i\gamma' z') \},$$

$$\mathbf{H}' = -i \, \mathrm{curl}' \{ \mathbf{\Pi}' \exp(i\gamma' z') \}, \tag{14}$$

and

$$\varepsilon' = 1 + |\mathbf{E}'|^2. \tag{15}$$

In the following we work with the normalized system of Eqs. (13)–(15) and omit the primes.

3. CHOICE OF POTENTIAL

Since the general solution of Maxwell's equations is described by only two independent functions,[10] the component Π_z can be set equal to zero. The field described by the component Π_x corresponds to a linearly polarized beam with its electric field oriented along the x axis in the case when the transverse dimensions of the beam are much greater than the wavelength. To relate a Maxwellian spatial soliton to a quasioptical one, we shall consider the case where the vector $\mathbf{\Pi}$ has only one nonzero component, e.g., $\mathbf{\Pi}=(\Pi(x,y),0,0)$. This can be done because, as shown by calculations, for solitonlike field structures the terms in Eq. (13) containing $\partial(\ln\varepsilon)/\partial x$ and $\partial(\ln\varepsilon)/\partial y$ are small and do not qualitatively change the solution over a wide range of energies. Then Eq. (13) becomes the scalar equation

$$\Delta_\perp \Pi - \gamma^2 \Pi + \varepsilon \Pi = 0, \tag{16}$$

and the field strengths are given in terms of Π by

$$\mathbf{E} = \frac{1}{\varepsilon} \left[\mathbf{i} \left(\gamma^2 \Pi - \frac{\partial^2 \Pi}{\partial y^2} \right) + \mathbf{j} \frac{\partial^2 \Pi}{\partial x \, \partial y} + \mathbf{k} \left(i\gamma \frac{\partial \Pi}{\partial x} \right) \right],$$

$$\mathbf{H} = \mathbf{j}\gamma \Pi + \mathbf{k} i \frac{\partial \Pi}{\partial y}. \tag{17}$$

If the potential Π is an even function in x and y, then the nonlinear permittivity ε and the x-component of the electric field strength are also even functions in x and y, E_y is odd in x and y, and E_z is odd in x and even in y. In addition, if the potential Π is real, then the components E_x and E_y are also real, while E_z is purely imaginary. These properties will be satisfied below. In this case, the expression in the square brackets in Eq. (2) has the form

$$(\mathbf{E}\cdot\mathbf{E})\mathbf{E}^* - (\mathbf{E}\cdot\mathbf{E}^*)\mathbf{E} = 2[\mathbf{i}E_z E_z E_x + \mathbf{j}E_z E_z E_y - \mathbf{k}(E_x E_x + E_y E_y) E_z]. \tag{18}$$

Note that on the right hand side of Eq. (18) the longitudinal component E_z of the electric field shows up as a factor. In the quasioptical limit ($\gamma \to 1$, power close to the critical self-focusing power), this component is small:

$$|E_z/E_x|^2 \sim (\lambda/w)^2 \ll 1$$

(λ is the wavelength of the light and w is the beam width). As calculations show (Sec. 5), as the propagation constant γ increases, the amplitude of the longitudinal component increases, but its relative fraction stabilizes at a low level, even in the case of extremely high powers. This justifies neglecting the anisotropy in the nonlinearity and introducing a scalar nonlinear permittivity.

We introduce the normalized power $P'=P/P_0$, where P_0 is the critical self-focusing power in the quasistatic approximation,[6]

$$P_0 = 11.7 \frac{c\varepsilon_0^{3/2}}{8\pi k^2 \varepsilon_2}. \tag{19}$$

Substituting Eqs. (19) and (17) in Eq. (11), we obtain an expression for the normalized power,

$$P' = \frac{\gamma}{11.7} \mathrm{Re} \left(\int_S \frac{1}{\varepsilon} \left(\gamma^2 \Pi^* - \frac{\partial^2 \Pi^*}{\partial y^2} \right) \Pi \, dS \right). \tag{20}$$

4. APPROXIMATE ASYMPTOTIC SOLUTION

We shall construct the asymptotic solution of these equations for high powers (compared to the critical self-focusing power) or for the limit $\gamma \gg 1$. Numerical calculations (Sec. 6) show that, for the chosen polarization, the electric field has predominantly an x-component, with the potential distribution being close to axisymmetric. In this case, Eq. (17) implies that

$$\mathbf{E} \approx E_x \mathbf{i} \approx \frac{\gamma^2}{\varepsilon} \Pi \mathbf{i},$$

so that, given Eq. (15), we have

$$\varepsilon = 1 + \frac{\gamma^4}{\varepsilon^2} |\Pi|^2. \tag{21}$$

At high intensities, we can neglect unity on the right hand side of Eq. (21). Then

$$\varepsilon = \gamma^{4/3} |\Pi|^{2/3}, \tag{22}$$

and Eq. (16) takes the form

$$\frac{1}{\gamma^2} \Delta_\perp \Pi - \Pi + \left| \frac{\Pi}{\gamma} \right|^{2/3} \Pi = 0. \tag{23}$$

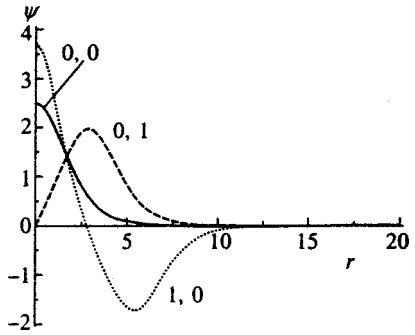

FIG. 1. Radial functions for the ''optical needle'' regime. The pairs of labels (n,m) on the radial profiles denote the radial and (n) and azimuthal (m) indices.

We eliminate the large parameter γ from Eq. (23) and introduce new scales for the transverse coordinates $\mathbf{r}_\perp$ and potential Π:

$$\tilde{\mathbf{r}}_\perp = \gamma \mathbf{r}_\perp, \quad \tilde{\Pi} = \Pi/\gamma. \tag{24}$$

Then we arrive at the following equation:

$$\tilde{\Delta}_\perp \tilde{\Pi} - \tilde{\Pi} + |\tilde{\Pi}|^{2/3}\tilde{\Pi} = 0. \tag{25}$$

Note that the form of the nonlinearity (a power of 2/3 in Eq. (25) instead of the power of 2 in the ordinary quasioptical approximation) indicates effective saturation of the nonlinearity (nonparaxiality effect).

We shall seek localized solutions (modes) of this equation that fall off sufficiently rapidly with distance from the beam axis. In general, Eq. (25) is valid only near the beam axis, where the field is large. At the periphery, we should use the linear equation (with a dielectric permittivity $\varepsilon \approx 1$) and match its solution (expressed in terms of cylindrical functions) with the solution of Eq. (25). However, this procedure leads to a correction of higher order in the small parameter γ^{-1}, so that in the rest of this section we shall examine the solution of Eq. (25) over the entire region of the transverse coordinates.

We seek a solution of Eq. (25) in polar coordinates (r,φ) of the form

$$\tilde{\Pi} = \Psi(r)\exp(im\varphi), \quad m = 0, \pm 1, \pm 2, \ldots \tag{26}$$

Now the real radial functions Ψ are determined by the equation

$$\frac{d^2\Psi}{dr^2} + \frac{1}{r}\frac{d\Psi}{dr} - \left(1 + \frac{m^2}{r^2}\right)\Psi + \Psi^{5/3} = 0 \tag{27}$$

with the boundary conditions

$$\Psi \sim r^{|m|} \ (r \to 0), \quad \Psi \to 0 \ (r \to \infty). \tag{28}$$

The modes are labeled by two integral indices: $n = 0,1,2,\ldots$ (radial index, equal to the number of zeroes in the function for finite r, $0 < r < \infty$) and $m = 0, \pm 1, \pm 2, \ldots$ (azimuthal index). Some of the first radial functions are plotted in Fig. 1. As the indices increase, the maximum amplitude and power of the corresponding beam increase. The power

$$P_{nm} = \gamma C_{nm} P_0,$$

where P_0 is the critical self-focusing power (19). The coefficients C_{nm} for the first modes, computed from a modified Eq. (20) including a contribution to the permittivity from only the x-component of the electric field strength, are

$$C_{0,0} = 3{,}9, \quad C_{1,0} = 23{,}6, \quad C_{2,0} = 59{,}0,$$

$$C_{3,0} = 110{,}2, \quad C_{4,0} = 177{,}2,$$

and

$$C_{0,1} = 13{,}3, \quad C_{1,1} = 42{,}4, \quad C_{0,2} = 24{,}4, \quad C_{1,2} = 63{,}0.$$

The lowest power (for a given value of γ) corresponds to the fundamental mode $n = m = 0$, which, accordingly, is the most stable. This type of self-channelling will be considered in the following calculations.

In the present approximation the fundamental mode is the axisymmetric part of the E_x component, $E_x^{(0)} \simeq \gamma \tilde{\Pi}^{1/3}$, while the other components of the field, E_y and E_z, and the asymmetric correction to the fundamental component, $E_x^{(1)}$, are expressed in terms of $\tilde{\Pi}$ in accordance with Eq. (17):

$$E_z = \frac{i\gamma}{\tilde{\varepsilon}}\frac{d\tilde{\Pi}}{dr}\cos\varphi, \quad E_y = \frac{\gamma}{2\tilde{\varepsilon}}Q(r)\sin(2\varphi),$$

$$E_x^{(1)} = \frac{\gamma}{2\tilde{\varepsilon}}[R(r) - Q(r)\cos(2\varphi)],$$

$$Q(r) = \frac{d^2\tilde{\Pi}}{dr^2} - \frac{1}{r}\frac{d\tilde{\Pi}}{dr} = -\frac{2}{r}\frac{d\tilde{\Pi}}{dr} + \tilde{\Pi} - \tilde{\Pi}^{5/3},$$

$$R(r) = \tilde{\Pi} - \tilde{\Pi}^{5/3}, \quad \tilde{\varepsilon} = \tilde{\Pi}^{2/3}. \tag{29}$$

As noted above, the components E_x and E_y of the electric field are real, while E_z is purely imaginary. According to Eq. (29), all the field components have a simple angular dependence. Given the bell shape of $\tilde{\Pi}(r) = \Psi_{0,0}(r)$, it is easy to confirm that the component iE_z has two extrema (a maximum and a minimum), while E_y has four extrema (two maxima and two minima). When the contribution of the components E_y and, especially, E_z to the permittivity is taken into account, a recalculation of the radiation power according to Eq. (20) yields a reduction in the coefficients C_{nm}. (For example, then $C_{0,0} = 3$.) The beam width satisfies $w \sim \gamma^{-1}$, so that for sufficiently high powers, arbitrarily narrow stationary beams (needles of light) are formed. The possibility of obtaining light beams narrower than the wavelength in a linear medium can be explained qualitatively by the fact that, for strong fields, the effective wavelength in a Kerr medium decreases owing to an increase in the nonlinear refractive index. The maximum (axial) power density $S_z^{(0)}$ and axial intensity $I^{(0)} = E_x^2(r=0)$ increase rapidly with the total power: $S_z^{(0)} \sim \gamma^3$ and $I^{(0)} \sim \gamma^2$. Note that all the field components (E_x, E_y, E_z) are of the same order in γ (first), so that with increasing γ, only the scale of the field structure varies.

5. NUMERICAL ALGORITHM

We shall solve the system of Eqs. (13)–(15) for a given power (20) iteratively. First, using the distribution $\varepsilon_{N-1}(r,\varphi)$ calculated in the previous iteration, where r,φ are the polar coordinates and $N=1,2,\ldots$ is the iteration number, we find the distribution $\mathbf{\Pi}_N(r,\varphi)$ from the following equation that corresponds to Eq. (13):

$$\Delta_\perp \Pi_{xN} - \gamma_N^2 \Pi_{xN} + \varepsilon_{N-1}\Pi_{xN} - \frac{\partial(\ln \varepsilon_{N-1})}{\partial x}\left(\frac{\partial \Pi_{xN}}{\partial x} + \frac{\partial \Pi_{yN}}{\partial y}\right) = 0,$$

$$\Delta_\perp \Pi_{yN} - \gamma_N^2 \Pi_{yN} + \varepsilon_{N-1}\Pi_{yN} - \frac{\partial(\ln \varepsilon_{N-1})}{\partial y}\left(\frac{\partial \Pi_{xN}}{\partial x} + \frac{\partial \Pi_{yN}}{\partial y}\right) = 0. \quad (30)$$

As an initial distribution $\varepsilon(r,\varphi)$ we chose the distribution for a soliton with the same energy in a medium with the saturated nonlinearity, calculated in the parabolic approximation. Equation (30) for a given $\varepsilon_{N-1}(r,\varphi)$ determines the discrete mode spectrum of an inhomogeneous (in r and φ) dielectric waveguide. As noted above, the terms that ''intermingle'' the components Π_x and Π_y of the vector are extremely small; hence, the eigenvalues γ_N^2 of the system of Eqs. (30) are essentially doubly degenerate. The degeneracy is easily removed by the terms $(\partial(\ln\varepsilon)/\partial x)(\partial\Pi_x/\partial x)$ and $(\partial(\ln\varepsilon)/\partial y)\times(\partial\Pi_y/\partial y)$ in the case where the distribution ε_{N-1} is not axisymmetric. In the first iteration, the modes are strictly doubly degenerate and we have the possibility of choosing the polarization state (e.g., to set $\Pi_y=0$). From the solutions of Eq. (30) we choose the fundamental mode, in accordance with Sec. 4, as the nodeless distribution $|\Pi_N(r,\varphi)|$ (from the pair of modes with the maximum γ_N) and with a polarization close to that of the preceding iteration. In the numerical algorithm for solving Eqs. (30), we represent the components of the vector $\mathbf{\Pi}$ in the form of an expansion in eigenvalues of the Sturn–Liouville problem for the linear wave equation inside some auxiliary cylinder.

Using Eqs. (14) and (15) we find the nonlinear permittivity ε_N from the values of $\mathbf{\Pi}_N$ calculated from Eq. (30).

This approach is easily generalized to the case of other forms of nonlinearity (e.g., including saturation).

6. COMPUTATIONAL RESULTS

One of the most important characteristics of self-channelling is the dependence of the radiation power on the propagation constant γ. Figure 2 shows a plot of this relationship with the power P normalized to the critical self-focusing power (19) and γ normalized to the wave number k of the radiation in a linear medium (12). The quasioptical approximation corresponds to the interval $\gamma-1\ll 1$, where the power is near critical. As γ increases, the power rises and, when the terms $\partial(\ln\varepsilon)/\partial x$ and $\partial(\ln\varepsilon)/\partial y$ are neglected, the dependence becomes linear for large γ in agreement with the analytic description of Sec. 4.

Figure 3 shows effective beam width w as a function of the propagation constant. The width was defined as

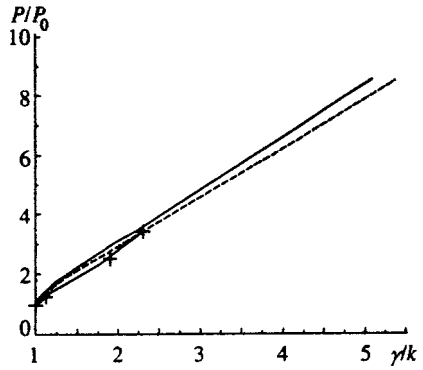

FIG. 2. Radiation power as a function of the propagation constant γ: the smooth curve is a calculation neglecting the terms containing $\partial(\ln\varepsilon)/\partial x$ and $\partial(\ln\varepsilon)/\partial y$; in the dashed curve the terms including $\partial(\ln\varepsilon)/\partial x$ and $\partial(\ln\varepsilon)/\partial y$ are included in first order perturbation theory; and, the points are the solution with full inclusion of the terms containing $\partial(\ln\varepsilon)/\partial x$ and $\partial(\ln\varepsilon)/\partial y$.

$$w = \frac{\int_S |\mathbf{E}|^2 r^2 dr\, d\varphi}{\int_S |\mathbf{E}|^2 r dr\, d\varphi}. \quad (31)$$

As the power is reduced on approaching the critical power (the quasioptical limit), the beam width increases without bound. As the power is increased, the beam width decreases.

The vector structure of the field is strongly dependent on the power (or is uniquely related to it by the propagation constant). In the quasioptical limit the components E_y and E_z are negligibly small (see Sec. 3). As the power increases, their relative contribution becomes larger. Here when the terms $\partial(\ln\varepsilon)/\partial x$ and $\partial(\ln\varepsilon)/\partial y$ are neglected, it becomes stable and in the limit of high powers, in accordance with Sec. 4, a universal field structure develops in which the maximum intensity of the longitudinal field component is roughly a factor of 10 smaller than that of the transverse component. Only its scales change with increasing γ. The approximate description of Sec. 4 correctly reflects its major characteristics. When the solutions that take the intermingling of the components of the vector potential into account are solved, the result is similar, but the ratio of the peak intensities of the longitudinal and transverse components increases to one third. This characteristic field structure is illustrated in Fig. 4.

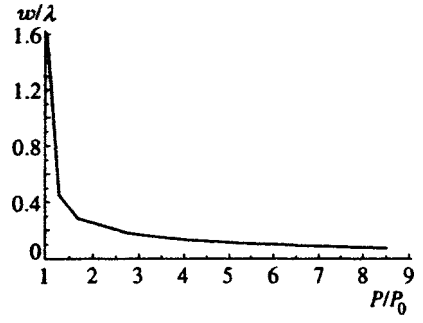

FIG. 3. The effective beam with w as a function of power (calculated neglecting the terms containing $\partial(\ln\varepsilon)/\partial x$ and $\partial(\ln\varepsilon)/\partial y$).

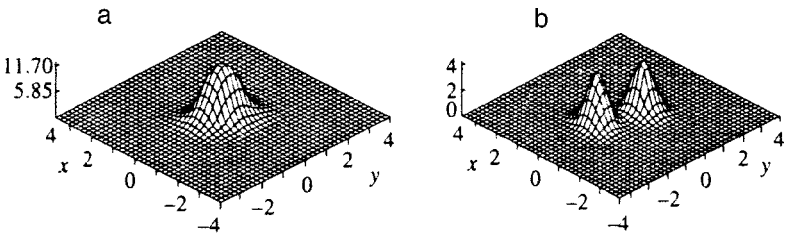

FIG. 4. Distributions of the electric field intensity over the transverse cross section of a beam in the ''needles of light'' approximation: a — the intensity of the transverse field component $I_\perp = |E_x|^2 + |E_y^2|$; b — intensity of the longitudinal field component $I_\parallel = |E_z|^2$, $P = 4P_0$.

These calculations confirm the existence of rather narrow light beams (optical needles) at powers several times greater than the critical self-focusing power in the quasioptical approximation.

7. CONCLUSION

By solving the complete system of Maxwell's equations for a transparent medium with a Kerr nonlinearity, we have thus determined the vector structure of the electromagnetic field during self-channelling regimes. It has been found that these regimes exist for powers exceeding the critical self-focusing power. As the power is increased, a universal (invariant) field structure develops in which only the scales change. In particular, self-channelling with a channel width smaller than the (linear) wavelength of the light or needles of light can occur. With their extremely high concentration of optical power, these regimes can be of both scientific and applied interest.

Of course, the results presented here were obtained under certain assumptions. The major limitation is the deviation of the nonlinearity of the medium from a Kerr nonlinearity (3) at high radiation intensities. In this case, saturation of the nonlinearity must be taken into account, for example, in the form

$$\varepsilon = \varepsilon_0 + \frac{\varepsilon_2 |\mathbf{E}|^2}{1 + |\mathbf{E}|^2 / I_S}, \tag{32}$$

where I_S is the saturation intensity. The Kerr nonlinearity (3) is obtained from Eq. (32) for $|\mathbf{E}|^2 \ll I_S$. The case $\varepsilon_2 |\mathbf{E}|^2 > \varepsilon_0$ is also possible for $\varepsilon_2 I_S \gg \varepsilon_0$, which determines the conditions for applicability of the preceding analysis. Another important question is the analysis of the stability of these regimes and the conditions for their development. Although the way the power depends on the propagation constant indicates that the regime is stable according to the Kolokolov–Vakhitov criterion,[11] it should be recalled that this has been proven only in the quasioptical approximation. In practice, power at a level exceeding the self-focusing critical power is not sufficient for the formation of needles of light. The initial beam profile is also important, since broad, high power beams of light break up into isolated ''filaments.''[7] More detailed study of the stability and features of self-channelling in media with saturation nonlinearity is required.

This work was partially supported by the Russian Fund for Fundamental Research (Grant No. 98-02-18202) and the International Scientific–Technical Foundation (Project No. 666).

*)E-mail: rosanov@ilph.spb.su

[1] G. A. Askar'yan, Zh. Éksp. Teor. Fiz. **42**, 1567 (1962) [Sov. Phys. JETP **15**, 1088 (1962)].
[2] S. A. Akhmatov, A. P. Sukhorukov, and R. V. Khokhlov, Usp. Fiz. Nauk **93**, 2 (1967) [Sov. Phys. Usp. **10**, 609 (1968)].
[3] V. N. Lugovoĭ and A. M. Prokhorov, Usp. Fiz. Nauk **111**, 203 (1973) [Sov. Phys. Usp. **15**, 519 (1973)].
[4] Y. R. Shen, Prog. Quantum Electron. **4**, 1 (1975).
[5] J. H. Marburger, Prog. Quantum Electron. **4**, 35 (1975).
[6] S. N. Vlasov and V. I. Talanov, *Self-Focussing of Waves* [in Russian], Institute of Applied Physics, Russian Academy of Sciences, Nizhny Novgorod (1997).
[7] N. N. Rozanov and V. A. Smirnov, Zh. Éksp. Teor. Fiz. **76**, 2060 (1976) [Sov. Phys. JETP **43**, 1075 (1976)].
[8] D. A. Kirsanov and N. N. Rozanov, Opt. Spektr. **86** (1999).
[9] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, 2nd ed., Pergamon, Oxford (1984) [Russian orig. Nauka, Moscow (1982)].
[10] J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, New York (1941).
[11] A. A. Kolokolov and N. G. Vakhitov, Izv. Vuzov. Radiofizika **16**, 1020 (1973).

Translated by D. H. McNeill

Mechanism of spontaneous radiation relaxation of an impurity atom in a photonic band-gap crystal

A. M. Basharov*)

Moscow State Institute of Engineering Physics, 115409 Moscow, Russia
(Submitted 2 February 1999)
Zh. Éksp. Teor. Fiz. **116**, 469–484 (August 1999)

On the base of the concept of radiative interaction between atoms of two different species, a novel electric-dipole mechanism of spontaneous radiating relaxation of an impurity atom in a medium with a gap in the photonic density of states about the transition frequency has been suggested. A kinetic equation for an impurity atom has been derived, and polarizations of the photonic band-gap crystal and an induced electromagnetic wave at the spontaneous transition frequency due to the suggested mechanism have been investigated. The parameters of the induced wave are determined by those of the suggested mechanism, and this dependence can be used in experimental investigations of these parameters. © *1999 American Institute of Physics.* [S1063-7761(99)00808-2]

1. INTRODUCTION

Recently a lot of attention has been focused on emission and propagation of electromagnetic waves under conditions when some structural and/or dynamic factors impede these processes in certain ranges of field parameters. For example, in structures formed by periodic dielectric layers, Bragg reflection[1–3] prevents penetration and propagation of weak electromagnetic waves of certain frequencies, whereas a sufficiently intense electromagnetic fields of the same frequencies can propagate in the regime of band-gap soliton waves.[4,5] A polariton branch in the spectrum and a gap in the excitation spectrum due to interaction between light and an optically dense medium[6–9] may also prevent penetration of weak electromagnetic waves of specific frequencies and lead to multistable reflection regimes for intense waves, their self-oscillations,[10] and penetration into the dense medium. There can be other causes leading to gaps in the density of states of electromagnetic waves. There is now available a fairly wide range of materials in which the propagation of optical waves of specific frequencies is forbidden in the classical sense.[11] We call such media photonic band-gap (PBG) crystals, although this term is usually applied to a more narrow range of periodic dielectric structures. We will disregard the geometrical factors which may lead to closing the gap in the density of states of electromagnetic waves that propagate in certain directions.

If an impurity atom is embedded in a PBG crystal so that the frequency of the interatomic transition coincides with the band gap of photon states, the lifetime of the excited state of such an impurity atoms is unusually long. Such a state of an excited atom in a PBG crystal is called either localized light or a bound state of the atom and photon.[1–3,12–14] Investigation of the dynamics of excited atoms in PBG crystals is of interest for various branches of physics and shows much promise in view of application to fabrication of various components of the quantum computer.[15]

Lifetimes of excited states of impurity atoms in PBG crystals were calculated by many researchers.[16–26] They usually considered situations when the transition frequency of an impurity atom is close to an edge of the band gap in a PBG crystal. In this case, the dynamics of the spontaneous light emission is anomalous and demonstrates a set of remarkable features[16–18] because the Markovian approximation cannot be used in calculations. Introduction of two identical interacting impurity atoms to such systems makes the problem the more interesting.[23] With a large number of impurity atoms, the band gap can contain a lot of impurity levels.[24] Another approach to the localized photon problem is demonstrated in Refs. 25 and 26. It is noteworthy that some authors, in addition to application of various approaches, took into account effects of external electromagnetic fields.[27]

As concerns the role of PBG crystal media in the analysis of the dynamics of an excited atom under conditions of photon localization, it is usually limited to the interaction between an atom and collective optical modes of the medium, which are related, in one way or another, to its excitation spectrum.[22–26] This paper focuses the reader's attention on a general mechanism of spontaneous relaxation of an impurity atom in a PBG crystal under conditions when the conventional optical relaxation is forbidden because of the gap in the density of photon states. Interaction between an impurity atom and atoms of the PBG crystal is treated locally, without taking into account collective processes and assuming that only neighboring atoms are involved, which are considered in isolation from the rest of the PBG crystal atoms. Undoubtedly, the dynamics of an excited impurity atom in a PBG crystal presents a very complicated problem. A self-consistent approach should take into account the entire hierarchy of interaction processes, separate the main contributors, and analyze their effect on the development of instabilities of various kinds. There is every reason to take into account the mechanism suggested in this paper in a self-consistent theory, especially in problems concerned with

 © 1999 American Institute of Physics

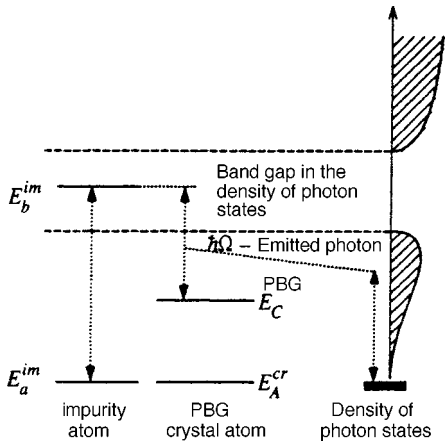

FIG. 1. Diagram of energy levels of an impurity atom and an atom of a photonic band-gap crystal essential for the mechanism under consideration, and the density of photon states near the transition frequency of the impurity atom.

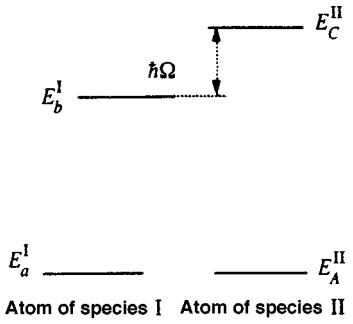

FIG. 2. An alternative configuration of levels involved in radiative processes.

additional action of external electromagnetic fields on an impurity atom.[27]

The underlying idea of the suggested mechanism of spontaneous radiation relaxation can be easily understood on the base of the following considerations (see Fig. 1). It is known that atoms of two sorts where electric-dipole optical transitions between their energy levels $E_b^{\rm im}\to E_a^{\rm im}$ (the impurity atom or atom of species I) and two-photon transitions $E_C^{\rm PBG}\to E_A^{\rm PBG}$ (the PBG crystal atom or atom of species II) are allowed interact with one another in an electromagnetic field of frequency Ω when the following condition is satisfied:

$$E_b^{\rm im}-E_a^{\rm im}\approx E_C^{\rm PBG}-E_A^{\rm PBG}+\hbar\Omega, \tag{1}$$

so that the excitation can be transferred from one atom to another and a photon of frequency Ω can be emitted or absorbed, depending on the initial populations of atomic levels. For example, if the impurity atom is in the excited state and the PBG crystal atom is in the ground state, the impurity atom is de-excited from the level $E_b^{\rm im}$ to $E_a^{\rm im}$ as a result of interaction between the atoms in the optical field, and the PBG crystal atom is excited from the lower level $E_A^{\rm PBG}$ to the higher level $E_B^{\rm PBG}$ as a result of the two-photon transition. In this process, a photon is emitted whose frequency Ω can be outside the band gap in the photonic density of states. These processes have been discovered and investigated in detail by researchers of atomic collisions, who call them radiative atomic collisions.[28–30] Such processes are observed in coherent optical fields of high amplitudes. Many different processes may also occur in addition to that discussed in this paper, depending on the configuration of atomic energy levels. (see Fig. 2 and Refs. 29 and 30).

In this paper, the radiative two-photon mechanism of excitation transfer applied to atomic collisions as a feasible mechanism of spontaneous radiative relaxation of an impurity atom in a PBG crystal. In contrast with the approach developed earlier,[28–30] in which the atomic dynamics was described in terms of the classical theory of depolarizing atomic collisions,[31–33] the relaxation dynamics of an impurity atom in a PBG crystal should be described in terms of radiative atomic collisions combined with a consistent quantum-mechanical description of the emitted electromagnetic field, and an adequate model of photonic heat reservoir, with allowance for the real density of photon states in the PBG crystal. The model system discussed here includes an impurity atom, an atom of a PBG crystal, and a photonic heat reservoir. Since the frequencies of emitted photons in the suggested mechanism are outside the band gap in the density of photon states in the PBG crystal, and since the widths of impurity atomic levels are much smaller than the characteristic frequencies over which the photonic density of states changes, the photonic heat reservoir in this problem can be treated as broad-band. The relaxation dynamics of an impurity atom can then be conveniently described using the Ito quantum-mechanical equation[34] and a unitary transformation changing the initial Hamiltonian of the problem to an effective Hamiltonian.[30] In this case, the parameters of the photonic density of states are incorporated in the coupling constant in the Ito equation. The basic parameter, which characterizes the relaxation dynamics of an impurity atom, also determines the polarization of a PBG crystal at the frequency of spontaneous radiation in the suggested model. This paper considers a coherent mechanism generating polarization of a PBG crystal under coherent pumping of impurity atoms and their environment. This model disregards the collective effects associated with transfer of excitations of atoms of PBG crystal through the bulk. These processes undoubtedly affect the parameters of the radiation mechanism under discussion, but such collective effects *per se* cannot block its action. They should be taken into account as the theory is developed further.

The reported work is a natural continuation of the previous work[35] involving the analysis of a two-photon mechanism of relaxation of atomic states through simultaneous emission and/or absorption of two photons, one of which belongs to an intense coherent wave described in classical terms and the other to a quantum-mechanical photonic heat reservoir. The relevant relaxation rate proved to be proportional to the coherent wave amplitude.[35] In the case studied in this paper, a photon from an intense coherent wave is replaced with excitation of an atom of a PBG crystal. Owing to Raman-like processes, the emitted photon is outside the band gap in the photonic density of states. The two-photon relaxation mechanism yields a very small coupling constant

of this spontaneous radiative mechanism, which is a function of the density of photon states and parameters of interaction between an impurity atom and neighboring atoms of the PBG crystal, which allows us to ignore this relaxation mechanism in usual situations. In a PBG crystal, however, where the density of photon states corresponding to the photon emitted in the usual transition between the excited and ground states is zero, the conventional relaxation mechanism is blocked. In this case the small relaxation rate and the mechanism that determines this process can be the decisive factors that determine lifetime of the excited state of the impurity atom and the channels through which the excess energy is dissipated. Probably, the calculated rate will be different when collective states of the PBG crystal are taken into account by this mechanism, but this should be a topic of a dedicated study.

The paper is organized as follows. Section 2 presents the initial and effective Hamiltonians of the problem. In Sec. 3, the Ito quantum equation, the kinetic equation for the impurity atom, and the relaxation operator are derived. Section 4 treats the polarization of the PBG crystal and emission of a coherent wave at the spontaneous radiation frequency due to coherent pumping of the material. In Conclusions, we will analyze factors that affect the relaxation rate of the impurity atom. Appendix describes the relation between the parameters of the unitary transformation that relates the Hamiltonians in the interaction and Schrödinger representations.

2. INITIAL AND EFFECTIVE HAMILTONIANS OF THE PROBLEM

The full Hamiltonian of the system, which includes an impurity atom, an atom of the PBG crystal, and a photonic thermostat, is written in the form

$$H=H_{\rm im}+H_{\rm PBG}+H_{\rm th}+V_{\rm im-PBG}+V_{\rm im}+V_{\rm PBG}, \tag{2a}$$

where $H_{\rm in}$, $H_{\rm PBG}$, and $H_{\rm th}$ are the Hamiltonians of the separate components of the system under consideration, respectively, the terms $V_{\rm im}$ and $V_{\rm PBG}$ describe interaction between the photonic thermostat on one side, and the impurity and PBG crystal atoms, respectively, on the other, whereas $V_{\rm im-PBG}$ determines the dipole–dipole interaction between the atoms. It is convenient to use the one-dimensional approximation for the photonic heat reservoir and express the components of the full Hamiltonian as follows:

$$H_{\rm im}=\sum_{\alpha} E_{\alpha}^{\rm im} a_{\alpha}^{+} a_{\alpha}, \quad H_{\rm PBG}=\sum_{\beta} E_{\beta}^{\rm PBG} b_{\beta}^{+} b_{\beta},$$

$$H_{\rm th}=\sum_{\lambda} \int d\omega\, \hbar\omega c_{\lambda\omega}^{+} c_{\lambda\omega},$$

$$V_{\rm im}=-i\hbar \sum \int d\omega\, K(\omega)(\mathbf{e}_{\lambda\omega}\cdot\mathbf{d}_{\alpha\alpha'}^{\rm im}) a_{\alpha}^{+} a_{\alpha'} c_{\lambda\omega}+{\rm H.c.},$$

$$V_{\rm PBG}=-i\hbar \sum \int d\omega\, K(\omega)(\mathbf{e}_{\lambda\omega}\cdot\mathbf{d}_{\beta\beta'}^{\rm PBG}) b_{\beta}^{+} b_{\beta'} c_{\lambda\omega}+{\rm H.c.},$$

$$V_{\rm im-PBG}=\sum_{\substack{\alpha\alpha'\\ \beta\beta'}} U_{\alpha\beta,\alpha'\beta'} a_{\alpha}^{+} a_{\alpha'} b_{\beta}^{+} b_{\beta'}$$

$$U_{\alpha,\beta,\alpha'\beta'}=((\mathbf{d}_{\alpha\alpha'}^{\rm im}\cdot\mathbf{d}_{\beta\beta'}^{\rm PBG})l^2-3(\mathbf{d}_{\alpha\alpha'}^{\rm im}\cdot\mathbf{l})(\mathbf{d}_{\beta\beta'}^{\rm PBG}\cdot\mathbf{l}))l^{-5}. \tag{2b}$$

Here the superscripts im and PBG label parameters that characterize the impurity atom and the PBG crystal atom, respectively. For example, $E_{\alpha}^{\rm im}$ is the energy of an isolated impurity atom in the state of rest characterized by quantum numbers α. The creation and annihilation operators for the impurity atom in a state with energy $E_{\alpha}^{\rm im}$ are denoted by a_{α}^{+} and a_{α}, whereas the similar operators for the PBG crystal atom are b_{β}^{+} and b_{β}. The creation and annihilation operators of a photon with frequency ω and polarization vector $\mathbf{e}_{\lambda\omega}$ are denoted by $c_{\lambda\omega}^{+}$ and $c_{\lambda\omega}$. The central frequency of the photonic heat reservoir is denoted by Ω. Assume that creation and annihilation operators satisfy the Bose commutation rules. As regards the atomic operators, this is a matter of convenience, because the final results are the same as those for the Fermi commutation rules. Further, $\mathbf{d}$ is the operator of the atomic dipole moment, and l is the separation between the impurity atom and the nearest atom of the photonic heat reservoir. The coupling constant $K(\omega)$ is proportional to the density of photon states $g(\omega)$: $K(\omega)=\sqrt{2\pi\omega/\hbar g(\omega)}$.

An important point is that, in contrast with the previously reported studies,[12–14,16–27] the excited level of the impurity atom $E_{b}^{\rm im}$, which coincides with the band gap of the photonic density of states, is set at a considerable distance from the gap edge, so that the corresponding detuning $\Delta_{g}^{\rm im}$ is markedly larger than the frequency shift of the transition under consideration due to interaction with gap-edge photons. In this case, the effect of the localized photon[12,13] on the impurity atom can be disregarded, and the relaxation mechanisms discussed in the earlier publications[12–14,16–27] are inessential. The separation l between the impurity atom and the PBG crystal atom is assumed to be much shorter than the characteristic wavelength, then the dipole–dipole coupling of the atoms of different sorts is determined by the longitudinal electromagnetic field and little affected by the gap in the density of states of transverse electromagnetic field. This is another feature distinguishing the situation under discussion from the resonant dipole–dipole interaction between identical atoms, whose efficiency depends on the smallness of $\Delta_{g}^{\rm im}$, i.e., the proximity of the excited impurity level to the band-gap edge in the density of photon states.[12,13,20–22] Finally, in the interaction operators $V_{\rm im}$ and $V_{\rm PBG}$, whose shape is similar to that of the operator of resonant dipole–dipole interaction, the central frequency Ω of the optical field differs from that of the impurity transition by the frequency of the two-photon transition in the PBG crystal atom and is far from the band-gap edge in the photonic density of states, so that the photonic ensemble can be treated as a broad-band heat reservoir. This approach is different from that suggested by John and Wang,[13] where the frequencies of photons re-emitted by impurity atoms are close to the band gap edge, and the proximity between frequency Ω and the impurity transition frequency is important ($\Delta_{g}^{\rm im}$ should be small). The resonant condition (1) for Ω relates transition frequencies in atoms of different species and does not impose a limitation on the position of Ω with respect to the band-gap edge. Thus, the conditions discussed in this paper are, in a sense, opposite to those introduced in the earlier

studies,[12,13,20–22] so that the relaxation mechanisms associated with the localization photon, ''dressing'' of the impurity atom by the localized photon field,[12,13] and the respective terms in Hamiltonian (2) can be neglected.

Let us consider the Schrödinger equation of the system:

$$i\hbar\frac{\partial}{\partial t}|\Psi\rangle = H|\Psi\rangle.$$

Following the conventional technique applied to resonant processes,[30] let us introduce, using the unitary operator e^{-iS}, the new state vector $|\tilde{\Psi}\rangle = e^{-iS}|\Psi\rangle$, which follows the equation

$$i\hbar\frac{\partial}{\partial t}|\tilde{\Psi}\rangle = \tilde{H}|\tilde{\Psi}\rangle$$

with the Hamiltonian

$$\tilde{H} = e^{-iS}He^{iS} - i\hbar e^{-iS}\frac{\partial}{\partial t}e^{iS}. \tag{3}$$

The unitary property of the transformation means that all observable quantities in the problem remain unchanged. Let us determine a transformation that would clearly emphasize the dominant role of resonant processes [condition (1)] in the nondiagonal matrix elements of the Hamiltonian. To this end, let us express S and $\tilde{H}$ as series in powers of constants that couple the atoms to the electric field of the photonic heat reservoir and to one another:

$$S = S^{(10)} + S^{(01)} + S^{(11)} + \ldots,$$

$$\tilde{H} = \tilde{H}^{(00)} + \tilde{H}^{(10)} + \tilde{H}^{(01)} + \tilde{H}^{(11)} + \ldots, \tag{4}$$

where the integers m and n in each pair of the indices denote the mth order in the coupling constant between the atom and the field (the first index) and the nth order in the coupling constant between the atoms (the second index). We have[30]

$$\tilde{H}^{(00)} = H_{\rm im} + H_{\rm PBG} + H_{\rm th},$$

$$\tilde{H}^{(10)} = V_{\rm im} + V_{\rm PBG} - i[S^{(10)}, \tilde{H}^{(00)}] + \hbar\frac{\partial}{\partial t}S^{(10)},$$

$$\tilde{H}^{(01)} = V_{\rm im-PBG} - i[S^{(01)}, \tilde{H}^{(00)}] + \hbar\frac{\partial}{\partial t}S^{(01)},$$

$$\tilde{H}^{(11)} = -\frac{i}{2}[S^{(01)}, V_{\rm im} + V_{\rm PBG}] - \frac{i}{2}[S^{(10)}, V_{\rm im-PBG}] - \frac{i}{2}[S^{(01)}, \tilde{H}^{(10)}] - \frac{i}{2}[S^{(10)}, \tilde{H}^{(01)}] - i[S^{(11)}, \tilde{H}^{(00)}] + \hbar\frac{\partial}{\partial t}S^{(11)},$$

$$\tilde{H}^{(20)} = -\frac{i}{2}[S^{(10)}, V_{\rm im} + V_{\rm PBG}] - \frac{i}{2}[S^{(10)}, \tilde{H}^{(10)}] - i[S^{(20)}, \tilde{H}^{(00)}] + \hbar\frac{\partial}{\partial t}S^{(20)},$$

$$\tilde{H}^{(02)} = -\frac{i}{2}[S^{(01)}, V_{\rm im-PBG}] - \frac{i}{2}[S^{(01)}, \tilde{H}^{(01)}] - i[S^{(02)}, \tilde{H}^{(00)}] + \hbar\frac{\partial}{\partial t}S^{(02)}. \tag{5}$$

Hereafter the brackets denote the commutator of two operators: $[A,B] = AB - BA$.

Since only resonant processes involving photons of the electromagnetic field and atomic excitations are efficient, we should set $\tilde{H}^{(10)} = \tilde{H}^{(01)} = 0$; hence we derive expressions for $S^{(10)}$ and $S^{(01)}$ under the condition that the electromagnetic field is turned on adiabatically:

$$S^{(10)} = \sum_{\substack{\alpha\alpha'\\ \beta\beta'\\ \lambda}} \int d\omega\, K(\omega)\left\{\frac{(\mathbf{e}_{\lambda\omega}\cdot\mathbf{d}^{\rm PBG}_{\beta\beta'})b^+_\beta b_{\beta'}}{\omega^{\rm PBG}_{\beta\beta'} - \omega} + \frac{(\mathbf{e}_{\lambda\omega}\cdot\mathbf{d}^{\rm im}_{\alpha\alpha'})a^+_\alpha a_{\alpha'}}{\omega^{\rm im}_{\alpha\alpha'} - \omega}\right\} c_{\lambda\omega} + \text{H.c.},$$

$$S^{(01)} = i\hbar^{-1}\sum_{\substack{\alpha\alpha'\\ \beta\beta'}} \frac{U_{\alpha\beta,\alpha'\beta'}}{\omega^{\rm PBG}_{\beta\beta'} + \omega^{\rm im}_{\alpha\alpha'}} a^+_\alpha a_{\alpha'} b^+_\beta b_{\beta'},$$

$$\omega^{\rm im}_{\alpha\alpha'} = (E^{\rm im}_\alpha - E^{\rm im}_{\alpha'})/\hbar, \quad \omega^{\rm PBG}_{\beta\beta'} = (E^{\rm PBG}_\beta - E^{\rm PBG}_{\beta'})/\hbar \tag{6}$$

(note that the interaction representation is more convenient for specific calculations (see Appendix)). Let us retain in $\tilde{H}^{(11)}$, in addition to the diagonal elements, only the terms responsible for the resonant processes [Eq. (1)] and the approximation of a rotating wave. The other terms determine operators $S^{(11)}$, $S^{(20)}$, etc., and they are inessential for the further consideration. Their expressions only confirm that the selected approach is self-consistent, because these terms do not have resonant denominators. This procedure is fully identical to that described in earlier publications.[30,35] As a result, the effective Hamiltonian can be expressed in the form

$$H^{\rm eff} = H_{\rm im} + H_{\rm PBG} + H_{\rm th} + \tilde{H}^{(11)} + \tilde{H}^{(02)}, \tag{7}$$

$$\tilde{H}^{(11)} = i\hbar\sum_\lambda \int d\omega\, K(\omega) R_+ c_{\lambda\omega} + \text{H.c.}, \tag{8a}$$

$$R_+ = h_\lambda(l) a^+_b a_a b^+_A b_C,$$

$$h_\lambda(l) = \sum_\beta \left\{\frac{U_{bA,a\beta}(\mathbf{e}_{\lambda\omega}\cdot\mathbf{d}^{\rm PBG}_{\beta C})}{\hbar(\omega^{\rm PBG}_{\beta A} - \omega^{\rm im}_{ba})} + \frac{(\mathbf{e}_{\lambda\omega}\cdot\mathbf{d}^{\rm PBG}_{A\beta})U_{b\beta,aC}}{\hbar(\omega^{\rm PBG}_{\beta C} + \omega^{\rm im}_{ba})}\right\}, \tag{8b}$$

$$\tilde{H}^{(02)} = \sum_{\alpha\beta} u_{\alpha\beta}(l) a^+_\alpha a_\alpha b^+_\beta b_\beta,$$

$$u_{\alpha\beta}(l) = \sum_{\alpha'\beta'} \frac{U_{\alpha\beta,\alpha'\beta'}U_{\alpha'\beta',\alpha\beta}}{\hbar(\omega^{\rm im}_{\alpha\alpha'} + \omega^{\rm PBG}_{\beta\beta'})}. \tag{9}$$

In deriving Eq. (8b), we have used the resonant condition (1) for the process under consideration and set in the intermediate formulas $\omega^{\rm PBG}_{\beta A} + \omega \approx \omega^{\rm PBG}_{\beta C} + \omega^{\rm im}_{ba}$ and $\omega^{\rm PBG}_{\beta C} - \omega \approx \omega^{\rm PBG}_{\beta A} - \omega^{\rm im}_{ba}$.

The term $\tilde{H}^{(02)}$ describes the broadening and shift of the impurity atom levels due to collisions. Wherever possible, it will be either neglected or expressed in the simplest form.

The term $\tilde{H}^{(11)}$ determines the mechanism of radiative relaxation under consideration. Here the main parameter is $h_\lambda(l)$, because it depends on the separation l between the atoms and on the configuration of energy levels of PBG crystal atoms. A felicitous selection of atoms in the model can, in principle, lead to considerably higher $h_\lambda(l)$, which is indicated by the denominators in the components of $h_\lambda(l)$. The dipole moments corresponding to transitions from the levels E_A^{PBG} and E_C^{PBG} to other levels are also large, especially those minimizing the denominators mentioned above. The spontaneous radiation relaxation rates can be calculated by standard rules with the help of Eq. (8b). A more important task, however, is derivation of the kinetic equation for the impurity atom, because versatile problems of the quantum optics can be solved on the basis of this equation. The radiative processes under consideration determine the relaxation operator in the kinetic equation and polarization of the PBG crystal.

3. ITO EQUATION AND RELAXATION OPERATOR

Assume that the photonic thermostat in our system is described by conventional equations[34–36]

$$\langle c_\lambda^+(t)c_{\lambda'}(t')\rangle = N_{\mathrm{ph}}\delta_{\lambda\lambda'}\delta(t-t'),$$

$$\langle c_\lambda(t)c_{\lambda'}^+(t')\rangle = (1+N_{\mathrm{ph}})\delta_{\lambda\lambda'}\delta(t-t'),$$

$$\langle c_\lambda^+(t)c_{\lambda'}^+(t')\rangle = \langle c_\lambda(t)c_{\lambda'}(t')\rangle = 0,$$

$$c_\lambda(t) = \frac{1}{\sqrt{2\pi}}\int d\omega \exp[-i\omega(t-t_0)]c_{\lambda\omega}, \tag{10}$$

where N_{ph} is the density of thermostat photons, t_0 is a certain initial moment of time with respect to which the photonic operators $c_{\lambda\omega}$ and $c_{\lambda\omega}^+$ are determined. The main assumption of this approximation is that the coupling constant does not depend on the frequency, but is equated to its value at the central frequency Ω of the photonic thermostat:

$$K(\omega) = K(\Omega) = \sqrt{\kappa/2\pi}. \tag{11}$$

This assumption is known as Markov's approximation.[34]

Let us write the Heisenberg equation for a certain operator A characterizing the complex of the impurity atom and the neighboring atom of the PBG crystal:

$$\dot{A} = -\frac{i}{\hbar}[A, H_{\mathrm{im}} + V + H_{\mathrm{PBG}} + \tilde{H}^{(02)}] + \sqrt{\kappa}[A, R_+]c(t) - \sqrt{\kappa}[A, R_+^+]c^+(t). \tag{12}$$

Here we have assumed that (1) operators of the photonic heat reservoir change with time as operators of a noninteracting system; (2) the Markovian approximation applies; (3) polarization effects can be neglected. In order to enable further application of the kinetic equation to the nonlinear optics of the impurity atoms in PBG crystal, we have added the operator V, which takes account of various possible interactions of the impurity atoms, e.g., with the external coherent fields, to the impurity atom operator.

Further, let us introduce Wiener's quantum-mechanical process:[34]

$$B(t,t_0) = \int_{t_0}^{t} dt' c(t'), \quad [B(t,t_0), B^+(t,t_0)] = t - t_0,$$

and the Ito conventional integration and differentiation rules. In order to keep some elements of notation used in the previous studies,[35,36] the same letter is used in the notation of Wiener's process. After standard transformations,[34–36] we obtain the Ito quantum-mechanical equation in the form

$$\begin{aligned} dA = &-\frac{i}{\hbar}[A, H_{\mathrm{im}} + V + H_{\mathrm{PBG}} + \tilde{H}^{(02)}]dt \\ &+ \sqrt{\kappa}[A, R_+]dB(t) - \sqrt{\kappa}[A, R_+^+]dB^+(t) \\ &+ \frac{\kappa}{2}\{(1+N_{\mathrm{ph}})(R_+[A, R_+^+] + [R_+, A]R_+^+) \\ &+ N_{\mathrm{ph}}(R_+^+[A, R_+] + [R_+^+, A]R_+)\}dt, \end{aligned} \tag{13}$$

where the Ito increments obey the algebra

$$dB^+(t)dB(t) = N_{\mathrm{ph}}dt, \quad dB(t)dB^+(t) = (1+N_{\mathrm{ph}})dt,$$

$$dB(t)dB(t) = dB^+(t)dB^+(t) = dB(t)dt = dt\,dB(t) = dB^+(t)dt = dt\,dB^+(t) = dt\,dt = 0$$

and, instead of the Leibnitz rule, the Ito rule for differentiation of products applies:

$$d(A_1A_2) = (dA_1)A_2 + A_1dA_2 + (dA_1)(dA_2).$$

Further manipulations with nonanticipating operators in front of the Ito increments lead to an equation for the two-particle density operator for a group consisting of the impurity atom and the neighboring PBG crystal atom:

$$\frac{d}{dt}\wp + \hat{\Im}\wp = \frac{i}{\hbar}[\wp, H_{\mathrm{im}} + V + H_{\mathrm{PBG}} + \tilde{H}^{(02)}] \tag{14}$$

with the relaxation operator of the form

$$\begin{aligned} \hat{\Im}\wp = &-\frac{\kappa}{2}(1+N_{\mathrm{ph}})(2R_+^+\wp R_+ - \wp R_+R_+^+ - R_+R_+^+\wp) \\ &-\frac{\kappa}{2}N_{\mathrm{ph}}(2R_+\wp R_+^+ - \wp R_+^+R_+ - R_+^+R_+\wp). \end{aligned} \tag{15}$$

The kinetic equation for the density operator ρ for the impurity atom is derived from Eq. (15) by calculating the trace with respect to the variables of the PBG crystal atom:

$$\frac{d}{dt}\rho + \hat{\Gamma}\rho = \frac{i}{\hbar}[\rho, \overline{H}_{\mathrm{im}} + V], \tag{16}$$

$$\hat{\Gamma}\rho = -\frac{\kappa}{2}(1+N_{\rm ph})\mathrm{Tr}_{\rm PBG}(2R_+^+\rho\otimes\rho_0^{\rm PBG}R_+ - \rho\otimes\rho_0^{\rm PBG}R_+R_+^+ - R_+R_+^+\rho\otimes\rho_0^{\rm PBG}) - \frac{\kappa}{2}N_{\rm ph}\,\mathrm{Tr}_{\rm PBG}(2R_+\rho\otimes\rho_0^{\rm PBG}R_+^+ - \rho\otimes\rho_0^{\rm PBG}R_+^+R_+ - R_+^+R_+\rho\otimes\rho_0^{\rm PBG}). \quad (17)$$

Here $\rho_0^{\rm PBG}$ is the density operator for the PBG crystal atoms in the neighborhood of the impurity atom. This density operator is determined at the initial moment and describes the equilibrium distribution of these atoms over their energies and a certain distribution of these atoms over the distance from the impurity atom. In the case where a more detailed analysis, which takes into account the polarization properties and degeneracy of atomic states, is needed, the operator $\rho_0^{\rm PBG}$ includes a distribution over mutual alignments of the impurity and PBG crystal atoms. The trace in Eq. (17) is calculated over the variables of PBG crystal atoms surrounding the impurity atom.

The Hamiltonian $\overline{H}_{\rm im}$ of impurity atoms takes into account the shifts of atomic levels due to interaction with PBG crystal atoms. This interaction results in a spread of the frequency of the $E_b^{\rm im}\to E_a^{\rm im}$ transition around a certain central frequency. This spread is ignored in what follows.

The matrix elements of the relaxation operator for the energy levels of the impurity atom the frequency of the transition between which coincides with the gap in the photonic density of states are given by the formulas

$$(\hat{\Gamma}\rho)_{ba} = \left\{\frac{\kappa}{2}(1+N_{\rm ph})\overline{|h(l)|^2}_A + \frac{\kappa}{2}N_{\rm ph}\overline{|h(l)|^2}_C\right\}\rho_{ba},$$

$$(\hat{\Gamma}\rho)_{aa} = -\kappa(1+N_{\rm ph})\overline{|h(l)|^2}_A\rho_{bb} + \kappa N_{\rm ph}\overline{|h(l)|^2}_C\rho_{aa},$$

$$(\hat{\Gamma}\rho)_{bb} = \kappa(1+N_{\rm ph})\overline{|h(l)|^2}_A\rho_{bb} - \kappa N_{\rm ph}\overline{|h(l)|^2}_C\rho_{aa}. \quad (18)$$

Here $\overline{|h(l)|^2}$ denotes the square of the radiative interatomic interaction parameter averaged over the interatomic distances:

$$\overline{|h(l)|^2}_A = \langle A||h(l)|^2\rho_0^{\rm PBG}|A\rangle,$$

$$\overline{|h(l)|^2}_C = \langle C||h(l)|^2\rho_0^{\rm PBG}|C\rangle.$$

The index denoting the dependence of the basic parameter $h(l)$ on the polarization of emitted photons is omitted because we have ignored polarization effects. The ket-vectors $|A\rangle$ and $|C\rangle$ denote states with energies $E_A^{\rm PBG}$ and $E_C^{\rm PBG}$.

The structure of the relaxation operator defined by Eq. (18) reflects the obvious fact that, if a PBG crystal atom is in the excited state ($\langle C|\rho_0^{\rm PBG}|C\rangle\neq 0$), then, concurrently with the process described above, there is a reciprocal process in which a photon from the heat reservoir is absorbed, the PBG crystal atom transfers to its ground state $E_A^{\rm PBG}$, and the impurity atom is driven to the excited state.

The coefficients in front of the matrix elements in Eq. (18) are determined by the probability $\gamma_{\rm im}$ that the excited impurity atom spontaneously emits a photon in accordance with the suggested mechanism:

$$\gamma_{\rm im} = \kappa\overline{|h(l)|^2}_A. \quad (19)$$

Here the photonic heat reservoir is assumed to be empty ($N_{\rm ph}=0$), and PBG crystal atoms are assumed to be in the ground state ($\langle C|\rho_0^{\rm PBG}|C\rangle = 0$).

4. POLARIZATION OF PBG CRYSTAL

Let us discuss the effect of the suggested mechanism of spontaneous radiation relaxation on the polarization of the PBG crystal. We assume for convenience that the atomic transitions $E_b^{\rm im}\to E_a^{\rm im}$ and $E_C^{\rm PBG}\to E_A^{\rm PBG}$ are driven in a coherent manner, i.e., the nondiagonal elements of the corresponding density matrices are nonzero:

$$\rho_{ba} = \rho_{ba}^{\rm in}\exp[i(\mathbf{k}\cdot\mathbf{r} - \omega_{ba}^{\rm im}t - \varphi_{\rm im})],$$

$$\rho_{CA} = \rho_{CA}^{\rm in}\exp[i(\mathbf{K}\cdot\mathbf{r} - \omega_{CA}^{\rm PBG}t - \varphi_{\rm PBG})], \quad (20)$$

where $\rho_{bd}^{\rm in}$ and $\rho_{CA}^{\rm in}$ are the density matrices at a certain moment of time $t_{\rm in}$, $\mathbf{k}$ and $\mathbf{K}$ are the wave vectors, and $\varphi_{\rm im}$ and $\varphi_{\rm PBG}$ are constant phases. We assume that, when a certain level of excitation is achieved at the time $t\geqslant t_{\rm im}$, external fields neither act on the PBG crystal nor propagate through it.

The $E_b^{\rm im}\to E_a^{\rm im}$ transition of the impurity atoms can be excited by the three-photon resonant interaction with a pulsed coherent wave of a carrier frequency $\omega_{ba}^{\rm im}/3$ if this frequency is outside the band gap in the photonic density of states. Similarly, the $E_C^{\rm PBG}\to E_A^{\rm PBG}$ transition can be excited in a coherent manner through the two-photon interaction with a coherent wave with the carrier frequency $\omega_{CA}^{\rm PBG}/2$, which is also outside the band gap. Anyway, utilization of multi-photon interactions with coherent waves of different frequencies, however, allows us to circumvent the difficulty associated with the coincidence between the photonic band gap and the frequencies of pumping optical waves. Here we do not discuss details of these coherent processes. Note only that our analysis is limited to the model in which such coherent processes have little effect on the assumed photonic density of states in the PBG crystal, and the pulsed envelopes of the coherent pumping waves are necessary to rule out alternative multiphoton processes that develop in PBG crystals when radiative relaxation through the main channels is forbidden because of the presence of the band gap in the photonic density of states.

Polarization $\mathbf{P}$ of the PBG crystal due to the mechanism of spontaneous radiative relaxation and optical coherence of the states of impurity atoms and their environment [Eq. (20)] is determined by the general formula

$$\mathbf{P} = \mathrm{Tr}(\wp(\mathbf{d}^{\rm im}+\mathbf{d}^{\rm PBG})).$$

Using a unitary transformation, we can rewrite this expression in the form

$$\mathbf{P} = \mathrm{Tr}(e^{-iS}\wp e^{iS}(e^{-iS}\mathbf{d}^{\rm im}e^{iS} + e^{-iS}\mathbf{d}^{\rm PBG}e^{iS})) \approx \mathrm{Tr}(\rho\otimes\rho^{\rm PBG}(\mathbf{d}^{\rm im} - i[S,\mathbf{d}^{\rm im}] - i[S,\mathbf{d}^{\rm PBG}] + \cdots)) \approx \mathrm{Tr}(\rho\otimes\rho^{\rm PBG}(\mathbf{d}^{\rm im}+\mathbf{d}^{\rm PBG} - i[S^{(01)},\mathbf{d}^{\rm im}] - i[S^{(01)},\mathbf{d}^{\rm PBG}])). \quad (21)$$

It is clear that, in addition to the usual polarization of impurity atoms in the PBG crystal, they are also polarized at the combined frequencies $\omega_{ba}^{\rm im}-\omega_{CA}^{\rm PBG}(\mathbf{P}^{(-)})$ and $\omega_{ba}^{\rm im}+\omega_{CA}^{\rm PBG}(\mathbf{P}^{(+)})$:

$$P^{(-)}=-\hbar^{-1}\rho_{ba}\rho_{CA}^{\rm PBG*}\sum_{\beta}\left(\frac{U_{aC,b\beta}d_{\beta A}^{\rm PBG}}{\omega_{\beta C}^{\rm PBG}+\omega_{ba}^{\rm im}}+\frac{d_{C\beta}^{\rm PBG}U_{a\beta,bA}}{\omega_{\beta A}^{\rm PBG}-\omega_{ba}^{\rm im}}\right)+\text{c.c.},\tag{22}$$

$$P^{(+)}=-\hbar^{-1}\rho_{ba}\rho_{CA}^{\rm PBG}\sum_{\beta}\left(\frac{U_{aA,b\beta}d_{\beta C}^{\rm PBG}}{\omega_{\beta A}^{\rm PBG}+\omega_{ba}^{\rm im}}+\frac{d_{A\beta}^{\rm PBG}U_{a\beta,bC}}{\omega_{\beta C}^{\rm PBG}-\omega_{ba}^{\rm im}}\right)+\text{c.c.}.\tag{23}$$

Hereafter we again disregard the polarization effects and do not include in an explicit form averaging over interatomic distance. Then Eq. (22) can be written in a more compact form using the notation (8b):

$$P^{(-)}=-\rho_{ba}\rho_{CA}^{\rm PBG*}h^{*}(l)+\text{c.c.}.$$

Since the frequency $\omega_{ba}^{\rm im}-\omega_{CA}^{\rm PBG}\equiv\Omega$ is assumed to be outside the band gap in the photonic density of states, after interaction with the coherent optical waves in the time interval $t\geqslant t_{\rm in}$, a coherent wave due to coherent matrix elements (20) propagates through the PBG crystal, whose electric field is given by the equation

$$E_{\rm ind}^{(-)}=\mathscr{E}^{(-)}(t-x/c)\times\exp\{i[(\mathbf{k}-\mathbf{K})\cdot\mathbf{r}-\Omega t-\varphi_{\rm im}+\varphi_{\rm PBG}]\}+\text{c.c.},\tag{24}$$

$$\mathscr{E}^{(-)}(t)=-2\pi i\chi L\rho_{ba}^{\rm in}\rho_{CA}^{\rm in*}h^{*}(l)\times\exp[-(\gamma_{\rm im}+\gamma_{\rm PBG})(t-t_{\rm in})/2].$$

The propagation of the induced wave is determined by the condition of spatial synchronism $|\mathbf{k}-\mathbf{K}|^2=\Omega^2/c^2$, where c is the phase velocity of the wave with carrier frequency Ω in the PBG crystal. The x-axis is aligned with the propagation direction of the induced wave, and L is the dimension (in the x-direction) of the region where the optical coherence described by Eq. (20) is generated.

It is clear that the amplitude of the induced wave is controlled by the parameter $h(l)$, which determines the rate of the spontaneous radiation relaxation by the mechanism under discussion.

The damping of the induced wave described by Eq. (24) depends on both the relaxation matrix of the optical coherence of the impurity atom by the mechanism under consideration ($\gamma_{\rm im}$) and the relaxation of the two-photon $E_C^{\rm PBG}\rightarrow E_A^{\rm PBG}$ transition (with the rate $\gamma_{\rm PBG}$). If there are no levels between $E_C^{\rm PBG}$ and $E_A^{\rm PBG}$ to which optical transitions from the level $E_C^{\rm PBG}$ are allowed, and if the other relaxation processes can be ignored, $\gamma_{\rm PBG}\sim\gamma_{\rm im}$, the decay rate of the induced wave allows us to estimate the time of spontaneous radiation by the suggested mechanism.

If there are other levels between $E_C^{\rm PBG}$ and $E_A^{\rm PBG}$ to which optical transitions from level $E_C^{\rm PBG}$ are allowed at a frequency outside all band gaps in the photonic density of states, we have $\gamma_{\rm PBG}\gg\gamma_{\rm im}$, and the decay of the induced wave is not controlled by the mechanism of spontaneous radiation suggested in this paper.

5. CONCLUSIONS

In this paper, we have discussed an important example of a multiphoton process which has become feasible due to the progress in modern techniques of fabrication of photonic band-gap materials, in which conventional one-photon processes are suppressed because of the presence of band gaps in the density of photon states. In contrast to the mechanism of the two-photon (quadrupole) relaxation of an impurity atom, which was suggested by John and Wang,[12,13] the two-photon mechanism of spontaneous relaxation described in this paper incorporates interaction between two atoms, so it is essentially different from the previous one.[12,13] Its efficiency essentially depends on the mean separation between an impurity atom and its nearest neighbor from the PBG crystal atoms (in this respect, the efficiency of the suggested mechanism is not lower than that of the quadrupole mechanism) and the configuration of energy levels of the atoms of PBG crystal. The most favorable, for the processes under discussion is when the PBG crystal atom has a level $E_\beta^{\rm PBG}>E_C^{\rm PBG}$ such that the denominator $\omega_{\beta A}^{\rm PBG}-\omega_{ba}^{\rm im}$ in the expression for the basic parameter $h(l)$ is minimal, so that this parameter increases sharply. A high polarizability of PBG crystal atoms (large dipole moments $d_{\beta C}^{\rm PBG}$ and $d_{\beta A}^{\rm PBG}$) is also favorable for the mechanism of radiative relaxation under discussion.

An increase in the spontaneous relaxation rate of an impurity atom should also take place when not one, but a set of levels $E_{C1}^{\rm PBG}$, $E_{C2}^{\rm PBG}$,... of PBG crystal atoms, to which two-photon transitions from the $E_A^{\rm PBG}$ level are allowed, satisfy condition (1) with different frequencies Ω of emitted photons:

$$E_b^{\rm im}-E_a^{\rm im}\approx E_{Cj}^{\rm PBG}-E_A^{\rm PBG}+\hbar\Omega_j,\quad j=1,2,\ldots$$

If the frequencies Ω_j, where $j=1,2,...,M$, are outside the band gap in the photonic density of states, and the differences $|\Omega_j-\Omega_{j'}|$ are larger than the widths of the corresponding spectral lines of both PBG crystal and impurity atoms, then, according to Lax,[37] the frequencies Ω_j can be treated as central frequencies of decoupled photonic heat reservoirs, which can be labeled by the same subscripts $j=1,2,...,M$. Then the matrix elements of the relaxation operator (18) and the rate of spontaneous relaxation (19) of an excited impurity atom can then be expressed as

$$(\hat{\Gamma}\rho)_{ba}=\rho_{ba}\sum_{j=1}^{M}\left\{\frac{\kappa_j}{2}(1+N_{{\rm ph}\,j})\overline{|h_j(l)|_A^2}+\frac{\kappa_j}{2}N_{{\rm ph}\,j}\overline{|h_j(l)|_{Cj}^2}\right\},$$

$$(\hat{\Gamma}\rho)_{aa}=-\rho_{bb}\sum_{j=1}^{M}\kappa_j(1+N_{\mathrm{ph}\,j})\overline{|h_j(l)|^2_A}+\rho_{aa}\sum_{j=1}^{M}\kappa_j N_{\mathrm{ph}\,j}\overline{|h_j(l)|^2_{Cj}},$$

$$(\hat{\Gamma}\rho)_{bb}=\rho_{bb}\sum_{j=1}^{M}\kappa_j(1+N_{\mathrm{ph}\,j})\overline{|h_j(l)|^2_A}-\rho_{aa}\sum_{j=1}^{M}\kappa_j N_{\mathrm{ph}\,j}\overline{|h_j(l)|^2_{Cj}},$$

$$\gamma_{\mathrm{im}}=\sum_{j=1}^{M}\kappa_j\overline{|h_j(l)|^2_A},$$

where the parameters labeled by j derive from similar parameters considered above as a result of an obvious generalization. The case where $\Omega_s\approx(E_b^{\mathrm{im}}-E_a^{\mathrm{im}})/\hbar$ and the transition frequency $(E_b^{\mathrm{im}}-E_a^{\mathrm{im}})/\hbar$ of the impurity atom is close to an edge of the photonic band gap deserve a special treatment.

Note once again that the aim of the reported study was to demonstrate what novel quantum-mechanical mechanisms can be realized in photonic band-gap crystals. We have set aside the issue of the effect of these processes on the density of photon states and neglected all collective effects that can be associated with this dependence. The objects treated in our analysis (an impurity atom and its closest neighbor from PBG crystal atoms) were considered in isolation from the PBG crystal. This approximation, however, has allowed us to give a clear description of the suggested mechanism of spontaneous radiation relaxation and generation of polarization in the PBG crystal.

In this paper, we have not discussed the processes leading to the optical coherence described by Eq. (20). The point is that the features of PBG crystals also determine the specific properties of the three-photon excitation of impurity atoms. In this case, the dominant relaxation mechanism is an alternative two-photon process studied in the earlier work.[35] The nonlinear optics of impurity atoms in PBG crystals under three-photon excitation deserves a dedicated investigation. In this context, the results concerning coherent control of PBG crystals should be revised.

This work was part of a project sponsored by the Russian Fund for Fundamental Research (Grant 98-02-17429).

I wish to express my gratitude to V. P. Yakovlev and A. I. Maimistov for illuminating discussions.

APPENDIX

Let us determine the relations between the Hermitian operators which determine the unitary transformation of the Hamiltonian leading to the effective Hamiltonian in the interaction and Schrödinger representations. In the representation of interaction, we have

$$|\overline{\Psi}\rangle=\exp\left(-\frac{iH_0t}{\hbar}\right)|\Psi\rangle,\quad i\hbar\frac{\partial}{\partial t}|\overline{\Psi}\rangle=\overline{H}|\overline{\Psi}\rangle,$$

$$\overline{H}=\exp\left(\frac{iH_0t}{\hbar}\right)(V_{\mathrm{im-PBG}}+V_{\mathrm{im}}+V_{\mathrm{PBG}})\times\exp\left(-\frac{iH_0t}{\hbar}\right)=\overline{V}_{\mathrm{im-PBG}}+\overline{V}_{\mathrm{im}}+\overline{V}_{\mathrm{PBG}},$$

$$H_0=H_{\mathrm{im}}+H_{\mathrm{PBG}}+H_{\mathrm{th}}.$$

If the Hamiltonian in the interaction representation is subjected to a unitary transformation with a view to separating the effective Hamiltonian:

$$|\dot{\tilde{\Psi}}\rangle=e^{-iQ}|\overline{\Psi}\rangle,\quad i\hbar\frac{\partial}{\partial t}|\dot{\tilde{\Psi}}=\tilde{\overline{H}}|\dot{\tilde{\Psi}}\rangle,$$

$$\tilde{\overline{H}}=e^{-iQ}(\overline{V}_{\mathrm{im-PBG}}+\overline{V}_{\mathrm{im}}+\overline{V}_{\mathrm{PBG}})e^{iS}-i\hbar e^{-iQ}\frac{\partial}{\partial t}e^{iQ},$$

and the conventional expansion in powers of the coupling constant is performed:

$$Q=Q^{(10)}+Q^{(01)}+Q^{(11)}+\cdots,\quad \tilde{\overline{H}}=\tilde{\overline{H}}^{(10)}+\tilde{\overline{H}}^{(01)}+\tilde{\overline{H}}^{(11)}+\cdots$$

using the obvious formulas

$$\tilde{\overline{H}}^{(10)}=\overline{V}_{\mathrm{im}}+\overline{V}_{\mathrm{PBG}}+\hbar\frac{\partial}{\partial t}Q^{(10)},$$

$$\tilde{\overline{H}}^{(01)}=\overline{V}_{\mathrm{im-PBG}}+\hbar\frac{\partial}{\partial t}Q^{(01)},$$

$$\tilde{\overline{H}}^{(11)}=-\frac{i}{2}[Q^{(01)},\overline{V}_{\mathrm{im}}+\overline{V}_{\mathrm{PBG}}]-\frac{i}{2}[Q^{(10)},\overline{V}_{\mathrm{im-PBG}}]-\frac{i}{2}[Q^{(01)},\tilde{\overline{H}}^{(10)}]-\frac{i}{2}[Q^{(10)},\tilde{\overline{H}}^{(01)}]+\hbar\frac{\partial}{\partial t}Q^{(11)},$$

$$\tilde{\overline{H}}^{(20)}=-\frac{i}{2}[Q^{(10)},\overline{V}_{\mathrm{im}}+\overline{V}_{\mathrm{PBG}}]-\frac{i}{2}[Q^{(10)},\tilde{\overline{H}}^{(10)}]+\hbar\frac{\partial}{\partial t}Q^{(20)},$$

$$\tilde{\overline{H}}^{(02)}=-\frac{i}{2}[Q^{(01)},\overline{V}_{\mathrm{im-PBG}}]-\frac{i}{2}[Q^{(01)},\tilde{\overline{H}}^{(01)}]+\hbar\frac{\partial}{\partial t}Q^{(02)},\ldots,$$

then, assuming that the equalities $\tilde{H}^{(m,n)}=\exp(-iH_0t/\hbar)\tilde{\overline{H}}^{(m,n)}\exp(iH_0t/\hbar)$ and $\tilde{H}^{(0,0)}=H_0$ hold, we obtain by comparing with Eq. (5) (or through direct calculation) an important simple relation

$$Q^{(m,n)}=\exp(iH_0t/\hbar)S^{(m,n)}\exp(-iH_0t/\hbar),$$

which allows us to obtain the effective Hamiltonian and the corresponding unitary transformation in the Schrödinger representation using the interaction representation.

*)E-mail: ashat@ashat.mephi.msk.su

[1] S. John, Phys. Rev. Lett. **53**, 2169 (1984).
[2] E. Yablonovitch, Phys. Rev. Lett. **58**, 2059 (1987).
[3] S. John, Phys. Rev. Lett. **58**, 2486 (1987).
[4] M. J. Stell and C. M. de Sterke, Phys. Rev. A **48**, 1625 (1993).
[5] B. I. Mantsyzov, Phys. Rev. A **51**, 4939 (1995).
[6] A. S. Davydov, *Theory of Molecular Excitons* [in Russian], Nauka, Moscow (1968).
[7] V. M. Agranovich and V. L. Ginzburg, *Crystal Optics with Spacial Dispersion and Theory of Excitons*, Springer Verlag, Berlin (1984).
[8] V. V. Zheleznyakov, V. V. Kocharovskii, and Vl. V. Kocharovskii, Usp. Fiz. Nauk **32**, 835 (1989).
[9] A. L. Ivanov, H. Haug, and L. V. Keldysh, Phys. Rep. **296**, 237 (1998).
[10] V. Malyshev and E. C. Jarque, J. Opt. Soc. Am. B **12**, 1868 (1995).
[11] *Photonic Gap Materials*, edited by C. M. Soukoulis, NATO ASI Ser. E, Vol. 315, Kluwer Academic, Dordrecht (1996).
[12] S. John and J. Wang, Phys. Rev. Lett. **64**, 2418 (1990).
[13] S. John and J. Wang, Phys. Rev. B **43**, 12772 (1991).
[14] C. Weibuch, *Localization of Light in Disordered and Periodic Dielectrics*, Plenum, New York (1995).
[15] *Fundamental Problems in Quantum Theory* (Ann. N. Y. Acad. Sci., Vol. 755, ed. by D. M. Greenberger, A. Zeilinger), New York Academy of Sciences, New York (1995).
[16] S. John and T. Quang, Phys. Rev. A **50**, 1764 (1994).
[17] R. F. Nabiev, P. Yeh, and J. J. Sanchez-Mondragon, Phys. Rev. A **47**, 3380 (1993).
[18] S. Bay, P. Lambropoulos, and K. Molmer, Opt. Commun. **132**, 237 (1996).
[19] V. I. Rupasov and M. Singh, J. Phys. A **29**, L205 (1996).
[20] V. I. Rupasov and M. Singh, Phys. Rev. Lett. **77**, 338 (1996).
[21] V. I. Rupasov and M. Singh, Phys. Lett. A **222**, 258 (1996).
[22] V. I. Rupasov and M. Singh, Phys. Rev. A **56**, 898 (1997).
[23] M. R. Singh and W. Lau, Phys. Lett. A **231**, 115 (1997).
[24] M. R. Singh and W. Lau, Phys. Status Solidi B **203**, 401 (1997).
[25] N. Vats and S. John, Phys. Rev. A **58**, 4168 (1998).
[26] H. Huang, X.-H. Lu, and S.-Y. Zhu, Phys. Rev. A **57**, 4945 (1998).
[27] T. Quang, M. Woldeyohannes, S. John, and G. S. Agarval, Phys. Rev. Lett. **79**, 5238 (1997).
[28] D. I. Gudzenko and S. I. Yakovlenko, Zh. Éksp. Teor. Fiz. **62**, 1686 (1972) [Sov. Phys. JETP **35**, 877 (1972)].
[29] S. I. Yakovlenko, *Collision Effects in Radiation* [in Russian], Énergoatomizdat, Moscow (1984).
[30] A. M. Basharov, *Photonics. Method of Unitary Transformation in Nonlinear Optics* [in Russian], Izdatel'stvo MIFI, Moscow (1990).
[31] M. I. D'yakonov and V. I. Perel', Zh. Éksp. Teor. Fiz. **48**, 345 (1965) [Sov. Phys. JETP **21**, 227 (1965)].
[32] A. P. Kazantsev, Zh. Éksp. Teor. Fiz. **51**, 1751 (1966) [Sov. Phys. JETP **24**, 1183 (1966)].
[33] A. M. Basharov, Zh. Prikl. Spektrosk. **51**, 106 (1989).
[34] C. W. Gardiner, *Quantum noise*, Springer, Berlin (1991).
[35] A. M. Basharov, Zh. Éksp. Teor. Fiz. **102**, 1126 (1992) [Sov. Phys. JETP **75**, 611 (1992)].
[36] A. M. Basharov, Zh. Éksp. Teor. Fiz. **111**, 25 (1997) [JETP **84**, 13 (1997)].
[37] M. Lax, Phys. Rev. **145**, 110 (1966).

Translation was provided by the Russian Editorial office.
Edited by S. J. Amoretty

Generation of sub-Poissonian light by a four-level microlaser with a high-Q cavity

Yu. M. Golubev*)

Scientific Research Institute of Physics, St. Petersburg State University, 198904 St. Petersburg, Russia

B.-G. Englert

Max-Planck University, Garching, Germany

Hwang Lee and M. O. Scully†)

Texas A&M University, USA

H. Walther†)

Ludwig-Maximilian University, Munich, Germany

(Submitted 10 March 1999)
Zh. Éksp. Teor. Fiz. **116**, 485–502 (August 1999)

We develop a kinetic theory that describes the behavior of a monatomic four-level laser when the atom is fixed inside a high-Q optical cavity. Such a statement of the problem is similar to that used in the experiment of G. M. Meyer, H.-J. Briegel, and H. Walther [Europhys. Lett. **37**, 317 (1997)]. The condition that the number of photons is large and the photon fluctuations are small is employed. We show that by selecting the parameters of the periodic electromagnetic pulses exciting the atom one can achieve regular pumping of the upper laser level and generate sub-Poissonian laser light. We also discuss the reasons why the statistical pattern of the radiation differs from the micromaser pattern with regular injection of atoms.
© *1999 American Institute of Physics.* [S1063-7761(99)00908-7]

1. INTRODUCTION

The term ''microlaser'' is usually used in connection with systems in which a small number of atoms (or even one atom) interact with the laser mode. There are two ways to ensure the effectiveness of this interaction. First, by fixing a small number of active atoms in one way or another (say, by using an atomic or ionic trap) in an optical cavity and exciting them by an external electromagnetic field. Second, by using atomic fluxes with low concentrations of atoms as the active medium. Both approaches were realized in experiments conducted by Walther and co-workers.[1,2] Since the problem ceased to be purely speculative, it requires thorough physical substantiation and detailed theoretical study.

All attempts to explain the behavior of microlasers (micromasers) by well-known theories have met with formal difficulties.[3–5] This, for instance, was well illustrated by Briegel and Englert,[6] who found that generalizations may lead to such difficulties as, say, the appearance of negative diagonal elements in the density matrix of the lasing field. This is to be expected, however, since an important requirement used in deriving the kinetic equations in these papers was that the number of atoms participating in lasing be large. Indeed, there, at a certain stage, the field matrix acquired an increment over a long time interval Δt (the minimum duration required by the theoretical analysis) owing to the interaction between the laser mode and the atoms excited to the laser state in the same interval. Then the contribution to this increment of the atoms excited outside the interval was ignored, which is physically justified only if there are many excited atoms in the time interval Δt and they are distributed along the time axis more or less uniformly. Note that the cited papers deal with a large numbers of different atoms (rather than with a single atom excited many times), since the object under investigation was a two-level system and the excitation involved atoms from an external source. But what will happen if there is only one atom (or a small number of atoms) and we must allow not only for two laser levels but also for auxiliary levels from which the atom goes to the laser states and to which it later returns? Proper theoretical analysis can answer this question.

A possible approach to this problem was developed in Ref. 7 for micromasers. The researchers used a new method which, we believe, may be especially useful for microlasing involving a small number of photons. In the present paper we assume, in describing a microlaser, that the cavity has a high Q-factor and that there is a substantial buildup of photons during steady-state lasing in the laser mode. This makes it possible to use the approximation of small photon fluctuations, so that employing the common method of the kinetic equation for the density matrix of the lasing field appears to be the optimum approach.

One reason for our investigation is that we believe that at present there is no theory of a microlaser. Another reason is that a microlaser with pulsed pumping may prove to be a promising source of sub-Poissonian light. The reader will recall that the greatest achievements in this area of quantum electronics were made with a semiconductor laser.[8] However, the experimental difficulties associated with the use of

 © 1999 American Institute of Physics

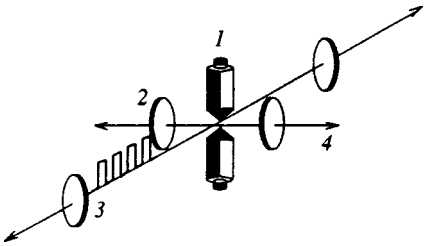

FIG. 1. The schematic of a thought experiment in which *1* is the ionic trap, *2* is the microlaser, *3* is the exciting pulsed laser, and *4* is the direction in which the laser light travels to a photodetector.

solid-state lasers are well known, so that other avenues of research that may prove promising must be investigated. In building the theory of a microlaser, we begin with the situation that was realized in the experiments described in Ref. 2. The researchers localized, via an ionic trap, a Ca^+ ion inside an optical cavity. Excitation was done by an external electromagnetic field. The Ca^+ ion can effectively by represented by a four-level ''atom.'' Our theory adopts all these assumptions, but we also assume that the excitation is pulsed, which makes to possible, by varying the pulse parameters, to realize different types of excitation of the active medium, from completely stochastic (Poissonian) to regular.

The physical situation discussed in this paper is similar to the one presented in Ref. 4: the object of investigation is the same, a four-level atom, and it is assumed that a powerful and short pulse of electromagnetic radiation excites the atom (or all the atoms) from the ground state through the upper state to the intermediate state, which is the upper laser state. However, the mathematic tools used in Ref. 4 and in the present paper are different. Here we examine the atom as a truly four-level atom, with not only the lasing field being incorporated into the picture but also the pump field. This allows us to meaningfully discuss the problem for any number of atoms, since any degree of excitation of the ground state can be taken into account.

At the same time, the four-level atom was needed in Ref. 4 only to justify the concept of regular excitation of the medium. After the regularity of pumping had been fully established, the two-level atom was taken as the base for calculations. Of course, it is well known that a four-level atom can be effectively reduced to a two-level atom with constant pumping to the upper laser level, but for this to be true we must assume that we can ignore the variations in the population of the ground state. It is clear from general considerations (and this is shown below) that in this case the excitation of the laser level can only be stochastic in nature. To ensure regularity we must require that all the atoms (or a single atom if there is only one atom, but with a probability equal to unity) simultaneously leave the ground state because of the pump pulse acting on them. And there is no way in which this situation can be reduced to that of a two-level atom under constant pumping.

2. A PHYSICAL MODEL OF A LASER

We assume that a single four-level atom is in a trap at the intersection of the axes of two optical cavities (Fig. 1).

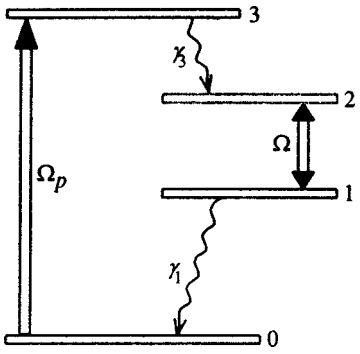

FIG. 2. Energy level diagram of the atom: Ω_p is the pump Rabi frequency, and Ω is the Rabi frequency of the output laser radiation.

One of the cavities belongs to the microlaser being studied, and the other is the cavity of an auxiliary pulsed laser used the excite the atom.

The level diagram of the atom is depicted in Fig. 2. The pulsed pump radiation with a Rabi frequency Ω_p is in resonance with the atomic transition 0–3. Together with spontaneous emission in the 3–2 transition that takes place with a rate γ_3, the pumping ensures that the upper laser level 2 is populated. When the threshold is reached, this ensures lasing in the 2–1 transition with the Rabi frequency Ω. Finally, spontaneous emission with a rate γ_1 again brings the atom to the ground state 0.

To simplify the situation mathematically, we adopt certain relationships that link the characteristic temporal parameters of our problem. First we require that t_p, the time it takes the pump pulse to transfer the atom from the ground state to the upper laser state, be the shortest time in the problem. Later we will show that this time obeys the formula

$$t_p=\frac{\gamma_3}{4\Omega_p^2}. \tag{1}$$

We also assume that the longest characteristic time interval is the lifetime of the laser mode, $\varkappa^{-1}$. It is this requirement that will enable us to set up the master equation for the density matrix of the lasing field.

The other characteristic times are T_p (the length of the pump pulse), T (the pulsed pumping period; see Fig. 3), Ω^{-1} (the Rabi period of lasing), and γ_1^{-1} (the time of spontaneous relaxation of the lower laser state).

Interacting with the external pulsed pump field and the vacuum fields that ensure spontaneous emission, the atom leaves (at a certain time) the ground state 0 and returns to

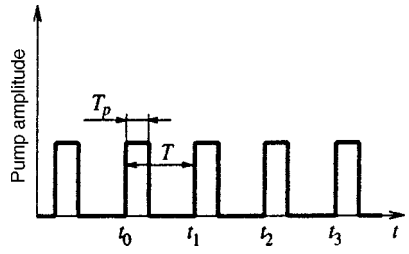

FIG. 3. The temporal pump pulse pattern: T_p is the length of the pulse, T is the pulsed pump period, and $t_k=t_0+kT$, where $k=0,\pm1,\pm2,\ldots$.

that state after a characteristic time has elapsed. We require here that this characteristic time be much shorter than the pump period T. This, in particular, means that immediately before the arrival of the next pump pulse the atom is sure to be in the ground state. At the same time, a requirement that plays an important role in the kinetic theory considered is that the evolution of the field in the cavity be much slower than the evolution of the atomic system. This means that the constant κ^{-1} must be much larger than all the characteristic atomic constants. More than that, to guarantee steady-state lasing under pulsed pumping, we require that

$$T \ll \kappa^{-1}. \tag{2}$$

If we also require that the pump pulse length T_p and the characteristic time t_p that it takes the upper laser level to become populated due to the pump pulse be shorter than the Rabi period Ω^{-1} of lasing,

$$T_p, t_p \ll \Omega^{-1}, \tag{3}$$

then in the course of these time intervals we can ignore all evolution of the atomic states related to the laser field. At the same time, for our specific purposes of sub-Poissonian lasing, we will assume that the laser field saturates the atomic transition and hence

$$\Omega^{-1} \ll \gamma_1^{-1}. \tag{4}$$

Collecting all these requirements, we arrive at the following inequalities:

$$T_p, t_p \ll \Omega^{-1} \ll \gamma_1^{-1} \ll T \ll \kappa^{-1}. \tag{5}$$

3. THE MASTER KINETIC EQUATION

According to the chosen physical conditions expressed by the inequalities (5), the evolution of the atomic variables proceeds much faster than that of the field. As noted earlier, this is achieved by selecting a high-Q optical cavity. Under these conditions the kinetic approximation is valid and, at least in principle, we can write the equation for the density matrix ρ of the lasing field. In general this equation can be written

$$\dot{\rho} = (\hat{L} + \kappa \hat{R})\rho. \tag{6}$$

The operator $\kappa \hat{R}$ is the damping of a field oscillator with a rate κ. In the Glauber representation, which we use in our theory, there is the well-know relationship

$$\hat{R} = \frac{\partial}{\partial u} u, \quad \alpha = \sqrt{u} \exp\{i\varphi\}. \tag{7}$$

Here α is the eigenvalue of the operator of annihilation of a laser photon:

$$a|\alpha\rangle = \alpha|\alpha\rangle, \quad [a, a^{\dagger}] = 1. \tag{8}$$

The operator $\hat{L}$ determines the laser evolution related to the interaction of the laser oscillator with the active medium, i.e., the single atom in our case. Strictly speaking, in steady-state lasing, i.e., after all transient processes have ceased, this operator is a periodic function of time: $\hat{L}(t) = \hat{L}(t+T)$. Since we agreed that the pump pulse is very short, we can ignore all evolution of the field at these moments and write

$$\hat{L} = \sum_k \mu(t - t_k)\hat{S}(t - t_k), \tag{9}$$

where t_k and $t_k + T_p$ $(k = \ldots,0,1,2,\ldots)$ is the beginning and end of the kth pump pulse (Fig. 3), and the function μ is defined as

$$\mu(t) = \begin{cases} 1 & \text{if } T_p < t < T, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

We study the physical situation involving many laser photons and ignore all field variations that occur during the pump period, since only one photon is added to the field during that time. This allows us to write

$$\overline{\rho} = \frac{1}{T} \int_{t-T/2}^{t+T/2} \rho(t')dt' = \rho(t). \tag{11}$$

Thus, after averaging Eq. (6) over the period T the equation becomes

$$\dot{\rho} = (\hat{\bar{S}} + \kappa \hat{R})\rho, \tag{12}$$

since

$$\hat{\bar{L}} = \frac{1}{T} \int_{t-T/2}^{t+T/2} \hat{L}(t')dt' = \frac{1}{T} \int_0^T \hat{S}(t')dt' = \hat{\bar{S}}. \tag{13}$$

Below we construct the operator $\hat{\bar{S}}$ explicitly.

4. THE EQUATION IN THE GLAUBER REPRESENTATION

What are the formal steps that must be taken to construct the operator $\hat{\bar{S}}$? First we must write, using the common approach, the starting equation for the density matrix $\hat{F}$ of the atom and the quantized lasing field. This equation must allow, among things, for the interaction between the atom and the pump field and for all important spontaneous relaxations (see Fig. 2).

Then we write the equation in the diagonal Glauber representation. As a result, the density matrix $\hat{F}$, still remaining an operator in relation to the atomic variables, becomes a c-number function in relation to the complex-valued amplitude α of the lasing field (and of the pump field, α_p): $\hat{F} = \hat{F}(\alpha_p, \alpha, t)$.

Next we factorize the density matrix: $\hat{F} = \hat{\sigma}\rho + \hat{\pi}$. Here $\hat{\sigma}$ is the density matrix of the atom interacting with the ''classical'' pump and lasing fields whose amplitudes are, respectively, α_p and α, ρ is the sought matrix of the lasing field, and π is the term responsible for the atomic–field correlation. We assume that the atomic matrix $\hat{\sigma}$ obeys a standard kinetic equation. All this makes it possible to write a complete system of equations for the matrices ρ and $\hat{\pi}$, which will depend on $\hat{\sigma}$ as a parameter.

The next step amounts to adopting a kinetic approximation with allowance for the fact that atomic motion in our conditions is much faster than field motion. More than that, in the approximation of a large number of photons we can also ignore the periodic dependence associated with the

pulsed pumping of the medium. Using the adiabaticity of the system and ignoring, in the first step, the dependence of ρ on t, we can use the equation for $\hat{\pi}$ to obtain the explicit dependence of $\hat{\pi}$ on ρ. Plugging this adiabatic solution into the equation for ρ and averaging the result over the period T, we obtain the sought expression for $\dot{\hat{S}}$.

Now we discuss each step in detail. A master equation that allows for two fields (a classical and a quantum) and spontaneous decays can be derived on the basis of very general ideas:

$$\dot{\hat{F}} = -i[\hat{V}, \hat{F}] - \hat{R}_{\text{at}}\hat{F}. \tag{14}$$

The Hamiltonian

$$\hat{V} = ig(a|1\rangle\langle 2| - a^{\dagger}|2\rangle\langle 1|) + ig_p(\alpha_p|0\rangle\langle 3| - \alpha_p^*|3\rangle\langle 0|) \tag{15}$$

allows for the interaction of the classical pump field with the amplitude α_p and the quantum lasing field represented by its photon operators a and $a^{\dagger}$ [Eq. (8)]. Here $|i\rangle\langle k|$ are the standard atomic projectors, and the operator $\hat{R}_{\text{at}}$ ensures atomic relaxation with rates γ_1 and γ_3 (see Fig. 2).

We use the Glauber representation in our theoretical discussions. Then Eq. (14) can be written

$$\dot{\hat{F}} = -i[\hat{V}_0, \hat{F}] - \hat{R}_{\text{at}}\hat{F} + g\hat{D}\hat{F}. \tag{16}$$

Now the first term on the right-hand side describes the interaction of the atom with only ''classical'' fields, since the Hamiltonian

$$\hat{V}_0 = ig(\alpha|1\rangle\langle 2| - \alpha^*|2\rangle\langle 1|) + ig_p(\alpha_p|0\rangle\langle 3| - \alpha_p^*|3\rangle\langle 0|) \tag{17}$$

contains the complex-value lasing amplitude α (and a^*) instead if the photon operators a and $a^{\dagger}$. The term

$$\hat{D}\hat{F} = \frac{\partial}{\partial\alpha}|1\rangle\langle 2|\hat{F} + \frac{\partial}{\partial\alpha^*}\hat{F}|2\rangle\langle 1| \tag{18}$$

is responsible for the formation of the statistical properties of lasing.

5. FACTORIZATION OF THE DENSITY MATRIX

Let us write the density matrix of the atom and the lasing field as follows:

$$\hat{F} = \hat{\sigma}\rho + \hat{\pi}, \tag{19}$$

where $\hat{\sigma}$ is the atomic density matrix, which describes the behavior of the atom in two ''classical'' fields with amplitudes α_p and α and allows for real spontaneous relaxations,

$$\dot{\hat{\sigma}} = -i[\hat{V}_0, \hat{\sigma}] - \hat{R}_{\text{at}}\hat{\sigma}, \tag{20}$$

and $\rho = \text{Tr}_{\text{at}}\hat{F}$ is the density matrix of the lasing field in the diagonal representation, which is the result of taking the trace of the total matrix. Since $\text{Tr}_{\text{at}}\hat{\sigma} = 1$, we have [according to (19)] the following relationship:

$$\text{Tr}_{\text{at}}\hat{\pi} = 0. \tag{21}$$

Combining all these formulas with the starting equation (16) yields

$$\dot{\rho} = g\frac{\partial}{\partial\alpha}(\sigma_{21}\rho + \pi_{21}) + \text{c.c.} \tag{22}$$

On the right-hand side we have terms of two types: those that are independent of $\hat{\pi}$ and describe semiclassical ''noiseless'' development of lasing, and those that are dependent on $\hat{\pi}$ and are responsible for the statistical aspect of the problem. Separating these terms, we can write Eq. (22) as

$$\dot{\rho} = g\overline{D}\rho + g\left(\frac{\partial}{\partial\alpha}\pi_{21} + \frac{\partial}{\partial\alpha^*}\pi_{12}\right). \tag{23}$$

Here $\overline{D} = \text{Tr}_{\text{at.}}(D\sigma)$ and, with allowance for the definition (18),

$$\overline{D}\rho = \hat{\delta}_+\frac{\sigma_{21}}{\alpha}\rho. \tag{24}$$

The operator $\hat{\delta}_+$ is defined as follows:

$$\hat{\delta}_+ = \frac{\partial}{\partial\alpha}\alpha + \frac{\partial}{\partial\alpha^*}\alpha^* = 2\frac{\partial}{\partial u}u. \tag{25}$$

In writing (23) we allowed for that fact that in our conditions $\sigma_{21}/\alpha = \sigma_{12}/\alpha^*$.

In the same way as we derived Eq. (23) we can derive an equation for $\hat{\pi}$:

$$\dot{\hat{\pi}} = -i[\hat{V}_0, \hat{\pi}] - \hat{R}_{\text{at}}\hat{\pi} + g(\hat{D} - \overline{D})\hat{\sigma}\rho + g[\hat{D}\hat{\pi} - \text{Tr}_{\text{at}}(\hat{D}\hat{\pi})\hat{\sigma}]. \tag{26}$$

6. THE ITERATION SERIES FOR THE CORRELATION MATRIX $\hat{\pi}$

To express $\hat{\pi}$ in terms of ρ, we must use Eq. (26). Since this equation cannot be solved directly, we employ an iteration procedure. To this end we assume that the principal terms in Eq. (26) are those that contain no derivatives with respect to the complex-valued amplitudes of the lasing field. The terms containing such derivatives are small compared to the principal terms. Bearing this smallness in mind, we can write $\hat{\pi}$ in the form of a series in the power of this ''smallness'':

$$\hat{\pi} = \hat{\pi}_0 + \hat{\pi}_1 + \hat{\pi}_2 + \cdots. \tag{27}$$

The zeroth approximation $\hat{\pi}_0$ obeys the equation

$$\dot{\hat{\pi}}_0 = -i[\hat{V}_0, \hat{\pi}_0] - R_{\text{at}}\hat{\pi}_0. \tag{28}$$

We can always require that initially there be no correlation between the atom and the lasing field. But then, in view of the homogeneity of the differential equation (28), the correlation matrix $\hat{\pi}_0(t) = 0$, provided that $\hat{\pi}_0(0) = 0$.

Allowing for all this, we can write the equation for the first iteration π_1:

$$\dot{\hat{\pi}}_1 = -i[\hat{V}_0, \hat{\pi}_1] - \hat{R}_{\text{at}}\hat{\pi}_1 + g(\hat{D} - \overline{D})\sigma\rho. \tag{29}$$

Similar equations can be obtained for all subsequent iterations, and more than that, there are no grounds for ignoring them at the given stage. Such ground could arise if we were certain that we are dealing with classical fields, whose statistical properties are described adequately even by phenom-

enological Maxwell equations with stochastic sources. In our case, however, among other things, we will attempt to describe situations in which the lasing field becomes sub-Poissonian. Hence we are forced to retain not only the first iteration, which contains the first derivatives with respect to the complex-valued amplitudes of the lasing field, but also the iterations of all orders, thus allowing for the derivatives of all orders with respect to the complex-valued amplitudes. Nevertheless, we will show that even for nonclassical fields the higher-order iterations contribute nothing to the spectrum of photocurrent fluctuations.

Let us analyze Eq. (29). Allowing for the fact that we are interested only in solution in the interval between the pulses that excite the active medium, $t_i+T_p<t<t_i+T$, since in our conditions

$$\pi_{ik}=\sum_i \mu(t-t_i)\,\pi(t-t_i-T_p), \tag{30}$$

we can write Eq. (29) in terms of the matrix elements:

$$\begin{aligned}
&\dot{\pi}_{22}=g(\alpha\pi_{12}+\pi_{21}\alpha^*)+gS_{22},\\
&\dot{\pi}_{11}=-\gamma_1\pi_{11}-g(\alpha\pi_{12}+\pi_{21}\alpha^*)+gS_{11},\\
&\dot{\pi}_{12}=-\frac{\gamma_1}{2}\pi_{12}-g(\alpha^*\pi_{22}-\pi_{11}\alpha^*)+gS_{12},\\
&\dot{\pi}_{21}=-\frac{\gamma_1}{2}\pi_{21}-g(\alpha\pi_{22}-\pi_{11}\alpha)+gS_{21}.
\end{aligned} \tag{31}$$

This system must be augmented by the normalization condition (21):

$$\pi_{00}+\pi_{11}+\pi_{22}+\pi_{33}=0. \tag{32}$$

The inhomogeneous terms

$$S_{ik}=[(D\sigma)_{ik}-\overline{D}\sigma_{ik}]\rho \tag{33}$$

can be written explicitly in the form

$$S_{00}=-\hat{\delta}_+\frac{\sigma_{00}\sigma_{21}}{\alpha}\rho,\quad S_{11}=\hat{\delta}_+\frac{(1-\sigma_{11})\sigma_{21}}{\alpha}\rho, \tag{34}$$

$$\begin{aligned}
&S_{22}=-\hat{\delta}_+\frac{\sigma_{22}\sigma_{21}}{\alpha}\rho,\quad S_{33}=-\hat{\delta}_+\frac{\sigma_{33}\sigma_{21}}{\alpha}\rho,\\
&S_{03}=-\hat{\delta}_+\frac{\sigma_{03}\sigma_{21}}{\alpha}\rho,\quad S_{12}=-\hat{\delta}_+\frac{\sigma_{12}\sigma_{21}}{\alpha}\rho+\frac{\partial}{\partial\alpha}\sigma_{22}\rho.
\end{aligned} \tag{35}$$

We see that to write the inhomogeneous terms in the final form, we must know the explicit expressions for the elements of the atomic density matrix, σ_{ik}, also in the time interval between the pump pulses. In Appendix A we find these elements, which have the following explicit dependence:

$$\begin{aligned}
\sigma_{11}(t-t_i-T_p)=\sigma_0\frac{4\Omega^2}{\omega^2}\sin^2\left[\frac{\omega}{2}(t-t_i-T_p)\right]\\
\times\exp\left[-\frac{\gamma_1}{2}(t-t_i-T_p)\right],
\end{aligned}$$

$$\begin{aligned}
\sigma_{22}(t-t_i-T_p)=\sigma_0\left\{\frac{4\Omega^2}{\omega^2}\cos^2\frac{\omega}{2}(t-t_i-T_p)\right.\\
\times\exp\left[-\frac{\gamma_1}{2}(t-t_i-T_p)\right]+\frac{\gamma_1}{2\omega}\\
\times\sin\omega(t-t_i-T_p)-\frac{\gamma_1^2}{4\omega^2}\cos^2\omega\\
\left.\times(t-t_i-T_p)\right\}\exp\left[-\frac{\gamma_1}{2}(t-t_i-T_p)\right],
\end{aligned} \tag{36}$$

$$\begin{aligned}
\sigma_{12}(t-t_i-T_p)=-\sigma_0\frac{g\alpha^*}{\omega}\left[\frac{\gamma_1}{\omega}\sin^2\frac{\omega}{2}(t-t_i-T_p)\right.\\
\left.+\sin\omega(t-t_i-T_p)\right]\\
\times\exp\left[-\frac{\gamma_1}{2}(t-t_i-T_p)\right].
\end{aligned}$$

The quantity

$$\sigma_0\equiv\sigma_{22}(T_p-t_i)=1-\exp\left(-\frac{T_p}{t_p}\right) \tag{37}$$

is population of the upper laser level by the time the pump pulse has ceased to act.

We see that the Rabi oscillations have the frequency

$$\omega=\frac{\gamma_1}{2}\sqrt{1+I},\quad I=\frac{16\Omega^2}{\gamma_1^2}. \tag{38}$$

In Appendix B the matrix elements π_{ik} are calculated explicitly under the assumption that the periodic function $\rho(t)$ is ignored. The matrix element π_{12} we are interested in has the form

$$\pi_{12}=\hat{\delta}_+\frac{1}{\alpha}F_1(t)\rho+\hat{\delta}_-\frac{1}{\alpha}F_2\rho. \tag{39}$$

In addition to the operator $\hat{\delta}_+$, we introduce another differential operator,

$$\hat{\delta}_-=\frac{\partial}{\partial\alpha}\alpha-\frac{\partial}{\partial\alpha^*}\alpha^*=-i\frac{\partial}{\partial\varphi}, \tag{40}$$

and the notation

$$\begin{aligned}
F_1(t)=\frac{\Omega}{2}\left\{\frac{(1-\sigma_{11})\sigma_{21}}{\alpha}\right\}\otimes\left\{\left(\sin\omega t-\frac{1}{\sqrt{I}}\sin^2\frac{\omega t}{2}\right)\right.\\
\left.\times\exp\left(-\frac{\gamma_1 t}{2}\right)\right\}+\frac{\Omega}{2}\left\{\frac{\sigma_{22}\sigma_{21}}{\alpha}\right\}\\
\otimes\left\{\left(\sin\omega t+\frac{1}{\sqrt{I}}\sin^2\frac{\omega t}{2}\right)\exp\left(-\frac{\gamma_1 t}{2}\right)\right\},
\end{aligned} \tag{41}$$

$$F_2(t)=\frac{g}{2}\{\sigma_{22}\}\otimes\left\{\exp\left(-\frac{\gamma_1 t}{2}\right)\right\}, \tag{42}$$

and define the convolution of two functions:

$$\{g_1(t)\}\otimes\{g_2(t)\}=\int_0^t g_1(t')g_2(t-t')dt'. \tag{43}$$

7. THE MASTER EQUATION AVERAGED OVER THE PERIOD

Now we can write Eq. (23) explicitly:

$$\dot\rho=\frac{\partial}{\partial u}u\left[\kappa+\sum_i \mu(t-t_i)g\,\sigma_{21}(t-t_i)/\alpha\right]\rho + \sum_i \mu(t-t_i)\left[-\left(\frac{\partial}{\partial u}\right)^2 2uF_1(t-t_i-T_p) + \frac{1}{u}\frac{\partial^2}{\partial\varphi^2}F_2(t-t_i-T_p)\right]\rho+\{\cdots\}. \tag{44}$$

Here $\{\cdots\}$ stands for terms containing high-order derivatives with respect to u and φ, which follow from the iteration series (27).

We see that the coefficients in Eq. (44) are functions of time, in accordance with the fact that the pumping of the resonant medium is a function of time. The reader will recall that we have limited ourselves to the approximation of large photon numbers in the lasing mode and thus can ignore all field variations over one pump period since at most only one photon is added to this field in this time interval. This means that Eq. (44) can be averaged over the pump period, in view of which it acquires the form

$$\dot\rho=\frac{\partial}{\partial u}u\left(\kappa+\frac{g\overline{\sigma}_{21}}{\alpha}\right)\rho-\left(\frac{\partial}{\partial u}\right)^2 2u\overline{F}_1(u)\rho + \frac{1}{u}\frac{\partial^2}{\partial\varphi^2}\overline{F}_2(u)\rho+\{\cdots\}. \tag{45}$$

In the case of strong lasing ($\Omega\gg\gamma_1$) we have

$$\overline{F}_1(u)=-\frac{\sigma_0}{2gT}\frac{1}{u}\left(\frac{n_s}{u}+\frac{\sigma_0}{2}\right),\quad n_s=\frac{\gamma_1^2}{4g^2}, \tag{46}$$

$$\overline{F}_2(u)=\frac{g\sigma_0}{\gamma_1^2 T}. \tag{47}$$

To calculate the functions

$$\overline{F}_i(u)=\frac{1}{T}\int_0^T F_i(t)\,dt,\quad i=1,2, \tag{48}$$

we must use the explicit expressions (41) and (42). Moreover, mathematically we can simplify matters if we bear in mind that the period T is much longer than atomic times. Then, in averaging, integration with respect to times from 0 to T can be replaced by integration from 0 to ∞, which yields

$$\overline{\{g_1(t)\}\otimes\{g_2(t)\}}=T\overline{g}_1\overline{g}_2. \tag{49}$$

Doing the necessary integrations, we arrive at

$$\overline{\sigma_{21}}=-\frac{2g\alpha\sigma_0}{\omega^2 T}\left(1-\frac{\gamma_1^2}{8\omega^2}\right)\xrightarrow{\Omega\gg\gamma_1}-\frac{g\alpha\sigma_0}{2\Omega^2 T},$$

$$\overline{(1-\sigma_{11})\sigma_{21}}=-\frac{2g\sigma_0}{\omega^2 T}\left[1-\frac{\gamma_1^2}{8\omega^2}-\sigma_0\frac{3\Omega^2}{2\omega^2}\left(1-\frac{5\gamma_1^2}{8\omega^2}\right)\right]\xrightarrow{\Omega\gg\gamma_1}-\frac{g\sigma_0}{2\Omega^2 T}\left(1-\frac{3}{8}\sigma_0\right),$$

$$\overline{\sigma_{22}\sigma_{21}}=-\frac{g\sigma_0^2}{\omega^2 T}\left[\frac{1}{4}+\frac{3\Omega^2}{\omega^2}+\frac{\gamma_1}{16\omega}\left(1+\frac{\gamma_1}{\omega}-\frac{2\Omega^2}{\omega^2}\frac{\gamma_1}{\omega}-\frac{7\gamma_1^2}{\omega^2}\right)\right]\xrightarrow{\Omega\gg\gamma_1}-\frac{g\sigma_0^2}{4\Omega^2 T},$$

$$\overline{\sigma}_{22}=\frac{\sigma_0}{\gamma_1 T}\left[\frac{4\Omega^2}{\omega^2}+\frac{\gamma_1^2}{\omega^2}\left(\frac{1}{2}+\frac{\Omega^2}{\omega^2}-\frac{\gamma_1^2}{8\omega^2}\right)\right]\xrightarrow{\Omega\gg\gamma_1}\frac{\sigma_0}{\gamma_1 T}. \tag{50}$$

8. THE APPROXIMATION OF SMALL PHOTON FLUCTUATIONS

Now, to simply the equation still further, we adopt the requirement that the photon fluctuations be small:

$$u=n+\varepsilon,\quad n\gg\varepsilon, \tag{51}$$

where n is the semiclassical steady-state solution of the laser problem for the number of photons. Clearly, this requirement can be met only if $n\gg1$.

The semiclassical truncated equation for n can be written as follows. In the master equation (45) we drop all derivatives with respect to the field variables except the first-order derivatives. Then we multiply the modified equation by u and integrate the results over all the values of u and φ. We get result is

$$\dot n=-(\kappa+g\overline{\sigma}_{21}/\alpha)n=0. \tag{52}$$

Thus, the condition for steady-state lasing is

$$\kappa=\frac{2g^2\sigma_0}{\omega^2 T}\left(1-\frac{\gamma_1^2}{8\omega^2}\right). \tag{53}$$

A useful condition in the case of lasers is the threshold condition

$$\frac{4g^2\sigma_0}{\gamma_1^2 T}=\kappa. \tag{54}$$

Now we can employ condition (51) and, keeping only the principal terms, write the master kinetic equation in a saturating laser field ($\Omega\gg\gamma_1$) in the form

$$\frac{\partial\rho}{\partial t}=\kappa\frac{\partial}{\partial\varepsilon}\varepsilon\rho+\kappa n\xi\frac{\partial^2\rho}{\partial\varepsilon^2}+\frac{\kappa}{4n_s}\frac{\partial^2\rho}{\partial\varphi^2}+\{\cdots\}, \tag{55}$$

where

$$\xi=-\frac{1}{2}\left(\frac{\gamma_1^2}{2\Omega^2}+\sigma_0\right) \tag{56}$$

is the statistical Mandel parameter, which is a characteristic of intracavity photon fluctuations: $\overline{\Delta n^2}=n(1+\xi)$.

9. THE PHOTOCURRENT SPECTRUM IN THE DETECTION OF MICROLASER RADIATION

In the photodetection of single-mode lasing, we can use a formula for the photocurrent spectrum in the Glauber representation in the approximation of small photon fluctuations:

$$i_\omega^{(2)} = i_{\text{shot}}^{(2)}\left(1 + \frac{2\kappa}{n}\,\text{Re}\int_0^\infty \overline{\varepsilon\varepsilon(t)}\,\exp\{i\omega t\}dt\right). \tag{57}$$

Applying standard mathematical techniques, we can use the master equation (55) to obtain the differential equation

$$\dot{\overline{\varepsilon\varepsilon(t)}} = -\kappa\overline{\varepsilon\varepsilon(t)}. \tag{58}$$

Its solution

$$\overline{\varepsilon\varepsilon(t)} = \overline{\varepsilon^2}\exp\{-\kappa t\} \tag{59}$$

contains the unknown quantity $\overline{\varepsilon^2}$, which can be found by solving an equation that follows from (55):

$$\dot{\overline{\varepsilon^2}} = -2\kappa\overline{\varepsilon^2} + 2\kappa\xi n = 0. \tag{60}$$

Note that the terms in Eq. (55) denoted by $\{\cdots\}$ and containing higher-order derivatives with respect to the complex variables of the field contribute nothing to (58) and (60), and this conclusion is independent of the type of mathematical approximation employed.

Performing the required mathematical manipulations, we can easily obtain an expression for the photocurrent spectrum:

$$i_\omega^{(2)} = i_{\text{shot}}^{(2)}\left(1 - 2|\xi|\,\frac{\kappa^2}{\kappa^2+\omega^2}\right). \tag{61}$$

From the qualitative point of view it is clear that there is a dip in the spectrum at near-zero frequencies against the background of shot noise, which is frequency-independent. As is known, this dip indicates the quantum nature of light, which manifests itself in a specific physical situation. The depth of the dip is determined by the Mandel parameter ξ, whose value (56) depends on the efficiency of the excitation of the atom to the upper laser level in the course of one pump pulse. We already know that the excitation efficiency is given by the quantity σ_0 of (37). Clearly, if $T_p \gg t_p$, then $\sigma_0 \approx 1$ and $\xi = -1/2$. Thus, if we select the parameters of the pump pulse so that by the end of the action of the pulse the atom is sure to be on the upper laser level, regular pumping of the laser medium is guaranteed, which leads to sub-Poissonian lasing with maximum suppression (down to zero) of shot noise at zero frequency. In the opposite case ($T_p \ll t_p$), pumping stops being reliable and the lasing becomes Poissonian, since suppression of the near-zero frequency shot noise can be ignored. These physical results are similar to those obtained in Ref. 4 for lasers. Thus, we have corroborated the assumption made in the Introduction that such a microlaser can serve as a source of sub-Poissonian light with maximum manifestation of the quantum properties.

At the same time we see (even at the level of the master quantum equation) that this system is in every respect unlike a maser,[5] which comes as a surprise if we reason on the basis of the purely corpuscular phenomenology of light. Indeed, a microlaser with periodic excitation of the atom and a maser in which the atom periodically arrives in the cavity do the same thing, i.e., periodically excite the laser mode by adding exactly one photon to it. From the viewpoint of corpuscular phenomenology, there should be no difference in the statistics of lasing. We see, however, that this is not the case. The reason is that photons are not simply small ''balls''—they are quantum mechanical particles described, say, by wave functions. In the case of a maser, when we can ignore relaxation from the laser levels and use the mathematical model of Jaynes and Cummings,[9] these wave functions differ from those in the case of a single atom, where relaxation is important. We see that not only the regularity with which new photons appear in the cavity is important in the formation of the correlation properties of lasing but so are the quantum mechanical states of these photons.

To be able to estimate the results correctly, we must understand at what values of the physical parameters will the conditions (5) be met. Clearly, the values $\kappa \sim 10^{15}\,\text{s}^{-1}$, $g \sim 10^8\,\text{s}^{-1}$, $T \sim 10^{-7}\,\text{s}$, and $\gamma_1 \sim 10\,\text{s}^{-1}$ ensure that the conditions (5) are met. When making the estimates, we must bear in mind that, according to (53), in the saturation regime ($g\sqrt{n} \gg \gamma_1$) we have $n \sim (\kappa T)^{-1}$. This physical situation is an improvement in the sense of the Q-factor of the cavity and the efficiency of the interaction between the laser wave and the medium compared to the case considered, say, in Ref. 10, where it is assumed that $k \sim 10^6\,\text{s}^{-1}$ and $g \sim 10^7\,\text{s}^{-1}$.

This work was made possible by Grants from INTAS-RFBR (Grant 95-0656), the Russian Fund for Fundamental Research (Grant 98-02-18129), and the State Committee for Institutions of Higher Learning (Grant 95-0-5.4-66).

APPENDIX A: BEHAVIOR OF A FOUR-LEVEL ATOM IN TWO CLASSICAL EXTERNAL FIELDS, THE PULSED PUMP FIELD AND THE LASING FIELD

Let us assume that we are dealing with a four-level atom whose energy level diagram is depicted in Fig. 2. Pulsed coherent pumping with a complex-valued amplitude α_p involves only the 0–3 transition of the atom, while the lasing field with the complex-valued amplitude α_p involves the 1–2 transition of the atom. All this is expressed by an interaction Hamiltonian written in the dipole approximation,

$$V_0 = ig(\alpha|1\rangle\langle 2| - \alpha^*|2\rangle\langle 1|) + ig_p(\alpha_p|0\rangle\langle 3| - \alpha_p^*|3\rangle\langle 0|), \tag{62}$$

where the $|i\rangle\langle k|$ are the standard projection operators.

Bearing in mind that there is spontaneous relaxation in the transitions 3–2 and 1–0, we can write the following system of equations for the elements of the atomic density matrix:

$$\dot{\sigma}_{00} = -g_p(\alpha_p^*\sigma_{30} + \alpha_p\sigma_{03}) + \gamma_1\sigma_{11},$$

$$\dot{\sigma}_{33} = g_p(\alpha_p^*\sigma_{30} + \alpha_p\sigma_{03}) - \gamma_3\sigma_{33},$$

$$\dot{\sigma}_{03} = -g_p\alpha_p^*(\sigma_{33} - \sigma_{00}) - \frac{\gamma_3}{2}\sigma_{03},$$

$$\begin{aligned}
\dot\sigma_{30}&=-g_p\alpha_p(\sigma_{33}-\sigma_{00})-\frac{\gamma_3}{2}\sigma_{30},\\
\dot\sigma_{11}&=-g(\alpha^*\sigma_{21}+\alpha\sigma_{12})-\gamma_1\sigma_{11},\\
\dot\sigma_{22}&=g(\alpha^*\sigma_{21}+\alpha\sigma_{12})+\gamma_3\sigma_{33},\\
\dot\sigma_{12}&=-g\alpha^*(\sigma_{22}-\sigma_{11})-\frac{\gamma_1}{2}\sigma_{12},\\
\dot\sigma_{21}&=-g\alpha(\sigma_{22}-\sigma_{11})-\frac{\gamma_1}{2}\sigma_{21}.
\end{aligned}\tag{63}$$

Since we have chosen the pumping to be pulsed with a period T, the steady-state solution of this system of equations will also be a periodic function of time with the same period. Hence we can find the solution in any one time interval in which the pump pulse acts and in any time interval between pump pulses. The complete solution of the problem can be written

$$\sigma_{ik}=\sum_i\,[(1-\mu(t-t_i))\sigma_{ik}(t-t_i)+\mu(t-t_i)\sigma_{ik}(t-t_i-T_p)],\tag{64}$$

where we have introduced the notation

$$\mu(t)=\begin{cases}1 & \text{if } T_p<t<T,\\ 0 & \text{otherwise.}\end{cases}\tag{65}$$

A1. Evolution of the atomic states during the action of a pump pulse ($0<t<T_p$)

Taking into account the initial inequalities (5), we can write the system of equations (63) as follows:

$$\begin{aligned}
\dot\sigma_{00}&=-g_p(\alpha_p^*\sigma_{30}+\alpha_p\sigma_{03}),\\
\dot\sigma_{33}&=g_p(\alpha_p^*\sigma_{30}+\alpha_p\sigma_{03})-\gamma_3\sigma_{33},\\
\dot\sigma_{03}&=-g_p\alpha_p^*(\sigma_{33}-\sigma_{00})-\frac{\gamma_3}{2}\sigma_{03},\\
\dot\sigma_{30}&=-g_p\alpha_p(\sigma_{33}-\sigma_{00})-\frac{\gamma_3}{2}\sigma_{30},\\
\dot\sigma_{11}&=0,\quad \dot\sigma_{22}=\gamma_3\sigma_{33},\quad \dot\sigma_{12}=0,\quad \dot\sigma_{21}=0.
\end{aligned}\tag{66}$$

We assume that during the action of a pump pulse there only be a transition (partial or complete) of the atom from the ground state to the upper laser state. Since our conditions (5) ensure that before each new pulse the atom finds itself in the ground state, we have the unique finite initial condition

$$\sigma_{00}(0)=1.\tag{67}$$

In our approximation of a fairly weak and short pump pulse, we can write the following finite solutions:

$$\sigma_{00}(t)=\exp\left\{-\frac{t}{t_p}\right\},\quad \sigma_{00}(t)=1-\exp\left\{-\frac{t}{t_p}\right\},\tag{68}$$

where

$$t_p=\frac{\gamma_3}{\Omega_p^2}\tag{69}$$

is the characteristic time in the course of which a pump pulse transfers the atom from the ground state to the upper laser state.

A2. Evolution of the atomic density matrix in the time interval between pump pulses ($T_p<t<T$)

Now the system of equations for the elements of the atomic density matrix assumes the form

$$\begin{aligned}
\dot\sigma_{11}&=-g(\alpha^*\sigma_{21}+\alpha\sigma_{12})-\gamma_1\sigma_{11},\\
\dot\sigma_{22}&=g(\alpha^*\sigma_{21}+\alpha\sigma_{12}),\\
\dot\sigma_{12}&=-g\alpha^*(\sigma_{22}-\sigma_{11})-\frac{\gamma_1}{2}\sigma_{12},\\
\dot\sigma_{21}&=-g\alpha(\sigma_{22}-\sigma_{11})-\frac{\gamma_1}{2}\sigma_{21,}
\end{aligned}\tag{70}$$

with the finite initial condition

$$\sigma_{22}(T_p)=\sigma_0.\tag{71}$$

We go over to the Laplace transformation:

$$x_{ik}(p)=\int_0^\infty dt\sigma_{ik}(t)\exp\{-pt\}.\tag{72}$$

For the quantities $x_{ik}(p)$ we can now write a system of algebraic equations,

$$\begin{aligned}
(\gamma_1+p)x_{11}&=-g(\alpha^*x_{21}+\alpha x_{12}),\\
-\sigma_0+px_{22}&=g(\alpha^*x_{21}+\alpha x_{12}),\\
(\gamma_1/2+p)x_{12}&=-g\alpha^*(x_{22}-x_{11}),\\
(\gamma_1/2+p)x_{21}&=-g\alpha(x_{22}-x_{11}),
\end{aligned}\tag{73}$$

which can be solved relatively easily:

$$\begin{aligned}
x_{11}(p)&=\sigma_0\frac{2\Omega^2}{(p+\gamma_1/2)(p+\gamma_1/2-i\omega)(p+\gamma_1/2+i\omega)},\\
x_{22}(p)&=\sigma_0\frac{2\Omega^2+(p+\gamma_1)(p+\gamma_1/2)}{(p+\gamma_1/2)(p+\gamma_1/2-i\omega)(p+\gamma_1/2+i\omega)},\\
x_{12}(p)&=x_{21}^*(p)\\
&=-\sigma_0\frac{g\alpha^*(p+\gamma_1)}{(p+\gamma_1/2)(p+\gamma_1/2-i\omega)(p+\gamma_1/2+i\omega)}.
\end{aligned}\tag{74}$$

As before, by

$$\omega=\sqrt{4\Omega^2-\frac{\gamma_1^2}{4}}\tag{75}$$

we have denoted the Rabi frequency, with which the matrix elements oscillate in a strong field when $4\Omega>\gamma_1$.

Using the inverse Laplace transformation, we obtain

$$\sigma_{11}(t)=\sigma_0\frac{4\Omega^2}{\omega^2}\sin^2\frac{\omega t}{2}\exp\left(-\frac{\gamma_1 t}{2}\right),$$

$$\sigma_{22}(t)=\sigma_0\left[\frac{4\Omega^2}{\omega^2}\cos^2\frac{\omega t}{2}\exp\left(-\frac{\gamma_1}{2t}\right)+\frac{\gamma_1}{2\omega}\sin\omega t\,\frac{\gamma_1^2}{4\omega^2}\cos\omega t\right]\exp\left(-\frac{\gamma_1 t}{2}\right), \tag{76}$$

$$\sigma_{12}(t)=\sigma_{21}^*(t)=-\sigma_0\frac{g\alpha^*}{\omega}\left(\frac{\gamma_1}{\omega}\sin^2\frac{\omega t}{2}+\sin\omega t\right)\times\exp\left(-\frac{\gamma_1 t}{2}\right).$$

APPENDIX B: AN ANALYTIC SOLUTION FOR THE CORRELATION MATRIX $\hat{\pi}$

We seek the solution of the system of equations (31) via the Laplace transformation:

$$x_{ik}(p)=\int_0^\infty \pi_{ik}(t)\exp\{-pt\}dt. \tag{77}$$

Allowing for the fact that all the initial data are zeros, in the Laplace representation we arrive at a system of algebraic equations:

$$\begin{aligned}
&px_{22}=g(\alpha x_{12}+x_{21}\alpha^*)+gS_{22}(p),\\
&(\gamma_1+p)x_{11}=-g(\alpha x_{12}+x_{21}\alpha^*)+gS_{11}(p),\\
&\left(\frac{\gamma_1}{2}+p\right)x_{12}=-g\alpha^*(x_{22}-x_{11})+gS_{12}(p),\\
&\left(\frac{\gamma_1}{2}+p\right)x_{21}=-g\alpha(x_{22}-x_{11})+gS_{21}(p).
\end{aligned} \tag{78}$$

The reader will recall that we use the adiabatic approximation and ignore $\rho(t)$. Here

$$S_{ik}(p)=\int_0^\infty S_{ik}(t)\exp\{-pt\}dt. \tag{79}$$

For our purposes we are interested only in the matrix element

$$x_{12}(p)=S_{12}(p)\frac{g}{p+\gamma_1/2}+g^2\alpha^*\{S_{11}(p)p-S_{22}(p)(p+\gamma_1)-2[\alpha S_{12}(p)+S_{21}(p)\alpha^*]\}Q^{-1}(p), \tag{80}$$

where

$$Q(p)=\left(p+\frac{\gamma_1}{2}\right)\left(p+\frac{\gamma_1}{2}-i\omega\right)\left(p+\frac{\gamma_1}{2}+i\omega\right). \tag{81}$$

Performing the inverse Laplace transformation, we arrive at the sought time dependence in form (39).

*)E-mail: yuri-golubev@pop.convey.ru
†)Max-Planck Institute, Garching, Germany

[1] D. Meschede, H. Walther, and G. Müller, Phys. Rev. Lett. **54**, 551 (1985).
[2] G. M. Meyer, H.-J. Briegel, and H. Walther, Europhys. Lett. **37**, 317 (1997).
[3] M. O. Scully and W. E. Lamb, Jr., Phys. Rev. A **159**, 208 (1967).
[4] Yu. M. Golubev and I. V. Sokolov, Zh. Éksp. Teor. Fiz. **87**, 408 (1984) [Sov. Phys. JETP **60**, 234 (1984)].
[5] P. Filipowicz, J. Javanainen, and P. Meystre, Opt. Commun. **58**, 327 (1986); Phys. Rev. A **34**, 3077 (1986); J. Opt. Soc. Am. B **3**, 906 (1986); L. A. Lugiato, M. O. Scully, and H. Walther, Phys. Rev. A **36**, 740 (1987).
[6] H.-J. Briegel and B.-G. Englert, Phys. Rev. A **52**, 2361 (1995).
[7] B.-G. Englert, T. Gantsog, A. Schenzle, Ch. Wagner, and H. Walther, Phys. Rev. A **53**, 4386 (1996).
[8] Y. Yamamoto, S. Machida, and O. Nilson, Phys. Rev. A **34**, 4025 (1986); W. H. Richardson, Y. Yamamoto, and S. Machida, Phys. Rev. Lett. **66**, 2867 (1991).
[9] E. T. Jaynes and F. W. Cummings, Proc. IEEE **51**, 89 (1963).
[10] Mu Yi and C. M. Savage, Phys. Rev. A **46**, 5944 (1992).

Translated by Eugene Yankovsky

Theory of the field-controlled scintillation detector

M. I. Ryazanov*)

Moscow State Engineering Physics Institute (Technical University), 115409 Moscow, Russia
(Submitted 10 March 1999)
Zh. Éksp. Teor. Fiz. **116**, 503–510 (August 1999)

It is shown theoretically that application of an external electromagnetic field with specially selected frequencies in a dielectric containing impurity atoms allows a scintillation process of emission from impurity atoms to be activated which does not involve phonons. The process is similar to that in a conventional scintillator except that the field is completely controlled. The behavior of the emission process can change radically as the amplitude and frequency of the external field are varied, thus providing a means for wide-range control of such a ''field scintillator.'' © *1999 American Institute of Physics.* [S1063-7761(99)01008-2]

1. INTRODUCTION

The near field of a radiating system of charges is known to be predominantly longitudinal, so that in dense dielectrics an excited atom transfers energy to an unexcited atom through a longitudinal field by dipole–dipole interaction rather than through the emission and absorption of transverse waves. For this reason an atom residing in the interior of a dense dielectric cannot emit transverse waves, and the atomic excitation energy migrates through the material for a long time. In a scintillation detector with excitonic energy transfer the excitation energy is usually transported to an emission center,[1,2] for example, an impurity atom whose excitation frequency is lower than the excitation frequency of an atom of the main (host) material and lies outside the absorption band, allowing radiation to propagate freely from any point of the dielectric to the outside. The excess energy is spent in the generation of phonons, greatly reducing the probability of the reverse process and leading to incoherent emission from different impurity atoms. The concentration of scintillating impurities must be small enough that the dipole–dipole transfer of excitation between impurities suppresses the emission process.

Since impurity atoms transport the excitation energy of a material this way only with phonon production, it is difficult to control the emission process. However, the requisite conditions can be created for instigation of another emission process similar to that in an ordinary scintillator, except that it is entirely field-controlled. For this it is sufficient for the scintillator to be exposed to an external electromagnetic field with a specially selected frequency. The new process differs from the conventional version in that the excess energy is not transported by phonons, but by the external field.

The nature of the emission process can change drastically as the amplitude and frequency of the external field are varied, permitting total control of the process.

The theory of the field scintillator and the assessment of its properties and the necessary conditions for its creations are of interest.

2. IMPURITY ATOM IN AN EXCITED MATERIAL PARTICLE

We consider a dielectric with a small concentration of impurity atoms, whose electronic states have energies $E_1 < E_2 < E_3$. Assume that 0–3 transitions between the ground state with energy E_0 and the E_3 state are forbidden and 1–2 transitions between E_1 and E_2 states are forbidden as well, and let 0–1, 1–3, 3–2, and 2–0 transitions be allowed. We choose impurity atoms such that the energy of the first excited level of a host atom coincides with the energy E_2 of the second excited level of an impurity atom. A high-energy particle transmitted through the dielectric excites atoms, the most populated level being the first excited level of the host atoms. The excitation energy then migrates through the material via resonant dipole–dipole interaction between atoms of the material.

The second excited energy level of an impurity atom participates in the migration of excitation energy to the outside together with host atoms and therefore becomes populated as a result of transmission of the high-energy particle. However, since the 2–1 transition is dipole-forbidden, the first excited level of an impurity atom does not become populated, and so transverse waves are not radiatedin the transition from the first excited state to the ground state. Let us now assume that an auxiliary external electromagnetic field propagates in the scintillator as the sum of two plane waves:

$$\mathbf{E}(\mathbf{r},t)=\mathbf{E}_2\cos(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)+\mathbf{E}_1\cos(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t), \qquad (1)$$

whose frequencies ω_1 and ω_2 are close to the resonant frequencies of the 1–3 and 2–3 transitions, $\omega_{31}=(E_3-E_1)/h$ and $\omega_{32}=(E_3-E_2)/h$, i.e., they obey the inequalities

$$|\omega_1-\omega_{31}|\ll\omega_{31}, \quad |\omega_2-\omega_{32}|\ll\omega_{32}.$$

The action of the field (1) induces transitions between the E_2 and E_3 states and between the E_3 and E_1 states. As a result of these transitions the excited level of an impurity atom with energy E_1 is populated, so that emission takes place at the frequency $\omega_{10}=(E_1-E_0)/h$, which is outside the absorption band, thereby ensuring the free escape of radiation from the material.

 © 1999 American Institute of Physics

Consequently, the excitation energy generated by the high-energy particle in the material goes over to radiation that can be utilized to detect the high-energy particle. It is helpful to bear in mind that the scintillation process discussed here occurs only under the influence of the electromagnetic field (1), so that the behavior of the process is dictated entirely by this field. We also note that in the choice of impurity atoms for the field scintillator the critical requirements are altogether different from those in the case of the conventional scintillator, and so the impurity atoms considered here, generally speaking, cannot support the operation of a scintillator in the conventional regime. It must therefore be assumed that the conventional scintillator operating regime is impossible in this medium.

We now consider the dielectric after a high-energy particle has left it, at which time the particle responsible for the excitation of its atoms has ceased to act, but the excitation energy is still migrating through the entire material. In this case the excitation energy is concentrated primarily in long-lived elementary excitations of the material, which are longitudinal electromagnetic plane waves.

The electric field of this wave is known to have the form

$$\mathbf{E}(\mathbf{r},t)=\mathbf{E}\exp\{i\mathbf{q}_0\cdot\mathbf{r}-i\omega_p t-i(\alpha/2)q_0^2 t\}, \tag{2}$$

where the frequency ω_p causes the dielectric function to vanish, $\varepsilon(\omega_p,q=0)=0$ (Ref. 3). In general the field in the excited material must be a superposition of longitudinal waves (2). We can assume in the first approximation that the same field $\mathbf{E}_0(\mathbf{R},t)$ acts on an atom at a point $\mathbf{R}$ in the interior of the medium, regardless of whether this atom belongs to the impurity or the host material. The population amplitude $c(\mathbf{R},t)$ of the first excited level of a host atom situated at the point $\mathbf{R}$ is related to the field by the equation

$$\begin{aligned} &ih\partial c(\mathbf{R},t)/\partial t=-\mathbf{d}^0\cdot\mathbf{E}_0(\mathbf{R},t)\exp(i\omega_{20}t),\\ &h\omega c(\mathbf{R},\omega)=-\mathbf{d}^0\cdot\mathbf{E}_0(\mathbf{R},\omega+\omega_{20}), \end{aligned} \tag{3}$$

where $\mathbf{d}_0$ is the matrix element of the transition from the ground state to the first excited state of a host atom. If, on the other hand, an impurity atom occupies the point $\mathbf{R}$, the population amplitude of the second excited level of this atom $a_2^0(\mathbf{R},t)$ without external fields is related to the field by the equation

$$ih\partial a_2(\mathbf{R},t)/\partial t=-\mathbf{d}_{20}\cdot\mathbf{E}^0(\mathbf{R},t)\exp(i\omega_{20}t), \tag{4}$$

where

$$\mathbf{d}_{mk}=\int dq\,\Psi_m^*(q)\mathbf{d}\Psi_k(q)$$

is the transition dipole moment in the impurity atom. It follows from Eqs. (3) and (4) that

$$a_2^0(\mathbf{R},t)=\chi c(\mathbf{R},t),\quad \chi=|\mathbf{d}_{20}|/|\mathbf{d}^0|.$$

It is a well-known fact that the local field $\mathbf{E}^{\rm loc}(\mathbf{R},t)$ acting on a material atom does not coincide with the average macroscopic field, so that if the macroscopic field (1) is given, an atom at the point $\mathbf{R}$ is acted upon by the local field[4–7]

$$\begin{aligned} \mathbf{E}^{\rm loc}(\mathbf{R},t)&=\mathbf{E}_2^{\rm loc}\cos(\mathbf{k}_2\cdot\mathbf{R}-\omega_2 t)+\mathbf{E}_1^{\rm loc}\cos(\mathbf{k}_1\cdot\mathbf{R}-\omega_1 t)\\ &=(1/3)\{[\varepsilon(\omega_2,k_2)+2]\mathbf{E}_2\cos(\mathbf{k}_2\cdot\mathbf{R}-\omega_2 t)\\ &\quad+[\varepsilon(\omega_1,k_1)+2]\mathbf{E}_1\cos(\mathbf{k}_1\cdot\mathbf{R}-\omega_1 t)\}. \end{aligned} \tag{5}$$

If the radiation energy is much smaller than the total excitation energy of the material, the population amplitude of the first excited energy level of host atoms $c(\mathbf{R},t)$ can be assumed not to change as a result of radiation, i.e., to be a given quantity. The total field acting on an impurity atom consists of the fields near the frequencies ω_{20}, ω_{32}, and ω_{31}. The field at ω_{20} is the field $\mathbf{E}_0(\mathbf{R},t)$ generated by excited host atoms; the other frequencies are associated with the auxiliary field (1), which produces resonant transitions between impurity levels 2–3 and 3–1. The field $\mathbf{E}^0(\mathbf{r},t)$ exerts a nonresonant influence on other transitions in an impurity atom, i.e., is much weaker and can be disregarded.

We denote by $a_i(\mathbf{R},t)$ the population amplitude of the ith level of an impurity atom occupying the point $\mathbf{R}$ under the influence of the auxiliary field (1). We assume that $a_2(\mathbf{R},t)\ll 1$; the population of the ground state is then close to unity, and to lowest order we can set $a_0(\mathbf{R},t)=1$. Retaining only resonant terms, we can now obtain a system of equations describing the behavior of the population amplitude $a_i(\mathbf{R},t)$ of the states of the impurity atom under the influence of the total field from the Schrödinger equation.

We first consider the case of exact resonance, $\omega_1=\omega_{31}$ and $\omega_2=\omega_{32}$. Introducing the notation

$$\begin{aligned} &V_{13}=V_{31}^*=(1/2)(\mathbf{E}_1^{\rm loc}\cdot\mathbf{d}_{13})\exp(-i\mathbf{k}_1\cdot\mathbf{R}),\\ &V_{23}=V_{32}^*=(1/2)(\mathbf{E}_2^{\rm loc}\cdot\mathbf{d}_{23})\exp(-i\mathbf{k}_2\cdot\mathbf{R}), \end{aligned} \tag{6}$$

we can obtain a system of equations for the amplitudes in the form

$$ih\partial a_3(\mathbf{R},t)/\partial t=-V_{32}a_2(\mathbf{R},t)-V_{31}a_1(\mathbf{R},t), \tag{7}$$

$$ih\partial a_2(\mathbf{R},t)/\partial t=\chi ih\partial c(\mathbf{R},t)/\partial t-V_{23}a_3(\mathbf{R},t), \tag{8}$$

$$h\partial a_1(\mathbf{R},t)/\partial t=-V_{13}a_3(\mathbf{R},t). \tag{9}$$

Fourier-transforming the amplitudes in time, from Eqs. (7)–(9) we can obtain a system of three linear inhomogeneous algebraic equations

$$h\omega a_1(\mathbf{R},\omega)+V_{13}a_3(\mathbf{R},\omega)=0, \tag{10}$$

$$h\omega a_2(\mathbf{R},\omega)+V_{23}a_3(\mathbf{R},\omega)=h\omega\chi c(\mathbf{R},\omega), \tag{11}$$

$$h\omega a_3(\mathbf{R},\omega)+V_{31}a_1(\mathbf{R},\omega)+V_{32}a_2(\mathbf{R},\omega)=0. \tag{12}$$

Solving this system of equations, we readily find the population amplitude of the first excited state of the impurity atom:

$$a_1(\mathbf{R},\omega)=\frac{(\mathbf{E}_1^{\rm loc}\cdot\mathbf{d}_{13})(\mathbf{E}_2^{\rm loc}\cdot\mathbf{d}_{32})\chi c(\mathbf{R},\omega)\exp\{i(\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{R}\}}{|(\mathbf{E}_1^{\rm loc}\cdot\mathbf{d}_{31})|^2+|(\mathbf{E}_2^{\rm loc}\cdot\mathbf{d}_{32})|^2-4(h\omega)^2}. \tag{13}$$

It is evident from Eq. (13) that the population of the first excited state of the impurity atom is directly proportional to the population of the first excited level of the host atom.

3. EMISSION FROM IMPURITY ATOMS IN EXACT RESONANCE

It follows from the preceding discussion that three excited states of the impurity atom are populated in the excited material under the influence of the auxiliary field. By assumption, 3–0 transitions are dipole-forbidden, and only 2–0 and 1–0 dipole transitions from excited states to the ground state are possible. The 2–0 transition causes dipole–dipole transfer of excitation energy to nearby atoms of the majority material, and the frequency of this transition lies in the absorption band. Consequently, radiation can escape from the material to the outside only by virtue of a 1–0 transition. The frequency of the 1–0 transition lies in the transparency range, so that the radiation interacts weakly with the material and is freely emitted from it if there are no other like impurity atoms in the vicinity (i.e., within a distance shorter than the wavelength) of the emitting impurity atom. If such atoms are present, competition arises between the dipole–dipole transfer of excitation energy to the nearest impurity atom and the emission of a transverse wave. The emission of a transverse wave by the impurity atom is the main avenue of deexcitation of the impurity if the number of impurity atoms per unit volume n_1 satisfies the inequality

$$n_1 < (\omega_{10}/c)^3. \tag{14}$$

On the other hand, the dipole–dipole transfer of excitation between host atoms plays the dominant role if the number of host atoms per unit volume n_0 ($n_0 \geqslant n_1$) obeys the inequality

$$n_0 > (\omega_{20}/c)^3. \tag{15}$$

When inequality (15) is satisfied, the 1–0 transition current at an impurity atom situated at the point $\mathbf{R}_a$ becomes a source of transverse waves. The total current is equal to the sum of the transition currents in each atom, and in the dipole approximation the Fourier transform of the total current has the form

$$\mathbf{j}_{01}(\mathbf{r},\omega) = -i\omega\mathbf{d}_{01}\sum_a a_1(\mathbf{R}_a,\omega-\omega_{10})\,\delta(\mathbf{r}-\mathbf{R}_a),$$

and after carrying out a coordinate Fourier transform we find

$$\mathbf{j}_{01}(\mathbf{k},\omega) = -i\omega\mathbf{d}_{01}(2\pi)^{-3}\sum_a a_1(\mathbf{R}_a,\omega-\omega_{10}) \times \exp(-i\mathbf{k}\cdot\mathbf{R}_a). \tag{16}$$

Assuming that the material is transparent at the frequency ω, we can easily obtain an equation for the angular and frequency distribution of the radiation:

$$\frac{d^2U}{d\Omega\,d\omega} = \frac{\omega^2}{\varepsilon^{1/2}(\omega)}|[\mathbf{k}\cdot\mathbf{d}_{01}]|^2\left\{\sum_a |a_1(\mathbf{R}_a,\omega-\omega_{10})|^2\right\} + \sum_a\sum_{b\neq a} a_1(\mathbf{R}_a,\omega-\omega_{10})a_1^*(\mathbf{R}_b,\omega-\omega_{10}) \times \exp\{-i\mathbf{k}\cdot(\mathbf{R}_a-\mathbf{R}_b)\}. \tag{17}$$

The summation in Eq. (17) is carried out over all impurity atoms. If inequality (14) holds, the double-sum term in (17) is small in comparison with the first term. For a uniform distribution of impurity atoms the summation over $\mathbf{R}_a$ can be replaced by integration. In view of this, the following relation can be obtained by means of Eq. (13):

$$\frac{d^2U}{d\Omega\,d\omega} = \frac{\omega^2}{\varepsilon^{1/2}(\omega)}\int d^3R \times\left|\frac{[\mathbf{k}\cdot\mathbf{d}_{01}]\chi V_{13}V_{23}c(\mathbf{R},\omega-\omega_{10})\exp\{i(\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{R}\}}{h^2(\omega-\omega_{10})^2-|V_{13}|^2-|V_{23}|^2}\right|^2. \tag{18}$$

If the denominator in Eq. (18) is not too small, we can disregard absorption at the frequencies ω_1 and ω_2. Equation (18) is then simplified:

$$\frac{d^2U}{d\Omega\,d\omega} = \frac{\omega^2|[\mathbf{k}\cdot\mathbf{d}_{01}]^2||\chi V_{13}V_{23}|^2\int d^3R|c(\mathbf{R},\omega-\omega_{10})|^2}{\varepsilon^{1/2}(\omega)\{h^2(\omega-\omega_{10})^2-|V_{13}|^2-|V_{23}|^2\}^2}. \tag{19}$$

In the frequency range where the denominator in (19) becomes small, the imaginary part of the permittivity for the frequencies ω_1 and ω_2 must be taken into account, and the more rigorous Eq. (18) must be used instead of (19).

4. SCINTILLATOR EMISSION INDUCED BY A LONGITUDINAL WAVE

As an example we consider radiation generated in the scintillator when the longitudinal plane wave (2) is transmitted through it. The local field acting on an atom of the material differs from (2) by the factor $[\varepsilon(\omega,q)+2]/3=2/3$:

$$\mathbf{E}_0^{\rm loc}(\mathbf{r},t) = (2/3)\mathbf{E}_0\exp\{i\mathbf{q}_0\cdot\mathbf{r}-i\omega_p t-i(\alpha/2)q_0^2 t\}. \tag{20}$$

Substitution of this expression into Eq. (3) gives

$$c(\mathbf{R},\omega) = -\mathbf{d}^0\cdot\mathbf{E}_0(1/h\omega)\exp(i\mathbf{q}_0\cdot\mathbf{R}) \times\delta(\omega-\omega_{20}-\omega_p-\alpha q_0^2/2). \tag{21}$$

Equation (21) can be used to transform the angular and frequency distribution of the scintillator radiation energy (19) to the form ($\mathscr{V}$ is the working volume of the scintillator)

$$\frac{d^2U}{d\omega\,d\Omega} = \frac{\mathscr{V}|[\mathbf{k}\cdot\mathbf{d}_{01}]|^2|V_{13}V_{23}|^2|\mathbf{d}_{20}\cdot\mathbf{E}_0|^2\delta(\omega-\omega_{20}-\omega_p-\alpha q_0^2/2)}{2\pi^3 9h^2\varepsilon^{1/2}(\omega)\{h^2(\omega-\omega_{10})^2-|V_{13}|^2-|V_{23}|^2\}}. \tag{22}$$

It is readily seen from the expression derived for the radiated energy distribution that the scintillator emission frequency does not depend on the direction of the longitudinal wave vector.

5. DISCUSSION OF THE RESULTS

It is important to note that, together with the process discussed here, the reverse process can also take place, where the field of a high-energy particle takes a bound electron of an impurity atom from the ground state to the first excited state, i.e., induces a 0–1 transition, and the auxiliary field (1) takes this electron in succession into the excited state 3 and then into the excited state 2 of the impurity atom. This process competes with the primary process under discussion, but it is much weaker, being proportional to the energy transferred by the particle to excitation of one first excited level of all the impurity atoms, whereas the intensity of the primary process is proportional to the energy from the high-energy particle that goes into excitation of all the host atoms. However, the impurity concentration must satisfy inequality (14), so that the forward process is many times stronger than the reverse process. Consequently, the existence of the reverse process can be ignored, and the resulting relative error is of order the density of the impurity atoms, n_1/n_0.

The first consideration in selecting a specific material for the field scintillator is the fact that in an impurity-free material the resonant excitation corresponding to the first excited level of a bound atomic electron should propagate through the entire sample essentially without any losses. This behavior can be confirmed experimentally by a comparatively simple procedure. In particular, therefore, we can rule out all substances in which emission centers are not bound to impurities. In selecting the impurity atom, it is necessary above all that the energy of its second excited level coincide with (or be sufficiently close to) the energy of the first excited level of the host atom. It is safe to say that the principal condition in selecting the impurity atom is that the first and third energy levels of bound electrons in the impurity atoms must be situated outside the absorption band and that dipole transitions between the ground and third excited levels of the atom must be forbidden, along with transitions between the first and second excited levels of the impurity atom.

In regard to the positions of the first and third levels of the impurity atom, once the principal condition has been met, the energy differences E_3-E_2 and E_2-E_1 are not so critical. For example, these energy differences can be in different frequency intervals. Other choices of impurity atoms with different positions of the electron energy levels are also possible. For example, the ''third'' excited level can be an energy level below the second level, i.e., below the first excited level of the host atom (this arrangement of the impurity levels has been used previously[8] in an investigation of the coherent part of the impurity radiation). In this case the resonant field will also induce emission from impurity atoms, but then emission will also be possible without a resonant field, thus undermining the controllability of the scintillator. When the above arrangement of the levels is used without a resonant field, emission from impurity atoms can take place only through the thermal excitation of a bound electron of the impurity from the second to the first excited state, and this process can be disregarded for $\omega_{32} \gg kT$. Another possible example is an arrangement in which the excited energy levels of the impurity atom are in positions such that the ''first'' excited level is above the second level, i.e., above the first excited level of the host atom. In this case emission from impurity atoms is possible only when a resonant field is applied, but then only part of the impurity radiation energy is supported by excitation of the host material, and part of the radiation energy emerges at the expense of a reduction in the energy of the resonant field. The accuracy of determination of the host excitation energy required for the acquisition of information about the properties of high-energy particles is diminished in this case. Such an arrangement of the impurity atom energy levels is therefore unsuitable for the field scintillator.

The foregoing discussion shows that the field scintillator is feasible over an extensive region. Consequently, the construction of such a scintillator is a realistic proposition, where the energy of electron degrees of freedom excited by a high-energy particle in a dielectric is converted into radiation emitted by impurity atoms in a process that is fully maintained and controlled by a specially selected external electromagnetic field. The process differs from a conventional scintillator in that the excess energy (the difference between the energy of excitation of an atom of the medium and the emitted photon energy) is not transported by phonons, but by the external field.

As a result, if the host material of the scintillator and the impurity atoms are fixed, the distribution of the scintillator-radiated energy (19) or (18) is determined entirely by the amplitudes and frequencies of the auxiliary external field driving the process. Shutting off the external field stops the scintillator from emitting, despite the conservation of excitation energy in the material. Varying the frequencies of the auxiliary external field causes the intensity and frequency spectrum of the radiation to change.

This work has received partial support from the Russian Fund for Fundamental Research (Grant 99-02-17015).

*)E-mail: ryazanov@theor.mephi.msk.su

[1] A. I. Abramov, Yu. A. Kazanskiĭ, and E. S. Matusevich, *Fundamentals of Experimental Methods in Nuclear Physics* [in Russian], Énergoatomizdat, Moscow (1985).
[2] V. K. Lyapidevskiĭ, *Radiation Detection Methods* [in Russian], Énergoatomizdat, Moscow (1987).
[3] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, 2nd ed. (rev. and enl., with L. P. Pitaevskii), Pergamon Press, Oxford-New York (1984).
[4] S. L. Adler, Phys. Rev. **126**, 413 (1962).
[5] E. G. Maksimov and I. I. Mazin, Solid State Commun. **27**, 527 (1978).
[6] J. E. Sipe, Can. J. Phys. **53**, 2095 (1978).
[7] M. I. Ryazanov, Zh. Éksp. Teor. Fiz. **108**, 1778 (1995) [JETP **81**, 974 (1995)].
[8] V. K. Lyapidevskiĭ and M. I. Ryazanov, JETP Lett. **32**, 496 (1980).

Translated by James S. Wood

Quasiclassical approximation with the centrifugal potential excluded

B. M. Karnakov*) and V. D. Mur

Moscow Engineering Physics Institute (Technical University), 115409 Moscow, Russia

V. S. Popov

Institute of Theoretical and Experimental Physics, 117259 Moscow, Russia

(Submitted 26 March 1999)
Zh. Éksp. Teor. Fiz. **116**, 511–525 (August 1999)

We develop a modification of the WKB method (the modified quantization method, or MQM) for finding the radial wave functions. The method is based on excluding the centrifugal potential from the quasiclassical momentum and changing correspondingly the phase in the Bohr–Sommerfeld quantization condition. MQM is used to calculate the asymptotic coefficients at zero and at infinity. We use the examples of power-law and funnel potentials to show that MQM not only dramatically broadens the possibilities of studying the energy spectrum and the wave functions analytically but also ensures accuracy to within a few percent even when one calculates states with a radial quantum number $n_r \sim 1$, provided that the angular momentum l is not too large. We also briefly discuss the possibility of generalizing MQM to the relativistic case (the spinless Salpeter equation). © *1999 American Institute of Physics.* [S1063-7761(99)01108-7]

1. The quasiclassical approximation (the WKB method) is one of the most effective methods of quantum mechanics and mathematical physics (see, e.g., Refs. 1–8). A modification of the WKB method was proposed in Refs. 9 and 10. The approach consisted in excluding the centrifugal potential from the quasiclassical momentum, $p(r) \to \tilde{p}(r)$, and taking this potential into account by changing the phase γ_l in the Bohr–Sommerfeld quantization rule:

$$\int_0^{r_0} \tilde{p}(r)\, dr \equiv \int_0^{r_0} \sqrt{2[E_{n_r l} - V(r)]}\, dr = \pi(n_r + \gamma_l). \quad (1)$$

Here and in what follows $\hbar = m = 1$, $n_r = 0,1,2,\ldots$ is the radial quantum number, r_0 is the turning point,[1)] and γ_l depends on the orbital angular momentum l and the potential's behavior at small distances. If the potential is attractive and satisfies $V(r) \propto r^\alpha$ ($\alpha > -2$, which excludes the fall-to-center case[2]), we have[9,10]

$$\gamma_l = \frac{2l+3}{4} \quad \text{if} \quad \alpha > 0, \quad (2a)$$

$$\gamma_l = \frac{2l+3+\alpha}{2(2+\alpha)} \quad \text{if} \quad -2 < \alpha < 0. \quad (2b)$$

In particular, for short-range potentials with a Coulomb singularity at zero ($\alpha = -1$, Yukawa, Hülthén, etc., potentials), the right-hand side of Eq. (1) contains $n_r + l + 1 \equiv n$, the principal quantum number, and the dependence of $E_{n_r l}$ on l exists only in the corrections to the quantization rule (1). Note that for potentials that are finite at zero ($\alpha \geqslant 0$) the quantization rule (1) with $\gamma_l = (2l+3)/4$ is given in Ref. 5.

The approach specified by (1) and (2) (we refer to it as the modified quantization method, or MQM) broadens the possibility of studying the energy spectrum analytically. In the present paper we generalize MQM to the case of confinement potentials, which have only a discrete spectrum of energy levels, and use it to calculate the radial wave function in the subbarrier region, including the asymptotic coefficient at zero distances (Sec. 2). A comparison with the exact and numerical solutions of the Schrödinger equation for power-law potentials shows that the MQM formulas are not only exact in the asymptotic limit $n_r \to \infty$ but retain their high accuracy even for states with $n_r \sim 1$, provided that the orbital angular momentum is not too large. In Sec. 3, MQM is applied to the funnel potential, often used in quantum chromodynamics,[11,12] which makes it possible to obtain equations for the energy spectrum and the asymptotic coefficient at zero analytically. Quasiclassical formulas for the asymptotic coefficients at infinity are derived in Sec. 4. Section 5 studies the possibility of generalizing MQM to the relativistic case. The results of our investigation are briefly discussed in Sec. 6. Finally, some details of the calculations, namely, the study of MQM (1) for the funnel potential and the general properties of the quantization integral in (1), are grouped in Appendices A and B.

2. The idea of excluding the centrifugal potential from the quasiclassical momentum is based on the fact that for highly excited states ($n_r \gg 1$), where the wave functions has many nodes, this potential is important (for $l \sim 1$) only at small distances, while in the main region of particle localization the potential acts as a perturbation $\propto \hbar^2$ and can be discarded in the quasiclassical approximation. Here the quasiclassical wave function can be written

$$\chi_{El}^{\mathrm{WKB}}(r) = \frac{C}{\sqrt{\tilde{p}(r)}} \sin\left\{ \int_0^r \tilde{p}(r)\, dr - \pi\delta_l \right\}, \quad (3)$$

 © 1999 American Institute of Physics

where for the normalized bound-state wave function we have[2,3]

$$C=\frac{2}{\sqrt{T_r}},\quad T_r=\frac{2\pi}{\omega_r}=2\int_0^{r_0}\frac{dr}{\tilde{p}(r)}. \tag{4}$$

The value of δ_l can be found by matching the quasiclassical wave function to the exact solution of the Schrödinger equation satisfying the boundary condition

$$\chi_{El}(r)=c_{El}r^{l+1},\quad r\to 0, \tag{5}$$

where c_{El} is the asymptotic coefficient at zero (i.e., at very small distances). The solution of the matching problem depends on the shape of the potential $V(r)$ at small distance (i.e., in the region where the classical principles break down), the energy E, and the orbital angular momentum l.

We begin with the case of high energies $E=k^2/2$. Here a potential that is regular at small distances can be interpreted as a perturbation. If we limit ourselves to first-order perturbation theory in the potential $V(r)$, it is convenient to write (3) in the form

$$\chi_{El}^{\mathrm{WKB}}(r)=\frac{C}{\sqrt{k}}[\sin\phi_0(r)+\phi_1(r)\cos\phi_0(r)], \tag{6}$$

where $\delta_l=\delta_l^{(0)}+\delta_l^{(1)}$,

$$\phi_0(r)=kr-\pi\delta_l^{(0)}, \tag{7}$$

$$\phi_1(r)=\int_0^r(\sqrt{k^2-2V(r)}-k)dr-\pi\delta_l^{(1)}$$

(specifying the factor kr in the phase $\phi_0(r)$ explicitly is equivalent to ignoring the potential V). On the other hand, in the absence of a potential,

$$\chi_{El}^{(0)}(r)=A\sqrt{kr}\,J_\nu(kr)\approx A\sqrt{\frac{2}{\pi}}\sin\left(kr-\frac{\pi l}{2}\right),$$

$$kr\gg\nu^2,\quad \nu=l+\frac{1}{2}. \tag{8}$$

The region with $r\gg\nu^2/k$ is where the exact and quasiclassical wave functions are matched. By comparing (6) and (7) with (8) we find that

$$\delta_l^{(0)}=\frac{l}{2},\quad A=\sqrt{\frac{\pi}{2k}}C. \tag{9}$$

Now we can use the asymptotic form (as $z\to 0$) of the Bessel function $J_\nu(z)$ and Eqs. (4) and find that

$$c_{El}^{(0)}=\frac{\sqrt{\omega_r}}{\Gamma(l+3/2)}\left(\frac{E}{2}\right)^{(2l+1)/4} \tag{10}$$

(the zeroth approximation). In this approximation, the energy spectrum of the particle is determined by the quantization rule (1) and (2a), in which

$$\gamma_l=\delta_l^{(0)}+\frac{3}{4}=\frac{2l+3}{4}.$$

We see that the quantization rule holds for all $\alpha>-2$.

Formula (10) was derived with $V(r)$ ignored at small distances. We allow for $V(r)$ by perturbation-theory techniques. To this end we use the Green's function of the radial Schrödinger equation for a free particle:

$$G_{E\nu}(r,r')=(-1)^l\pi\sqrt{r_<r_>}\,J_\nu(kr_<)J_{-\nu}(kr_>),$$

where $r_<$ ($r_>$) is the smaller (larger) of the numbers r and r'. The first-order correction (in the potential) to the unperturbed solution (8) is

$$\chi_{El}^{(1)}(r)=-\int_0^\infty G_{E\nu}(r,r')V(r')\chi_{El}^{(0)}(r')dr'$$

$$=(-1)^{l+1}A\pi\sqrt{kr}\left\{J_\nu(kr)\int_r^\infty rV(r)J_\nu(kr)\right.$$

$$\left.\times J_{-\nu}(kr)dr+J_{-\nu}(kr)\int_0^r rV(r)J_\nu^2(kr)dr\right\}, \tag{11}$$

with $\chi_{El}=\chi_{El}^{(0)}+\chi_{El}^{(1)}$.

For $r\to 0$, only the first term in the braces, which is proportional to r^{l+1} [as in (5)], is important. This leads to renormalization of the asymptotic coefficient at zero:

$$c_{El}^{\mathrm{WKB}}=c_{El}^{(0)}\{1+d_l(E)\}, \tag{12}$$

$$d_l(E)=(-1)^{l+1}\pi\int_0^\infty rV(r)J_\nu(kr)J_{-\nu}(kr)dr.$$

In view of the rapid oscillations of the Bessel functions, the asymptotic behavior of $d_l(E)$ at high energies is determined by the behavior of the potential as $r\to 0$. For potentials with a power-law asymptotic behavior, $V(r)=gr^\alpha/\alpha$, we have[2)]

$$d_l(E)=(-1)^{l+1}g\frac{\pi\Gamma(-\alpha-1)\Gamma(1+\alpha/2)}{2\alpha\Gamma(-\alpha/2)\Gamma(-\nu-\alpha/2)\Gamma(\nu-\alpha/2)}$$

$$\times\left(\frac{2}{k}\right)^{\alpha+2} \tag{13}$$

(see formula 6.574.2 in Ref. 13). In particular,

$$d_l(E)=g\frac{\pi}{2(2E)^{1/2}},\quad -g\frac{\pi(\nu^2-1/4)}{4(2E)^{3/2}},\quad -g\frac{\nu^3-\nu}{12E^2} \tag{14}$$

for $\alpha=-1$, 1, and 2, respectively. For the logarithmic potential $V(r)=g\ln r$, the coefficient $d_l(E)\propto(\ln E+b_l)/E$ can be obtained analytically by passing to the limit $\alpha\to 0$ in (13). Note that the larger the exponent α, the less important is the potential $V(r)$ at small distances and the faster the correction (14) decreases with increasing energy: $d_l(E)\propto 1/E^{1+\alpha/2}$.

The case of attractive power-law potentials $V(r)=-\zeta/r^\beta$ $(0<r<\infty)$ with $0<\beta\equiv-\alpha<2$, where there is level crowding as $E\to-0$, can be examined along similar lines. When solving the Schrödinger equation at small distances (in the region where the classical principles do not work), we can discard the term with the energy,

$$\chi_{El}^{(0)}(r)=A\sqrt{r}\,J_\mu\left(\frac{\sqrt{8\zeta}}{2-\beta}r^{1-\beta/2}\right),\quad \mu=\frac{2l+1}{2-\beta}. \tag{15}$$

TABLE I. The accuracy of MQM calculations of the asymptotic coefficients at zero for power-law potentials $V(r)=r^{\alpha}/\alpha$.

l	α	$n_r=0$	1	2	3	10
0	2	1.50225	1.83987	2.05704	2.22186	2.88969
		2.8(−3)	1.1(−4)	2.0(−5)	5.8(−6)	9(−8)
	−1	2.04124(−1)	1.20962(−1)	8.06872(−2)	5.84237	1.59817(−2)
		1.0(−2)	1.3(−3)	5.2(−4)	2.1(−4)	6.1(−6)
1	1	9.23722(−1)	1.20480	1.39708	1.54805	2.19768
		6.4(−3)	2.8(−3)	1.6(−3)	1.1(−3)	2.0(−4)
	2	1.22658	1.93940	2.56558	3.14218	6.53295
		−3.0(−4)	−3.5(−5)	−8.5(−6)	−3.0(−6)	−6(−8)
	−1	9.01601(−3)	6.98771(−3)	5.35462(−3)	4.21685(−3)	1.40121(−3)
		2.8(−2)	6.3(−3)	2.2(−3)	1.0(−3)	4.1(−5)
2	1	4.36675(−1)	6.77368(−1)	8.81660(−1)	1.06500	2.08230
		−7.9(−4)	−2.6(−3)	−1.9(−3)	−1.3(−3)	−3.0(−4)
	2	7.75759(−1)	1.45131	2.17697	2.94763	9.23898
		5.5(−3)	8.1(−4)	2.3(−4)	8.6(−5)	2.3(−6)
	−1	2.20092(−4)	2.03986(−4)	1.73900(−4)	1.46814(−4)	5.84389(−5)
		1.4(−2)	2.2(−3)	6.0(−4)	2.1(−4)	2.6(−6)
3	1	1.66280(−1)	2.93371(−1)	4.19592(−1)	5.45822(−1)	1.43173
		−4.6(−2)	−2.6(−2)	−1.7(−2)	−1.2(−2)	−2.5(−3)
	2	4.14661(−1)	8.79628(−1)	1.45870	2.14715	9.69891
		1.6(−2)	2.9(−3)	9.2(−4)	3.8(−4)	1.2(−5)

Note: For given n_r and l, the first row gives the exact value of the coefficient $c_{n_r l}$ and the second, the error $\eta_{n_r l}$ from (20). Notation: $a(b)\equiv a\times 10^{b}$.

Hence

$$c_{n_r l}^{(0)}=\frac{2^{(\mu+1)/2}\zeta^{\mu/2}\omega_r^{1/2}}{\Gamma(\mu+1)(2-\beta)^{\mu+1/2}}, \tag{16}$$

with the frequency ω_r given by the same expression in (4).

Allowing for the term with the energy by perturbation techniques, we arrive at formula (12), but now

$$d_l(E)=\frac{4\pi E}{\sin\pi\mu}\,\frac{(2-\beta)^{\beta/(2-\beta)}}{(8\zeta)^{2/(2-\beta)}}\int_0^{\infty} z^{(2+\beta)/(2-\beta)} \times J_{\mu}(z)J_{-\mu}(z)\,dz.$$

In particular, for the Coulomb potential $V(r)=-Z/r$ we have

$$c_{nl}^{\mathrm{WKB}}=Z^{l+3/2}\frac{2^{l+1}}{(2l+1)!\,n^{3/2}}\left\{1-\frac{l(l+1)(2l+1)}{12n^2}\right\}, \tag{17}$$

which coincides with the expansion of the exact coefficient c_{nl} at zero (see Ref. 2).

We illustrate the accuracy of the quasiclassical results by the example of power-law attractive potentials, $V(r)=r^{\alpha}/\alpha$. For $\alpha>0$ we use Eqs. (1), (2a), and (10) to find the quasiclassical expressions for the energy levels and the asymptotic coefficients at zero:

$$E_{n_r l}^{\mathrm{WKB}}=\left[\epsilon_{\alpha}\left(n_r+\frac{1}{4}(2l+3)\right)\right]^{2\alpha/(2+\alpha)}, \tag{18}$$

$$\epsilon_{\alpha}=\sqrt{2\pi}\,\alpha^{1-1/\alpha}\frac{\Gamma((2+3\alpha)/2\alpha)}{\Gamma(1/\alpha)},$$

$$c_{n_r l}^{(0)}=\sqrt{\frac{\alpha\epsilon_{\alpha}}{2+\alpha}}\,\frac{(E_{n_r l}^{\mathrm{WKB}})^{\beta_l}}{2^{(2l-1)/4}\Gamma(l+3/2)},\quad \beta_l=\frac{l+1-\alpha^{-1}}{2}, \tag{19}$$

where we have allowed for the fact that $\omega_r=\partial E_{n_r l}/\partial n_r$; the corrections to $c_{n_r l}^{\mathrm{WKB}}$ are given by formulas (13) and (14).

A comparison of the results of calculating $c_{n_r l}^{\mathrm{WKB}}$ and the cases with $\alpha=-1$, 1, and 2 is done in Table I, which lists the values of the exact asymptotic coefficients $c_{n_r l}$ and of the ratio

$$\eta_{n_r l}=\frac{c_{n_r l}^{(0)}}{c_{n_r l}}\{1+d_l(E_{n_r l}^{\mathrm{WKB}})\}-1. \tag{20}$$

For s states in the case of Coulomb and linear potentials, the asymptotic coefficients (12) coincide with the exact coefficients,

$$c_{n_r 0}=\frac{2}{(n_r+1)^{3/2}}\ (\alpha=-1),\quad c_{n_r 0}=\sqrt{2}\ (\alpha=1),$$

and therefore are not listed in Table I.

Table I shows that even for states with $n_r\sim 1$, the error of the MQM formulas is usually no more than a few percent and rapidly decreases with increasing n_r. Note the important role of the correction $d_l(E)$. Formulas (10) and (15) of the zeroth approximation are asymptotically exact in the limit $n_r\to\infty$, but for $n_r\sim 1$ and $l\neq 0$ they produce large errors and can only be used for making estimates [see, e.g., the contribution of d_l to (17)].

Several remarks concerning the quantization rule (1) and (2a) are in order. The quantization rule suggests that levels with the same values of $N=2n_r+l$ are, in a certain sense, degenerate;[3)] this effect is similar to the well-known accidental degeneracy of the energy spectrum of a spherical oscillator.[2,3,5] Figure 1 illustrates this fact, depicting the energy spectrum $E_{n_r l}$ of the lower states in the case of power-law potentials with the exponents $\alpha=1$ and 4. Clearly, for a given value of N, the proximity of levels with different values l is violated only as the orbital angular momentum gets larger (but the s and d levels and the energies of the p and f

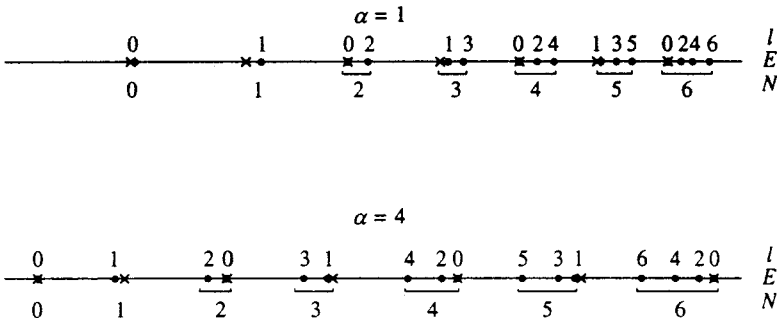

FIG. 1. The energy spectrum $E_{n_r l}$ of the lower states in the potential $V(r)=r^\alpha/\alpha$ with $\alpha=1$ and 4. Dots depict the exact energy values, with the respective values of the orbital angular momentum l and the quantum number $N=2n_r+l$ placed near the dots. The symbol $\times$ denotes the quasiclassical value provided by (18).

states are always close to each other; as N increases, the levels with larger values of l begin to move closer to the these levels).

Now let us study the corrections to the quantization rule (1) and (2a) that arise when we allow for $V(r)$ at small distances. Note that in the quasiclassical region $r \gg \nu^2/k$ (but where $V(r)$ is still much smaller than k^2) only the second term in the braces in (11) is important. This term corresponds to the correction to the phase $\phi_0(r)$ that arises because the potential $V(r)$ has been taken into account in (6) and (7). Equating (6) and (11) yields

$$\delta_l^{(1)}=\frac{1}{\pi}\int_0^r\{\pi r V(r)J_\nu^2(kr)+\sqrt{k^2-2V(r)}-k\}dr \qquad (21)$$

for $r \gg \nu^2/k$. In the case of power-law potentials with $\alpha<1$, the integral becomes a constant, which is independent of r and modifies the quantization rule (1) and (2a):

$$\int_0^{r_0}\sqrt{2[E_{n_r l}^{\mathrm{WKB}}-V(r)]}dr = \pi\left[n_r+\frac{l}{2}+\frac{3}{4}+\delta_l^{(1)}(E_{n_r l}^{\mathrm{WKB}})\right]. \qquad (22)$$

For instance, for potentials with a Coulomb singularity $V(r)=-Z/r$ as $r\to 0$, the integration in (21) can be extended to $r=\infty$. We obtain

$$\delta_l^{(1)}(E)=\frac{1}{\pi k a_B}[\ln k a_B+1+\psi(l+1)], \quad k a_B \gg 1, \quad (23)$$

where $a_B=1/Z$ is the Bohr radius, and $\psi(z)$ is the logarithmic derivative of the gamma function.

3. For the funnel potential

$$V(r)=-\frac{Z}{r}+gr, \quad 0<r<\infty, \qquad (24)$$

the quantization integral (1) can be calculated analytically, which makes it possible to easily study the energy spectrum $E_{n_r l}$ (see Appendix A).

For states with $E_{n_r l}>0$, the asymptotic coefficients at zero can be calculated by the formula

$$c_{n_r l}^{\mathrm{WKB}}=c_{n_r l}^{(0)}(1+d_l^{(C)}+d_l^{(L)}), \qquad (25)$$

where the coefficients $c_{n_r l}^{(0)}$ have been defined in (10), the corrections $d_l^{(C)}$ and $d_l^{(L)}$ refer to the Coulomb and linear (in r) parts of the potential (24) and are given by (14), and the energy $E_{n_r l}^{\mathrm{WKB}}$ is determines by the quantization rule (22) with phase (23).

However, for potentials with a Coulomb singularity at zero, formulas (12) and (25) can modified even more by allowing exactly for the effect of this potential at any energy values $E=k^2/2>0$. Here the solution of the radial Schrödinger equation is given by the regular Coulomb function $F_l(-Z/k,kr)$; see Ref. 14. Using the asymptotic expression of this function as $r\to 0$, we can show that if we allow for the Coulomb potential at small distances Eqs. (12) become ($q=1/ka_B$)

$$c_{El}=c_{El}^{(0)}\left\{\frac{2\pi q}{1-\exp(-2\pi q)}\Big/\prod_{s=1}^{l}\left(1+\frac{q^2}{s^2}\right)\right\}^{1/2} \qquad (26)$$

(for $l=0$ the product must be set to unity). Replacing $c_{n_r l}^{(0)}(1+d_l^{(C)})$ in (25) with (26), we arrive at the modified expression for the asymptotic coefficient $c_{n_r l}^{\mathrm{WKB}}$.

A comparison of the results of MQM calculations of the energy levels and the asymptotic coefficients for the funnel potential with the exact values obtained by solving the Schrödinger equation numerically is done in Table II for $l=0$, 1, and 2. Here we have taken the values $Z=0.68812$ and $g=1/2$, which are used in QCD to describe the states of charmonium $c\bar{c}$ (see Ref. 12). The accuracy of the calculations is characterized by two quantities,

TABLE II. The results of MQM calculations for the funnel potential (24); $Z=0.68812$ and $g=1/2$.

			$l=0$		
n_r	0	1	2	3	4
$E_{n_r l}$	0.49018	1.61443	2.42105	3.10454	3.71656
$\varepsilon_{n_r l}$	−0.132	−5.0(−3)	−1.6(−3)	−7.9(−4)	−4.7(−4)
$c_{n_r l}$	2.08332	1.66335	1.53617	1.46909	1.42580
$\eta_{n_r l}$	4.0(−2)	5.1(−3)	2.3(−3)	1.4(−3)	9.8(−4)

l		1			2	
n_r	0	1	2	0	1	2
$E_{n_r l}$	1.30557	2.14835	2.85424	1.84780		
$\varepsilon_{n_r l}$	−3.8(−2)	−1.4(−2)	−7.2(−3)	−4.6(−2)		
$c_{n_r l}$	0.89695	1.07009	1.18071	0.31027	0.45700	0.57479
$\eta_{n_r l}$	1.1(−2)	9.0(−3)	6.2(−3)	6.8(−2)	3.4(−2)	2.2(−2)

$$\epsilon_{n_r l}=\frac{E^{\mathrm{WKB}}_{n_r l}}{E_{n_r l}}-1, \quad \eta_{n_r l}=\frac{c^{\mathrm{WKB}}_{n_r l}}{c_{n_r l}}-1. \tag{27}$$

Table II shows that the quasiclassical values E^{WKB} and c^{WKB} are accurate to within a few percent (with the exception of the energy of the ground state, for which $ka_B=1.44$, while, strictly speaking, the quantization rule (22) and (23) can be applied only if $ka_B \gg 1$). Thus, for the funnel potential, as in the case of power-law potentials, MQM provides acceptable accuracy in calculating the energy levels and the asymptotic coefficients at zero.

4. We consider the bound states of a particle for the case of potentials that vanish at infinity: $V(r)=-Z/r+o(1/r^2)$ as $r\to\infty$ with $Z\geqslant 0$. Here the asymptotic expression for the normalized wave function is

$$\psi_{\kappa lm}(\mathbf{r})\approx 2C_{\kappa l}\kappa^{3/2}(\kappa r)^{\nu-1}e^{-\kappa r}Y_{lm}\left(\mathbf{r}/r\right), \tag{28}$$

where $\kappa=\sqrt{-2E}$, E is the energy level, and the parameter $\nu=Z/\kappa$ is the effective principal quantum number (also often denoted by n^*). The asymptotic coefficients $C_{\kappa l}$ (as well as the asymptotic coefficients at zero [Eq. (5)] are often encountered in quantum mechanics and its applications. The quasiclassical approximation for these coefficients was studied in Ref. 15. We have found it convenient to write the quasiclassical formulas for the $C_{\kappa l}$ and the quantization rule proper in a more general form.

Let

$$V(r)=-\frac{g}{2R^2}v(x), \quad x=\frac{r}{R}, \tag{29}$$

where R is the characteristic radius, g is the dimensionless coupling constant, and the function $v(x)$ determines the shape of the interaction potential. Then the quantization condition and the expression for the asymptotic coefficient become[4)]

$$\sqrt{g}\,Q(z)=N, \quad N=n_r+\gamma_l, \tag{30}$$

$$C^{\mathrm{WKB}}_{\kappa l}=\left[-\frac{8\pi N z Q'(z)}{Q(z)}\right]^{-1/2}\left[N\frac{zx_0}{Q(z)}\right]^{-\sigma N} \times\exp\left\{N\frac{J(z)}{Q(z)}-\frac{(l+1/2)^2}{N}J_1(z)Q(z)\right\}, \tag{31}$$

where $\sigma=v_\infty/2zQ(z)$, $z=\kappa R/\sqrt{g}$, $E^{\mathrm{WKB}}_N=-gz^2(2R^2)^{-1}$, $v_\infty=\lim_{x\to\infty}xv(x)=2ZR/g$, and we have introduced the functions

$$Q(z)=\frac{1}{\pi}\int_0^{x_0(z)}\sqrt{v(x)-z^2}\,dx, \quad v(x_0)=z^2, \tag{32}$$

$$J(z)=zx_0(z)-\int_{x_0(z)}^{\infty}\left[\sqrt{z^2+v(x)}-z+\frac{v_\infty}{2zx}\right]dx, \tag{33}$$

$$J_1(z)=\frac{1}{2}\int_{x_0(z)}^{\infty}\frac{dx}{x^2\sqrt{z^2-v(x)}}, \tag{34}$$

which are determine by $v(x)$ (the shape of the potential); see Appendix B.

TABLE III. The accuracy of MQM calculations of the asymptotic coefficient at infinity.

l	$n_r=0$	1	2	3	10
0	−4.30(−2)	−2.09(−2)	−1.39(−2)	−1.04(−2)	−3.8(−3)
1	2.85(−2)	−7.5(−4)	−5.0(−3)	−5.6(−3)	−3.3(−3)
2	0.147	4.53(−2)	1.82(−2)	7.9(−3)	−1.8(−3)
3	0.306	0.118	5.88(−2)	3.33(−2)	1.5(−3)

Note: This table lists the values of $\tilde{\eta}_{n_r l}$ in the case of a Coulomb potential.

These formulas provide an easy way of calculating the energy spectrum and the asymptotic coefficients. Let us consider several examples in which the calculations can be done analytically.

(1) For the attractive Coulomb potential, $v(x)=1/x$, $\gamma_l=l+1$, $N\equiv n=n_r+l+1$ is the principal quantum number, and $Z=g/2R$. A simple calculation by the above formulas yields

$$Q(z)=\frac{1}{2z}, \quad x_0(z)=\frac{1}{z^2}, \quad \sigma=1,$$

$$J(z)=\frac{1+2\ln 2}{2z}, \quad J_1(z)=z.$$

This implies that the exact spectrum of the Coulomb levels is $E_n=-Z^2/2n^2$ and that the MQM formula for the asymptotic coefficient is

$$C^{\mathrm{WKB}}_{nl}=\frac{1}{\sqrt{8\pi n}}\left(\frac{2e}{n}\right)^n\exp\left\{-\frac{(l+1/2)^2}{2n}\right\}. \tag{35}$$

A comparison with exact values of the coefficients C_{nl} taken from Ref. 2 is done in Table III, which lists the values of $\tilde{\eta}_{n_r l}=C^{\mathrm{WKB}}_{nl}/C_{nl}-1$. Note that

$$\frac{C^{\mathrm{WKB}}_{nl}}{C_{nl}}=1-\frac{n}{24[n^2-(l+1/2)^2]}+\cdots, \quad n\to\infty. \tag{36}$$

We now turn to short-range potentials, for which $Z=v_\infty=0$.

(2) For the Hülthén potential, $v(x)=(e^x-1)^{-1}$, $\gamma_l=l+1$, $N\equiv n=n_r+l+1=1,2,\ldots$ (as for the Coulomb potential), $x_0(z)=\ln(1+z^{-2})$, and

$$Q(z)=\sqrt{1+z^2}-z, \quad J(z)=2\left(\frac{\tanh^{-1}y}{y}-\ln 2z\right), \tag{37}$$

where $y=z/\sqrt{1+z^2}$, $0<z<\infty$. The quantization condition (30) yields

$$z=\frac{1}{2}\left(\sqrt{\frac{g}{g_n}}-\sqrt{\frac{g_n}{g}}\right), \quad E^{\mathrm{WKB}}_n=-\frac{(g-n^2)^2}{8R^2n^2} \tag{38}$$

for $n<\sqrt{g}$ (the values of g equal to $g_n=n^2$ correspond to occurrence of a bound state). For s levels, the quasiclassical MQM spectrum coincides with the exact spectrum.[9]

The functions Q and J can also be calculated analytically for the potentials $v(x)=1/\cosh^2 x$ and $\exp(-x)$. For further details see Appendix B.

5. Let us briefly discuss the application of MQM to the relativistic case. We limit ourselves to the Salpeter equation

for two relativistic quarks in a state with the orbital angular momentum l in the center-of-mass reference frame:[5)]

$$\{\sqrt{\mathbf{p}^2+m_1^2}+\sqrt{\mathbf{p}^2+m_2^2}+V(r)\}\psi_n=M_n\psi_n. \tag{39}$$

The quasiclassical mass spectrum M_n is determined by the quantization rule (1), in which with allowance for relativistic kinematics we have

$$\tilde{p}^2(r)=\frac{1}{4M^2}[M^2-(m_1+m_2)^2][M^2-(m_1-m_2)^2], \tag{40}$$

where $M=M_n-V(r)$, and $V(r_0)=M_n-m_1-m_2$, with r_0 the turning point. For the confinement potential $V(r)=\sigma r$, the quantization integral (1) becomes

$$J=\int_0^{r_0}\tilde{p}(r)dr=\frac{(m_1+m_2)^2}{4\sigma}\int_1^{\epsilon^2}\frac{dx}{x}\sqrt{(x-1)(x-\mu^2)},$$

$$x=\left[\frac{M_n-\sigma r}{m_1+m_2}\right]^2,$$

and can be calculated analytically:

$$J=\frac{(m_1+m_2)^2}{4\sigma}\left\{(\epsilon^2-1)\xi-(1+\mu^2)\ln(1+\xi)\right.$$
$$\left.-\frac{1}{2}(1+\mu)^2\ln\frac{\epsilon^2-1}{1-\mu^2}+2\mu\ln\frac{\epsilon}{\xi-\mu}\right\},$$
$$\epsilon=\frac{M_n}{m_1+m_2},\quad \mu^2=\left(\frac{m_1-m_2}{m_1+m_2}\right)^2,\quad \xi=\sqrt{\frac{\epsilon^2-\mu^2}{\epsilon^2-1}}, \tag{41}$$

with $\epsilon>1\geqslant\mu\geqslant 0$ and $\xi>1$. The equation $J=\pi(n_r+l/2+3/4)$ for M_n can easily be solved numerically.

If $m_1+m_2\ll M_n$, then (41) yields

$$M_n^2-2(m_1^2+m_2^2)\left(\ln\frac{2M_n}{m_1+m_2}+\frac{1}{2}\right)$$
$$+2\left[\left(\frac{m_1}{m_1+m_2}\right)^2\ln\frac{2m_1}{m_1+m_2}\right.$$
$$\left.+\left(\frac{m_2}{m_1+m_2}\right)^2\ln\frac{2m_2}{m_1+m_2}\right]+O\left(\frac{(m_1^2+m_2^2)^2}{M_n^2}\right)$$
$$=4\pi\sigma\left(n_r+\frac{l}{2}+\frac{3}{4}\right), \tag{42}$$

where n_r is the radial quantum number. Note that the constant γ in (42) corresponds to the value adopted in MQM for potentials that are finite at zero [see Eq. (2a)].

The above formulas become much simpler in the case of quarks with equal masses and also in the limit $m_1/m_2\to 0$. Thus, for $m_1=m_2=m$ we have $\tilde{p}=\sqrt{M^2-4m^2}/2$, and

$$M_n\sqrt{M_n^2-4m^2}-4m^2\cosh^{-1}\frac{M_n}{2m}=(4n_r+2l+3)\pi\sigma \tag{43}$$

(see Refs. 20 and 21). In particular, at $m=0$ (massless quarks), the quasiclassical spectrum can be found explicitly:

$$M_n=M_n^{(0)}=\sqrt{(4n_r+2l+3)\pi\sigma},\quad n=0,1,2,\ldots. \tag{44}$$

A comparison of (44) with the results of numerical calculations done by Cea *et al.*[20] and Basdevant and Boukraa[21] shows that the accuracy of the quasiclassical approximation is impressive: even for the lowest state with $n_r=l=0$ the error δ does not exceed 3% and rapidly decreases with increasing n, while although the accuracy decreases with increasing l, for $l\leqslant 3$ and $n_r>0$ it is still satisfactory ($\delta\leqslant 5\%$). Thus, MQM may be of use also in studying relativistic wave equations.

6. Here are some final remarks.

(1) In MQM, the quasiclassical momentum is written in the same way as in the one-dimensional case, but the constant γ in the quantization rule changes. By excluding the centrifugal potential from the quasiclassical momentum one can obtain analytic formulas for the energy spectrum and the wave functions (in particular, for the asymptotic coefficients at zero and infinity) for states with finite orbital angular momenta.

(2) The MQM formulas are asymptotically exact for $n_r\gg 1,l$. To make then valid in the region of small values of the radial quantum number, $n_r\sim 1$, one must allow for corrections that depend on the behavior of the potential at small distances.

(3) Such corrections are especially important in calculating the asymptotic coefficients at zero. In the case of smooth potentials, allowing for these corrections ensures a accuracy to within a few percent in calculations of nodeless p, d, and f states.

The authors are grateful to Yu. A. Simonov for discussions of the results. The work was partially supported by a grant from the Russian Fund for Fundamental Research (Grant 98-02-17007).

APPENDIX A

In the case of the funnel potential (24), the quantization integral (1) can be calculated analytically,[6)] which yields an equation for the energy spectrum,

$$\frac{Z}{\kappa}{}_2F_1\left(\frac{1}{4},\frac{3}{4};2;-\frac{16gZ}{\kappa^4}\right)=N, \tag{A1}$$

where $N=n_r+\gamma_l$, and $\kappa=\sqrt{-2E}$. Let us examine two extreme cases.

(a) For deep levels, which are determined primarily by the Coulomb potential ($E<0$, and $N\equiv n$ is the principal quantum number), Eq. (A1) yields

$$\kappa_n=\frac{Z}{n}\left(1-\frac{3}{2}\lambda-\frac{1}{4}\lambda^2-\frac{39}{16}\lambda^3-\cdots\right),\quad \lambda\to 0, \tag{A2}$$

where $\lambda=n^4g/Z^3$ is the effective (dimensionless) coupling constant. The perturbation-theory series for screened Coulomb potentials of the general form

$$V(r)=\frac{Z}{r}\sum_{k=0}^{\infty}v_kr^k$$

has been constructed by McEnnan *et al.*[24] and Grant and Lai.[25] Their results imply that

$$\kappa_{nl}=\frac{Z}{n}\left[1-\frac{1}{2}(3-\rho^2)\lambda-\frac{1}{4}\left(1-3\rho^2+2\rho^4-\frac{5}{2n^2}\right)\lambda^2 - \frac{1}{16}\left(39-20\rho^2+11\rho^4-14\rho^6+\frac{60+5\rho^2}{n^2}\right)\lambda^3-\cdots\right], \tag{A3}$$

with $\rho=\sqrt{l(l+1)}/n$. For nodeless ($\rho^2=1-n^{-1}$) states this expansion agrees with the one in Ref. 26. A comparison of (A2) and (A3) shows that the quasiclassical equation (A1) is a good approximation for $n\gg 1, l$ and especially for s states.

(b) The other extreme case, $Z\to 0$, $E>0$, and $N=n-(2l+1)/4$, requires an analytic continuation of Eq. (A1) to positive energy values $E\equiv\epsilon/2>0$, which can be done using formula 15.3.21 of Ref. 14:

$$\frac{Z}{(\epsilon^2+16gZ)^{1/4}}\,{}_2F_1\left(\frac{1}{2},\frac{1}{5};2;\frac{1}{2}\left[1+\frac{\epsilon}{(\epsilon^2+16gZ)^{1/2}}\right]\right)=N. \tag{A4}$$

Hence

$$\epsilon=\epsilon_0\{1+2\zeta[\ln\zeta-(2\ln 2+1)]+O(\zeta^2\ln\zeta)\}, \tag{A5}$$

$$\epsilon_0=(3\pi gN)^{2/3},\quad \zeta=\frac{Z}{(2\pi N)^{4/3}g^{1/3}}=\frac{1}{[(3\pi)^4\lambda]^{1/3}}, \tag{A6}$$

with ϵ_0 coinciding (at $l=0$) with the quasiclassical spectrum in the linear potential. The expansion (A5) contains the term $\zeta\ln\zeta$, which is nonanalytic in ζ. However, allowing for the correction (23), which in our case is equivalent to replacing N with

$$N\left\{1-\frac{Z}{\pi N\epsilon_0^{1/2}}\left[\ln\frac{Z}{\epsilon_0^{1/2}}-1-\psi(l+1)\right]\right\},$$

balances the effect of this term, and the final result is

$$\epsilon=\epsilon_0\{1-2\zeta[\ln(3\pi N)+2\ln 2-\psi(l+1)]+O(\zeta^2)\}. \tag{A7}$$

Thus, the correction (23) to the quantization rule not only yields qualitative agreement between the quasiclassical and exact spectra but also ensures a correct analytic dependence of the energy levels on the parameters that determine the behavior of the potential at small distances.

APPENDIX B

Let us examine the general properties of the quantization integral $Q(z)$. Suppose that in (29) we have $v(x)\approx ax^{-\rho}\to\infty$ as $x\to 0$, $0<\rho<2$ (we assume that $v(x)>0$ and $v'(x)<0$ for $0<x<\infty$). Then we have $x_0(z)\propto z^{-2/\rho}$ as $z\to\infty$ and

$$Q(z)\approx a_1 z^{-(2-\rho)/\rho},\quad a_1=a^{1/\rho}\Gamma\frac{\left(\frac{2-\rho}{\rho}\right)}{2\sqrt{\pi}\Gamma(\rho^{-1})}. \tag{B1}$$

But if the potential is finite at zero, $v(0)=v_0$, the only values of z that Eqs. (30)–(34) contain are those from the interval $(0<z<\sqrt{v_0})$.

Another limiting case is $v(x)\approx\exp(-bx^\sigma)$ as $x\to\infty$. Here $z\to 0$ and $x_0(z)\propto(-\ln z)^{1/\sigma}$, so that

$$Q(z)=Q(0)-b_1 z(-\ln z)^{2(1-\sigma)/\sigma}+\cdots, \tag{B2}$$

$$Q(0)=\frac{1}{\pi}\int_0^\infty\sqrt{v(x)}dx,\quad b_1=\frac{1}{\sigma b^{1/\sigma}},$$

$$J(z)=c_1 z(-\ln z)^{1/\sigma}+\cdots,\quad c_1=\left(\frac{b}{2}\right)^{-1/\sigma}. \tag{B3}$$

We illustrate these formulas using the examples of the Yukawa and Hülthén potentials. In the first case $v(x)=(e^x-1)^{-1}$, $\rho=\sigma=a=b=1$, and $a_1=1/2$, so that

$$Q(z)=\begin{cases}1-z+\dfrac{1}{2}z^2+\cdots, & z\to 0,\\ \dfrac{1}{2z}-\dfrac{1}{8z^3}+\cdots, & z\to\infty,\end{cases} \tag{B4}$$

and in the second $v(x)=e^{-x}/x$, so that

$$Q(z)=\begin{cases}\sqrt{\dfrac{2}{\pi}}-z+\cdots, & z\to 0,\\ \dfrac{1}{2z}-\dfrac{1}{4z^3}+\cdots, & z\to\infty.\end{cases} \tag{B5}$$

Note that the analytic representation of $Q(z)$ in the case of the Yukawa potential is complicated.[9]

Equation (30) determines z and the energy levels $E=-gz^2/2R^2$ as functions of the coupling constant g. In particular, when the nl-level appears,

$$g_{nl}^{\mathrm{WKB}}=Q^{-2}(0)N^2. \tag{B6}$$

Finally, we note that for attractive potentials with a power-law decrease at infinity, $v(x)=ax^{-\rho}$ as $x\to\infty$, $0<\rho<2$, the law governing level crowding to the edge of the continuum ($n\to\infty$) can easily be derived:

$$E_{nl}\approx-\frac{A}{2R^2}g^{2/(2-\rho)}N^{-2\rho/(2-\rho)},\quad A=a_1^{2\rho/(2-\rho)} \tag{B7}$$

[the notation is the same is in (B1)]. In the limit $\rho\to 2$ the levels in the energy spectrum move farther apart.

Note added in proof (May 20, 1999): The quasiclassical quantization conditions for the Salpeter equation in the case of two massless quarks and for a system consisting of a light quark and a heavy antiquark were discussed by V. L. Morgunov, A. V. Nefediev, and Yu. A. Simonov (submitted to Phys. Lett. B) in connection with studies of a rotating string and the meson spectrum in QCD.

*)E-mail: karnak@theor.mephi.msk.su

1)In other words, $\tilde{p}(r_0)=0$. Note that for attractive potentials monotonically increasing with r, the left turning point in (1) is always $r=0$.

2)Usually the integral in (12) needs to be regularized, and this can be conveniently done by an analytic continuation in the parameter α.

3)Of course, this result is approximate and the degeneracy is lifted in the higher orders of the quasiclassical approximation.

4)The first equation can be found in Ref. 9.

[5)]We ignore spin in our calculations (for a derivation of Eq. (39) see Refs. 16 and 17). The WKB method for fermions obeying the Dirac equation (including states lying near the edge of the lower continuum) was developed in Refs. 18 and 19.

[6)]Here we have used formula 15.3.1 from Ref. 14 and the quadratic Kummer transformation 15.3.19. Similar transformations have proved useful in the theory of the Stark effect in a strong field.[22,23]

[1] L. I. Schiff, *Quantum Mechanics*, 3rd ed., McGraw-Hill, New York (1968).

[2] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-relativistic Theory*, 3rd ed., Pergamon Press, Oxford (1977).

[3] A. I. Baz', Ya. B. Zel'dovich, and A. M. Perelomov, *Scattering, Reactions, and Decay in Non-relativistic Quantum Mechanics* [in Russian], Nauka, Moscow (1975) [English translation of earlier Russian edition: NASA Techn. Transl. F-510 (1969)].

[4] A. B. Migdal, *Qualitative Methods in Quantum Theory*, Addison-Wesley, Reading, Mass. (1977).

[5] *Functional Analysis* [in Russian] (Mathematical Reference Library), Nauka, Moscow (1964), Chap. 7, $ 3.

[6] S. Flügge, *Practical Quantum Mechanics*, 2 vols., Springer-Verlag, Berlin (1971).

[7] M. V. Berry and K. E. Mount, Rep. Prog. Phys. **35**, 315 (1972).

[8] V. M. Galitskiĭ, B. M. Karnakov, and V. I. Kogan, *Problems in Quantum Mechanics* [in Russian], Nauka, Moscow (1992) [English translation of earlier Russian edition: Prentice–Hall, Englewood Cliffs, NJ (1963)].

[9] M. S. Marinov and V. S. Popov, Zh. Éksp. Teor. Fiz. **67**, 1250 (1974) [Sov. Phys. JETP **40**, 621 (1975)].

[10] M. S. Marinov and V. S. Popov, J. Phys. A **8**, 1575 (1975).

[11] E. Eichten, K. Gottfried, T. Kinoshita, K. D. Lane, and T.-M. Yan, Phys. Rev. D **17**, 3090 (1978).

[12] A. M. Badalyan, D. I. Kitoroage, and D. S. Pariĭskiĭ, Yad. Fiz. **46**, 226 (1987) [Sov. J. Nucl. Phys. **46**, 139 (1987)].

[13] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Sums, Series and Products*, Academic Press, New York (1980).

[14] *Handbook of Mathematical Functions*, M. Abramowitz and I. A. Stegun (Eds.), National Bureau of Standards Applied Mathematics Series 55, Washington, D.C. (1964).

[15] B. M. Karnakov, V. D. Mur, and V. S. Popov, Yad. Fiz. **61**, 481 (1998) [Phys. At. Nucl. **61**, 420 (1998)].

[16] J. Finger, D. Horn, and J. E. Mandula, Phys. Rev. D **20**, 3253 (1979).

[17] A. Yu. Dubin, A. B. Kaidalov, and Yu. A. Simonov, Phys. Lett. B **323**, 41 (1994); Yad. Fiz. **56**(12), 213 (1993) [Phys. At. Nucl. **56**, 1745 (1993)].

[18] V. S. Popov, V. L. Eletsky, V. D. Mur, and D. V. Voskrersensky, Phys. Lett. B **80**, 68 (1978); V. S. Popov, D. V. Voskresenskiĭ, V. L. Eletskiĭ, and V. D. Mur, Zh. Éksp. Teor. Fiz. **76**, 431 (1979) [Sov. Phys. JETP **49**, 216 (1979)].

[19] V. D. Mur and V. S. Popov, Yad. Fiz. **28**, 837 (1978) [Sov. J. Nucl. Phys. **28**, 429 (1978)].

[20] P. Cea, P. Colangelo, G. Nardulli, G. Paiano, and G. Preparata, Phys. Rev. D **26**, 1157 (1982); P. Cea, G. Nardulli, and G. Paiano, Phys. Rev. D **28**, 2291 (1983).

[21] J. L. Basdevant and S. Boukraa, Z. Phys. C **28**, 413 (1985).

[22] V. M. Vaĭnberg, V. D. Mur, V. S. Popov, and A. V. Sergeev, Zh. Éksp. Teor. Fiz. **93**, 450 (1987) [Sov. Phys. JETP **66**, 258 (1987)].

[23] V. S. Popov, V. D. Mur, A. V. Sergeev, and V. M. Vaĭnberg, Phys. Lett. A **149**, 418 (1990).

[24] J. McEnnan, L. Kissel, and R. H. Pratt, Phys. Rev. A **13**, 532 (1976).

[25] M. Grant and C. S. Lai, Phys. Rev. A **20**, 718 (1979).

[26] V. L. Eletsky, V. S. Popov, and V. M. Weinberg, Phys. Lett. A **84**, 235 (1981).

Translated by Eugene Yankovsky

Supermolecular structures in nematic–cholesteric mixtures

B. I. Lev*) and S. B. Chernyshuk

Institute of Physics, Ukrainian Academy of Sciences, 252650 Kiev, Ukraine
(Submitted 28 September 1998)
Zh. Éksp. Teor. Fiz. **116**, 526–542 (August 1999)

A phenomenological free-energy density for a mixture of a nematic with a cholesteric liquid crystal is constructed on the basis of symmetry considerations. The possible spatially nonuniform distribution of the individual components of the medium is taken into account. From a microscopic description of a mixture of two cholesteric liquid crystals and a nematic with a cholesteric liquid crystal, the dependence of the reciprocal of the pitch of the cholesteric spiral on the concentration of the cholesteric liquid crystal is determined, and the results agree with all experimental data. It is shown that at mesophase temperatures a spinodal decomposition, or stratification, of the mixture is possible and the possible resulting supermolecular structures are determined. Specifically, it is found that the spinodal decomposition in nematic–cholesteric mixtures could be responsible for the distortion of the ideal spiral structure and for the appearance of a second period in the director distribution. © *1999 American Institute of Physics.* [S1063-7761(99)01208-1]

1. INTRODUCTION

Practical applications of liquid crystals are based on the variation of the space–time distribution of the director under the action of external parameters — electric and magnetic fields, light, mechanical stresses, and temperature. Cholesteric liquid crystals for information displays have found greatest applications in optics and thermography. The required working parameters can also be attained by varying the internal characteristics of the mesophase — by adding external impurities, whose effect at the microscopic level gives the required macroscopic behavior. The best known and best studied such media are induced cholesterics. The introduction of chiral impurities into a nematic liquid crystal leads to spiral twisting of the mesophase with the pitch of the spiral inversely proportional to the impurity concentration. When the concentration of the chiral additives is sufficiently high, we have nematic–cholesteric mixtures. Such mixtures have found wide applications for information displays.[1–8]

The phenomenological theory of cholesteric liquid crystals has been well-developed a long time ago[3–5] and is in good agreement with experimental data. The situation is worse for the microscopic description of spiral twisting mechanisms in nematic–cholesteric mixtures, where the character and strength of the interaction between the molecules of the nematic and chiral impurities must be taken into account.[9–22] The production of a supermolecular structure of an induced cholesteric liquid crystal, investigated by de Gennes,[2,3] presupposes that the chiral mixture embedded in the nematic liquid crystal is capable of distorting the long-range order. Since a nematic liquid crystal transfers static torques under the combined action of the impurities, a new supermolecular spiral structure is induced. This approach presupposes that the regions of distortion of the orientational order overlap and that it is more suitable for describing supermolecular structures induced by macroscopic impurities. The induction of gyrotropy would be of a threshold character with respect to the concentration of impurities, since it would be determined by the efficiency of the interaction between the macroimpurities via the distortion of the director field.

A systematic analysis of the behavior of all macroscopic impurities as a whole embedded in a nematic liquid crystal was performed in Ref. 23, where the effective interaction energy between them via the elastic field of the distortions of the director is determined and the possible supermolecular structures induced by the macroimpurities are described. To describe the behavior of the microimpurities embedded in the mesophase and also mixtures of individual liquid crystals, where the sizes of the molecules of the individual subsystems do not differ much, it is necessary to have a theory that takes into account the character and strength of the intermolecular interaction. This is the difference between the approach that we propose and the previously developed approach.[2]

The objective of our work is to construct a systematic phenomenological theory of nematic–cholesteric mixtures which take into account the possible induced supermolecular structures that are determined by the character of the intermolecular interaction. The well-developed statistical methods for describing solid interstitial or substitution solutions,[24–26] with allowance for the orientational order, can be used for this purpose and the most general dependence of the elastic constants of a mixture on the concentrations of the components can be determined. This makes it possible to describe from a unified standpoint all existing experimental data on the concentration dependence of the reciprocal of the pitch of the cholesteric spiral in nematic–cholesteric and cholesteric–cholesteric mixtures.[8,27–30] The phenomenological expression for the free energy of a nematic–cholesteric mixture can also be written on the basis of general symmetry considerations, which take into account the character of the physical quantities describing the medium.[15] It is only nec-

 © 1999 American Institute of Physics

essary to take into account the fact that in mixtures there exists, together with the director, another vector that characterizes the state of the medium — the gradient of the concentration of individual components. Then, writing out all possible invariants, we obtain the phenomenological free-energy density of the mixture. The possible supermolecular structures induced in the mixture can be determined by minimizing the free energy obtained.

The introduction of chiral impurities induces a spiral twist of the nematic liquid crystal. This twist is of a nonthreshold character, and for low cholesteric concentrations the pitch of the spiral is inversely proportional to the concentration of the cholesteric liquid crystal. Analysis of the free energy of nematic–cholesteric mixtures shows that, together with the induction of gyrotropy, the conditions for a spatially nonuniform distribution of the concentration of individual components of the system can be realized in the medium. The existence of a diffusion mode in nematic–cholesteric mixtures was indicated some time ago by Kats.[31] The existence of a spatially nonuniform distribution of impurities in the induced cholesteric liquid crystal, in turn, can lead to a distortion of the ideal spiral structure and could be responsible for the appearance of a second, incommensurably large, period in the distribution in the angle of emergence of the director, which has already been observed experimentally.[32–34] In the present paper, the conditions for the appearance of new supermolecular structures in nematic–cholesteric mixtures are determined and the temperature and concentration required for the individual subsystems in order to observe these structures are estimated.

2. MICROSCOPIC APPROACH TO THE DESCRIPTION OF LIQUID-CRYSTAL MIXTURES

In this section we shall briefly consider how the equation for the phenomenological free-energy density for a mixture of two liquid-crystal substances is derived from first principles. This will enable us to establish the general dependence of the Frank elastic constants on the concentration of one component in the mixture. Let us consider a mesophase with a combination of impurities, whose physical properties are different from those of the liquid-crystal molecules. For a microscopic description of such a medium, we shall employ the well-developed approach[24–26] in the theory of multicomponent subsitution solution with allowance for an additional degree of freedom — the orientation of the molecules. We shall therefore assume that initially the centers of gravity of the rod-shaped molecules were located at the sites of a lattice, whose form does not play a fundamental role upon switching to the continuum approximation. This model is completely identical to the model of a multicomponent alloy, the only difference being that each site is occupied by a different kind of molecule, each characterized by a continuous series of orientational states. The configurational Hamiltonian of such a system, taking into account the orientation of the molecules, can be written in the form

$$H=\frac{1}{2}\sum_{\alpha,\alpha',\mathbf{R},\mathbf{R}'}\int d\Omega_{m,m'}V_{\alpha,\alpha'}(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\times c_{\alpha}(\mathbf{m},\mathbf{R})c_{\alpha'}(\mathbf{m}',\mathbf{R}'), \tag{1}$$

where the index α determines the kind of molecule, the vector $\mathbf{m}'$ characterizes the orientation of the molecule, $\mathbf{R}$ is the radius vector of the position of the molecule, $\int d\Omega_{m,m}=\int d\Omega_m\int d\Omega_{m'}$ is the integral over possible orientations of the molecules, $V_{\alpha,\alpha'}(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')$ is the pair interaction energy between molecules of the α and α' kinds with orientations $\mathbf{m}$ and $\mathbf{m}'$ located at the points $\mathbf{R}$ and $\mathbf{R}'$, $c_{\alpha}(\mathbf{m},\mathbf{R})$ is a random function taking on the values 1 or 0, depending on whether or not a particle of the α species with orientation $\mathbf{m}$ is located at the point $\mathbf{R}$. The condition of conservation of the total number of particles assumes the form

$$N=\sum_{\alpha,\mathbf{R}}\int d\Omega_m c_{\alpha}(\mathbf{m},\mathbf{R}), \tag{2}$$

where the summation extends over all sites of the initial lattice. The partition function of the system is described by the Hamiltonian (1) with a fixed number of particles:

$$Z=\delta(N-N(c_{\alpha}))\sum_{\{c_{\alpha}=0\}}^{\{1\}}\exp\left\{-\frac{H}{\theta}\right\}.$$

The sum is calculated taking into account all possible configurations of the particles. The summation over all possible distributions of the particles in the system corresponds to a path integral over the realizations of the distribution $c_{\alpha}(\mathbf{m},\mathbf{R})$. Following Refs. 25 and 26, the partition function of the canonical ensemble can be represented as

$$Z=Z_0\exp\left\{-\frac{1}{2\theta}\sum_{\alpha,\alpha',\mathbf{R},\mathbf{R}'}\int d\Omega_{m,m'}V_{\alpha,\alpha'}\times(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\langle c_{\alpha}(\mathbf{m},\mathbf{R})\rangle\langle c_{\alpha'}(\mathbf{m}',\mathbf{R}')\rangle\right\}\times\left\langle\exp\left\{-\frac{1}{2\theta}\sum_{\alpha,\alpha',\mathbf{R},\mathbf{R}'}\int d\Omega_{m,m'}V_{\alpha,\alpha'}\times(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\Delta c_{\alpha}(\mathbf{m},\mathbf{R})\Delta c_{\alpha'}(\mathbf{m}',\mathbf{R}')\right\}\right\rangle, \tag{3}$$

where $\langle c_{\alpha}(\mathbf{m},\mathbf{R})\rangle$ is the average occupation number of molecules of the α kind with orientation $\mathbf{m}$ at the site $\mathbf{R}$, and $\Delta c_{\alpha}(\mathbf{m},\mathbf{R})=c_{\alpha}(\mathbf{m},\mathbf{R})-\langle c_{\alpha}(\mathbf{m},\mathbf{R})\rangle$. The symbol $\langle\ldots\rangle$ means averaging over an ensemble of noninteracting particles, and Z_0 is the partition function that determines the entropy component of the free energy. The free energy corresponding to (3) is[25,26]

$$F=\theta\sum_{\alpha,\mathbf{R}}\int d\Omega_m\langle c_{\alpha}(\mathbf{m},\mathbf{R})\rangle\ln\langle c_{\alpha}(\mathbf{m},\mathbf{R})\rangle+\frac{1}{2}\sum_{\alpha,\alpha',\mathbf{R},\mathbf{R}'}\int d\Omega_{m,m'}V_{\alpha,\alpha'}(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\times\langle c_{\alpha}(\mathbf{m},\mathbf{R})\rangle\langle c_{\alpha'}(\mathbf{m}',\mathbf{R}')\rangle-\theta$$

$$\times\ln\left\langle \exp\left\{ -\frac{1}{2\theta}\sum_{\alpha,\alpha',\mathbf{R},\mathbf{R}'}\int d\Omega_{m,m'}V_{\alpha,\alpha'} \times(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\Delta c_{\alpha}(\mathbf{m},\mathbf{R})\Delta c_{\alpha'}(\mathbf{m}',\mathbf{R}')\right\}\right\rangle. \quad (4)$$

In the mean-field approximation, if we use the thermodynamic perturbation theory, we can separate explicitly the free energy F_c and the correlation energy $F_{\rm cor}$ as

$$F=F_c-F_{\rm cor},$$

where

$$\begin{aligned} F_c=\theta\sum_{\mathbf{R}}\ &\{\langle c_1(\mathbf{m},\mathbf{R})\rangle\ln\langle c_1(\mathbf{m},\mathbf{R})\rangle \\ &+\langle c_2(\mathbf{m},\mathbf{R})\rangle\ln\langle c_2(\mathbf{m},\mathbf{R})\rangle\} \\ &+\frac{1}{2}\sum_{\mathbf{R},\mathbf{R}'}\int d\Omega_{m,m'}V_{1,1}(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\langle c_1(\mathbf{m},\mathbf{R})\rangle \\ &\times\langle c_1(\mathbf{m}',\mathbf{R}')\rangle+2V_{1,2}(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\langle c_1(\mathbf{m},\mathbf{R})\rangle \\ &\times\langle c_2(\mathbf{m}',\mathbf{R}')\rangle+V_{2,2}(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\langle c_2(\mathbf{m},\mathbf{R})\rangle \\ &\times\langle c_2(\mathbf{m}',\mathbf{R}')\rangle, \end{aligned} \quad (5)$$

$$F_{\rm cor}=\theta\sum_{n=1}^{\infty}\frac{M_n(x)}{(-\theta)^n n!}, \quad (6)$$

$$\begin{aligned} x&=\frac{1}{2}\int d\Omega_{m.m'}V_{\alpha,\alpha'}(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')\Delta c_{\alpha} \\ &\quad\times(\mathbf{m},\mathbf{R})\Delta c_{\alpha'}(\mathbf{m}',\mathbf{R}'), \\ M_1(x)&=\langle x\rangle,\quad M_2(x)=\langle x^2\rangle-\langle x\rangle^2, \end{aligned}$$

and $M_n(x)$ are semi-invariants of order n. In what follows, we shall stay within the mean-field approximation and switch to a continuum description of the system by replacing summation by integration

$$\sum_R\ \to\rho\int dV,\quad \rho=\frac{N}{V}.$$

In addition, we assume that the spatial and orientational distributions of the particles are independent of one another. Then the total distribution function can be represented as a product of the spatial concentration and orientational distribution function as

$$\langle c_1(\mathbf{m},\mathbf{R})\rangle=(1-c(\mathbf{R}))f(\mathbf{m},\mathbf{R}),$$

$$\langle c_2(\mathbf{m},\mathbf{R})\rangle=c(\mathbf{R})g(\mathbf{m},\mathbf{R}),$$

where $c(\mathbf{R})$ is the concentration of one component, and $f(\mathbf{m},\mathbf{R})$ and $g(\mathbf{m},\mathbf{R})$ are the orientational distribution functions of the liquid-crystal and impurity molecules, respectively. We assume $c(\mathbf{R})$ to be the concentration, or relative fraction, of the cholesteric component in a nematic–cholesteric mixture. For our further arguments we assume that the orientational distribution functions of the nematic liquid crystal and the impurity are identical, so that only the spatial distribution functions of the concentration differ at each point. This means that the nematic liquid crystal and the impurity molecules are characterized by the same order parameter. This assumption is valid if it is assumed that the molecules of the impurity and the nematic liquid crystal are indistinguishable with respect to geometric dimensions and the character of the intermolecular interaction. All these assumptions are valid for typical nematic–cholesteric mixtures. Averaging the intermolecular interaction potential over the orientational distribution functions causes the appearance of a product of the squared order parameter and an effective interaction, which depends on the director vector at different points of the medium:[9–18]

$$s^2\tilde{V}_{ij}(\mathbf{r},\mathbf{n}(\mathbf{R}),\mathbf{n}(\mathbf{R}+\mathbf{r}))=\int d\Omega_{m,m'}V_{ij}(\mathbf{m},\mathbf{R},\mathbf{m}',\mathbf{R}')f \times(\mathbf{m},\mathbf{R})f(\mathbf{m}',\mathbf{R}'). \quad (7)$$

Taking this fact into account and setting $f(\mathbf{m},\mathbf{R})=g(\mathbf{m},\mathbf{R})$, we can write the free energy of a liquid crystal with an impurity in the form

$$\begin{aligned} F_c=\theta\rho\int d\mathbf{R}c(\mathbf{R})\ln c(\mathbf{R})+(1-c(\mathbf{R}))\ln(1-c(\mathbf{R})) \\ +\frac{\rho^2s^2}{2}\int d\mathbf{R}\int d\mathbf{r}(\mathbf{1}-\mathbf{c}(\mathbf{R})) \\ \times(1-c(\mathbf{R}'))\tilde{V}_{1,1}(\mathbf{r},\mathbf{n},\mathbf{n}')+2c(\mathbf{R}') \\ \times(1-c(\mathbf{R}))\tilde{V}_{1,2}(\mathbf{r},\mathbf{n},\mathbf{n}')c(\mathbf{R})c(\mathbf{R}')\tilde{V}_{2,2}(\mathbf{r},\mathbf{n},\mathbf{n}'). \end{aligned} \quad (8)$$

Analysis of this formula makes it possible to understand the dependence of the elastic constants of the liquid-crystal mixture on the concentration of one component. It can be concluded immediately that this dependence has the form

$$K_{ij}(c)=(1-c)^2K_{ij}^{\alpha}+2c(1-c)K_{ij}^{\alpha\beta}+c^2K_{ij}^{\beta}, \quad (9)$$

where K_{ij}^{α} and K_{ij}^{β} are the elastic constants of the first and second components, respectively; they are due to the intermolecular interaction of molecules of the same kind. The elastic constants $K_{ij}^{\alpha\beta}$ are due to the interaction between molecules of different kinds. In the most general case the average intermolecular interaction potential can be written as[11–18]

$$\begin{aligned} V(\mathbf{n},\mathbf{n}',\mathbf{r})=&I(\mathbf{n}\cdot\mathbf{n}',\mathbf{n}\cdot\mathbf{r},\mathbf{n}'\cdot\mathbf{r},r) \\ &+J(\mathbf{n}\cdot\mathbf{n}',\mathbf{n}\cdot\mathbf{r},\mathbf{n}'\cdot\mathbf{r},r)\cdot(\mathbf{n}\times\mathbf{n}')\cdot\mathbf{r}. \end{aligned}$$

Here the first term describes the orientational order of the molecules, and the second term is responsible for the appearance of the Lifshitz invariant $\mathbf{n}\times\operatorname{curl}\mathbf{n}$ or for inducing a cholesteric spiral. The second term is absent in nematic liquid crystals, and therefore for nematic–cholesteric mixtures we can immediately write the dependence of the reciprocal of the pitch of the spiral on the concentration of the cholesteric liquid crystal as

$$q_{\text{nem-chol}}=\frac{K_2(c)}{2K_{22}(c)}$$

$$=\frac{2c(1-c)K_2^{\alpha\beta}+c^2K_2^{\beta}}{(1-c)^2K_{22}^{\alpha}+2c(1-c)K_{22}^{\alpha\beta}+c^2K_{22}^{\beta}}, \qquad (10)$$

where K_2 is the elastic constant in the Lifshits invariant. This formula describes the entire diversity of experimental data with the corresponding parameters.[8,27–30] For low concentrations the linear dependence of the reciprocal of the pitch of the spiral on the concentration of the cholesteric liquid crystal is determined by the chiral interaction between molecules of cholesteric and nematic liquid crystals (CLCs and NLCs). If it is assumed that the constant in the denominator varies very little, $K_{22}(c)\approx$ const, then a simple analysis shows that when the signs of the constants $K_{22}^{\alpha\beta}$ and K_{22}^{β} are different, i.e., the signs of the chiral interactions CLC–NLC and CLC–CLC are different, the dependence $q(c)$ has a maximum ($K_{22}^{\alpha\beta}>0$) at the concentration

$$c_{\max}=\frac{K_2^{\alpha\beta}}{2K_2^{\alpha\beta}-K_{22}}. \qquad (11)$$

When the concentration is doubled, $c_0=2c_{\max}$, a transition of the mixture into the nematic phase is observed, and at high concentrations the twist changes sign. This situation occurs for mixtures of right-hand cholesterics with nematic liquid crystals, for example, for a mixture of MBBA with cholesteryl chloride.[27] However, if the signs of the constants $K_{22}^{\alpha\beta}$ and K_{22}^{β} are the same but $|K_2^{\alpha\beta}|>|K_{22}|$, then the function $q(c)$ possesses a maximum at the point $c_{\max}$ which does not vanish anywhere. This corresponds to a mixture of cholesteryl propionate with butoxybenzylidene-*n*-butylaniline.[8] For $|K_2^{\alpha\beta}|<|K_{22}|$ the reciprocal of the pitch increases monotonically with the cholesteric concentration and reaches its maximum value at the reciprocal of the pitch in the pure cholesteric liquid crystal. This is observed for a mixture of cholesterylpropioanate with butoxybenzylideneaminobenzonitrile.[8]

A similar formula can be written for the reciprocal of the pitch of the spiral in a mixture of two cholesteric liquid crystals:

$$q_{\text{chol-chol}}=\frac{K_2(c)}{2K_{22}(c)}$$

$$=\frac{K_2^{\alpha}(1-c)^2+K_2^{\alpha\beta}2c(1-c)+K_2^{\beta}c^2}{K_{22}^{\alpha}(1-c)^2+K_{22}^{\alpha\beta}2c(1-c)+K_{22}^{\beta}c^2}. \qquad (12)$$

The denominator in this formula is always positive, and apparently it is essentially independent of the concentration, since it determines the orientational order of the molecules in a preferred direction. The numerator is determined by the chiral interaction between molecules in a manner so that the constants K_2^{α}, K_2^{β}, and $K_2^{\alpha\beta}$ can have different signs. Depending on the magnitude of these constants, various experimental situations, which are described in detail in Ref. 8, can arise. For example, for $K_2^{\alpha}-K_2^{\alpha\beta}>0$ and $K_2^{\beta}-K_2^{\alpha\beta}>0$ the concentration dependence of the reciprocal of the pitch of the spiral has a minimum. If the conditions $K_2^{\alpha},K_2^{\beta}>0$ and $K_2^{\alpha\beta}<0$ are satisfied, then for sufficiently large $|K_2^{\alpha\beta}|$ the dependence of the reciprocal of the pitch vanishes twice in a manner so that the mixture is left-handed at low and high concentrations and right-handed at intermediate concentrations. This situation occurs for a mixture of cholesteryl-2-(2-ethoxy-ethoxy)ethyl carbonate with amyl-*n*-(4-cyanobenzylideneamino) cinnamate.[8]

3. SUPERMOLECULAR STRUCTURES IN NEMATIC–CHOLESTERIC MIXTURES

We now give a phenomenological description of a nematic–cholesteric mixture and find the possible supermolecular structures in it. Using the long-wavelength expansions of the director and concentration in the form

$$n_{\alpha}(\mathbf{R}')=n_{\alpha}(\mathbf{R}+\mathbf{r})=n_{\alpha}(\mathbf{R})+r_{\beta}\partial_{\beta}n_{\alpha}(\mathbf{R})+\frac{1}{2}r_{\beta}r_{\gamma}\partial_{\beta}\partial_{\gamma}n_{\alpha}(\mathbf{R}),$$

$$c(\mathbf{R}')=c(\mathbf{R}+\mathbf{r})=c(\mathbf{R})+r_{\beta}\partial_{\beta}c(\mathbf{R})+\frac{1}{2}r_{\beta}r_{\gamma}\partial_{\beta}\partial_{\gamma}c(\mathbf{R}),$$

we obtain from Eq. (8), after some simple transformations, the final expression for the free-energy density of a nematic with a cholesteric liquid crystal, with allowance for the possible spatially nonuniform distributions of the concentration and director and all symmetry invariants of the system:

$$\begin{aligned}F_c=&M(\nabla\alpha)^2-N\alpha^2+D\alpha^4+(K_{11}+w_1\alpha)(\operatorname{div}\mathbf{n})^2\\&+(K_{22}+w_2\alpha)(\mathbf{n}\times\operatorname{curl}\mathbf{n})^2+(K_{33}+w_3\alpha)\\&\times(\mathbf{n}\times\operatorname{curl}\mathbf{n})^2+(K_2+v\alpha)(\mathbf{n}\times\operatorname{curl}\mathbf{n})\\&+\mathbf{A}_1\operatorname{div}\mathbf{n}(\mathbf{n}\cdot\nabla\alpha)\\&+A_2(\mathbf{n}\times\operatorname{curl}\mathbf{n})\cdot\nabla\alpha+A_3(\mathbf{n}\cdot\nabla\alpha)^2. \qquad (13)\end{aligned}$$

Here α is the deviation of the cholesteric concentration from a uniform background:

$$c(\mathbf{R})=c_0+\alpha(\mathbf{R}).$$

The spatially nonuniform distribution of individual components presumes that the medium is separated into sections with different concentrations of the chiral (or nematic) molecules. In this stratification there are no sharp boundaries between regions with different concentrations. This is a smooth spatial redistribution of the concentration of individual components in the sample. Both components mix well, but for a certain character and strength of the intermolecular interaction it is energetically more favorable for them to produce a nonuniform distribution of the individual constituents of the medium. This process proceeds purely by diffusion, but it is caused by the internal intermolecular interactions. A similar phenomenon is observed in solid interstitial and substitution solutions,[24] the only difference being that in the latter there is spatial order, i.e., a crystal lattice, while in liquid-crystal mixtures there is no spatial order but there is orientational order of long molecules around a preferred direction — the director vector.

The quantities w_i and v are derivatives of the corresponding Frank constants at the equilibrium concentration:

$$w_i=\left(\frac{dK_i(c)}{dc}\right)_{c_0} \quad \text{and} \quad v=\left(\frac{dK_2(c)}{dc}\right)_{c_0}.$$

All information about the induction of gyrotropy is contained in the constant $K_2(c)$, which for a mixture depends on the concentration more strongly than all other Frank constants. Therefore it can be asserted that $|w_i| \ll |v|$. In what follows we assume that $w_i=0$.

The expression (13) is the Landau free energy of a nematic–cholesteric mixture at temperatures below the spinodal decomposition temperature of the mixture. It contains two order parameters — the director and the concentration nonuniformity. The phenomenological constants M, A_i, and K_{ii} are determined in terms on the microscopic interaction potentials and, in order of magnitude, are proportional to U_0/λ_0, where U_0 is the characteristic intermolecular interaction energy, and λ_0 is the average distance between the molecules. For example, for the interaction $V(r)(\mathbf{m}\cdot\mathbf{m}')^2$ we obtain the following expression in the one-constant approximation:

$$K=-\frac{4\pi}{3}\rho^2 s^2\int_{\lambda_0}^{\infty}V(r)r^4dr.$$

Similarly, using the chiral interaction potential $W(r)\times(\mathbf{m}\cdot\mathbf{m}')(\mathbf{m}\times\mathbf{m}')\cdot\mathbf{r}$, we obtain the constant K_2 in the form

$$K_2=-\frac{4\pi}{3}\rho^2 s^2\int_{\lambda_0}^{\infty}W(r)r^4dr.$$

From physical considerations the constant N should have the form

$$N=a(T^*-T),$$

where T^* is the spinodal decomposition temperature of the given mixture in the isotropic phase. In other words, the temperature T^* would be the spinodal decomposition temperature if it were higher than the temperature of the nematic–cholesteric mixture — isotropic liquid transition. This temperature does not depend on the orientational order parameter s and is determined exclusively by the part of the microscopic interaction potential between two molecules that does not depend on the orientation of the molecules. Using the second concentration derivative of the first entropy term in Eq. (8), we estimate this temperature to be

$$T^*\simeq\frac{U_0}{k_b}c(1-c).$$

The characteristic molecular interaction energy at a distance equal to the average range of the molecular forces is $U_0\sim 0.01-0.1$ eV, which gives the estimate $T^*\sim(10^2-10^3)c(1-c)$ K. It should be noted that the phenomenological expression for the free-energy (13) could be written immediately without invoking the microscopic theory. The free energy should contain invariants with derivatives of the director and concentration of degree no higher than second. Here we have restricted the analysis to terms that are not total derivatives. However, invoking the microscopic theory, we can put in a more concrete form the dependence of the elastic constants and the spinodal decomposition temperature on the concentration of the cholesteric liquid crystal. Now we can study directly the supermolecular structures that are possible in nematic–cholesteric mixtures.

3.1. Supermolecular structures along the axis of the spiral

Let us consider the expression (13) in the one-constant approximation for the case where all variables depend only on the coordinate z along the axis of the spiral. The director can be represented as

$$\mathbf{n}=(\cos\varphi\cos\theta,\ \sin\varphi\cos\theta,\ \sin\theta).$$

Then the free energy can be rewritten as

$$F_c=M\dot{\alpha}^2-N\alpha^2+D\alpha^4+K(\dot{\theta}^2+\cos^2\theta\cdot\dot{\phi}^2)-(K_2+v\alpha)\times\cos^2\theta\cdot\dot{\phi}+\frac{\Delta A}{2}\sin(2\theta)\cdot\dot{\theta}\dot{\alpha}+A_3\sin^2\theta\cdot\dot{\alpha}^2, \qquad (14)$$

$\Delta A=A_1-A_2$.

We see that the twist $\dot{\phi}$ enters by itself as an independent variable. For this variable the Euler–Lagrange equation can therefore be written in the form

$$\frac{\partial F_c}{\partial\dot{\phi}}=0.$$

This leads to the following relation between the local twist at the point z and the concentration at this point:

$$\dot{\phi}=\frac{K_2+v\alpha}{2K}. \qquad (15)$$

This relation shows that the concentration and director distributions are related to one another, and to determine equilibrium concentration distribution this expression for the twist must be substituted into the initial functional (14). The free energy density will then depend on the two variables θ and α. The appearance of the concentration nonuniformity α below the spinodal decomposition temperature can cause the director to leave the nematic–cholesteric plane. We assume that the angle θ of emergence of the director from the cholesteric plane is small. Then, expanding in powers of θ up to the quadratic term and substituting the expression (15) for the twist, we can write the free-energy density of the mixture in the form

$$F_c-F_0=M\dot{\alpha}^2-\tilde{N}\alpha^2+D\alpha^4+K\dot{\theta}^2+\frac{(K_2+v\alpha)^2}{4K}\theta^2+\Delta A\,\theta\dot{\theta}\dot{\alpha}+A_3\theta^2\alpha^2. \qquad (16)$$

Here $F_0=-Kq^2$ is the free-energy density of an ideal spiral; we took into account the normalization condition $\int\alpha(z)dz=0$. As one can see from this expression, the local relation between the concentration and the twist (15) leads to renormalization of the coefficient N and a decrease of the spinodal decomposition temperature T^* of the isotropic liquid:

$$\tilde{N}=N+q_c'^2K=a(T_c-T),$$

$$T_c=T^*-\frac{q_c'^2K}{a}.$$

Here T_c is the spinodal decomposition temperature of the nematic–cholesteric mixture, which already depends on the orientational order parameter s via the elastic constants. Since the order parameter is itself temperature-dependent, the following equation can be obtained for the spinodal decomposition temperature:

$$T_c = \frac{c(1-c)}{k_b}(U_0 - s^2(T_c)V_0), \tag{17}$$

where U_0 and V_0 are of the same order of magnitude and determine the characteristic intermolecular interaction energy, where U_0 determines the orientation–independent part of the interaction, and V_0 describes the chiral interaction between the CLC and NLC molecules. Since the order parameter is always less than 1, the spinodal decomposition temperature T_c differs very little from the temperature T^*, appearing in front of α^2 in the Landau expansion, and it can also be estimated as

$$T_c \sim (10^2 - 10^3)c(1-c)K,$$

which falls into the temperature range of the mesophase. It is maximum for concentration $c = 1/2$, i.e., for an equal number of nematic and cholesteric molecules in the nematic–cholesteric mixture.

On the basis of the Onsager kinetic equations it can be shown[24] that for spinodal decomposition the wave with wave number

$$g = \frac{1}{\sqrt{2}}\sqrt{\frac{\tilde{N}}{M}} \sim \frac{1}{\lambda_0}\sqrt{\frac{T_c - T}{T_c}},$$

where λ_0 is the average distance between the molecules ($\lambda_0 = 50$ Å), grows most rapidly. The wavelength of the concentration wave can then be determined by the relation

$$\lambda_c \sim \lambda_0 \sqrt{\frac{T_c}{T_c - T_0}}. \tag{18}$$

We assume that spinodal decomposition already occurs in the liquid-crystal phase that has been formed; i.e., the temperature T_c is less than the temperature of the transition into the isotropic phase. We shall consider the situation where the temperature is only negligibly less than the temperature T_c and we can assume that $T_c/(T_c - T) \sim 100$. The wavelength of the concentration wave is then estimated to be $\lambda_c \sim 500$ Å. The time required to establish such a concentration wave can be estimated assuming that this is a purely diffusion process. If it is assumed that the size of the nonuniformity is $\lambda \sim 500$ Å, and that the typical diffusion coefficient of the liquid-crystal molecules[3,5] is $D \approx 4 \times 10^{-6}$ cm^2/s, then the time required to establish the nonuniformity is $\tau \approx \lambda^2/D \approx 6 \times 10^{-6}$ s. For this reason, we shall consider a concentration wave to be established, and we assume that the nonuniform twist corresponding to this concentration wave is formed in accordance with Eq. (15). We write the concentration wave in the form

$$\alpha = \alpha_0 \cos(gz),$$

and we shall examine the conditions under which this concentration nonuniformity will lead additionally to the emergence of the director from the cholesteric plane. Both α and θ are zero in the ideal-spiral state, and therefore below the spinodal decomposition temperature both variables are infinitesimals of the same order of magnitude:

$$\theta \sim \varepsilon \quad \text{and} \quad \alpha \sim \varepsilon.$$

For a systematic analysis we shall first restrict the analysis to the terms proportional to the third power of ε inclusively, i.e., we drop in Eq. (16) terms proportional to α^4 and $\theta^2\alpha^2$:

$$\delta F_\theta = K\dot{\theta}^2 + (b_0 + b_1\alpha)\dot{\theta}^2 + \Delta A\,\theta\dot{\theta}\dot{\alpha},$$

$$b_0 = Kq^2, \quad b_1 = qv.$$

We shall seek the solution for the angle $\theta(z)$ in the form

$$\theta(z) = \theta_0 \sin(\mu z - \chi(z)), \tag{19}$$

where the function $\chi(z)$ varies much more slowly than the linear term in the phase, $|\chi'(z)| \ll \mu$. Substituting this ansatz, we obtain the standard functional[35–39]

$$\frac{\delta F_\theta}{\theta_0^2} = \frac{K}{8}\{(g - \dot{\psi})^2 - 2\xi\cos\psi + 4q^2\}, \tag{20}$$

which describes the incommensurate phase in the distribution of the angle $\theta(z)$. Here we took into account that $g = 2\mu$ and $\psi(z) = 2\chi(z)$, and the constant ξ has the form

$$\xi = \frac{\alpha_0}{K}\left(Kq^2 + \frac{\Delta A g^2}{2}\right).$$

For a nontrivial solution to exist the condition $\delta F_\theta < 0$ must be satisfied. The condition of an extremum for the functional (20) has the form of the sine-Gordon equation

$$\frac{d^2\psi}{d^2z} - \xi\sin\psi = 0$$

with the solution

$$z - z_0 = \int_0^\psi \frac{dx}{\sqrt{C - 2\xi\cos x}}, \tag{21}$$

where $C > 2\xi$ is the integration constant. For the class of solutions (21) the free energy is a function of the parameter C, which is determined by minimizing the free energy. Substituting into Eq. (20) the solution obtained and using the relation

$$\frac{d\psi}{dz} = (C - 2\xi\cos\psi)^{1/2},$$

we obtain the final expression for the free energy as a function of the parameter C:

$$\frac{\delta F_\theta}{\theta_0^2} = \frac{K}{8}\left\{C + 4q^2 + g^2 - \frac{4g\pi}{l} - \frac{4\xi}{l}\int_0^{2\pi}\frac{dx\cos x}{\sqrt{C - 2\xi\cos x}}\right\}, \tag{22}$$

where l is the period of variation of $\psi(z)$, which can be expressed in terms of the elliptic functions as

$$l = \int_0^{2\pi}\sqrt{C - 2\xi\cos x}\,dx = \frac{2\gamma}{\sqrt{\xi}}K(\gamma), \quad \gamma^2 = \frac{4\xi}{C + 2\xi}.$$

Finally, the free energy can be expressed in terms of the elliptic functions as

$$\frac{\delta F_\theta}{\theta_0^2}=\frac{K}{8}\left\{C+4q^2+g^2-\frac{2g\pi\sqrt{\xi}}{\gamma K(\gamma)}+4\xi\left[1-\frac{2}{\gamma^2}+\frac{2E(\gamma)}{\gamma^2K(\gamma)}\right]\right\}. \tag{23}$$

We can investigate the asymptotic behavior of this expression. As $C\to 2\xi$ $(\gamma\to 1)$, we have

$$E(\gamma)\to 1, \quad K(\gamma)\to\ln\left(\frac{4}{\sqrt{1-\gamma^2}}\right).$$

It can be concluded on the basis of this asymptotic behavior that the free energy (22) possesses a minimum with respect to C, i.e., $\psi\neq 0$ and an incommensurate phase exists if

$$\xi<\xi_c=\frac{\pi^2g^2}{16}. \tag{24}$$

If this condition is not satisfied, then the expression (22) possesses a minimum at $C=2\xi$, $\psi=0$. In this case the energy is

$$\frac{\delta F_\theta}{\theta_0^2}=\frac{K}{8}\{4q^2+g^2-2\xi\}. \tag{25}$$

The condition $\delta F_\theta<0$ requires that the angle θ be nonzero if the following relation is satisfied:

$$4q^2+g^2<2\xi=2\frac{\alpha_0}{K}\left(Kq^2+\frac{\Delta Ag^2}{2}\right). \tag{26}$$

This relation relates the reciprocal q of the pitch of the cholesteric spiral, the wave number g of the concentration wave, and the amplitude α_0 of the concentration wave. If this relation is satisfied, then the concentration nonuniformity (the concentration wave) will lead not only to nonuniform twisting but also to periodic emergence of the angle θ from the cholesteric plane with a period twice that of the concentration wave. In addition, if $2\xi<\pi^2g^2/8$ or, equivalently,

$$g>\frac{2q}{\sqrt{\frac{\pi^2}{8}-1}}\approx 4q,$$

then a phase $\psi(z)\neq 0$, incommensurate with respect to the period of the concentration wave and the pitch of the cholesteric spiral, will appear in the distribution $\theta(z)$ of the angle. This relation between the wave vectors of the concentration wave and the cholesteric spiral actually always holds, since estimates for the period of the concentration wave give a value of the order of 500 Å , which is much smaller than the pitch of the cholesteric spiral, $P\sim 5000$ Å. Therefore it can be assumed that $g\gg q$. We can then rewrite the necessary energy condition (26), under which the angle θ is nonzero:

$$\Delta A>\frac{K}{\alpha_0}. \tag{27}$$

If for a definite amplitude α_0 of the concentration wave the difference $\Delta A=A_1-A_2$ is a factor of $1/\alpha_0$ greater than the Frank elastic constant K, then the director will periodically leave the cholesteric plane in the direction of the axis of the spiral. The angle of emergence will have the form

$$\theta=\theta_0\sin\left(\frac{gz}{2}-\chi(z)\right).$$

At short distances this is a periodic function with a ''fast'' period $4\pi/g$, twice that of the concentration wave. The function $\chi(z)$ changes very little at such distances. However, at large distances the function $\chi(z)$ has the form of a ''devil's staircase'' and changes periodically by π with the period $l\gg 4\pi/g$. This is how the existence of the incommensurate phase is manifested in the distribution $\theta(z)$ of the angle.

However, if the relation between the elastic constants and the amplitude of the concentration wave $\Delta A>K/\alpha_0$ does not hold as the amplitude α_0 of the wave increases, then the spinodal decomposition of the mixture will be accompanied only by self-consistent adjustment of the local twist to the given concentration nonuniformity without the director emerging from the cholesteric plane (15). For large amplitudes of the concentration wave the terms proportional to α^4 in the Landau expansion can no longer be ignored. These terms lead to a bound on the amplitude of the concentration wave and establish an equilibrium configuration in the concentration distribution $\alpha(z)$.

Let us now consider a configuration such that the director does not leave the cholesteric plane. We write the Landau free energy up to terms α^4 as

$$F_c-F_0=M\dot{\alpha}^2-\tilde{N}\alpha^2+D\alpha^4. \tag{28}$$

Here $F_0=-Kq^2$ is the free-energy density of an ideal spiral, $\tilde{N}=a(T_c-T)>0$. The free energy is a functional similar to the Lagrangian $L=T-U$, where the variable z plays the role of time. The potential energy is a quartic polynomial $U=\tilde{N}\alpha^2-D\alpha^4$. It has a maximum at the point $\alpha^2_{\max}=\tilde{N}/2D$. The conserved quantity is the energy

$$H=T+U=M\dot{\alpha}^2+\tilde{N}\alpha^2-D\alpha^4.$$

Different periodic distributions $\alpha(z)$ correspond to different oscillations in the potential U with total energy H. The maximum energy at which a periodic solution $\alpha(z)$ still exists corresponds to the total energy

$$H=U_{\max}=\frac{\tilde{N}^2}{4D}.$$

The concentration distribution corresponding to this energy can be found from the law of conservation of energy

$$M\dot{\alpha}^2+\tilde{N}\alpha^2-D\alpha^4=\frac{\tilde{N}^2}{4D},$$

which gives

$$\alpha=\alpha_{\max}=\coth\sqrt{\frac{\tilde{N}}{2M}}z. \tag{29}$$

This formula describes the separation of the mixture at temperatures $T<T_c$ into two regions with a different fraction of the cholesteric component. Correspondingly, each component will have its own twist (or pitch of the spiral). The total decrease of the free energy of the nematic–cholesteric mixture, as compared with the free energy of an ideal spiral for a given configuration, has the form

$$\mathfrak{F}_c-\mathfrak{F}_0=-V\frac{\tilde{N}^2}{4D}<0, \tag{30}$$

where V is the volume of the system. A similar separation into two phases was studied by de Gennes and Brochard,[2] who investigated a suspension of macroscopic magnetic impurities in a nematic in the presence of a magnetic field. They showed that a magnetic field, parallel to the director, applied to a suspension with compensated macroscopic magnetic moment leads to separation of the suspension into two parts with oppositely directed magnetic moments.

We can therefore draw certain conclusions. A concentration nonuniformity leads to a nonuniformity of the twisting of the spiral along the z axis in accordance with Eq. (15), so that the nonuniformity of twisting is directly proportional to the concentration nonuniformity. If the spinodal decomposition temperature T_c is less than the temperature of the transition into the isotropic phase, then for $T<T_c$ two scenarios are possible. In the first one, if the relation (27) is satisfied, then a concentration wave will cause the director vector to leave the cholesteric plane periodically. The distribution $\theta(z)$ of the angle of emergence of the director has a large period that is incommensurate with the period of the concentration wave. In the second scenario, if the relation (27) is not satisfied, then the concentration nonuniformity will grow and an equilibrium stratification of the entire sample into two phases with different contents of the cholesteric liquid crystal in each one will be established. In addition, a different pitch of the cholesteric spiral will be established in each phase. The thickness of the transitional region is of the order of

$$\delta l\sim\lambda_0\sqrt{\frac{T_c}{T_c-T}},$$

where λ_0 is the distance between the molecules. Now, we shall examine separately the case where a concentration wave appears in a direction perpendicular to the axis of the spiral.

3.2. Supermolecular structure in a direction perpendicular to the axis of the spiral

Let us consider an arbitrary direction perpendicular to the axis of the cholesteric spiral. We denote it by x. We return to the most general functional (13), which takes into account the spatially nonuniform distribution of the components in the nematic–cholesteric mixtures. As shown above, we can set $w_i=0$ and take into account only the cofactor, which is associated with the nonuniformity of the concentration, in the Lifshitz invariant. We shall use the one-constant approximation. We also set $A_1=A_2=A$, so that $\Delta A=0$. We can then rewrite the free-energy density of the mixture in the form

$$\begin{aligned}F_c=&M(\nabla\alpha)^2-N\alpha^2+K((\mathrm{div}^2\,\mathbf{n})+(\mathrm{curl}^2\,\mathbf{n}))\\&+(K_2+\nu\alpha)(\mathbf{n}\times\mathrm{curl}\,\mathbf{n})+A(\mathrm{div}\,\mathbf{n})(\mathbf{n}\cdot\nabla\alpha)\\&+(\mathbf{n}\times\mathrm{curl}\,\mathbf{n})\cdot\nabla\alpha).\end{aligned} \tag{31}$$

We write the director in the form $\mathbf{n}=(\cos\varphi\cos\theta,\sin\varphi\cos\theta,\sin\theta)$, where θ is the angle between the director and the cholesteric plane. We set $\alpha=\alpha(x)$, $\varphi_x=0$, $\theta=\theta(x,z)$, and $\varphi_z=q$. Actually, we wish to find for a wave propagating in a direction perpendicular to the axis of the spiral at least one solution with energy less than that of the undistorted spiral. We assume the twist along the axis of the spiral to be constant and equal to q. We shall take into account only the terms which are second-order infinitesimals in the angle θ and the concentration nonuniformity α. We can then rewrite the free-energy density in the form

$$\begin{aligned}&F_c-F_0=M\alpha_x^2-N\alpha^2+\delta F_\theta,\\&\delta F_\theta=K(\theta_x^2+\theta_z^2+\theta^2q^2)+A\alpha_x\theta_z\cos(qz)\\&\qquad+Aq\alpha_x\theta\sin(qz)-\nu\theta_x\alpha\sin(qz).\end{aligned} \tag{32}$$

Here δF_θ is the elastic part of the free energy, which is associated with the director. The combinations $\theta_x\alpha\sin(qz)$, $\theta_z\alpha_x\cos(qz)$, and $\theta\alpha_x\sin(qz)$ are encountered in the last three terms. They all have the same form if we write the ansatz for the concentration and angle in the form

$$\alpha=\alpha_0\cos(gx),\quad \theta=\theta_0\sin(gx)\sin(qz). \tag{33}$$

The elastic part of the free energy will then assume the form

$$\delta F_\theta=\frac{K\theta_0^2}{4}(g^2+2q^2)-\frac{g\theta_0\alpha_0}{4}(2qA+\nu).$$

From the condition $\partial\delta F_\theta/\partial\theta_0=0$ for a minimum of the free energy we obtain a relationship between the amplitude of the concentration wave and the amplitude of the angle of emergence of the director from the cholesteric plane:

$$\theta_0=\alpha_0\frac{(\nu+2qA)g}{2K(g^2+2q^2)}. \tag{34}$$

The free-energy density for such a structure will then be determined by the relation

$$\delta F_\theta=-\alpha_0^2\frac{(\nu+2qA)^2g^2}{16K(g^2+2q^2)}<0. \tag{35}$$

In summary, we conclude that a concentration nonuniformity in a direction perpendicular to the axis of the spiral leads to a two-period structure in the orientation of the director. The angle of emergence of the director from the cholesteric plane possesses a period along the axis of the spiral that equals exactly the pitch of the cholesteric spiral and it possesses an additional period in a perpendicular direction, equal to the period of the concentration nonuniformity in this direction.

Physically, this means that at locations where the concentration of the cholesteric liquid crystal is higher, the rate of twist is also higher. This applies for directions parallel and perpendicular to the spiral.

4. DISCUSSION

In our study we have systematically determined the phenomenological free-energy density for an arbitrary mixture of two liquid-crystal substances. We used a procedure and methods similar to the microscopic theory of multicomponent interstitial and substitution solid solutions[24] but with allowance for the orientational interaction of individual molecules.[25,26] We have thus obtained the dependence of the Frank elastic constants on the concentration of individual components. Special attention was devoted to a mixture of a nematic and a cholesteric liquid crystal. A general expression in agreement with all possible experimental situations was obtained for the dependence of the reciprocal of the pitch of the spiral on the concentration of the cholesteric liquid crystal.[8,27–30] It was shown that spinodal decomposition, i.e., spatial concentration separation of the mixture, is possible in a mixture of a nematic with a cholesteric liquid crystal at mesophase temperatures. The period of the spatially modulated distribution of individual components exhibits critical behavior as the spinodal decomposition temperature is approached. The most general expression for the free-energy density, which takes into account the concentration separation of the mixture, was obtained, and the possible supermolecular structures in the distribution of the director and concentration were analyzed. Two cases of a concentration wave were examined systematically — in the direction of the pitch of the spiral and in a perpendicular direction. Two possible scenarios of spinodal decomposition were constructed for a concentration wave along the spiral. First, as the amplitude of the wave increases, the director adjusts self-consistently, which necessarily leads to a nonuniform twist along the axis of the spiral. In addition, the director can leave the cholesteric plane periodically. The distribution $\theta(z)$ possesses a period that is incommensurate with the period of the concentration wave. Second, if the condition (15) for emergence of the director is not satisfied, then spinodal decomposition in the direction of the spiral leads to further separation of the medium into two phases with a different fraction of the cholesteric liquid crystal in each one. Correspondingly, each phase has its own pitch of the cholesteric spiral. Both scenarios lead to the existence of a second period in the distribution of the director in the direction of the axis of the spiral for a nematic–cholesteric mixture. The appearance of a second period in the director distribution was observed experimentally in Refs. 32–34.

We have shown that a concentration wave in a direction perpendicular to the axis of the spiral leads to double-periodic emergence of the director from the cholesteric plane. The period of variation of the angle of emergence along the axis of the spiral equals the pitch of the spiral, while the period of variation of the angle of emergence in a perpendicular direction equals the period of the concentration wave in this direction.

We are deeply grateful to P. M. Tomchuk for numerous discussions and V. M. Pergamenshchik for suggesting possible invariants associated with concentration gradients. One of us (B.I.L.) acknowledges financial support from the CRDF Foundation (Grant No. UEI-310).

*)E-mail: lev@iop.kiev.ua

[1] S. Chandrasekhar, in *Cholesteric Liquid Crystals* (New York, 1984), p. 213.
[2] F. Brochard and P. G. de Gennes, J. de Phys. **31**, 691 (1970).
[3] P. de Gennes, *The Physics of Liquid Crystals* 2nd edition (Clarendon Press, Oxford, 1993) [Russian translation, Mir, Moscow, 1977].
[4] G. C. Chilaya and L. N. Lisetskiĭ, Usp. Fiz. Nauk **134**, 279 (1981) [Sov. Phys. Usp. **24**, 496 (1981)].
[5] L. M. Blinov, *Electro- and Magneto-Optics of Liquid Crystals* (Wiley, New York, 1983) [Russian original, Nauka, Moscow, 1978].
[6] S. A. Pikin, *Structural Transformations in Liquid Crystals* (Gordon and Breach, New York, 1991) [Russian original, Nauka, Moscow, 1981].
[7] S. Chandrasekhar, *Liquid Crystals* (Cambridge Univeristy Press, New York, 1977) [Russian translation, Mir, Moscow, 1980].
[8] V. A. Belyakov and A. S. Sonin, *Optics of Cholesteric Liquid Crystals* (Nauka, Moscow, 1982) p. 165.
[9] G. S. Chilaya and L. N. Lisetski, Mol. Cryst. Liq. Cryst. **140**, 243 (1986).
[10] D. N. Keating, Mol. Cryst. Liq. Cryst. **8**, 315 (1960).
[11] W. J. A. Goossens, Mol. Cryst. Liq. Cryst. **12**, 327 (1971).
[12] A. Wulf, J. Chem. Phys. **59**, 1487 (1973).
[13] Y. R. Lin-Yiu, Y. M. Shin, G. W. Woo, and H. T. Tan, Phys. Rev. A **14**, 445 (1976).
[14] Y. R. Lin-Yiu, Y. M. Shin, and G. W. Woo, Phys. Rev. A **15**, 2550 (1977).
[15] H. Sandler and A. Wyler, Phys. Rev. Lett. **74**, 3073 (1994).
[16] B. V. Vander Meer, G. Vertogen, A. J. Dekker, and J. G. J. Ypta, J. Chem. Phys. **65**, 3935 (1976).
[17] W. J. A. Gossens, J. de Phys. Coll. **40**, 158 (1979).
[18] B. I. Lev and P. M. Tomchuk, Teor. Mat. Fiz. **60**, 101 (1977).
[19] Q. W. Hornreich and S. Shtrikman, Phys. Rev. A **29**, 3444 (1984).
[20] H. Kazawaguchi and M. Wada, Mol. Cryst. Liq. Cryst. **44**, 97 (1978).
[21] K. C. Lim and J. T. Ho, Phys. Rev. Lett. **40**, 944 (1978).
[22] H. Schröder, J. Chem. Phys. **72**, 3271 (1980).
[23] B. I. Lev and P. M. Tomchuk, Phys. Rev. E **59**, 591 (1999).
[24] A. G. Khachaturyan, *The Theory of Phase Transformations and the Structure of Solid Solutions* (Nauka, Moscow, 1974).
[25] A. G. Khachaturyan and D. A. Badalyan, Fiz. Tverd. Tela (Leningrad) **14**, 657 (1975) [Sov. Phys. Solid State **14**, 2270 (1975)].
[26] D. A. Badalyan, Kristallografiya **27**, 20 (1982) [Sov. Phys. Crystallogr. **27**, 10 (1982)].
[27] H. Kazawaguchi and M. Wada, Jpn. J. Appl. Phys. **14**, 657 (1975).
[28] J. E. Adams and W. E. L. Haas, Mol. Cryst. Liq. Cryst. **15**, 27 (1971).
[29] J. E. Adams, W. E. L. Haas, and J. J. Wysocki, Phys. Rev. Lett. **22**, 92 (1969).
[30] T. Nakagiri, H. Kodama, and K. K. Kobayashi, Phys. Rev. Lett. **23**, 540 (1969).
[31] E. I. Kats, Zh. Éksp. Teor. Fiz. **74**, 2320 (1978) [Sov. Phys. JETP **47**, 1205 (1978)].
[32] E. D. Belotskiĭ, I. P. Il'chishin, B. I. Lev, A. V. Tolmachev, P. M. Tomchuk, and M. T. Shpak, JETP Lett. **51**, 245 (1990).
[33] I. P. Il'chishin, M. T. Shpak, and E. A. Tikhonov, Ukr. Fiz. Zh. **33**, 10 (1988).
[34] B. I. Lev, A. I. Ovcharenko, A. V. Tolmachev, P. M. Tomchuk, and E. D. Chesnokov, Ukr. Fiz. Zh. **35**, 1197 (1990).
[35] L. N. Bulaevskiĭ and D. I. Khomskiĭ, Zh. Éksp. Teor. Fiz. **74**, 1864 (1978) [Sov. Phys. JETP **47**, 971 (1978)].
[36] V. A. Pokrovskiĭ and A. P. Talanov, Zh. Éksp. Teor. Fiz. **75**, 1151 (1978) [Sov. Phys. JETP **48**, 579 (1978)].
[37] W. L. McMillan, Phys. Rev. B **12**, 2042 (1975).
[38] W. L. McMillan, Phys. Rev. B **14**, 1496 (1976).
[39] W. L. McMillan, Phys. Rev. B **16**, 643 (1977).

Translated by M. E. Alferieff
Edited by S. J. Amoretty

Threshold effects in homeotropically oriented nematic liquid crystals in an external electric field

V. P. Romanov*) and G. K. Sklyarenko

St. Petersburg State University, 198904 St. Petersburg, Russia
(Submitted 25 February 1999)
Zh. Éksp. Teor. Fiz. **116**, 543–550 (August 1999)

The electric field-induced orientational transition in a homeotropically oriented nematic liquid crystal cell is investigated. The interaction with the field as a result of anisotropy of the permittivity and flexoelectric polarization is taken into account. For an arbitrary energy of interaction of the nematic with the substrate simple relations are derived for determining the threshold characteristics of the phase transition. It is shown that, in contrast with planar orientation, in fields above a critical value a periodic structure can occur only by virtue of the flexoelectric effect. The resulting dependences for the threshold parameters in the given geometry are exceptionally useful for determining experimentally the surface energy and the difference in the flexoelectric coefficients. © *1999 American Institute of Physics.* [S1063-7761(99)01308-6]

1. INTRODUCTION

One of the remarkable properties of liquid crystals is the ability to change their orientational structure under the influence of comparatively weak external fields. This particular sensitivity is most conspicuous in the Kerr and Cotton–Mouton effects near the transition from the nematic to the isotropic liquid phase, in the flexoelectric effect in nematics, and in the Fréedericksz effect.[1–3] These effects are currently drawing considerable attention in connection with the widespread practical application of liquid crystals in optical devices for imaging data. As a rule, effects of this kind are observed in small samples, so that their energy of interaction with the substrate has a significant influence on the structure of the liquid crystal. It is essential, therefore, that the true values of the corresponding parameters be taken into account in solving problems associated with the reorientation of liquid crystal cells.

The best-known orientational phase transition in nematic liquid crystals is the Fréedericksz effect, where the uniform distribution of the director changes in the presence of a certain threshold field.[1–3] In 1985 it was discovered[4] that the structure resulting from such a transition for an initially planar-oriented nematic and rigid boundary conditions can also be periodic, depending on the ratio $r=K_2/K_1$ of the Frank elastic constants: when r exceeds a certain critical value r_c, the Fréedericksz transition induces an aperiodic distortion of the layer, and for $r<r_c$ it induces a periodic distortion. Simóes *et al.*[5,6] have conducted a theoretical and experimental study of periodic structures occurring in such a geometry in fields above the Fréedericksz threshold in lyotropic nematic cells. It is important to note that the approximation of rigid anchoring to the substrate is often used in the theoretical description of nematic liquid crystals, because the solution of the problem becomes enormously complicated when the finiteness of the surface energy is taken into account. It has been shown[7,8] that the type of structure formed depends on the anchoring energy with the substrate, but the form of this dependence cannot be established analytically.

The possible existence of a periodic distribution of the direction in a planar-oriented nematic cell with rigid boundary conditions has been investigated previously by Pikin *et al.*[9,10] in a study of the flexoelectric effect. This effect, which was discovered by Meyer[11] in 1969 and is associated with anisotropy of the configuration of nematic liquid crystal molecules, is linear in the field, whereas the Fréedericksz transition, which is caused by anisotropy of the dielectric permittivity, is a quadratic function of the field. Galatola *et al.*[12] have published a generalized pattern of the behavior of a nematic layer in an external electric field, taking both of these effects into account, for planar orientation and rigid anchoring to the substrate; in the paper they also discuss the fluctuations of the director in such a system and have derived a phase diagram for the two possible types of distortions. The problem is exceedingly involved in this case and is not amenable to analytical solution.

Because of the threshold character of the effects in question, they provide a very useful tool for measuring the parameters of nematics, but the achievement of this goal requires simple analytical expressions for the threshold characteristics of the system.

In this paper we investigate the behavior of nematic liquid crystals in an external electric field for an initially homeotropically oriented director in the most general case of arbitrary surface energy and various Frank elastic constants. In this geometry a dispersion relation between the field and the period of the structure can be derived in a very simple form, greatly simplifying the calculations and permitting the critical parameters of a nematic to be calculated as functions of the anchoring energy and the thickness of the cell. We show that the resulting static structure can be either periodic or aperiodic, depending on the ratio between the elastic constant K_2 and the electric constants of the nematic. It is significant that a periodic structure can emerge in such a system only when flexoelectric polarization is present in it.

 © 1999 American Institute of Physics

The results of the study offer good possibilities for the experimental measurement of such important characteristics of nematics as the surface energy and the difference in the flexoelectric coefficients.

2. PERIODIC FLEXOELECTRIC STRUCTURE IN A NEMATIC LIQUID CRYSTAL

We consider a homeotropically oriented nematic liquid crystal cell contained between plane-parallel plates in an external electric field. We introduce a Cartesian coordinate system with origin at the center of the cell and with the z axis normal to the plates. Let $\mathbf{E}=(E,0,0)$ denote the electric field vector. The director vector can be written in the form $\mathbf{n}=\mathbf{n}_0+\delta\mathbf{n}=(n_x(y,),n_y(y,z),1)$, where $\delta\mathbf{n}=(n_x(y,z),n_y\times(y,z),0)$ is a small deviation of the director from its equilibrium orientation $\mathbf{n}_0=(0,0,1)$. We assume that the structure formed in the cell under the influence of the field is homogeneous in x, since the creation of such a distribution of the director requires the smallest field.[8,12,13] The corresponding increment of the free energy, to within second-order terms in $\delta\mathbf{n}$, has the form[1,2,12,13]

$$\Delta F=\frac{1}{2}\int d\mathbf{r}\{K_1(\partial_y n_y)^2+K_2(\partial_y n_x)^2+K_3[(\partial_z n_x)^2+(\partial_z n_y)^2]-\epsilon E^2 n_x^2-2eEn_x\partial_y n_y\}, \tag{2.1}$$

where K_i $(i=1,2,3)$ denote the Frank constants, $\epsilon=\varepsilon_a/4\pi$, $\varepsilon_a=\varepsilon_\parallel-\varepsilon_\perp$ is the anisotropy of the permittivity, $\varepsilon_\parallel$ and $\varepsilon_\perp$ are the permittivities along and across the axis of the nematic, respectively, $e=e_1-e_3$ is the difference between the flexoelectric coefficients, and the symbol ∂_j $(j=x,y,z)$ is the partial derivative with respect to the corresponding coordinate. For definiteness we assume that $\varepsilon_a>0$.

The condition for the existence of an equilibrium orientational structure stipulates that the variational derivatives $\delta\Delta F/\delta n_x$ and $\delta\Delta F/\delta n_y$ are equal to zero; after integration by parts this condition yields the system of equations

$$K_2\partial_y^2 n_x+K_3\partial_z^2 n_x+eE\partial_y n_y+\epsilon E^2 n_x=0,$$
$$K_1\partial_y^2 n_y+K_3\partial_z^2 n_y-eE\partial_y n_x=0. \tag{2.2}$$

Taking into account the symmetry of the system (2.2), we look for a solution in the form

$$n_x=\sin(qy)\,\theta_1(z),$$
$$n_y=\cos(qy)\,\theta_2(z), \tag{2.3}$$

where q is the wave number and $\theta_{1,2}(z)$ are certain functions. The system of equations (2.2) can now be rewritten in the form

$$\hat{A}\begin{pmatrix}\theta_1(z)\\ \theta_2(z)\end{pmatrix}=0, \tag{2.4}$$

where the matrix $\hat{A}$ is equal to

$$\hat{A}=\begin{pmatrix}-K_2q^2+\epsilon E^2+K_3\partial_z^2 & -eEq\\ -eEq & -K_1q^2+K_3\partial_z^2\end{pmatrix}. \tag{2.5}$$

To solve Eq. (2.4), we diagonalize the matrix $\hat{A}$ by the transformation $\hat{U}^{-1}\hat{A}\hat{U}$; the matrix $\hat{U}$ is formed from the eigenvectors of $\hat{A}$ and has the form

$$\hat{U}=\begin{pmatrix}2eEq & 2eEq\\ f-g & f+g\end{pmatrix}, \tag{2.6}$$

where

$$f=(K_1-K_2)q^2+\epsilon E^2,$$
$$g=\sqrt{f^2+(2eEq)^2}. \tag{2.7}$$

Multiplying Eq. (2.4) on the left by $\hat{U}^{-1}$ and transforming to the new variables

$$\begin{pmatrix}\theta_1'(z)\\ \theta_2'(z)\end{pmatrix}=\hat{U}^{-1}\begin{pmatrix}\theta_1(z)\\ \theta_2(z)\end{pmatrix}, \tag{2.8}$$

we obtain

$$\begin{pmatrix}-Q^2+\partial_z^2 & 0\\ 0 & -P^2+\partial_z^2\end{pmatrix}\begin{pmatrix}\theta_1'(z)\\ \theta_2'(z)\end{pmatrix}=0, \tag{2.9}$$

where

$$P=\sqrt{\frac{1}{2K_3}[(K_1+K_2)q^2-\epsilon E^2+g]}, \tag{2.10}$$

$$Q=\sqrt{\frac{1}{2K_3}[(K_1+K_2)q^2-\epsilon E^2-g]}. \tag{2.11}$$

The solution of the system (2.9) has the form

$$\begin{pmatrix}\theta_1'(z)\\ \theta_2'(z)\end{pmatrix}=\begin{pmatrix}a_1e^{Qz}+a_2e^{-Qz}\\ b_1e^{Pz}+b_2e^{-Pz}\end{pmatrix}, \tag{2.12}$$

where b_i and c_i $(i=1,2)$ are constants of integration. Taking (2.3), (2.6), (2.8), and (2.12) into account, we obtain expressions for the functions $n_x(y,x)$ and $n_y(y,z)$:

$$n_x(y,z)=2eEq(a_1e^{Qz}+a_2e^{-Qz}+b_1e^{Pz}+b_2e^{-Pz})\sin(qy),$$
$$n_y(y,z)=[(f+g)(a_1e^{Qz}+a_2e^{-Qz})+(f-g)(b_1e^{Pz}+b_2e^{-Pz})]\cos(qy). \tag{2.13}$$

The coefficients a_i and b_i in Eqs. (2.13) are determined from boundary conditions imposed on the director $\mathbf{n}(x,y,z)$ to describe the surface energy W.

3. INTRODUCTION OF BOUNDARY CONDITIONS

To analyze the influence of the nematic-substrate anchoring energy W on the formation of the flexoelectric structure, we consider boundary conditions in the form[3]

$$W\theta_1\pm K_3\partial_z\theta_1=0,\quad z=\pm L/2,$$
$$W\theta_2\pm K_3\partial_z\theta_2=0,\quad z=\pm L/2 \tag{3.1}$$

(we assume that the surface energy is identical on different plates). Substituting Eqs. (2.13) into (3.1), we obtain a sys-

tem of four equations for evaluating the coefficients a_i and b_i. Setting its determinant equal to zero, we have

$$\det \hat{S}=0, \tag{3.2}$$

where the matrix $\hat{S}$ has the form

$$\hat{S}=\begin{pmatrix} \hat{S}_Q & \hat{S}_P \\ (f+g)\hat{S}_Q & (f-g)\hat{S}_P \end{pmatrix},$$

$$\hat{S}_Q=\begin{pmatrix} (W+K_3Q)\exp(QL/2) & (W-K_3Q)\exp(-QL/2) \\ (W-K_3Q)\exp(-QL/2) & (W+K_3Q)\exp(QL/2) \end{pmatrix},$$

$$\hat{S}_P=\begin{pmatrix} (W+K_3P)\exp(PL/2) & (W-K_3P)\exp(-PL/2) \\ (W-K_3P)\exp(-PL/2) & (W+K_3P)\exp(PL/2) \end{pmatrix}.$$

It is readily confirmed that Eq. (3.2) reduces to the condition

$$\det \hat{S}_Q \det \hat{S}_P=0,$$

which can be rewritten in the form

$$\left[\tanh(QL)+\frac{2\xi Q}{1+\xi^2Q^2}\right]\left[\tanh(PL)+\frac{2\xi P}{1+\xi^2P^2}\right]=0, \tag{3.3}$$

where $\xi=K_3/W$ is a characteristic length representing the influence of the bounding solid surface on the orientation of the nematic in a certain wall region.

It is evident from Eqs. (2.7) and (2.10) that P is real-valued for any values of E and q. In light of expression (3.3) this means that the system of equations (3.1) has a nontrivial solution only under the condition

$$\tan\rho=\frac{2h\rho}{\rho^2-h^2}, \tag{3.4}$$

where

$$h=L/\xi, \quad \rho=iQL, \tag{3.5}$$

which in conjunction with (2.7) and (2.11) yields a dispersion relation for $E(q)$:

$$E(q)=\sqrt{\frac{\left(\frac{\rho^2}{L^2}K_3+K_2q^2\right)\left(\frac{\rho^2}{L^2}K_3+K_1q^2\right)}{\frac{\rho^2}{L^2}\epsilon K_3+(\epsilon K_1+e^2)q^2}}.$$

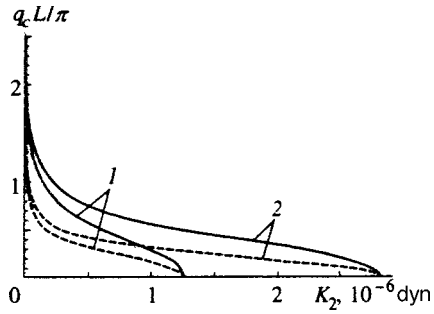

FIG. 1. Dependence of the critical wave number q_c on the elastic constant K_2 for different values of the difference in the flexoelectric coefficients: (*1*) $e_1-e_3=0.33\times10^{-11}$ C/m; (*2*) 0.5×10^{-11} C/m. The solid curves correspond to $W=\infty$ (rigid anchoring), and the dashed curves to $W=0.001$ erg/cm^2 (compliant boundary conditions). Typical parameters of nematic liquid crystals have been used in the calculations: $\varepsilon_a=0.1$, $K_1=0.7\times10^{-6}$ dyn, $K_3=0.5\times10^{-6}$ dyn, $L=10^{-3}$ cm.

Minimizing the latter expression, we obtain the critical wave number and critical field corresponding to the surface energy W:

$$q_c=\frac{\rho}{L}\sqrt{\frac{\mu K_3}{e^2+\epsilon K_1}},$$

$$E_c=\frac{\rho}{L}\sqrt{\frac{K_3}{\epsilon}}\delta, \tag{3.6}$$

where

$$\mu=\frac{e}{\sqrt{K_1K_2}}\sqrt{e^2+\epsilon(K_1-K_2)}-\epsilon,$$

$$\delta=\sqrt{\frac{\epsilon}{\mu+\epsilon}\left(1+\frac{\mu K_2}{e^2+\epsilon K_1}\right)\left(1+\frac{\mu K_1}{e^2+\epsilon K_1}\right)}, \tag{3.7}$$

and $\rho=\rho(W,L,K_3)$ is the smallest root of Eq. (3.4), the one that determines the threshold field. We now examine limiting cases.

For rigid anchoring of the director to the substrate, i.e., for $W=\infty$, the boundary conditions (3.1) assume the form

$$\begin{pmatrix} \theta_1(z=\pm L/2) \\ \theta_2(z=\pm L/2) \end{pmatrix}=0. \tag{3.8}$$

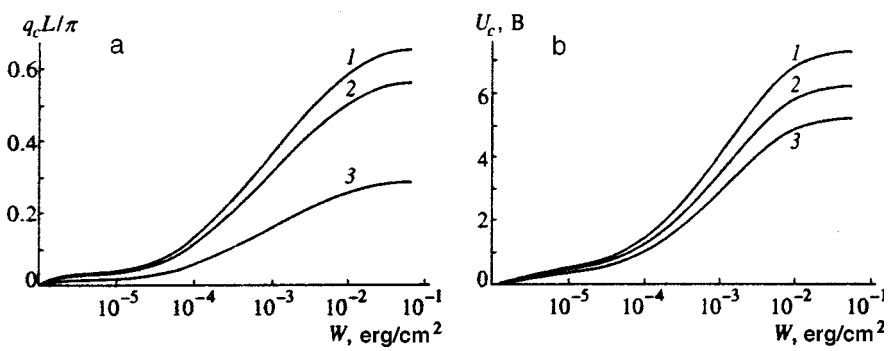

FIG. 2. Dependence of the threshold characteristics of a nematic q_c (a) and U_c (b) on the anchoring energy with the substrate W for various values of the difference in the flexoelectric coefficient: (*1*) $e_1-e_3=0.33\times10^{-11}$ C/m; (*2*) 0.5×10^{-11} C/m; (*3*) 0.67×10^{-11} C/m. The same typical nematic parameters as in Fig. 1 plus $K_2=10^{-6}$ dyn have been used in the calculations.

From Eq. (3.4) we find $\rho=\pi$. In the one-constant approximation, i.e., for $K_1=K_2=K_3$, Eqs. (3.6) coincide with the results of Ref. 13. The asymptotic form of the critical field in the limit $W\to\infty$ has the form

$$E_c\approx E_c^{\infty}\left(1-\frac{2K_3}{WL}\right),$$

where

$$E_c^{\infty}=\frac{\pi}{L}\sqrt{\frac{K_3}{\epsilon}}\delta$$

is the threshold value for rigid boundary conditions.

For an unbounded sample, which corresponds to $W=0$, Eq. (3.4) yields $\rho=0$, which in conjunction with definition (3.5) leads to a linear function $E(q)$:

$$E=q\sqrt{\frac{K_1K_2}{\epsilon K_1+e^2}}.$$

In this case, therefore, the effect does not have a threshold character, i.e., $E_c^0=0$, and every value of the field E corresponds to its own wave number q, i.e., its own period of the flexoelectric structure.[11,13] For low energies W the solution has the form

$$E_c\approx\sqrt{\frac{2W}{\epsilon L}}\,\delta.$$

As W increases, obviously, ρ increases monotonically from $\rho|_{W=0}=0$ to $\rho|_{W=\infty}=\pi$.

4. DISCUSSION

We now analyze the results. It is evident from Eqs. (3.6) and (3.7) that the type of structure formed in a homeotropically oriented nematic liquid crystal cell under the influence of an external electric field is dictated by the sign of μ, which depends on the parameters of the nematic: e, ϵ, K_1, and K_2. For $\mu>0$ the function $E(q)$ has a minimum at a point $q_c\neq 0$, and for $\mu\leqslant 0$ the minimum field corresponds to $q=0$. It is significant that this sign does not depend on the Frank constant K_1.

In the case $\epsilon K_2/e^2<1$, at $E=E_c$ a phase transition takes place in the nematic liquid crystal cell from the homogeneous homeotropic state to a periodic flexoelectric structure (periodic Fréedericksz effect) with wave number q_c, where the existence of flexoelectric polarization, i.e., the requirement $e\neq 0$, is a necessary condition for such a transition. This is the way in which the behavior of a homeotropically oriented nematic cell in an external electric field differs fundamentally from that of a planar-oriented cell, because in the latter the director can acquire a periodic distribution even without the flexoelectric effect.[4–8] The fact that the behavior of the periodic structure is controlled specifically by the Frank constant K_2 is attributable to the resulting twist distortion described by this constant in such a transition.

For $\epsilon K_2/e^2\geqslant 1$ the ordinary (aperiodic) Fréedericksz effect occurs at $E=E_c$. This criterion does not depend on the nematic-substrate anchoring energy.

The dependence of the threshold characteristics of a nematic on the surface energy W is shown in Fig. 2. We note that for $W>0.01$ erg/cm^2 these characteristics do not change appreciably. It is evident that the function $q_c(W)$ is more sensitive than $U_c(W)$ to the difference in the flexoelectric coefficients.

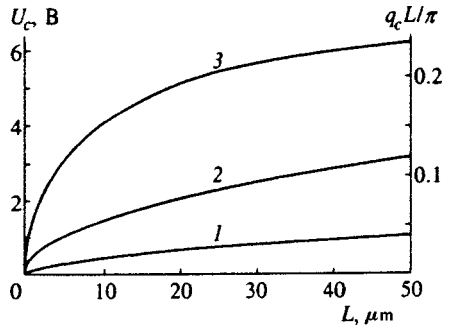

FIG. 3. Dependence of the threshold characteristics on the cell thickness L for $e_1-e_3=[1]0.33\times10^{-11}$ C/m and various values of the surface energy: (*1*) $W=0.001$ erg/cm^2; (*2*) $W=0.01$ [1] erg/cm^2; (*3*) $W=0.1$ erg/cm^2 (all other parameters of the nematic are the same as in Fig. 2. Curves *1–3* simultaneously describe the behavior of the functions $U_c(L)$ and $q_c(L)$.

Figure 1 shows the dependence of the critical wave number q_cL/π on the Frank elastic constant K_2 for different values of the difference in the flexoelectric coefficients in connection with rigid and compliant boundary conditions. The existence of the critical value K_2^0 defined by the equation $\epsilon K_2^0/e^2=1$ is evident in the graph: The periodic flexoelectric effect occurs for $K_2<K_2^0$, and the aperiodic Fréedericksz transition occurs for $K_2\geqslant K_2^0$.

Figure 3 shows graphs of U_c and q_c as functions of the cell thickness L for various energies of interaction of the nematic with the substrate. Notably, the critical parameters are highly sensitive to the thickness of the cell, and the resulting graphs can be used for the experimental determination of the nematic-substrate anchoring energy and the difference in the flexoelectric coefficients.

*)E-mail: vadim.romanov@pobox.spbu.ru

[1] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed., Clarendon Press, Oxford (1993).
[2] S. Chandrasekhar, *Liquid Crystals*, Cambridge Univ. Press, New York (1977); paperback (1980).
[3] S. A. Pikin, *Structural Transformations in Liquid Crystals* [in Russian], Nauka, Moscow (1981).
[4] F. Lonberg, and R. B. Meyer, Phys. Rev. Lett. **55**, 718 (1985).
[5] M. Simões, A. J. Palangana, L. R. Evangelista *et al.*, Phys. Rev. E **54**, 3765 (1996).
[6] A. J. Palangana, M. Simões, and L. R. Evangelista, Phys. Rev. E **56**, 4282 (1997).
[7] C. Oldano, Phys. Rev. Lett. **56**, 1098 (1986).
[8] E. Miraldi, C. Oldano, and A. Strigazzi, Phys. Rev. A **34**, 4348 (1986).
[9] Yu. P. Bobylev and S. A. Pikin, Zh. Éksp. Teor. Fiz. **72**, 369 (1977) [Sov. Phys. JETP **45**, 195 (1977)].
[10] Yu. P. Bobylev, V. G. Chigrinov, and S. A. Pikin, J. Phys. Colloq. **40**, C3-331 (1979).
[11] R. B. Meyer, Phys. Rev. Lett. **22**, 918 (1969).
[12] P. Galatola, C. Oldano, and M. Rajteri, Phys. Rev. E **49**, 1458 (1994).
[13] V. P. Romanov and G. K. Sklyarenko, Zh. Éksp. Teor. Fiz. **112**, 1675 (1997) [JETP **85**, 914 (1997)].

Translated by James S. Wood

Properties of shock-compressed liquid krypton at pressures of up to 90 GPa

V. D. Glukhodedov, S. I. Kirshanov, T. S. Lebedeva, and M. A. Mochalov

Russian Federal Nuclear Center — All-Russian Scientific-Research Institute of Experimental Physics, 607190 Sarov, Nizhny Novgorod Region, Russia

(Submitted 10 March 1999)
Zh. Éksp. Teor. Fiz. **116**, 551–562 (August 1999)

The following quantities of shock-compressed liquid krypton are measured behind a plane shock front at pressures up to 90 GPa: compressibility up to densities of 7 g/cm^3, brightness (color) temperatures of 6000–24000 K, and electrical conductivities of 40–60000 $(\Omega \cdot \mathrm{m})^{-1}$. $X-t$ diagram methods are used to estimate sound speeds of up to 5.5 km/s at pressures of 30–75 GPa. The optical absorption coefficients in the violet and red (30–300 cm^{-1}) are measured at pressures of 20–90 GPa from the rise in brightness of the shock front luminosity. The optical reflection coefficient of the shock front (~13%) at a pressure of 76.1 GPa is measured for the first time. © *1999 American Institute of Physics.* [S1063-7761(99)01408-0]

1. INTRODUCTION

Compressed inert gases are convenientsubjects for studies of the properties of close-packed substances at extreme pressures and temperatures. When the initial densities of the liquids are high it is easy to reach pressures of 100–1000 GPa and densities of 10–20 g/cm^3 in physical experiments,[1,2] while the optical transparency of the liquid makes it possible to record the luminosity of a shock front through the layer material that is not yet compressed.[2,3] The quasi-isentropic compressibility of liquid argon has been measured[1] to densities of 7 g/cm^3 at pressures up to 500 GPa. Measurements have been made[2] of the quasi-isentropic compressibility of liquid xenon to densities of 13 g/cm^3 at a pressure of 200 GPa, as well as of its compressibility, brightness (color) temperatures, and electrical conductivity during shock compression to 70 GPa. Studies have been made of the compressibility and brightness temperatures[3] of liquid argon up to 70 GPa, as well as of its electrical conductivity.[4]

Simultaneous measurements of the compressibility and temperatures make it possible to obtain thermodynamically complete information on the state of the shock-compressed material, which is important for creating new theoretical models and verifying existing models. Measurements of the electrical conductivity yield additional information on the behavior of the band gap with increasing temperature and on the density of the shock-compressed material.

The experimental studies of shock-compressed liquid krypton reported here fill a gap in research on the properties of the inert gases, since krypton has not been studied at high pressures and temperatures before.

2. EXPERIMENTAL RESULTS

Shock adiabat

As in the studies of liquid argon and xenon,[2,3] in the present experiments we used a cell with which it was possible simultaneously to record the shock speed and the radiation from its front. The initial state of the liquid krypton corresponds to a temperature of $T_0 = 120$ K and density of $\rho_0 = 2.41$ g/cm^3. For correct calculation of the shock-compression pressures, both the isentropic expansion of the Al, Cu, or Fe shields through which the shock emerges into the liquid krypton caused by the differences in their dynamic impedances and the rise in the initial density of the shielding materials as they were cooled to $T = 120$ K were taken into account.

The experimentally measurable parameter in this work is the shock speed D in the liquid krypton, with a mean square error in the average of less than $\pm(1-1.5)\%$. From this value and the parameters of the shock wave in the shield, the ''reflection'' method[5] gave the parameters of the shock, viz., the pressure P and the post-shock mass velocity U.

The experimental values of the shock speed D are plotted in Fig. 1 as a function of the mass velocity U, along with the best-fit curve to the experimental data. Each value was obtained by averaging the velocities measured by four or five pairs of probes in four or five independent experiments. The highest measured speed, $D = 8.49 \pm 0.13$ km/s at $U = 4.84$ km/s, corresponds to a pressure $P = 87.6 \pm 1.8$ GPa and density $\rho = 6.81 \pm 0.38$ g/cm^3. The sound speed in liquid krypton under standard conditions, $c_s = 690$ m/s, was taken into account in the statistical analysis of the experimental data. The approximation (fit) has the form

$$D = (0.700 \pm 0.042) + (1.943 \pm 0.034)U - (0.1125 \pm 0.0064)U^2. \quad (1)$$

Measurements of the kinetic parameters of shock waves, their speed D and mass velocity U, have been the principal source of information on the properties of shock-compressed materials up to now. Studies of the shock compressibility of liquid krypton at pressures up to $\simeq 90$ GPa and densities $\simeq 7$ g/cm^3 have convincingly demonstrated that the rise in the shock speed is reduced as the shock amplitude increases, which is equivalent to an increase in the compressibility.

 © 1999 American Institute of Physics

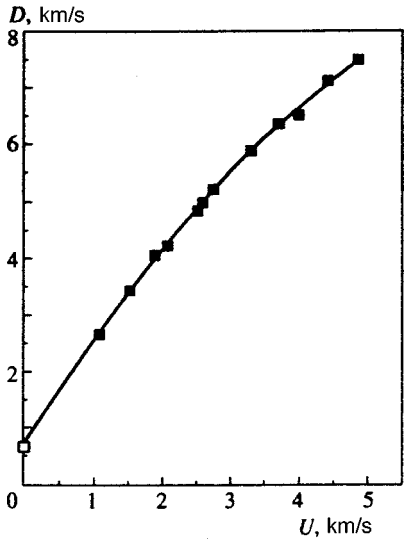

FIG. 1. Shock Adiabat of liquid krypton: ■ — experimental results from this paper, □ — sound speed; curve — approximation.

Brightness temperatures

A semiempirical equation of state with parameters based solely on shock adiabat data relating the pressure and density in a compressed state cannot provide a correct description of the experimental data over a wide range of $P-\rho-T$ states. Parameters calibrated in a singlestudy do not adequately describe other physical properties of a material. Thus, experimental measurements of the temperatures are important, as they make it possible to verify and refine the methods for describing the thermal components.

In this paper the temperatures are measured by recording the radiation from the shock front and comparing it with the emission from a standard source with a known temperature. This method makes it possible to measure the brightness temperature T_b of an emitting layer of material with the same monochromatic brightness at a given wavelength as a test body at temperature T, that is,

$$I_P(\lambda,T_b)=I(\lambda,T)=\varepsilon(\lambda,T)I_P(\lambda,T), \tag{2}$$

where $\varepsilon(\lambda,T)$ is the emissivity of the body and $I_P(\lambda,T)=C_1\lambda^{-5}/[\exp(C_2/\lambda T)-1]$ is the Planck function. Here the constants have the values $C_1=1.19\times10^{-16}$ W·m^2·sr^{-2} and $C_2=0.0144$ m·K, λ is the wavelength of the radiation, and T is the temperature.

This method was originated by Model'[6] and has been developed subsequently for studies of temperature in shock-compressed condensed media, including ionic crystals.[7–11] These papers show convincingly that it is possible to determine the temperatures of transparent, condensed, shock-compressed substances from their thermal radiation, and their various optical characteristics have been studied extensively: emission has been observed and, at high pressures ($P>100$ GPa), nonequilibrium emission from shock-compressed specimens of the ionic crystals NaCl, KCl, CsBr, and LiF. In addition, the absorption coefficients of NaCl, CsBr and other ionic crystals were measured under shock compression conditions for the first time.

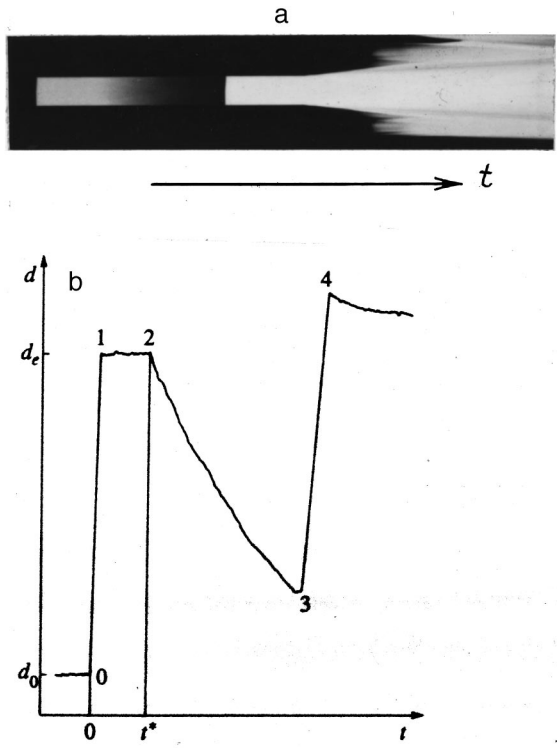

FIG. 2. (a) A streak photo of the luminosity of a shock front in liquid krypton and (b) a densitometer scan of the film blackening.

The first data on the measurement of brightness temperatures in shock fronts in liquid argon and nitrogen were published in Ref. 12. These data[12] have been extended[3] to 70 GPa in liquid argon. Data have been published[2] on measurements of brightness temperatures in the red at shock fronts in liquid xenon. Experiments on the break in the $T-P$ curves near 8000-10000 K for liquid argon and xenon revealed an effect that has been related[2,3] to the kinetics of electronic excitation in the conduction band that were first observed in studies of ionic crystals.[10,11] Recent papers have reported temperatures of 4000–14000 K at pressures of 18–90 GPa in liquid nitrogen[13–15] and temperature measurements in shock-compressed liquid xenon.[16]

In our present work a photographic method[2,3] is used to detect the emission from shock waves in liquid krypton at wavelengths $\lambda=670$ and 430 nm. In order to increase the information yield from the experiments, we have also measured the brightness temperatures of shock fronts in liquid krypton in the green ($\lambda=500$ nm) using a high-speed photomultiplier with a known photocathode sensitivity.

Figure 2a shows a streak photo of a shock front in liquid krypton in the red region of the spectrum at a pressure of 63.2 GPa, and Fig. 2b shows a densitometer scan of the blackening density of the film in this experiment. The streak photo and densitometer scan shown here are typical of our experiments with liquid krypton. Four characteristic regions can be identified in the scan of Fig. 2b: (1) a sharp rise in the emission brightness to a maximum over a period of 100–200 ns occurs in segment 0–1; (2) in segment 1–2, after an optical thickness ~1 for the emitting layer is attained, the brightness remains constant until the shock is overtaken by a

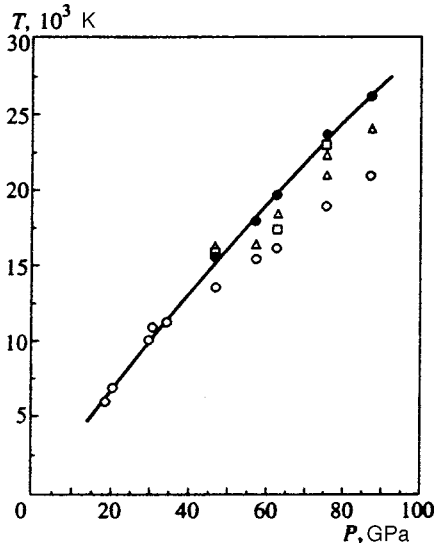

FIG. 3. Temperature of shock-compressed liquid krypton: ○ — in the red ($\lambda=670$ nm), □ — green ($\lambda=500$ nm), and △ — violet ($\lambda=430$ nm) regions of the spectrum; ● — true temperature ($T=15600$ K for $\varepsilon=0.9$, $T=18000$ K for $\varepsilon=0.9$, $T=19800$ K for $\varepsilon=0.85$, $T=23400$ K for $\varepsilon=0.87$, $T=26200$ K for $\varepsilon=0.85$); curve — approximation.

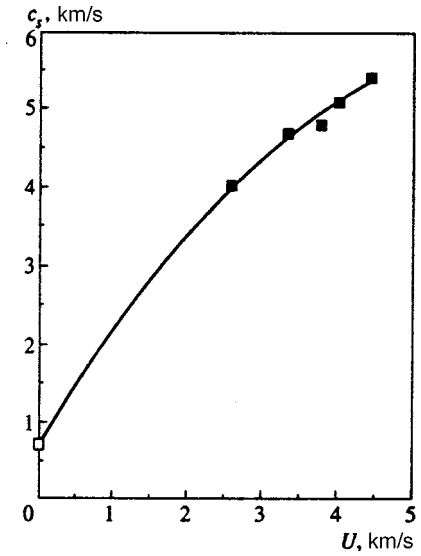

FIG. 4. Post-shock speed: ■ — experimental data, curve — approximation; □ — sound speed under normal conditions.

release wave from the direction of the impact plate (striker) at point 2; (3) from that time on in segment 2–3 the shock compression pressure falls and the brightness of the front decreases; and (4) after the shock wave emerges from the liquid, it is still strong enough to make the air radiate, as can be seen in segment 3–4. The time resolution in these experiments is $\simeq 1.5\times 10^{-7}$ s. The brightness temperature was estimated from the amount of blackening of the film in the constant-brightness segment (1–2). The duration of this segment was determined by the time the liquid existed in a compressed state and decreased as the shock-compression pressure was raised.

The results of the present work are plotted as a function of pressure in Fig. 3. Also shown there are the gasdynamic temperatures, including the emissivity of a shock front in liquid krypton and a fit curve. (See below.)

The assumption that the drop in the emitted intensity at time t^* (see Fig. 2b) is associated with a release wave which catches up with the shock front makes it possible to obtain additional information on the sound speed in the shock-compressed liquid krypton by the X–t diagram method.[5] The resulting values of the sound speed are plotted with an accuracy of $\pm 6\%$ as a function of the mass velocity of the particles in Fig. 4, together with an approximate fit curve for the following form within the region studied:

$$C=(0.70\pm 0.10)+(1.60\pm 0.10)U-(0.12\pm 0.02)U^2. \tag{3}$$

In choosing an approximation, we have taken into account the sound speed in liquid krypton under normal conditions.

Absorptivity of shock-compressed liquid krypton

The emissivity $\varepsilon(\lambda,T)$ [Eq. (2)] must be estimated in some way or other in order to determine the true temperature from the measured brightness. In practice, $\varepsilon(\lambda,T)$ is determined experimentally for the direction normal to the surface of the emitting body and (for a weak wavelength dependence) this yields

$$\varepsilon(\lambda,T)=\varepsilon(T)=\text{const}. \tag{4}$$

Estimating $\varepsilon(T)$ under shock-compression conditions is often difficult. Thus, the shock front temperature in liquid nitrogen[13–15] and xenon[16] has been estimated by choosing the value of ε in Eq. (2). An analysis of the results of Refs. 13–16 indicates that ε behaves irregularly as the temperatures and pressures increase and enters as a purely mathematical quantity chosen to provide the best fit to the experimental data. Thus, the question of the relationship between the fit parameter ε and the absorptivity ε^* determined from the basic equation

$$\varepsilon^*+\tau+R=1 \tag{5}$$

must be answered in order to fit the experimental temperatures.

Since ε and ε^* are equivalent in a state of thermodynamic equilibrium, the emissivity ε can be estimated using Eq. (5) if we measure the reflectivity R and transparency τ of the emitting body. The transparency of a substance during shock compression has been estimated in studies of gases[6] and of NaCl.[9] The transmission of the radiation in Refs. 6 and 9 was determined in accordance with the Bouguer formula:

$$\tau=e^{-\alpha l}, \tag{6}$$

where α is the linear absorption coefficient for the radiation in a layer of thickness l.

In the present experiments we have also studied the transmission of a layer of shock-compressed liquid Kr using the pyrometer described in Ref. 9. The emission was detected in the violet and red regions of the spectrum using interference filters with transmissions of $\sim 50\%$ at wavelengths of $\lambda_v=407$ nm and $\lambda_r=605$ nm. A typical oscilloscope trace of the increasing emission intensity during an

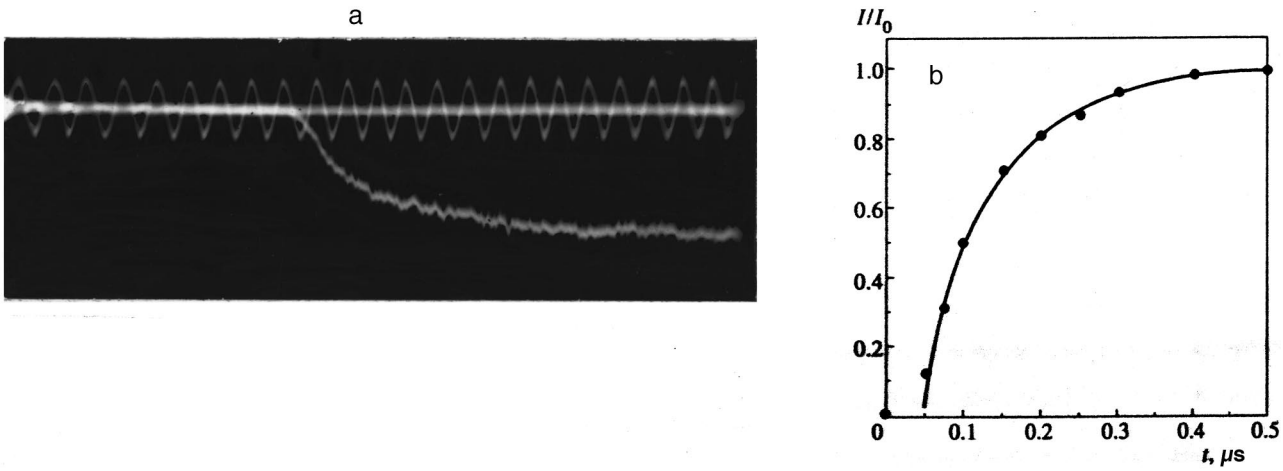

FIG. 5. Brightness as a function of time for a pressure of 18 GPa: a — oscilloscope trace (markers 40 ns); b —results of data processing: ● — experimental values; curve — approximation according to Eq. (7) for $\alpha = 61$ cm^{-1}.

experiment ($P = 18$ GPa) is shown in Fig. 5a. The increasing brightness, as in Refs. 6 and 9, is attributed to the growing layer of material compressed by the shock wave with an absorption coefficient (averaged) of α.

The intensity of the emission from a plane layer in the direction normal to it is given by

$$I = I_0[1 - e^{-\alpha l}] = I_0[1 - e^{-\alpha(D-U)t}]. \quad (7)$$

Here I_0 is the intensity of an optically dense layer, $l = (D - U)t$ is the thickness of the layer of shock-compressed, heated material, and t is the time for the shock wave to move through the material. It is easy to obtain a formula for estimating α from Eq. (7):

$$\alpha = -\frac{1}{(D-U)t} \ln\left(1 - \frac{I}{I_0}\right), \quad (8)$$

where the change in intensity I/I_0 with time was determined from oscilloscope traces similar to those in Fig. 5a, while D and U were obtained from the compressibility experiments. The results of an experiment at a pressure of 18 GPa are shown in Fig. 5b.

The measured optical absorption coefficients of liquid krypton in the red and violet ranged over 10 to 300 cm^{-1} for pressures of up to 80 GPa. Our results are plotted in Fig. 6. Also shown there is an approximat fit to all the data:

$$\ln \alpha = (6.94 \pm 0.18) - (2.116 \pm 0.205) \times 10^4/T. \quad (9)$$

The absorptivity of the shock-compressed layer, neglecting reflection, was calculated from the expression

$$\varepsilon \cong 1 - e^{-\alpha(D-U)t}. \quad (10)$$

Estimates show that for the measured values of the absorption coefficient, even at a level of 70% of the maximum amplitude of the light pulse, the absorptivity of the shock-compressed liquid krypton is high and can differ from unity mainly because of the reflectivity.

The method for measuring the reflection coefficient R at the shock front in transparent dielectrics has been described elsewhere.[8] It has been used by Kormer in studies of R in ionic crystals and liquids, which thus far are still unique.

In the present experiments we study the reflective properties of shock fronts in liquid krypton for the first time. As in Ref. 10, the reflection was measured using the intrinsic luminosity of the shock. The arrangement for studying the reflection of light from a shock front in liquid krypton is shown in Fig. 7a.

The emission from the shock front *1* reflected at an angle of $\sim 10°$ from a copper mirror *2* located in the liquid volume *3* was directed onto an external mirror *4* and, further, onto the inlet objective of a high speed camera. At the same time, the emission of the shock front itself was recorded. It is clear from the figure that the brightness in the segment k–m is determined by the emission from the front and by the light reflected from the copper mirror *2*, while that in the segment m–n is determined only by the emission from the front. The intensity of the total emission in the segment k–m can be estimated using the formula

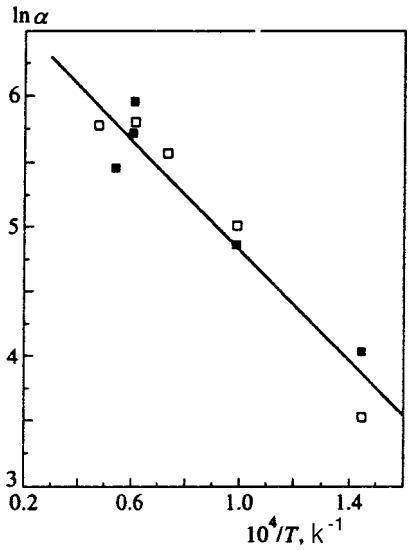

FIG. 6. Optical absorption coefficient in shock-compressed liquid krypton. Experimental data: □ — in the red ($\lambda_r = 605$ nm) and ■ — violet ($\lambda_v = 407$ nm) regions of the spectrum ; straight line — approximation.

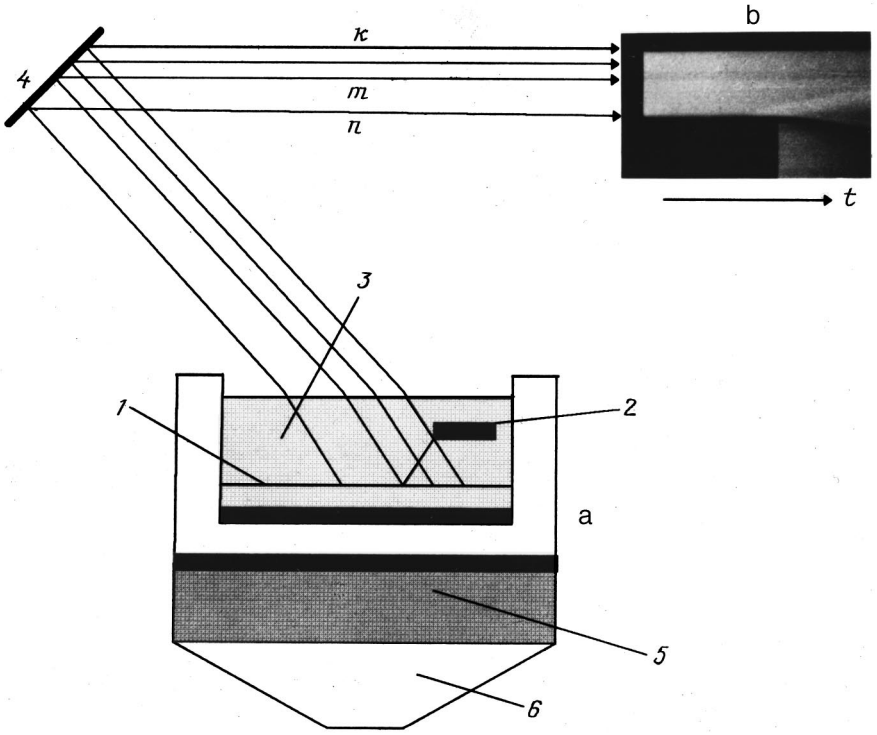

FIG. 7. Setup for measuring the reflection of light from a shock front in liquid krypton: (a) experimental apparatus, (b) a streak photo.

$$I = I_0[1+\xi\eta](1-R), \tag{11}$$

where I_0 is the intrinsic intensity of the shock front and ξ, η and R are the reflection coefficients of the copper mirror, shock front, and free liquid–air boundary, respectively. An experiment for liquid krypton was done according to this scheme for a wavelength $\lambda = 430$ nm and a pressure of 76.1 GPa. A streak photo of this experiment is shown in Fig. 7b and a densitometer scan of the film blackening, in Fig. 8. The jump in intensity associated with reflection from the shock front shows up clearly. The image shown in Fig. 7b was scanned photometrically perpendicular to the film plate. An analysis of the experimental data and an estimate using Eq. (11) yield $\eta \simeq 13\%$. The reflectivity ξ of the copper mirror at small angles of incidence ($\simeq 10°$) was estimated to be $\simeq 70\%$.[17] The densitometer scan of a control experiment at a pressure of 76.1 GPa without a copper mirror reveals no intensity jump. The present estimates show that when the reflection is taken into account, the absorptivity of shock-compressed krypton at 76.1 GPa is $\simeq 0.87$.

Measuring three brightness temperatures in our experiments made it possible to estimate the true temperature of a shock front in liquid krypton. It was assumed that the absorptivity is independent of wavelength, i.e.,

$$\varepsilon_{\text{viol}} = \varepsilon_{\text{green}} = \varepsilon_{\text{red}} = \varepsilon.$$

The spectral density of the emitted brightness which best described the brightness density for the measured brightness temperatures was calculated from Eq. (2) by fitting ε and T. The absorptivity $\varepsilon_{\text{viol}} \simeq 0.87$ in the violet measured in the present experiments served as a control for estimating the true temperatures. As an example, Fig. 9 shows the spectral density of the brightness at 76.1 GPa calculated using the Planck formula for $\varepsilon = 0.87$ and $T = 23400$ K, which was taken as the true temperature. The true temperatures for other shock compression pressures in liquid krypton were deter-

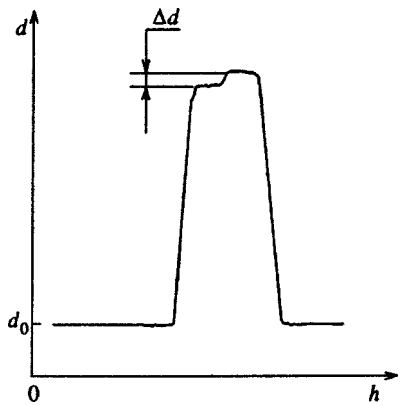

FIG. 8. Densitometer trace: d_0 is the film background and Δd is the jump in blackening associated with the reflection of light from the shock front. The photometric scan was made perpendicular to the film.

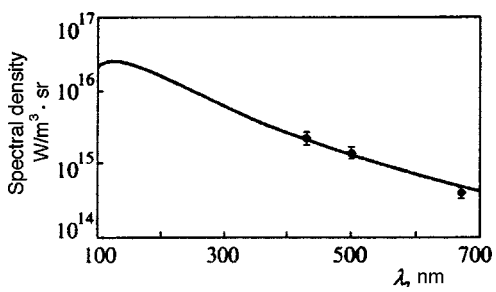

FIG. 9. Spectral density of the emission brightness at a pressure of 76.1 GPa: ●—experimental data; curve, calculated using the Planck formula for $T = 23400$ K and $\varepsilon = 0.87$.

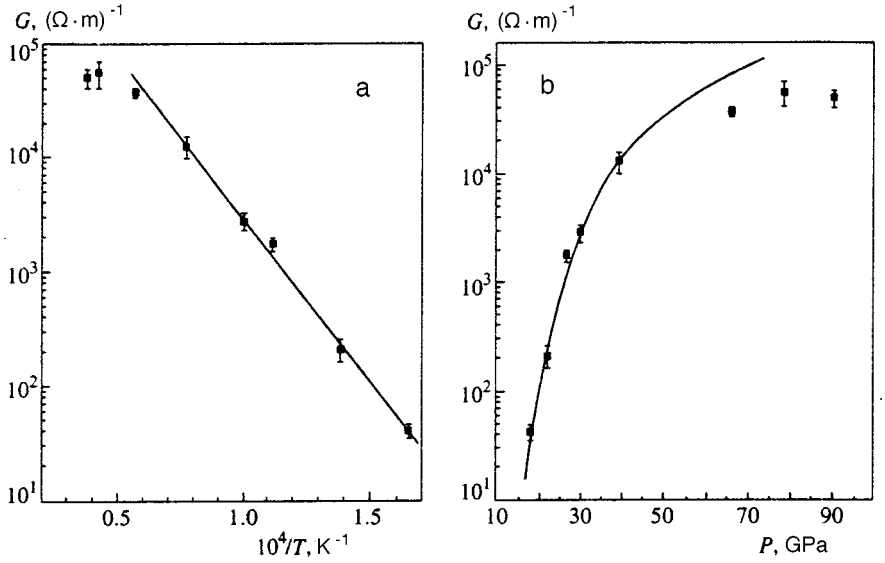

FIG. 10. Electrical conductivity of shock-compressed liquid krypton as a function of temperature (a) and pressure (b): ■ — experimental data; the straight line is a fit to the four lowest points.

mined in similar fashion and are shown in Fig. 3. Also shown there is an approximation curve for these data as a function of pressure (in GPa):

$$T = -315 + 366P - 0.7P^2. \tag{12}$$

As can be seen from Fig. 3, the difference between the brightness temperatures decreases as the pressure is reduced. Thus, in choosing an approximation, it was assumed that up to pressures of $\simeq 35$ GPa and temperatures $\simeq 12000$ K the measured and true temperatures coincide.

Our measurements of the optical absorption and reflection permitted an experimental estimate of the emissivity, which is a more sensitive parameter for verifying the various theoretical models than the shock pressure when determining the true temperature in the shock fronts. Studies of these parameters are important for evaluating the structure (smoothness) of a shock front.

Electrical conductivity

Additional information on the behavior of liquid krypton at high pressures and temperatures was obtained by measuring the electrical conductivity, which is directly related to the free electron density and band gap.

In our experiments we used a method[18] for measuring the conductivity that has been used before in studies of xenon[2] and argon.[4] Our results are shown in Fig. 10a. In the experimental plot of $G(1/T)$ one can distinguish a segment where the electrical conductivity of krypton has a dependence characteristic of amorphous semiconductors, i.e.,

$$G = G_0 \exp[-E(T,V)/2kT]. \tag{13}$$

Here $E(T,V)$ is the energy gap between the valence and conduction bands, $V = 1/\rho$, where ρ is the density, and G_0 depends weakly on the temperature. The experimental points in the initial segment of $\log G(1/T)$ were fit by least squares to a linear curve with 90% confidence, which yielded a band gap $E = 11.3 \pm 0.4$ eV in shock-compressed liquid krypton. As can be seen from this figure, this experimental value is characteristic only of the initial part of the $G(1/T)$ curve, and as the temperature rises the gap shrinks and the conductivity evidently saturates as a function of temperature. Similar behavior has been observed before in liquid xenon.[2] For liquid argon within the range studied here up to $G \simeq 2 \times 10^4$ $(\Omega \cdot \mathrm{m})^{-1}$, the conductivity typically depends linearly on temperature.[4]

The behavior of the experimental data as a function of the amplitude of a shock wave in liquid krypton is shown in Fig. 10b. The figure shows clearly the saturation in the conductivity, which approaches a limit of $\simeq 6 \times 10^4$ $(\Omega \cdot \mathrm{m})^{-1}$ at $\simeq 70$ GPa. In experiments with liquid nitrogen, the tendency of the conductivity to saturate[15] at $\simeq 100$ GPa has been identified with a dissociative phase transition in the liquid nitrogen in a shock wave. Similar behavior of the conductivity has been observed in experiments with liquid hydrogen,[19–21] for which saturation at $\simeq 140$ GPa has been attributed to metallization of hydrogen in the shock wave.

Shock compression presents unique opportunities for studying the transition of dielectrics into a metallic state, since the density increase is accompanied by a substantial rise in temperature and increase in the number of free electrons. This smears out the dielectric–metal transition, which creates the preconditions for observing a continuous transition of dielectrics to the metallic state.[22] In this case, the rise in the number of free electrons causes a rise in the optical reflectivity from the shock front, as we have observed experimentally at a pressure of 76.1 GPa.

3. CONCLUSION

Various properties of shock-compressed liquid krypton at pressures up to 100 GPa have been studied extensively for the first time in this paper, including compressibility on the Hugoniot adiabat, post-shock sound speed and electrical conductivity, optical absorption coefficient of the layer of shock-compressed material, and the temperatures and optical reflectivities at the shock front.

The set of experimental data presented here can be used to create an exact equation of state for krypton at high pressures and temperatures.

[1] V. D. Urlin, M. A. Mochalov, and O. L. Mikhaĭlova, Zh. Éksp. Teor. Fiz. **111**, 2099 (1997) [JETP **84**, 1145 (1997)].
[2] V. D. Urlin, M. A. Mochalov, and O. L. Mikhailova, High Press. Res. **8**, 595 (1992).
[3] F. V. Grigor'ev, S. B. Kormer, O. L. Mikhaĭlova, M. A. Mochalov, and V. D. Urlin, Zh. Éksp. Teor. Fiz. **88**, 1271 (1985) [Sov. Phys. JETP **81**, 751 (1985)].
[4] L. A. Gatilov, V. D. Glukhodedov, F. V. Grigor'ev, S. B. Kormer, L. V. Kuleshova, and M. A. Mochalov, Prikl. Mekh. Tekh. Fiz. **1**, 99 (1985).
[5] L. V. Al'tshuler, Usp. Fiz. Nauk. **85**, 197 (1965) [Sov. Phys. Usp. **8**, 52 (1965)].
[6] I. Sh. Model', Zh. Éksp. Teor. Fiz. **32**, 714 (1957) [Sov. Phys. JETP **5**, 589 (1957)].
[7] S. B. Kormer, M. V. Sinitsyn, G. A. Kirillov, and V. D. Urlin, Zh. Éksp. Teor. Fiz. **48**, 1033 (1965) [Sov. Phys. JETP **21**, 689 (1965)].
[8] S. B. Kormer, Usp. Fiz. Nauk **94**, 641 (1968) [Sov. Phys. Usp. **11**, 229 (1968)].
[9] S. B. Kormer, M. V. Sinitsyn, G. A. Kirillov, and L. T. Popova, Zh. Éksp. Teor. Fiz. **49**, 135 (1965) [Sov. Phys. JETP **22**, 97 (1966)].
[10] S. B. Kormer, M. V. Sinitsyn, and A. I. Kuryapin, Zh. Éksp. Teor. Fiz. **55**, 1626 (1968) [Sov. Phys. JETP **28**, 852 (1969)].
[11] Ya. B. Zel'dovich, S. B. Kormer, and V. D. Urlin, Zh. Éksp. Teor. Fiz. **55**, 1631, (1968) [Sov. Phys. JETP **28**, 855 (1969)].
[12] I. M. Voskoboĭnikov, M. F. Gogulya, and Yu. A. Dolgoborodov, DAN SSSR**246**, 579 (1979) [Sov. Phys. Dokl. **24**, 375 (1979)].
[13] H. B. Radousky, W. J. Nellis, M. Ross, D. C. Hamilton, and A. C. Mitchell, Phys. Rev. Lett. **57**, 2419 (1986).
[14] H. B. Radousky and M. Ross, High Press. Res. **1**, 39 (1988).
[15] H. B. Radousky, W. J. Nellis, M. Ross, D. C. Hamilton, and A. C. Mitchell, Phys. Rev. Lett. **57**, 2419 (1986).
[16] H. Radousky and M. Ross, Phys. Lett. A **129**, 43 (1988).
[17] M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford (1969).
[18] L. A. Gatilov and L. V. Kuleshova, Prikl. Mekh. Tekh. Fiz. **1**, 136 (1981).
[19] W. J. Nellis, S. T. Weir, and A. C. Holmes, Science **76**, 1860 (1996).
[20] W. J. Nellis, S. T. Weir, and A. C. Mitchell, Science **273**, 936 (1996).
[21] S. T. Weir, A. C. Mitchell, and W. J. Nellis, Phys. Rev. Lett. **76**, 1860 (1996).
[22] Ya. B. Zel'dovich, *Selected Works, Chemical Physics and Hydrodynamics*, Princeton Univ. Press, Princeton, NJ (1971) [Russian orig. Nauka, Moscow (1984), p. 130].

Translated by D. H. McNeill

Motion of magnetic flux lines in magnetohydrodynamics

V. P. Ruban*)

L. D. Landau Institute of Theoretical Physics, Russian Academy of Sciences, 117334 Moscow, Russia
(Submitted 8 April 1999)
Zh. Éksp. Teor. Fiz. **116**, 563–585 (August 1999)

A gauge-free description of magnetohydrodynamic flows of an ideal incompressible fluid, which takes into account the freezing-in of the magnetic field and the presence of cross invariants containing the vorticity, is obtained. This description is an extension of the canonical formalism well-known in ordinary hydrodynamics to the dynamics of frozen-in flux lines. Magnetohydrodynamics is studied as the long-wavelength limit of the two-fluid model of a plasma, in which the existence of two frozen-in fields – curls of the generalized momenta of the electron and ion fluids – follows from the symmetry of each component with respect to relabeling of the Lagrangian labels. The cross invariants in magnetohydrodynamics are limits of special combinations of topological invariants of the two-fluid model. A variational principle is formulated for the dynamics of frozen-in magnetic flux lines, and the Casimir functionals of the noncanonical Poisson brackets are found. © *1999 American Institute of Physics.* [S1063-7761(99)01508-5]

1. INTRODUCTION

It is well known that the law of freezing-in of flux lines follows from the Euler equations for an ideal liquid (provided the flow is isentropic). A consequence of this law is Kelvin's theorem on the conservation of the circulation of the velocity along an arbitrary streamline.[1] From the standpoint of the Lagrangian formalism, the conservation of these quantities is due to the special symmetry of the equations of ideal hydrodynamics.[2–6] In a Lagrangian description each fluid particle is labeled by a three-dimensional vector $\mathbf{a}$. The dynamics of the fluid is determined by indicating the positions $\mathbf{x}(\mathbf{a},t)$ of each fluid particle at an arbitrary moment in time t. The equations of motion for mapping $\mathbf{r}=\mathbf{x}(\mathbf{a},t)$ follow from the variational principle

$$\delta\int \mathscr{L}\{\mathbf{x}(\mathbf{a},t),\dot{\mathbf{x}}(\mathbf{a},t)\}dt=0.$$

The Lagrangian $\mathscr{L}$ of the fluid admits an infinitely parametric symmetry group — it assumes the same value on all mappings $\mathbf{x}(\mathbf{a},t)$ for which the Eulerian characteristics of the flow are the same — the density $\rho(\mathbf{r},t)$ and the velocity $\mathbf{v}(\mathbf{r},t)$. Such mappings differ from one another only by a relabeling of the Lagrangian labels $\mathbf{a}$, which is why in English the symmetry group is called the relabeling group. All conservation laws for the vorticity are a consequence of this symmetry of the Lagrangian with respect to relabeling of the labels (according the Noether's theorem). The most general formulation of these laws postulates the existence of a local vector Lagrangian invariant — the Cauchy invariant.[7]

A similar situation occurs in ideal magnetohydrodynamics (MHD). In contrast to ordinary hydrodynamics, in MHD the flux lines of the magnetic field $\mathbf{B}$ are frozen-in,[8] but the lines of the vorticity field $\boldsymbol{\Omega}$ are not. Nonetheless, in the course of the motion the vorticity cannot reach arbitrary states, since in addition to the topological invariants of the magnetic field, the so-called ''cross'' invariants containing the vorticity are also conserved. We call topological invariants $\mathscr{I}\{\mathbf{S}\}$ of a certain solenoidal field $\mathbf{S}$ functionals whose variation is zero if the variation of the field has the ''frozen-in'' form

$$\delta\mathbf{S}=\mathrm{curl}[\,\delta\mathbf{x}\times\mathbf{S}].$$

Such a variation shifts the flux lines of the field $\mathbf{S}$ by the small amount $\delta\mathbf{x}(\mathbf{r})$ without changing the topological characteristics of the field. Hence follows a criterion for a given functional to be topological:

$$[\,\mathrm{curl}(\,\delta\mathscr{I}/\delta\mathbf{S})\times\mathbf{S}]=0.$$

Specifically, the integral of the magnetic helicity

$$I_h\{\mathbf{B}\}=\int(\mathbf{B}\,\mathrm{curl}^{-1}\,\mathbf{B})d\mathbf{r}$$

is a topological invariant of the magnetic field. From the geometric standpoint, I_h is measure of the linkage of magnetic flux lines,[9] and for smooth mappings, to within a constant factor it is the same as the Hopf invariant.

The best known cross invariant of MHD is, probably, the cross-helicity integral

$$I_c=\int(\boldsymbol{\Omega}\,\mathrm{curl}^{-1}\,\mathbf{B})d\mathbf{r}=\int(\mathbf{v}\cdot\mathbf{B})d\mathbf{r},$$

which characterizes the number of linkages between the vortex and magnetic flux lines.[9] The freezing-in of the magnetic flux lines and the existence of cross invariants can be easily explained from the standpoint of the relabeling symmetry.. The point is that ordinary MHD can be regarded as the long-wavelength (or low-frequency) limit of the two-fluid model of plasma, where the ion and electron fluids interact with one another via the self-consistent electromagnetic field (see, for example, Ref. 10 and a recently published paper[6]). As is well known,[11] for each component there is a law of freezing-in of

 © 1999 American Institute of Physics

the curl of the generalized momentum in its own substance. For this reason, before passage to the MHD limit, in a two-fluid system two frozen-in fields are present explicitly:

$$\mathbf{\Omega}_1 = \operatorname{curl} \mathbf{v}_i + \frac{e}{m_i c}\mathbf{B}$$

and

$$\mathbf{\Omega}_2 = \frac{m_e}{m_i}\operatorname{curl} \mathbf{v}_e - \frac{e}{m_i c}\mathbf{B},$$

each with its own combination of topological invariants. In the limit of interest to us the dimensional parameter $\alpha = m_i c/e$ is small compared to the ratio $|\mathbf{B}|/|\mathbf{\Omega}_1 + \mathbf{\Omega}_2|$ of the characteristic values. Formally, the passage to the MHD limit means that the parameter α must be set equal to zero, preserving the order of magnitude of the quantities $\mathbf{B} \approx -(m_i c/e)\mathbf{\Omega}_2$ and $\mathbf{\Omega} = \mathbf{\Omega}_1 + \mathbf{\Omega}_2 \approx \operatorname{curl} \mathbf{v}_i$. The approximate equalities are used because the electronic inertia is ignored and the ion velocity $\mathbf{v}_i$ and electron velocity $\mathbf{v}_e$ are assumed to be the same in the leading order. Then the two fields $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ degenerate into one field, whose lines are the same as the flux lines of the magnetic field. To avoid the loss of half of the topological integrals of motion on passing to the limit $\alpha \to 0$, we must examine the limit for each topological invariant $\mathscr{T}\{\mathbf{B}\}$ of the magnetic field:

$$\mathscr{C}_{\mathscr{T}} = \lim_{\alpha \to 0} \frac{\mathscr{T}\{\mathbf{B} + \alpha\mathbf{\Omega}\} - \mathscr{T}\{\mathbf{B}\}}{\alpha} = \int \left(\frac{\delta \mathscr{T}}{\delta \mathbf{B}}\right) \cdot \mathbf{\Omega} d\mathbf{r},$$

which is an integral of motion of ideal MHD. The set of all $\mathscr{C}_{\mathscr{T}}$ is the set of cross invariants. Specifically, using this formula the cross-helicity integral I_c can be obtained from the magnetic-helicity integral I_h.

In this paper we report the results of a study of the motion of frozen-in fields in ideal MHD. It describes the flows that fix all labeling integrals of motion. The analogous representation in ordinary hydrodynamics — the Weber transformation[7] — has been known for a long time; a particular case of the latter is the parameterization of flow by the Clebsch variables. One of the basic results of this paper is an extension of Weber's representation of the case of ideal MHD. This new representation contains a complete vector field — a Lagrangian invariant. In ordinary hydrodynamics this invariant ensures conservation of the Cauchy invariant. As a result, it ensures all conservation laws for the vorticity.[5] In MHD this vector field ensures conservation of the cross invariants. In contrast to hydrodynamics, it cannot be expressed explicitly in terms of the observed fields. It is shown that the analog of the Weber representation for MHD can be obtained by passing to the limit $\alpha \to 0$ from a pair of ordinary Weber representations which parameterize the generalized momenta in the two-fluid model. Such a limit procedure shows the meaning of all quantities appearing in the result. The same answer is obtained by direct integration of the MHD equations in the Lagrangian description.

Another result of this work is that the parameterization found helps to formulate a variational principle for the dynamics of magnetic flux lines in incompressible MHD. The approach employed is based on the integral representation of frozen-in solenoidal fields, which fixes all of their topological properties. The magnetic flux lines are the new dynamical objects. Each flux line is enumerated by its Lagrangian label and admits a gauge freedom in its parameterization. Such a description is intermediate between the Lagrangian and Eulerian descriptions.[12–14] The gauge-invariant equations of motion obtained for the form of the lines determine only the transverse dynamics on which the observed quantities depend. The condition of conservation of the volume of each magnetic tube is not imposed in the form of a constraint, but rather it is an integral of motion that, according to Noether's theorem, follows from the symmetry of the new Lagrangian with respect to relabeling of the magnetic flux lines. This formulation of ideal incompressible MHD is similar to the representation of incompressible flows in ordinary hydrodynamics by means of frozen-in vortex lines.[13,15] The new representation of incompressible MHD flows describes best the fields with a complicated topology, as well as localized structures such as magnetic filaments.

This paper is organized as follows. In Sec. 1 the required information about the canonical formalism for general hydrodynamic systems is given and the Weber representation for generalized momenta is obtained. In Sec. 2, a model intermediate between the two-fluid model of plasma and magnetohydrodynamics — the so-called Hall MHD or MHD with dispersion[16] — is studied. In Sec. 3 the Weber representation for ordinary MHD is obtained from this model by passage to a limit. In Sec. 4 a gauge-free description of magnetic and vortex fields in incompressible MHD systems is introduced, the equations of motion are written out, and a variational principle is formulated for the dynamics of the new objects. The question of Casimir functionals of noncanonical Poisson brackets is also discussed in Sec. 4 (Ref. 17). In Sec. 5, a method for choosing a gauge, which gives canonically conjugate dynamical quantities in the new description, is demonstrated for two-dimensional MHD.

2. THE CANONICAL FORMALISM FOR A FLUID

For simplicity, we consider one-fluid systems first, and then we make obvious generalizations to multicomponent models. We shall be interested in the mechanical motion of a liquid as a continuous medium, and therefore we shall examine Lagrangian systems for which the elements of configuration space are the mappings $\mathbf{r} = \mathbf{x}(\mathbf{a})$ of the three-dimensional **a**-space of the labels of ''fluid particles'' into the physical **r**-space. The description of motion by giving the dependence of the mappings **x** on the time $\mathbf{r} = \mathbf{x}(\mathbf{a}, t)$ is generally called the Lagrangian description. This mapping determines the density field $\rho(\mathbf{r}, t)$ and the velocity field $\mathbf{v}(\mathbf{r}, t)$, and it is always possible to choose a labeling so that the density is expressed in terms of the Jacobian of the mapping

$$J_{\mathbf{x}} = \det \| \partial \mathbf{x} / \partial \mathbf{a} \|$$

without an additional factor which depends on **a**:

$$\rho(\mathbf{r}, t) = \left. \frac{1}{\det \| \partial \mathbf{x} / \partial \mathbf{a} \|} \right|_{\mathbf{a} = \mathbf{x}^{-1}(\mathbf{r}, t)}, \quad \mathbf{v}(\mathbf{r}, t) = \dot{\mathbf{x}}(\mathbf{a}, t)\big|_{\mathbf{a} = \mathbf{x}^{-1}(\mathbf{r}, t)}. \tag{1}$$

The dependences of all physical quantities $(\rho,\mathbf{v})$ and also the fields f external to the medium on the coordinates $\mathbf{r}$ and time t give the Eulerian description of flow.

We distinguish among all possible Lagrangians $\mathscr{L}\{\mathbf{x}(\mathbf{a}),\dot{\mathbf{x}}(\mathbf{a}),\ f(r),\ f_t(\mathbf{r})\}$ the ones that possess symmetry with respect to relabeling of the labels $\mathbf{a}$ (relabeling group). This means that the indicated functionals are determined completely by the Eulerian characteristics of the flow: the density, the velocity, and the external fields. The variation of the external fields that is required to construct the equations of motion of the fields is done directly in the Eulerian representation. For this reason, we shall first study Lagrangians of the form $\mathscr{L}=L\{\rho,\mathbf{v}\}$. Lagrangians of many physically important systems possess the indicated symmetry. In this case a closed system of equations in an Eulerian representation can be written for the quantities $\mathbf{v}$ and ρ. An equation for the density is obtained from the kinematic relation between the variation of the density and the small displacement $\delta\mathbf{x}(\mathbf{a})$ (the argument t is dropped for brevity):

$$\delta\rho(\mathbf{r})=-\nabla(\rho(\mathbf{r})\,\delta\mathbf{x}(\mathbf{a}(\mathbf{r}))). \tag{2}$$

This relation follows from Eq. (1) and gives the continuity equation

$$\rho_t+\nabla\cdot(\rho\mathbf{v})=0. \tag{3}$$

To write the dynamical Euler–Lagrange equation

$$\frac{d}{dt}\left(\left.\frac{\delta\mathscr{L}}{\delta\dot{\mathbf{x}}(\mathbf{a})}\right|_{\mathbf{x}(\mathbf{a})}\right)=\frac{\delta\mathscr{L}}{\delta\mathbf{x}(\mathbf{a})}^{\dot{\mathbf{x}}(\mathbf{a})}, \tag{4}$$

in the Eulerian representation it is necessary to use the formula

$$\delta\mathbf{v}(\mathbf{r})=\delta\dot{\mathbf{x}}(\mathbf{a}(\mathbf{r}))-(\delta\mathbf{x}(\mathbf{a}(\mathbf{r}))\cdot\nabla)\mathbf{v}(\mathbf{r}),$$

which follows from the definition (1), together with Eq. (2) for substituting into the equality of the variations

$$\int\left(\frac{\delta\mathscr{L}}{\delta\dot{\mathbf{x}}}\cdot\delta\dot{\mathbf{x}}+\frac{\delta\mathscr{L}}{\delta\mathbf{x}}\cdot\delta\mathbf{x}\right)d\mathbf{a}=\int\left(\frac{\delta\mathscr{L}}{\delta\mathbf{v}}\cdot\delta\mathbf{v}+\frac{\delta\mathscr{L}}{\delta\rho}\delta\rho\right)d\mathbf{r}.$$

A simple integration by parts gives

$$\left.\frac{\delta\mathscr{L}\{\rho,\mathbf{v}\}}{\delta\mathbf{x}(\mathbf{a})}\right|_{\dot{\mathbf{x}}(\mathbf{a})}=\left.\left(\nabla\left(\frac{\delta\mathscr{L}}{\delta\rho}\right)-\frac{1}{\rho}\left(\frac{\delta\mathscr{L}}{\delta v_m}\right)\nabla v_m\right)\right|_{\mathbf{r}=\mathbf{x}(\mathbf{a},t)}, \tag{5}$$

$$\mathbf{p}(\mathbf{a},t)\equiv\left.\frac{\delta\mathscr{L}\{\rho,\mathbf{v}\}}{\delta\dot{\mathbf{x}}(\mathbf{a})}\right|_{\mathbf{x}(\mathbf{a})}=\left.\left(\frac{1}{\rho}\left(\frac{\delta\mathscr{L}}{\delta\mathbf{v}}\right)\right)\right|_{\mathbf{r}=\mathbf{x}(\mathbf{a},t).} \tag{6}$$

Here $\mathbf{p}(\mathbf{a},t)$ is the generalized momentum, which, in general, is different from the velocity. This circumstance is very important for many models.

The equation of motion (4) is now (the generalized Euler equation)

$$(\partial_t+\mathbf{v}\cdot\nabla)\left(\frac{1}{\rho}\frac{\delta\mathscr{L}}{\delta\mathbf{v}}\right)=\nabla\left(\frac{\delta\mathscr{L}}{\delta\rho}\right)-\frac{1}{\rho}\left(\frac{\delta\mathscr{L}}{\delta v_m}\right)\nabla v_m. \tag{7}$$

The equations (3) and (7) completely determine the motion.

As an example, we shall consider the relativistically invariant expression for the action of a barotropic liquid (in dimensionless variables):

$$\mathscr{T}_r=\int\mathscr{L}_r dt=-\int dt\int d\mathbf{r}\,\varepsilon(\rho\sqrt{1-\mathbf{v}^2}), \tag{8}$$

where ρ is the number density of the conserved particles in the laboratory reference frame, $\rho\sqrt{1-\mathbf{v}^2}=\tilde{\rho}$ is the density in a locally comoving reference frame, and $\varepsilon(\tilde{\rho})$ is the relativistic internal energy density of the liquid, including the rest energy. It can be easily shown that the extension of the Euler equation in the well-known textbook by Landau and Lifshitz (see Ref. 1, Eq. (134.10) to relativistic isentropic flows is identical to Eq. (7), if we substitute in it the given Lagrangian and take into account the fact that the relativistic enthalpy can be expressed as $w=\partial\varepsilon/\partial\tilde{\rho}$:

$$(\partial_t+\mathbf{v}\cdot\nabla)\left(\frac{\mathbf{v}w(\tilde{\rho})}{\sqrt{1-\mathbf{v}^2}}\right)=-\nabla w(\tilde{\rho})\sqrt{1-\mathbf{v}^2}. \tag{9}$$

We underscore that the continuity equation must be written for ρ and not $\tilde{\rho}$. It can then be represented as

$$\partial_t\left(\frac{\tilde{\rho}(w)}{\sqrt{1-\mathbf{v}^2}}\right)+\nabla\cdot\left(\frac{\tilde{\rho}(w)\mathbf{v}}{\sqrt{1-\mathbf{v}^2}}\right)=0, \tag{10}$$

which makes it possible to treat the pair $(w,\mathbf{v})$ as dynamical quantities. The generalized momentum $\mathbf{p}$ can be determined from the formula

$$\mathbf{p}=w(\rho\sqrt{1-\mathbf{v}^2})\frac{\mathbf{v}}{\sqrt{1-\mathbf{v}^2}}. \tag{11}$$

We see that in relativistic hydrodynamics the generalized momentum depends in a nonuniversal manner on the velocity. The relation between $\mathbf{p}$ and $\mathbf{v}$ is unique for each equation of state, and it contains the density ρ. In the nonrelativistic limit $v\ll1$, $\rho\ll1$, $\tilde{\rho}\approx\rho$, $w(\tilde{\rho})\approx1+w_{nr}(\rho)$, we have, as might be suggested, the equations of ordinary hydrodynamics in which $w_{nr}(\rho)$ is the nonrelativistic enthalpy.

Next, using the standard rules, we switch from the Lagrangian to the Hamiltonian formalism. In this case, we use instead of the generalized velocity $\dot{\mathbf{x}}=\mathbf{v}$ a different variable — the generalized momentum $\mathbf{p}=\delta\mathscr{L}/\delta\dot{\mathbf{x}}$. In the Eulerian representation it is expressed by Eq. (6). We determine the Hamiltonian functional as

$$\mathscr{H}=\int(\mathbf{p}(\mathbf{a})\cdot\dot{\mathbf{x}}(\mathbf{a}))d\mathbf{a}-\mathscr{L}=\int\left(\frac{\delta\mathscr{L}}{\delta\mathbf{v}}\cdot\mathbf{v}\right)d\mathbf{r}-\mathscr{L}. \tag{12}$$

Here the velocity $\mathbf{v}$ must be expressed in terms of $\mathbf{p}$ using Eq. (6). It is easily verified that it can be determined, using the known Hamiltonian, by the formula

$$\mathbf{v}=\frac{1}{\rho}\left(\frac{\delta\mathscr{H}}{\delta\mathbf{p}}\right)\Bigg|_{\rho}. \tag{13}$$

The following relation holds between the derivatives:

$$\left.\frac{\delta\mathscr{L}}{\delta\rho}\right|_{\mathbf{v},\sigma}=\frac{\mathbf{p}}{\rho}\left(\frac{\delta\mathscr{H}}{\delta\mathbf{p}}\right)\Bigg|_{\rho}-\left.\frac{\delta\mathscr{H}}{\delta\rho}\right|_{\mathbf{p}}. \tag{14}$$

In the variables $\{\rho,\mathbf{p}\}$ the dynamical system corresponding to a certain Hamiltonian $\mathscr{H}\{\rho,\mathbf{p}\}$ assumes the form

$$\rho_t+\nabla\cdot\left(\frac{\delta\mathscr{H}}{\delta\mathbf{p}}\right)=0, \tag{15}$$

$$\mathbf{p}_t=\left[\left(\frac{\delta\mathscr{H}}{\delta\mathbf{p}}\right)\times\frac{\operatorname{curl}\mathbf{p}}{\rho}\right]-\nabla\left(\frac{\delta\mathscr{H}}{\delta\rho}\right), \tag{16}$$

as can be verified using Eqs. (5), (6), (13), and (14). Calculating the time derivative of a certain functional $F\{\rho,\mathbf{p}\}$, we obtain an expression of the form $\dot F=\{F,\mathscr{H}\}$, where the right-hand side is a bilinear antisymmetric form, determined for any two functionals $F\{\rho,\mathbf{p}\}$ and $G\{\rho,\mathbf{p}\}$:

$$\{F,G\}=\int\left(\frac{\operatorname{curl}\mathbf{p}}{\rho}\left[\frac{\delta F}{\delta\mathbf{p}}\times\frac{\delta G}{\delta\mathbf{p}}\right]\right)d\mathbf{r}+\int\left(\frac{\delta G}{\delta\rho}\nabla\cdot\left(\frac{\delta F}{\delta\mathbf{p}}\right)-\frac{\delta F}{\delta\rho}\nabla\cdot\left(\frac{\delta G}{\delta\mathbf{p}}\right)\right)d\mathbf{r}. \tag{17}$$

It can be verified that this formula determines the hydrodynamic Poisson bracket.[17] Besides the indicated antisymmetry property this form also satisfies the Jacobi identity

$$\{\{F,G\},H\}+\{\{H,F\},G\}+\{\{G,H\},F\}=0.$$

We shall now consider another example — a nonrelativistic barotropic fluid in a fixed external potential field. The Lagrangian is

$$\mathscr{L}_{nr}=\int\left(\rho\frac{\mathbf{v}^2}{2}-\varepsilon(\rho)-\rho\phi(\mathbf{r})\right)d\mathbf{r},$$

where $\varepsilon(\rho)$ is the ordinary nonrelativistic internal energy, and $\phi(\mathbf{r})$ is the potential of the external force, for example, the gravitational potential. The generalized momentum in this case is the velocity $\mathbf{p}=\mathbf{v}$, and the Hamiltonian is a sum of the kinetic, internal, and potential energies:

$$\mathscr{H}_{nr}=\int\left(\rho\frac{\mathbf{p}^2}{2}+\varepsilon(\rho)+\rho\phi(\mathbf{r})\right)d\mathbf{r}.$$

The equations of motion obtained from Eqs. (15) and (16) have the well-known form

$$\rho_t+\nabla\cdot(\rho\mathbf{v})=0,\quad \mathbf{v}_t=[\mathbf{v}\times\operatorname{curl}\mathbf{v}]-\nabla\left(\frac{\mathbf{v}^2}{2}+w(\rho)+\phi(\mathbf{r})\right).$$

As already mentioned, hydrodynamic-type systems possess an infinite number of integrals of motion. To demonstrate this fact we shall examine the time derivative of the quantity $u_\mu(\mathbf{a},t)=p_k(\partial x_k/\partial a_\mu)$. Using the Euler–Lagrange equation and Eqs. (5) and (6), we obtain

$$\dot u_\mu=\dot p_k\frac{\partial x_k}{\partial a_\mu}+p_k\frac{\partial\dot x_k}{\partial a_\mu}=\frac{\partial x_k}{\partial a_\mu}\left(\frac{\partial}{\partial x_k}\left(\frac{\delta\mathscr{L}}{\delta\rho}\right)-p_m\frac{\partial v_m}{\partial x_k}\right)+p_k\frac{\partial v_k}{\partial a_\mu}=\frac{\partial}{\partial a_\mu}\left(\frac{\delta\mathscr{L}}{\delta\rho}\right).$$

Integrating this relation with respect to time, we obtain the so-called Weber representation[7]

$$p_k(\mathbf{a},t)\frac{\partial x_k(\mathbf{a},t)}{\partial a_\mu}=u_{0\mu}(\mathbf{a})+\frac{\partial\varphi(\mathbf{a},t)}{\partial a_\mu}, \tag{18}$$

where the scalar potential φ satisfies the equation

$$\dot\varphi(\mathbf{a},t)=\frac{\delta\mathscr{L}}{\delta\rho}=\left(\frac{\mathbf{p}}{\rho}\cdot\left(\frac{\delta\mathscr{H}}{\delta\mathbf{p}}\right)-\frac{\delta\mathscr{H}}{\delta\rho}\right).$$

Conservation of the ''transverse'' component of the field $\mathbf{u}(\mathbf{a},t)$ follows from Eq. (18). This is equivalent to the assertion

$$\epsilon_{\alpha\beta\gamma}\frac{\partial p_k(\mathbf{a},t)}{\partial a_\beta}\frac{\partial x_k(\mathbf{a},t)}{\partial a_\gamma}=(\operatorname{curl}_{\mathbf{a}}\mathbf{u}_0(\mathbf{a}))_\alpha=\Omega_{0\alpha}(\mathbf{a}). \tag{19}$$

The equality (19) is, precisely, a formulation of the conservation laws, since its right-hand side is time-independent. In this form the integrals of motion are expressed in terms of the local characteristics of the flow and are simply a Cauchy invariant.[7] A different (and equivalent) formulation involves choosing an arbitrary solenoidal field $\mathbf{g}(\mathbf{a})$ and studying the integral

$$I_{\mathbf{g}}=\int p_k(\mathbf{a},t)\frac{\partial x_k(\mathbf{a},t)}{\partial a_\mu}g_\mu(\mathbf{a})d\mathbf{a}. \tag{20}$$

Substituting here the Weber representation (18) and integrating the term with φ by parts, we conclude that $I_{\mathbf{g}}=\text{const}$.

Choosing the field $\mathbf{g}(\mathbf{a})$ in a special form such that the field is concentrated on a closed contour gives Kelvin's theorem on the conservation of the circulation

$$\Gamma=\oint\mathbf{p}\cdot d\mathbf{r}$$

along a streamline.[1] We note that a direct application of Noether's theorem also gives the integrals of motion (20). Let us consider a one-parameter group of transformations of the Lagrangian labels $\mathbf{a}^\tau_{\mathbf{g}}(\mathbf{a})$, which is the solenoidal vector field $\mathbf{g}(\mathbf{a})$ (the solenoidal nature of the field follows from the constancy of the density). Let us consider an arbitrary mapping $\mathbf{x}(\mathbf{a},t)$. Irrespective of the value of the group parameter τ, all mappings $\mathbf{x}^\tau_{\mathbf{g}}(\mathbf{a},t)=\mathbf{x}(\mathbf{a}^\tau_{\mathbf{g}}(\mathbf{a}),t)$ describe the same flow $\{\rho(\mathbf{r},t),\mathbf{v}(\mathbf{r},t)\}$. For this reason, the Lagrangian assumes one value on such mappings. Noether's theorem states that in this case there exists a conservation law[18]

$$\int\mathbf{p}\cdot\left(\frac{\delta\mathbf{x}^\tau_{\mathbf{g}}}{\delta\tau}\right)\Bigg|_{\tau=0}d\mathbf{a}=\text{ const}.$$

Using for small τ the formulas

$$\mathbf{a}^\tau_{\mathbf{g}}(\mathbf{a})=\mathbf{a}+\tau\mathbf{g}(\mathbf{a})+O(\tau^2),$$

$$\mathbf{x}^\tau_{\mathbf{g}}(\mathbf{a},t)=\mathbf{x}(\mathbf{a}+\tau\mathbf{g}(\mathbf{a})+O(\tau^2),t)=\mathbf{x}(\mathbf{a},t)+\tau(\mathbf{g}(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{x}(\mathbf{a},t)+O(\tau^2),$$

we immediately obtain the expression (20) for $I_{\mathbf{g}}$.

The formula (19) can be rewritten in the following form, which determines a relation between the vorticity $\boldsymbol{\Omega}(\mathbf{r},t)=\operatorname{curl}\mathbf{p}(\mathbf{r},t)$ of the flow and a Cauchy invariant:

$$\boldsymbol{\Omega}(\mathbf{r},t)=\frac{(\boldsymbol{\Omega}_0(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{x}(\mathbf{a},t)}{\det(\partial\mathbf{x}(\mathbf{a},t)/\partial\mathbf{a})}\Bigg|_{\mathbf{a}=\mathbf{x}^{-1}(\mathbf{r},t)}=\int\delta(\mathbf{r}-\mathbf{x}(\mathbf{a},t))\times(\boldsymbol{\Omega}_0(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{x}(\mathbf{a},t)d\mathbf{a}. \tag{21}$$

The formula (21) shows that the lines of initial solenoidal field $\boldsymbol{\Omega}_0$ are deformed by the mapping $\mathbf{x}(\mathbf{a},t)$, while preserv-

ing all topological properties of the field in the process. Specifically, the degree of linkage of the vortex lines, which is determined by the helicity integral $\int \mathbf{p}\cdot\mathbf{\Omega}d\mathbf{r}$, remains unchanged.[9] It is easy to see that this conservation law applies to the family (20) and corresponds to the choice $\mathbf{g}(\mathbf{a})=\mathbf{\Omega}_0(\mathbf{a})$.

All calculations performed above admit an obvious generalization to N-fluid hydrodynamics. In multicomponent hydrodynamic models, where the state of system is described by several fields ρ_j and $\mathbf{v}_j$, the curl of each generalized momentum is frozen into its own substance with label $\mathbf{a}_j$. Weber's representations consisting of N formulas parameterize each generalized momentum. We also note that in certain models, it is more convenient to use instead of the densities ρ_j the quantities to which they are proportional, for example, the concentrations n_j of particles of a definite kind. This is the situation in the two-fluid model of plasma studied below.

3. FROZEN-IN FIELDS IN MAGNETOHYDRODYNAMICS

It is well-known that in ideal MHD the magnetic field is frozen into the fluid. The laws of conservation of the topological invariants of the magnetic field, for example, the measure of linkage of the magnetic flux lines,[19,11,20] are related to this fact. Besides these integrals of motion, there also exist so-called ''cross'' topological invariants, whose expressions contain the curl of the velocity. To understand their origin we recall that the MHD equations can be obtained in the low-frequency (or long-wavelength) limit from a two-fluid model of plasma, where electrons and ions are considered to be two separate fluids. Each component admits a labeling symmetry, and therefore the system contains two independent frozen-in fields,[21] i.e., two series of topological integrals of motion. As will be shown below, the long-wavelength, or low-frequency, limit means neglecting electronic inertia. The cross integrals of MHD in this case correspond to the limits of the special combinations of invariants of the two-fluid model.

Let us consider the two-fluid model of plasma in which the state of the system is described by the following fields: n_1 and n_2 — ion and electron densities; $\mathbf{v}_1$ and $\mathbf{v}_2$ — velocities of the ion and electron fluids; and $\mathbf{A}$ — the vector potential of the electromagnetic field. Let the scalar potential be zero, which fixes the gauge. We shall ignore the electron mass as compared with the ion mass m and assume that the conductivity of the medium is infinite, i.e., one fluid can flow without friction through the other. Let the velocities be nonrelativistic and the electric field much smaller than the magnetic field. Accordingly, the action of the fluid and the field can be written as

$$\mathscr{T}_{2f}=\int\left(mn_1\frac{\mathbf{v}_1^2}{2}+\frac{e}{c}(n_1\mathbf{v}_1-n_2\mathbf{v}_2)\cdot\mathbf{A}-\frac{1}{8\pi}(\operatorname{curl}\mathbf{A})^2-\varepsilon(n_1,n_2)\right)d\mathbf{r}dt, \tag{22}$$

where $\varepsilon(n_1,n_2)$ is the internal energy of the fluids. Variation of the action according to the scheme (7) gives the continuity equations

$$\operatorname{curl}\operatorname{curl}\mathbf{A}=\frac{4\pi e}{c}(n_1\mathbf{v}_1-n_2\mathbf{v}_2), \tag{23}$$

$$(\partial_t+\mathbf{v}_1\cdot\nabla)m\mathbf{v}_1=\frac{e}{c}(-\mathbf{A}_t+[\mathbf{v}_1\times\operatorname{curl}\mathbf{A}])-\nabla\frac{\partial\varepsilon}{\partial n_1}, \tag{24}$$

$$0=-\frac{e}{c}(-\mathbf{A}_t+[\mathbf{v}_2\times\operatorname{curl}\mathbf{A}])-\nabla\frac{\partial\varepsilon}{\partial n_2}. \tag{25}$$

To these equations we must add two continuity equation

$$n_{1,t}+\nabla\cdot(n_1\mathbf{v}_1)=0,\quad n_{2,t}+\nabla\cdot(n_2\mathbf{v}_2)=0. \tag{26}$$

Now, since by definition

$$\operatorname{curl}\mathbf{A}=\mathbf{B},$$

and also in the gauge chosen

$$-\frac{1}{c}\mathbf{A}_t=\mathbf{E},$$

Eq. (23) is easily recognized as the equation for the quasistationary magnetic field,[8] Eq. (24) is the equation of motion for the ions, and from Eq. (25) we obtain an equation for the freezing of the magnetic field into the electronic component:

$$\mathbf{B}_t=\operatorname{curl}[\mathbf{v}_2\times\mathbf{B}].$$

Applying the operation div to Eq. (23) gives, using the continuity equations, the relation

$$n_{1,t}-n_{2,t}=0.$$

Integrating over time gives the condition of electric neutrality

$$n_1=n_2=n.$$

We note that a different choice of the arbitrary integration function $[n_1-n_2=C(\mathbf{r})]$ is physically unjustified, because this would make the model under study, where quasineutrality is assumed at the outset, non-self-consistent. Eliminating n_2 and $\mathbf{v}_2$ from the equations written down and making the substitution $\mathbf{v}_1=\mathbf{v}$ gives the system (MHD with dispersion[16] or Hall MHD[6])

$$(\partial_t+\mathbf{v}\cdot\nabla)m\mathbf{v}=-\nabla w(n)+\frac{1}{4\pi n}[\operatorname{curl}\mathbf{B}\times\mathbf{B}],$$

$$w(n)=\frac{\partial}{\partial n}\varepsilon(n,n),$$

$$n_t+\nabla\cdot(n\mathbf{v})=0,\quad \mathbf{B}_t=\operatorname{curl}\left[\left(\mathbf{v}-\frac{c}{4\pi en}\operatorname{curl}\mathbf{B}\right)\times\mathbf{B}\right]. \tag{27}$$

We note that ordinary magnetohydrodynamics is obtained from these equations in the limit of high concentrations and small gradients (the MHD limit), where in the last equation the term $(c/4\pi en)\operatorname{curl}\mathbf{B}$ can be ignored compared to $\mathbf{v}$. At the same time, the vector potential $\mathbf{A}$ of magnetic field must be large compared with the characteristic values of $(mc/e)\mathbf{v}$, so that the magnetic and inertial terms in Euler's equation would be at least of the same order of magnitude. From these requirements follows a necessary condition for applicability of ideal MHD

$$nL^2 \gg mc^2/e^2,$$

where L is the spatial scale.

It is noteworthy that if the indicated term is not disregarded, hiding thereby the symmetry of the initial system, we obtain two different frozen-in fields

$$\mathbf{\Omega}_2 = -\frac{e}{mc}\mathbf{B},$$

which is frozen into the electronic component, and the field

$$\mathbf{\Omega}_1 = \mathrm{curl}\left(\mathbf{v} + \frac{e}{mc}\mathbf{A}\right) = \mathbf{\Omega} - \mathbf{\Omega}_2,$$

which is frozen into the ionic component. Therefore there are two series of topological invariants $\mathscr{T}\{\mathbf{\Omega}_1\}$ and $\mathscr{T}\{\mathbf{\Omega}_2\}$. We call topological invariants of a solenoidal field all possible quantitative characteristics that remain unchanged under variations of the field that reduce to a continuous deformation of the flux lines of the field. In the present paper, therefore, a frozen-in solenoidal field is imagined to be a continuous set of lines that do not start or end anywhere, they are not created and do not vanish with time, and they do not intersect one another in the course of the motion.

In the MHD limit the sum of the frozen-in fields (i.e., the field $\mathbf{\Omega}$) is ''infinitely'' small compared with each field, so that the corresponding invariants $\mathscr{T}\{\mathbf{B}+(mc/e)\mathbf{\Omega}\}$ and $\mathscr{T}\{\mathbf{B}\}$ are almost equal to one another. To preclude a loss of half of the invariants, we must consider the following limit for each $\mathscr{T}\{\mathbf{B}\}$:

$$\mathscr{C}_{\mathscr{T}} = \lim_{\alpha\to 0}\frac{\mathscr{T}\{\mathbf{B}+\alpha\mathbf{\Omega}\} - \mathscr{T}\{\mathbf{B}\}}{\alpha} = \int\left(\frac{\delta\mathscr{T}}{\partial\mathbf{B}}\cdot\mathbf{\Omega}\right)d\mathbf{r}, \qquad (28)$$

which is an integral of motion of ideal MHD. The set of all $\mathscr{C}_{\mathscr{T}}$ replaces the second series of topological invariants of the system (22) in the MHD limit. Therefore, in ordinary MHD each topological integral, which contains only the magnetic field, corresponds a ''cross'' invariant (28). For example, from the magnetic-helicity integral

$$I_h\{\mathbf{B}\} = \int (\mathbf{A}\cdot\mathbf{B})d\mathbf{r}$$

we obtain the cross-helicity integral

$$I_c = \int (\mathbf{A}\cdot\mathbf{\Omega})d\mathbf{r},$$

which characterizes the number of linkages of the vortex and magnetic flux lines.

Following the formalism expounded above, we shall switch to the Hamiltonian description of the system (22). The generalized momenta are determined in this case by the formulas

$$\mathbf{p}_1 = m\mathbf{v}_1 + \frac{e}{c}\mathbf{A}, \quad \mathbf{p}_2 = -\frac{e}{c}\mathbf{A}. \qquad (29)$$

The Hamiltonian has the form

$$\mathscr{H}_{2f} = \int\left(\frac{n_1(\mathbf{p}_1+\mathbf{p}_2)^2}{2m} + \frac{c^2}{8\pi e^2}(\mathrm{curl}\,\mathbf{p}_2)^2 + \varepsilon(n_1,n_2)\right)d\mathbf{r}. \qquad (30)$$

For generality, we shall consider all two-fluid Hamiltonians of the special form $\mathscr{H}\{\mathbf{p}_1+\mathbf{p}_2, -\alpha\,\mathrm{curl}\,\mathbf{p}_2,\ n_1,\ n_2\}$, where α is a parameter. We introduce the notation

$$\mathbf{p} = \mathbf{p}_1 + \mathbf{p}_2, \quad \mathbf{B} = -\alpha\,\mathrm{curl}\,\mathbf{p}_2.$$

All such systems possess the integral of motion

$$n_1 - n_2 = C(\mathbf{r}).$$

We are interested in the case $n_1 = n_2 = n$, for which the equations of motion can be rewritten as

$$\mathbf{p}_t = \left[\left(\frac{\delta\mathscr{H}}{\partial\mathbf{p}}\right)\times\frac{\mathrm{curl}\,\mathbf{p}}{n}\right] - \nabla\left(\frac{\delta\mathscr{H}}{\delta n}\right) + \left[\mathrm{curl}\left(\frac{\delta\mathscr{H}}{\delta\mathbf{B}}\right)\times\frac{\mathbf{B}}{n}\right], \qquad (31)$$

$$\mathbf{B}_t = \mathrm{curl}\left[\left(\frac{\delta\mathscr{H}}{\delta\mathbf{p}} - \alpha\,\mathrm{curl}\left(\frac{\delta\mathscr{H}}{\delta\mathbf{B}}\right)\right)\times\frac{\mathbf{B}}{n}\right], \qquad (32)$$

$$n_t + \nabla\left(\frac{\delta\mathscr{H}}{\delta\mathbf{p}}\right) = 0. \qquad (33)$$

It can be easily verified that the Poisson brackets for such systems have the form

$$\begin{aligned}\{F,G\} = \int &\left(\frac{\mathbf{B}}{n}\left(\left[\mathrm{curl}\frac{\delta F}{\delta\mathbf{B}}\times\frac{\delta G}{\delta\mathbf{p}}\right] - \left[\mathrm{curl}\frac{\delta G}{\delta\mathbf{B}}\times\frac{\delta F}{\delta\mathbf{p}}\right]\right)\right)\\ &\times d\mathbf{r} + \int\left(\frac{\mathrm{curl}\,\mathbf{p}}{n}\left[\frac{\delta F}{\delta\mathbf{p}}\times\frac{\delta G}{\delta\mathbf{p}}\right]\right)d\mathbf{r}\\ &- \alpha\int\left(\frac{\mathbf{B}}{n}\left[\mathrm{curl}\frac{\delta F}{\delta\mathbf{B}}\times\mathrm{curl}\frac{\delta G}{\delta\mathbf{B}}\right]\right)d\mathbf{r}\\ &+ \int\left(\frac{\delta G}{\delta n}\nabla\left(\frac{\delta F}{\delta\mathbf{p}}\right) - \frac{\delta F}{\delta n}\nabla\left(\frac{\delta G}{\delta\mathbf{p}}\right)\right)d\mathbf{r}. \qquad (34)\end{aligned}$$

If α is set equal to zero, these brackets become the Poisson brackets of ordinary MHD.[17] This limit corresponds to an infinite increase in the absolute values of each of the generalized momenta $\mathbf{p}_1$ and $\mathbf{p}_2$ with their sum remaining finite.

4. WEBER'S REPRESENTATION FOR MAGNETOHYDRODYNAMICS

We shall now determine into what Weber's representation transforms for an ion fluid in the limit $\alpha\to 0$. As a result, the flow velocities of the electronic and ion fluids become the same. As long as α is finite, we have two different, though slightly different from one another, mappings $\mathbf{x}_1(\mathbf{a})$ and $\mathbf{x}_2(\mathbf{c})$, where $\mathbf{a}$ is the label of the ion fluid, and $\mathbf{c}$ is the label of the electron fluid. We write the corresponding Weber representations for each generalized momentum:

$$\mathbf{p}_1 = \nabla a_\mu u_{0\mu}^{(1)}(\mathbf{a}) + \nabla\varphi^{(1)}, \quad \mathbf{p}_2 = \nabla c_\mu u_{0\mu}^{(2)}(\mathbf{c}) + \nabla\varphi^{(2)}. \qquad (35)$$

The field dependences $u_{0\mu}^{(1)}(\mathbf{a})$ and $u_{0\mu}^{(2)}(\mathbf{c})$ are such that if they are reduced to the same argument, then their sum will be finite but small compared with each term separately:

$$u_{0\mu}^{(1)}(\mathbf{a})+u_{0\mu}^{(2)}(\mathbf{a})=u_{0\mu}(\mathbf{a}).$$

Since $\mathbf{x}_1(\mathbf{a})$ and $\mathbf{x}_2(\mathbf{c})$ are almost the same, we can assert that in the expression (where, for simplicity, we omit the argument t)

$$\mathbf{c}(\mathbf{r})=\mathbf{a}(\mathbf{r})+\mathbf{d}(\mathbf{a}(\mathbf{r}))$$

the quantity $\mathbf{d}[\mathbf{a}(\mathbf{r})]$ is small compared with the other terms and is an infinitesimal of order α. Since the ion density equals the electron density, there is the additional condition

$$\det\|\partial\mathbf{a}/\partial\mathbf{r}\|=\det\|\partial\mathbf{c}/\partial\mathbf{r}\|.$$

This condition implies that $\nabla_{\mathbf{a}}\cdot\mathbf{d}=0$ for small values of $\mathbf{d}$. Let us now take the sum of the two expressions from Eq. (35) and retain only terms that are not small for small α:

$$\mathbf{p}=\nabla a_\mu u_{0\mu}(\mathbf{a})+\nabla c_\mu u_{0\mu}^{(2)}(\mathbf{c})-\nabla a_\mu u_{0\mu}^{(2)}(\mathbf{a})+\nabla\widetilde{\varphi}$$
$$\approx\nabla a_\mu u_{0\mu}(\mathbf{a})+\nabla d_\mu u_{0\mu}^{(2)}(\mathbf{a})+\nabla a_\mu d_\lambda u_{0\mu,\lambda}^{(2)}(\mathbf{a})+\nabla\widetilde{\varphi}.$$

Now we shall no longer distinguish between $\mathbf{a}$ and $\mathbf{c}$. Introducing the notation

$$\varphi=d_\mu u_{0\mu}^{(2)}+\widetilde{\varphi},\quad d_\lambda/\alpha=-\epsilon_{\lambda\alpha\beta}S_{\beta,\alpha}(\mathbf{a},t)$$

for the solenoidal field $\mathbf{d}(\mathbf{a},t)/\alpha$ and using the equality

$$\mathbf{B}_0(\mathbf{a})=-\alpha\,\mathrm{curl}_{\mathbf{a}}\,\mathbf{u}_0^{(2)}(\mathbf{a}),$$

we rewrite this relation in the form

$$p_k\frac{\partial x_k}{\partial a_\mu}=u_{0\mu}(\mathbf{a})+\varphi_{,\mu}(\mathbf{a},t)+B_{0\lambda}(\mathbf{a})(S_{\lambda,\mu}(\mathbf{a},t)-S_{\mu,\lambda}(\mathbf{a},t)).\tag{36}$$

This formula is an extension of Weber's representation to one-fluid hydrodynamics in the case where the system contains a frozen-in magnetic field.[14] We shall show that it can also be obtained directly by partial integration of the MHD equations written in the Lagrangian representation. The MHD Lagrangian has the form

$$\mathscr{L}_*=\int\left(\rho\frac{\mathbf{v}^2}{2}-\rho\widetilde{\varepsilon}(\rho)-\frac{\mathbf{B}^2}{8\pi}\right)d\mathbf{r}.$$

Here $\widetilde{\varepsilon}(\rho)$ is the specific internal energy. The pressure can be expressed in terms of this energy as follows:

$$p(\rho)=\rho^2(\partial\widetilde{\varepsilon}/\partial\rho),$$

and the enthalpy can be expressed as

$$w(\rho)=\partial(\widetilde{\varepsilon}\rho)/\partial\rho.$$

The first term is the kinetic energy, the second term is the internal energy, and the last term is the magnetic energy. The Lagrangian $\mathscr{L}_*$ can be written in terms of the mapping $\mathbf{x}(\mathbf{a},t)$ as follows:[22]

$$\mathscr{L}_*=\int\frac{\dot{\mathbf{x}}^2}{2}d\mathbf{a}-\int\widetilde{\varepsilon}(J_{\mathbf{x}}^{-1}(\mathbf{a}))d\mathbf{a}-\frac{1}{8\pi}$$
$$\times\int\left(\frac{\mathbf{B}_0(\mathbf{a})\cdot\nabla_{\mathbf{a}}\mathbf{x}}{J_{\mathbf{x}}(\mathbf{a})}\right)^2J_{\mathbf{x}}(\mathbf{a})d\mathbf{a}.\tag{37}$$

Here a formula similar to Eq. (21) was used to represent the frozen-in magnetic field $\mathbf{B}(\mathbf{r},t)$ in terms of Lagrangian quantities. The generalized Euler equation for one-fluid models with a frozen-in magnetic field can be obtained similarly to Eq. (7) and has the form

$$(\partial_t+\mathbf{v}\cdot\nabla)\left(\frac{1}{\rho}\frac{\delta\mathscr{L}}{\delta\mathbf{v}}\right)=\left(\frac{1}{\rho}\left[\mathbf{B}\times\mathrm{curl}\frac{\delta\mathscr{L}}{\delta\mathbf{B}}\right]+\nabla\left(\frac{\delta\mathscr{L}}{\delta\rho}\right)\right.$$
$$\left.-\frac{1}{\rho}\left(\frac{\delta\mathscr{L}}{\delta\nu_m}\right)\nabla\nu_m\right),\tag{38}$$

which in this case gives

$$(\partial_t+\mathbf{v}\cdot\nabla)\mathbf{v}=-\nabla w(\rho)+\frac{1}{4\pi\rho}[\mathrm{curl}\,\mathbf{B}\times\mathbf{B}].\tag{39}$$

Examining the time derivative of $p_k(\partial x_k/\partial a_\mu)$, we obtain without any special difficulties the representation (36), and the equations of motion for the quantities φ and $\mathbf{S}$ appearing there are

$$\dot{\varphi}(\mathbf{a},t)=\left(\frac{\mathbf{p}}{\rho}\left(\frac{\delta\mathscr{H}}{\partial\mathbf{p}}\right)-\frac{\delta\mathscr{H}}{\partial\rho}\right)\Bigg|_{\mathbf{r}=\mathbf{x}(\mathbf{a},t)},\tag{40}$$

$$\dot{S}_\mu(\mathbf{a},t)=-\frac{\partial\mathbf{x}}{\partial a_\mu}\cdot\left(\frac{\delta\mathscr{H}}{\partial\mathbf{B}}\right)\Bigg|_{\mathbf{r}=\mathbf{x}(\mathbf{a},t)}.\tag{41}$$

Here it is assumed that the corresponding representation for the frozen-in magnetic field [a formula of the type (21)] and the representation (36) for the momentum field are substituted into these equations. Together with the equation for $\mathbf{x}(\mathbf{a})$

$$\dot{\mathbf{x}}(\mathbf{a},t)=\frac{1}{\rho}\left(\frac{\partial\mathscr{H}}{\partial\mathbf{p}}\right)\Bigg|_{\mathbf{r}=\mathbf{x}(\mathbf{a},t)}.\tag{42}$$

Eqs. (40) and (41) form a closed system that corresponds to a fixed flow topology determined by the fields $\mathbf{B}_0$ and $\mathbf{u}_0$. Here the Hamiltonian $\mathscr{H}\{\rho,\mathbf{p},\mathbf{B}\}$ written in the Eulerian representation is obtained from the Lagrangian $\mathscr{L}\{\rho,\mathbf{v},\mathbf{B}\}$ according to the standard formula (12), but it depends not only on the momentum and density but also on the magnetic field. Thus, the MHD Hamiltonian has the form

$$\mathscr{H}_*=\int\left(\rho\frac{\mathbf{p}^2}{2}+\rho\widetilde{\varepsilon}(\rho)+\frac{\mathbf{B}^2}{8\pi}\right)d\mathbf{r},$$

and the equations of motion (42), (41), and (40) are

$$\dot{x}_k(\mathbf{a})=p_k=\frac{\partial a_\mu}{\partial x_k}\left(u_{0\mu}+\frac{\partial\varphi}{\partial a_\mu}+B_{0\alpha}\left(\frac{\partial S_\alpha}{\partial a_\mu}-\frac{\partial S_\mu}{\partial a_\alpha}\right)\right),\tag{43}$$

$$-\dot{S}_\alpha(\mathbf{a})=\frac{1}{4\pi}\frac{\partial x_k(\mathbf{a})}{\partial a_\alpha}\frac{\partial x_k(\mathbf{a})}{\partial a_\mu}\frac{B_{0\mu}(\mathbf{a})}{J_{\mathbf{x}}(\mathbf{a})},\tag{44}$$

$$\dot{\varphi}(\mathbf{a})-\frac{p_k(\mathbf{a})p_k(\mathbf{a})}{2}+w(J_{\mathbf{x}}^{-1}(\mathbf{a}))=0.\tag{45}$$

If the transformation law for the components of the vector $\mathbf{S}$ for transforming to the Eulerian representation is given as

$$S_k(\mathbf{r}) = \frac{\partial a_\mu}{\partial x_k} S_\mu(\mathbf{a}),$$

then the following parameterization of the velocity field in MHD is easily obtained from Eq. (43):

$$\mathbf{v}(\mathbf{r}) = u_{0\mu}(\mathbf{a})\nabla a_\mu + \nabla\varphi(\mathbf{r}) + \frac{[\mathbf{B}(\mathbf{r})\times \operatorname{curl}\mathbf{S}(\mathbf{r})]}{\rho(\mathbf{r})}.$$

The presence of a term with $\mathbf{u}_0(\mathbf{a})$ ensures that the values of the topological cross invariants will be nontrivial. In the particular case $\mathbf{u}_0 = 0$ this formula is identical to the result obtained in Ref. 23, where the question of introducing canonical variables into MHD was considered and it was shown that (for $\mathbf{u}_0 = 0$) the fields $(\mathbf{B},\mathbf{S})$ and (ρ,φ) form canonically-conjugate pairs.

5. MHD-TYPE SYSTEMS IN THE INCOMPRESSIBLE LIMIT

We shall now go over to incompressible MHD fluids. First, for simplicity, we consider the case of finite α, so that there are two different solenoidal fields $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$, each being frozen into its own substance. The entire dynamics then occurs in the class of isorotational fields,[24] and from the isovorticity follows topological equivalence,[19] which can be expressed by formulas similar to Eq. (21):

$$\mathbf{\Omega}_1(\mathbf{r},t) = \int \delta(\mathbf{r}-\mathbf{R}_1(\mathbf{a},t))(\mathbf{\Omega}_1^0(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{R}_1(\mathbf{a},t)\,d\mathbf{a}, \tag{46}$$

$$\mathbf{\Omega}_2(\mathbf{r},t) = \int \delta(\mathbf{r}-\mathbf{R}_2(\mathbf{c},t))(\mathbf{\Omega}_2^0(\mathbf{c})\cdot\nabla_{\mathbf{c}})\mathbf{R}_2(\mathbf{c},t)\,d\mathbf{c}, \tag{47}$$

Here the mappings $R_1(\mathbf{a},t)$ and $\mathbf{R}_2(\mathbf{c},t)$, in contrast to $\mathbf{x}_1(\mathbf{a},t)$ and $\mathbf{x}_2(\mathbf{c},t)$, do not carry any information about density. Specifically, the corresponding Jacobians $J_1 = \det\|\partial\mathbf{R}_1/\partial\mathbf{a}\|$ and $J_2 = \det\|\partial\mathbf{R}_2/\partial\mathbf{c}\|$ need not necessarily be equal to unity.[12] Therefore $\mathbf{a}$ and $\mathbf{c}$ are no longer labels of fluid particles, but rather they are more formal quantities. All information about topological properties is contained in the time-independent solenoidal fields $\mathbf{\Omega}_1^0(\mathbf{a})$ and $\mathbf{\Omega}_2^0(\mathbf{c})$. This representation of the fields $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ possesses a gauge freedom, since the same field can be parameterized by different mappings $\mathbf{R}$. Indeed, it is possible to switch to coordinates $\nu_1(\mathbf{a})$, $\nu_2(\mathbf{a})$, and $\xi(\mathbf{a})$ in Eq. (46) such that this formula can be rewritten as

$$\mathbf{\Omega}_1(\mathbf{r},t) = \int_{\mathscr{N}_1} d^2\nu \int \delta(\mathbf{r}-\mathbf{R}_1(\nu,\xi,t))\frac{\partial\mathbf{R}_1(\nu,\xi,t)}{\partial\xi}d\xi \tag{48}$$

and similarly for $\mathbf{\Omega}_2$. Here $\nu\in\mathscr{N}_1$ is the label of a vortex line lying in a fixed two-dimensional manifold $\mathscr{N}_1$, and ξ a parameter along this line. This form of expression holds locally for an arbitrary frozen-in solenoidal field; it holds globally only if the flux lines are closed. The gauge freedom is connected with the possibility of choosing a parameter along the lines and with the identity of the vortex lines themselves, which makes it possible to rename the labels ν. Gauge transformations of the first type exist for any field topology. Transformations of the second type, however, exist only if $\mathbf{\Omega}_1^0(\mathbf{a})$ possesses globally determined vortex surfaces.

The variational derivatives with respect to the new variables are determined by the formulas

$$\frac{\delta\mathscr{H}}{\partial\mathbf{R}_1} = \left[(\mathbf{\Omega}_1^0\cdot\nabla_{\mathbf{a}})\mathbf{R}_1\times\operatorname{curl}_{\mathbf{r}}\left(\frac{\delta\mathscr{H}}{\partial\mathbf{\Omega}_1(\mathbf{R}_1)}\right)\right], \tag{49}$$

$$\frac{\delta\mathscr{H}}{\partial\mathbf{R}_2} = \left[(\mathbf{\Omega}_2^0\cdot\nabla_{\mathbf{c}})\mathbf{R}_2\times\operatorname{curl}_{\mathbf{r}}\left(\frac{\delta\mathscr{H}}{\partial\mathbf{\Omega}_2(\mathbf{R}_2)}\right)\right], \tag{50}$$

A characteristic property of the new variational derivatives is that they are transverse with respect to the direction of the corresponding vortex lines. It is easy to see that only transverse displacements $\delta\mathbf{R}_1$ and $\delta\mathbf{R}_2$ give rise to variations of the fields and, correspondingly, vary the Hamiltonian.

The equations of motion for the fields $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$, which are curls of the generalized momenta, follow from the noncanonical Poisson brackets

$$\{F,G\} = \int\left(\mathbf{\Omega}_1\cdot\left[\operatorname{curl}\frac{\delta F}{\delta\mathbf{\Omega}_1}\times\operatorname{curl}\frac{\delta G}{\delta\mathbf{\Omega}_1}\right] + \mathbf{\Omega}_2\cdot\left[\operatorname{curl}\frac{\delta F}{\delta\mathbf{\Omega}_2}\times\operatorname{curl}\frac{\delta G}{\delta\mathbf{\Omega}_2}\right]\right)d\mathbf{r}. \tag{51}$$

These equations are

$$\mathbf{\Omega}_{1,t} = \operatorname{curl}[\operatorname{curl}(\delta\mathscr{H}/\delta\mathbf{\Omega}_1)\times\mathbf{\Omega}_1],$$
$$\mathbf{\Omega}_{2,t} = \operatorname{curl}[\operatorname{curl}(\delta\mathscr{H}/\delta\mathbf{\Omega}_2)\times\mathbf{\Omega}_2]. \tag{52}$$

Substituting here the representations (46) and (47), performing a series of intermediate calculations (similar to Ref. 12), and taking account of Eqs. (49) and (50), the equations of motion for $\mathbf{R}_1(\mathbf{a},t)$ and $\mathbf{R}_2(\mathbf{c},t)$ can be written as

$$[(\mathbf{\Omega}_1^0(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{R}_1\times\mathbf{R}_{1,t}] = \delta\mathscr{H}/\delta\mathbf{R}_1,$$
$$[(\mathbf{\Omega}_2^0(\mathbf{c})\cdot\nabla_{\mathbf{c}})\mathbf{R}_2\times\mathbf{R}_{2,t}] = \delta\mathscr{H}/\delta\mathbf{R}_2. \tag{53}$$

These equations determine only the transverse velocities. Velocities along vortex lines can be chosen arbitrarily, which has no effect on the dynamics of the fields and is due to the longitudinal gauge freedom. These equations follow from a variational principle with the Lagrangian

$$L = (1/3)\int d\mathbf{a}([\mathbf{R}_{1,t}\times\mathbf{R}_1]\cdot(\mathbf{\Omega}_1^0(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{R}_1) + (1/3)\int d\mathbf{c}([\mathbf{R}_{2,t}\times\mathbf{R}_2](\mathbf{\Omega}_2^0(\mathbf{c})\cdot\nabla_{\mathbf{c}})\mathbf{R}_2) - \mathscr{H}\{\mathbf{\Omega}_1\{\mathbf{R}_1\},\mathbf{\Omega}_2\{\mathbf{R}_2\}\}. \tag{54}$$

We underscore that in constructing the equations of motion by varying the expression (54) arbitrary variations $\delta\mathbf{R}_1$ and $\delta\mathbf{R}_2$, including variations that do not conserve the total volumes enclosed by some vortex surfaces, are admissible. It is easy to verify that the conservation laws for all such quantities, which should hold because the incompressibility of both liquids, in reality are integrals of motion of the system (53) and follow, according to Noether's theorem, from the symmetry of the Lagrangian with respect to the relabeling of the vortex lines. Therefore, for $\alpha\neq 0$ the situation is completely similar to that occurring in ordinary one-fluid hydrodynamics in the description of incompressible fluids by

frozen-in vortex lines.[12,15] The entire difference consists in the presence of two frozen-in fields instead of one. We write out here the specific two-fluid Hamiltonian, which on substitution into Eq. (54) corresponds in the limit $\alpha\to 0$ to ordinary incompressible MHD:

$$\mathscr{H}^{\alpha}_{\mathrm{MHD}}=\frac{\alpha^2}{8\pi}\int\frac{((\mathbf{\Omega}_2^0(\mathbf{c})\cdot\nabla_{\mathbf{c}})\mathbf{R}_2(\mathbf{c}))^2}{\det\|\partial\mathbf{R}_2/\partial\mathbf{c}\|}d\mathbf{c}+\frac{1}{8\pi}\times\int\int\frac{((\mathbf{\Omega}_2^0(\mathbf{c}_1)\cdot\nabla_1)\mathbf{R}_2(\mathbf{c}_1)\cdot(\mathbf{\Omega}_2^0(\mathbf{c}_2)\cdot\nabla_2)\mathbf{R}_2(\mathbf{c}_2))}{|\mathbf{R}_2(\mathbf{c}_1)-\mathbf{R}_2(\mathbf{c}_2)|}\times d\mathbf{c}_1d\mathbf{c}_2+\frac{1}{4\pi}\times\int\int\frac{((\mathbf{\Omega}_1^0(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{R}_1(\mathbf{a})\cdot(\mathbf{\Omega}_2^0(\mathbf{c})\cdot\nabla_{\mathbf{c}})\mathbf{R}_2(\mathbf{c}))}{|\mathbf{R}_1(\mathbf{a})-\mathbf{R}_2(\mathbf{c})|}\times d\mathbf{a}d\mathbf{c}+\frac{1}{8\pi}\times\int\int\frac{((\mathbf{\Omega}_1^0(\mathbf{a}_1)\cdot\nabla_1)\mathbf{R}_1(\mathbf{a}_1)\cdot(\mathbf{\Omega}_1^0(\mathbf{a}_2)\cdot\nabla_2)\mathbf{R}_1(\mathbf{a}_2))}{|\mathbf{R}_1(\mathbf{a}_1)-\mathbf{R}_1(\mathbf{a}_2)|}\times d\mathbf{a}_1d\mathbf{a}_2. \tag{55}$$

The first term in this expression is the magnetic energy

$$\mathscr{M}=\frac{1}{8\pi}\int\mathbf{B}^2d\mathbf{r},$$

and the last three terms are the kinetic energy of the incompressible liquid

$$\mathscr{H}=-\frac{1}{2}\int\mathbf{\Omega}\cdot\Delta^{-1}\mathbf{\Omega}d\mathbf{r}.$$

If $\mathbf{\Omega}_2^0(\mathbf{c})=0$, we have here the Hamiltonian of ordinary hydrodynamics.

The limit $\alpha\to 0$ can be taken in the representation of the fields, in the equations of motion, and in the Lagrangian similarly to the manner in which Eq. (36) was obtained. As a result, we obtain the following parameterization of the fields $\mathbf{\Omega}=\mathbf{\Omega}_1+\mathbf{\Omega}_2$ and $\mathbf{B}$:

$$\mathbf{\Omega}(\mathbf{r},t)=\int\delta(\mathbf{r}-\mathbf{R}(\mathbf{a},t))((\mathbf{\Omega}_0(\mathbf{a})+\mathrm{curl}_{\mathbf{a}}[\mathbf{B}_0(\mathbf{a})\times\mathbf{U}(\mathbf{a},t)])\cdot\nabla_{\mathbf{a}})\mathbf{R}(\mathbf{a},t)d\mathbf{a}, \tag{56}$$

$$\mathbf{B}(\mathbf{r},t)=\int\delta(\mathbf{r}-\mathbf{R}(\mathbf{a},t))(\mathbf{B}_0(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{R}(\mathbf{a},t)d\mathbf{a}. \tag{57}$$

Here the field $\mathbf{U}(\mathbf{a},t)$ is not necessarily solenoidal for the reason mentioned above that the Jacobians of the mappings $\mathbf{R}_1$ and $\mathbf{R}_2$, which appear in Eqs. (46) and (47), are arbitrary. The relation between the variational derivative $\delta\mathscr{H}/\delta\mathbf{U}$ and the old variational derivatives with respect to $\mathbf{\Omega}$ and $\mathbf{B}$ is given by the formula

$$-\frac{\delta\mathscr{H}}{\delta U_\mu}=\frac{\partial\mathbf{R}}{\partial a_\mu}\left[(\mathbf{B}_0\cdot\nabla_{\mathbf{a}})\mathbf{R}\times\mathrm{curl}_{\mathbf{r}}\frac{\delta\mathscr{H}}{\delta\mathbf{\Omega}(\mathbf{R})}\right]. \tag{58}$$

The analogous formula for $\delta\mathscr{H}/\delta\mathbf{R}$ is

$$\frac{\delta\mathscr{H}}{\delta\mathbf{R}}=\left[(\mathbf{B}_0\cdot\nabla_{\mathbf{a}})\mathbf{R}\times\mathrm{curl}_{\mathbf{r}}\frac{\delta\mathscr{H}}{\delta\mathbf{B}(\mathbf{R})}\right]+\left[((\mathbf{\Omega}_0+\mathrm{curl}_{\mathbf{a}}[\mathbf{B}_0\times\mathbf{U}])\cdot\nabla_{\mathbf{a}})\mathbf{R}\times\mathrm{curl}_{\mathbf{r}}\left(\frac{\delta\mathscr{H}}{\delta\mathbf{\Omega}(\mathbf{R})}\right)\right]. \tag{59}$$

The Poisson brackets for the incompressible MHD systems can be obtained from Eq. (51) by making the substitutions $\mathbf{\Omega}_1=\mathbf{\Omega}+\alpha^{-1}\mathbf{B}$ and $\mathbf{\Omega}_2=-\alpha^{-1}\mathbf{B}$ and taking the limit $\alpha\to 0$. It has the form[17]

$$\{F,G\}=\int\left(\mathbf{\Omega}\cdot\left[\mathrm{curl}\frac{\delta F}{\delta\mathbf{\Omega}}\times\mathrm{curl}\frac{\delta G}{\delta\mathbf{\Omega}}\right]\right)d\mathbf{r}+\int\left(\mathbf{B}\cdot\left(\left[\mathrm{curl}\frac{\delta F}{\delta\mathbf{B}}\times\mathrm{curl}\frac{\delta G}{\delta\mathbf{\Omega}}\right]-\left[\mathrm{curl}\frac{\delta G}{\delta\mathbf{B}}\times\mathrm{curl}\frac{\delta F}{\delta\mathbf{\Omega}}\right]\right)\right)d\mathbf{r}. \tag{60}$$

Substituting the relations (56) and (57) into the equations of motion for $\mathbf{\Omega}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$

$$\mathbf{\Omega}_t=\mathrm{curl}[\mathrm{curl}(\delta\mathscr{H}/\delta\mathbf{\Omega})\times\mathbf{\Omega}]+\mathrm{curl}[\mathrm{curl}(\delta\mathscr{H}/\delta\mathbf{B})\times\mathbf{B}], \tag{61}$$

$$\mathbf{B}_t=\mathrm{curl}[\mathrm{curl}(\delta\mathscr{H}/\delta\mathbf{\Omega})\times\mathbf{B}], \tag{62}$$

which follow from Eq. (60), with allowance for Eqs. (58) and (59), gives the equations of motion for $\mathbf{U}$ and $\mathbf{R}$

$$[(\mathbf{B}_0\cdot\nabla_{\mathbf{a}})\mathbf{R}\times\mathbf{R}_t](\partial\mathbf{R}/\partial a_\lambda)=-\delta\mathscr{H}/\delta U_\lambda, \tag{63}$$

$$[((\mathbf{\Omega}_0+\mathrm{curl}_{\mathbf{a}}[\mathbf{B}_0\times\mathbf{U}])\cdot\nabla_{\mathbf{a}})\mathbf{R}\times\mathbf{R}_t]-[(\mathbf{B}_0\cdot\nabla_{\mathbf{a}})\mathbf{R}\times(\mathbf{U}_t\cdot\nabla_{\mathbf{a}})\mathbf{R}]=\delta\mathscr{H}/\delta\mathbf{R}. \tag{64}$$

It is easily verified that the limiting Lagrangian is

$$L=\int d\mathbf{a}([(\mathbf{B}_0\cdot\nabla_{\mathbf{a}})\mathbf{R}\times(\mathbf{U}\cdot\nabla_{\mathbf{a}})\mathbf{R}]\cdot\mathbf{R}_t)+(1/3)\times\int d\mathbf{a}([\mathbf{R}_t\times\mathbf{R}]\cdot(\mathbf{\Omega}_0\cdot\nabla_{\mathbf{a}})\mathbf{R})-\mathscr{H}\{\mathbf{\Omega}\{\mathbf{R},\mathbf{U}\},\mathbf{B}\{\mathbf{R}\}\}, \tag{65}$$

and the Hamiltonian $\mathscr{H}_{\mathrm{MHD}}$ of incompressible MHD is

$$\mathscr{H}_{\mathrm{MHD}}=\frac{1}{8\pi}\int\frac{((\mathbf{B}_0(\mathbf{a})\cdot\nabla_{\mathbf{a}})\mathbf{R}(\mathbf{a}))^2}{\det\|\partial\mathbf{R}/\partial\mathbf{a}\|}d\mathbf{a}+\frac{1}{8\pi}\times\int\int\frac{((\mathbf{\Omega}(\mathbf{a}_1)\cdot\nabla_1)\mathbf{R}(\mathbf{a}_1)\cdot(\mathbf{\Omega}(\mathbf{a}_2)\cdot\nabla_2)\mathbf{R}(\mathbf{a}_2))}{|\mathbf{R}(\mathbf{a}_1)-\mathbf{R}(\mathbf{a}_2)|}\times d\mathbf{a}_1d\mathbf{a}_2, \tag{66}$$

where, for brevity, we have introduced the notation

$$\mathbf{\Omega}(\mathbf{a})=\mathbf{\Omega}_0(\mathbf{a})+\mathrm{curl}_{\mathbf{a}}[\mathbf{B}_0(\mathbf{a})\times\mathbf{U}(\mathbf{a},t)].$$

Thus, we have a variational principle for MHD-type Hamiltonian dynamics of two solenoidal vector fields whose topological properties are fixed by $\mathbf{\Omega}_0(\mathbf{a})$ and $\mathbf{B}_0(\mathbf{a})$. We underscore the difference between the system of equations (40), (41), and (42), which can describe an incompressible

liquid only as a limiting and technically ''inconvenient'' case, and Eqs. (63) and (64), in whose derivation incompressibility was assumed at the outset and actively used. As result, all dynamical quantities in them are free of couplings. The result of solving the system (40), (41), and (42) is the real motion $\mathbf{x}(\mathbf{a},t)$ of fluid particles with conservation of the volumes, while by solving the system (63) and (64) it is possible to determine the fields $\mathbf{B}(\mathbf{r},t)$ and $\mathbf{\Omega}(\mathbf{r},t)$ from Eqs. (57) and (56) and even the motion of frozen-in magnetic flux lines, but not the real particle distribution along these lines. In other words the Lagrangian (65) makes it possible to obtain more information about the flow than the Eulerian representation provides, but less information than the complete Lagrangian description.

We note that the equations of motion have a gauge-invariant form, and they can be solved for the time derivatives only because of the gauge invariance of the Hamiltonian. The two arbitrary functions arising in the process can be chosen on the basis of convenience. For example, given the form of the magnetic flux lines in various coordinate systems, they can be parameterized by one of the three coordinates, while the component of the vector $\mathbf{U}$ along $\mathbf{B}_0$ can be set to zero.

We now discuss the question of the Casimir functionals of the Poisson brackets (60). These brackets are the $\alpha\to 0$ limit of the two-fluid brackets (51), where $\mathbf{\Omega}_1=\mathbf{\Omega}+\alpha^{-1}\mathbf{B}$ and $\mathbf{\Omega}_2=-\alpha^{-1}\mathbf{B}$. By definition the Casimir functionals $C\{\mathbf{\Omega}_1,\mathbf{\Omega}_2\}$ of the bracket (51) satisfy the equations

$$\mathrm{curl}[\,\mathrm{curl}(\,\delta C/\mathbf{\Omega}_1)\times\mathbf{\Omega}_1]=0,$$

$$\mathrm{curl}[\,\mathrm{curl}(\,\delta C/\delta\mathbf{\Omega}_2)\times\mathbf{\Omega}_2]=0. \tag{67}$$

The geometric meaning of the conditions (67) is that the Casimir functionals C are constants on a set of isorotational fields:

$$\delta C=0 \quad \text{if} \quad \delta\mathbf{\Omega}_1=\mathrm{curl}[\,\delta\mathbf{x}_1\times\mathbf{\Omega}_1],$$

$$\delta\mathbf{\Omega}_2=\mathrm{curl}[\,\delta\mathbf{x}_2\times\mathbf{\Omega}_2],$$

and the small displacements $\delta\mathbf{x}_1$ and $\delta\mathbf{x}_2$ preserve volumes

$$(\nabla\cdot\delta\mathbf{x}_1)=0 \quad \text{and} \quad (\nabla\cdot\delta\mathbf{x}_2)=0.$$

From the equalities (67) it follows that

$$[\,\mathrm{curl}(\,\delta C/\delta\mathbf{\Omega}_1)\times\mathbf{\Omega}_1]=\nabla\Psi_C^{(1)},$$

$$[\,\mathrm{curl}(\,\delta C/\delta\mathbf{\Omega}_2)\times\mathbf{\Omega}_2]=\nabla\Psi_C^{(2)},$$

where $\Psi_C^{(1)}$ and $\Psi_C^{(2)}$ are scalar functions, and their constant-level sets are vortex surfaces, as is obvious from these formulas. It makes sense to make a special effort to single out the Casimir functionals C^t, for which $\Psi_{C^t}^{(1)}\equiv 0$ and $\Psi_{C^t}^{(2)}\equiv 0$. Such functionals remain unchanged even when the displacements $\delta\mathbf{x}_1$ and $\delta\mathbf{x}_2$ do not preserve volumes. Therefore it is natural to identify C^t with the topological invariants of the fields $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$. The parameterization (46) and (47) fixes C^t, since the Jacobians of the mappings $\mathbf{R}_1$ and $\mathbf{R}_2$ are not necessarily 1. All other Casimir functionals of the brackets (51), which are related to the preservation of the volumes of the vortex tubes, are integrals of motion of the dynamical system (53). Therefore the labeling symmetry of the fluids is not completely exhausted, in general, by switching to a description of the flows by means of $\mathbf{R}_1$ and $\mathbf{R}_2$ — symmetry with respect to relabeling of the vortex lines remains in the Lagrangian (54). Nonetheless, it is necessary to take into account that the presence of globally determined vortex surfaces in the fields $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ and especially the closure of the vortex lines are an additional symmetry. Fields with such properties are atypical. For this reason, flows of general form in which the vortex lines are entangled in a complicated manner are described completely adequately even by Eqs. (46) and (47) and hardly require any different representation.

In the limit $\alpha\to 0$ we have the parameterization (56) and (57), which fixes all topological Casimir functionals C^t, which in this case degenerate into topological invariants of the magnetic field and cross invariants. Both satisfy the conditions

$$[\,\mathrm{curl}(\,\delta C^t/\delta\mathbf{\Omega})\times\mathbf{\Omega}]+[\,\mathrm{curl}(\,\delta C^t/\delta\mathbf{B})\times\mathbf{B}]=0,$$

$$[\,\mathrm{curl}(\,\delta C^t/\delta\mathbf{B})\times\mathbf{\Omega}]=0.$$

All other Casimir functionals of the brackets (60), for example, the volumes of the magnetic tubes in the presence of global magnetic surfaces, are dynamical conservation laws of the system (63) and (64).

In concluding this section we point out that the Poisson brackets (60) can be directly transformed to new variables by means of Eqs. (58) and (59). Introducing the notation

$$\mathbf{h}=(\mathbf{B}_0.\nabla_{\mathbf{a}})\mathbf{R}, \quad \mathbf{w}=((\mathbf{\Omega}_0+\mathrm{curl}_{\mathbf{a}}[\mathbf{B}_0\times\mathbf{U}])\cdot\nabla_{\mathbf{a}})\mathbf{R},$$

$$\frac{\delta^*F}{\delta\mathbf{U}}=\frac{\delta F}{\delta U_\mu}\frac{\partial a_\mu}{\partial\mathbf{R}}$$

they can be written as

$$\{F,G\}=\int\left(\mathbf{h}\cdot\left[\frac{\delta^*G}{\delta\mathbf{U}}\times\frac{\delta^*F}{\delta\mathbf{U}}\right]\right)\frac{(\mathbf{h}\cdot\mathbf{w})}{|\mathbf{h}|^4}d\mathbf{a}$$
$$-\int\left(\frac{\mathbf{h}}{|\mathbf{h}|^2}\cdot\left(\left[\frac{\delta F}{\delta\mathbf{R}}\times\frac{\delta^*G}{\delta\mathbf{U}}\right]-\left[\frac{\delta G}{\delta\mathbf{R}}\times\frac{\delta^*F}{\delta\mathbf{U}}\right]\right)\right)d\mathbf{a}. \tag{68}$$

It is important that the gauge invariance of the functionals be used in the calculations. This gives the identities

$$\left(\mathbf{h}\cdot\frac{\delta^*F}{\delta\mathbf{U}}\right)=0, \quad \left(\mathbf{w}\cdot\frac{\delta^*F}{\delta\mathbf{U}}\right)-\left(\mathbf{h}\cdot\frac{\delta F}{\delta\mathbf{R}}\right)=0.$$

We note that variational derivatives with respect to $\mathbf{\Omega}_0$ and $\mathbf{B}_0$ were not required to transform the brackets directly to new variables. This confirms that these fields carry information about the Casimir functionals.

6. TWO-DIMENSIONAL INCOMPRESSIBLE MAGNETOHYDRODYNAMICS

In two-dimensional incompressible MHD the magnetic flux line formulism can be put into a simpler form. A magnetic field in a plane is given by a single scalar function $A(x,y,t)$ (the z component of the vector potential), whose contour lines are the magnetic lines and are transported by the flow. The (a_1,a_2) plane divides into regions separated

by separatrices — the contour lines of the function $A(a_1,a_2)$ that pass through its saddle points. Curvilinear coordinates (ξ,A) can be introduced in each such region. The coordinate ξ parameterizes the magnetic flux lines and admits a gauge invariance. The vortex field $\mathbf{\Omega}$ and the vector $\mathbf{U}$ also have a single nonzero z component. The parameterization of the fields for two-dimensional flows can be rewritten as

$$\Omega(\mathbf{r},t)=\int \delta(\mathbf{r}-\mathbf{R}(\xi,A,t))(\Omega_0(\xi,A)-U_{,\xi})d\xi dA, \tag{69}$$

$$\mathbf{B}(\mathbf{r},t)=\int \delta(\mathbf{r}-\mathbf{R}(\xi,A,t))\mathbf{R}_{,\xi}\,d\xi dA, \tag{70}$$

where $\mathbf{R}=(x,y)$. The equations of motion for U and $\mathbf{R}$ are

$$x_{,\xi}y_t-x_ty_{,\xi}=-\delta\mathscr{H}/\delta U, \tag{71}$$

$$(\Omega_0(\xi,A)-U_{,\xi})y_t+U_ty_{,\xi}=-\delta\mathscr{H}/\delta x, \tag{72}$$

$$(\Omega_0(\xi,A)-U_{,\xi})x_t+U_tx_{,\xi}=-\delta\mathscr{H}/\delta y. \tag{73}$$

It is convenient to introduce a dynamical variable $\Phi(\xi,A,t)$ defined to within an arbitrary additive function $f(A)$ as

$$\Phi_{,\xi}=\Omega_0(\xi,A)-U_{,\xi}. \tag{74}$$

In general, Φ is a multivalued function. The change in Φ around a closed magnetic flux line is determined by the integral

$$\Delta(A)=\oint \Omega_0(\xi,A)d\xi.$$

This change depends on the label of the line A, but it does not depend on the time and is therefore a topological integral of motion. It is easy to see that the conserved quantities $\Delta(A)$ are closely related to the cross invariants — Casimir functionals of the Poisson brackets (60), which in the two-dimensional case can be written as

$$\mathscr{C}_F=\int \Omega F(A)dxdy.$$

It is easy to see that the Lagrangian for the dynamical system (71)–(73) can be written in the form

$$\mathscr{L}_{MHD2D}=\int (x_{,\xi}y_t-x_ty_{,\xi})\Phi d\xi dA-\mathscr{H}\{x,y,\Phi\}. \tag{75}$$

We see here that the kinematic part of the Lagrangian of two-dimensional MHD (for an individual magnetic flux line) is identical to that for the dynamics of the free surface of an ideal liquid in two-dimensional potential flows.[25] Preservation of the areas bounded by each magnetic flux line follows from the symmetry of the Hamiltonian with respect to the group of transformations

$$\Phi(\xi,A)\to\Phi(\xi,A)+\tau f(A).$$

Locally, one of the Cartesian coordinates can be taken as the parameter ξ. Let the function $y(x,A,t)$ give the form of the magnetic flux lines. The variables y and Φ in this case are canonically conjugate quantities that satisfy the equations of motion

$$y_t=\delta\mathscr{H}/\delta\Phi, \quad \Phi_t=-\delta\mathscr{H}/\delta y.$$

As a very simple example, we write out the Hamiltonian for nonlinear Alfvén waves against the background due to a uniform magnetic field $\mathbf{B}_0=(1,0)$:

$$\mathscr{H}_{A2D}=\frac{1}{8\pi}\int\left(\frac{1+y_{,x}^2}{y_{,A}}\right)dxdA - \frac{1}{8\pi}\int\int \Phi_{,x_1}(x_1,A_1)\Phi_{,x_2}(x_2,A_2) \times\ln((x_1-x_2)^2+(y_1-y_2)^2)dx_1dA_1dx_2dA_2. \tag{76}$$

The solution $\Phi=0$, $y=A$ corresponds to a rest state with the field (1, 0). The part of this Hamiltonian that is quadratic in the small perturbations $\Phi(x,A,t)$ and $\eta(x,A,t)=y(x,A,t)-A$ has the form

$$\mathscr{H}^{(2)}_{A2D}=\int\frac{d^2\mathbf{k}}{(2\pi)^2}\left(\frac{k_1^2}{k_1^2+k_2^2}\frac{|\Phi_{\mathbf{k}}|^2}{2}+\frac{k_1^2+k_2^2}{8\pi}|\eta_{\mathbf{k}}|^2\right). \tag{77}$$

This expression gives the well-known dispersion relatio for low-amplitude Alfvén waves $\omega^2(\mathbf{k})=k_1^2/4\pi$.

7. CONCLUSIONS

Let us sum up the results. The main results of this work are contained, first, in Eq. (36), which is an extension of Weber's representation to the case where a magnetic field frozen into the liquid is present and, second, in Eqs. (56), (57), and (63)–(66), which describe incompressible MHD flows with fixed topology and establish a variational principle for the dynamics of new objects $\mathbf{U}$ and $\mathbf{R}$. It was also shown that incompressible flows can be described by means of frozen-in vortex fields for any number N of components in N-fluid hydrodynamic models [since Eqs. (46), (47), (53), and (54) admit an obvious generalization].

I wish to thank E. A. Kuznetsov for attention to this work and for helpful suggestions. This work was supported by the Russian Fund for Fundamental Research (Grant No. 97-01-00093), the Programs for State Support for Leading Science Schools (Grant No. 96-15-96093), and a Landau Scholarship grant.

*)E-mail: ruban@itp.ac.ru

[1] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics* (Pergamon Press, New York) [Russian original, Nauka, Moscow, 1988].
[2] R. Salmon, AIP Conf. Proc. **88**, 127 (1982).
[3] R. Salmon, Annu. Rev. Fluid Mech. **20**, 225 (1988).
[4] N. Padhye and P. J. Morrison, Fiz. Plazmy **22**, 960 (1996) [Plasma Phys. Rep. **22**, 869 (1996)].
[5] V. E. Zakharov and E. A. Kuznetsov, Usp. Fiz. Nauk **167**, 1137 (1997).
[6] V. I. Il'gisonis and V. P. Lakhin, Fiz. Plazmy **25**, 64 (1999) [Plasma Phys. Rep. **25**, 58 (1999)].
[7] H. Lamb, *Hydrodynamics*, 6th edition (Cambridge University Press, Cambridge, 1932) [Russian translation, Gostekhizdat, Moscow, 1947].
[8] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media* (Pergamon Press, New York) [Russian original, Nauka, Moscow, 1982].
[9] H. K. Moffatt, J. Fluid Mech. **35**, 117 (1969).
[10] K. Elsässer, Phys. Plasmas **1**, 3161 (1994).

[11] V. V. Yan'kov, Zh. Éksp. Teor. Fiz. **107**, 414 (1995) [JETP **80**, 219 (1995)].
[12] E. A. Kuznetsov and V. P. Ruban, JETP Lett. **67**, 1076 (1998).
[13] E. A. Kuznetsov and V. P. Ruban, Zh. Éksp. Teor. Fiz. **115**, 894 (1999) [JETP **88**, 492 (1999)].
[14] E. A. Kuznetsov and V. P. Ruban, in *MHD Waves and Turbulence, Lecture Notes in Physics*, edited by T. Passot and P.-L. Sulem (Springer-Verlag, New York, 1999).
[15] V. Berdichevsky, Phys. Rev. E **57**, 2885 (1998).
[16] V. I. Karpman, *Nonlinear Waves in Dispersive Media* (Nauka, Moscow, 1973).
[17] P. J. Morrison and J. M. Greene, Phys. Rev. Lett. **45**, 790 (1980).
[18] B. A. Dubrovin, S. P. Novikov, and A. T. Fomenko, *Modern Geometry* (Nauka, Moscow, 1979).
[19] M. I. Monastyrskiĭ and P. V. Sasorov, Zh. Éksp. Teor. Fiz. **93**, 1210 (1987) [Sov. Phys. JETP **66**, 683 (1987)].
[20] S. S. Moiseev, R. Z. Sagdeev, A. V. Tur, and V. V. Yanovskiĭ, Zh. Éksp. Teor. Fiz. **83**, 215 (1982) [Sov. Phys. JETP **56**, 117 (1982)].
[21] B. N. Kuvshinov, Fiz. Plazmy **22**, 971 (1996) [Plasma Phys. Rep. **22**, 878 (1996)].
[22] V. I. Il'gisonis and V. P. Pastukhov, Fiz. Plazmy **22**, 228 (1996) [Plasma Phys. Rep. **22**, 208 (1996)].
[23] V. E. Zakharov and E. A. Kuznetsov, Dokl. Akad. Nauk SSSR **194**, 1288 (1970) [Sov. Phys. Dokl. **15**, 913 (1970)].
[24] V. I. Arnol'd, *Mathematical Methods of Classical Mechanics*, 2nd edition (Springer-Verlag, New York, 1989) [Russian original, Nauka, Moscow, 1974].
[25] A. I. D'yachenko, V. E. Zakharov, and E. A. Kuznetsov, Fiz. Plazmy **22**, 916 (1996) [Plasma Phys. Rep. **22**, 829 (1996)].

Translated by M. E. Alferieff
Edited by S. J. Amoretty

Collective defect formation in the high-temperature superconductor $YBa_2Cu_3O_7$ under the influence of adsorbed water molecules

B. M. Gorelov*)

Institute of Surface Chemistry, National Academy of Sciences of Ukraine, 252022 Kiev, Ukraine
(Submitted 4 November 1998)
Zh. Éksp. Teor. Fiz. **116**, 586–603 (August 1999)

The mechanisms for defect formation stimulated by the adsorption of water molecules in the surface of $YBa_2Cu_3O_7$ ceramic are studied, together with the types of defects and their distributions. It is found that a water layer physically bound to the surface reduces the rates of annihilation and capture of positrons, changes the amount of barium and copper on the surface by a factor of two, and inhibits diffusive jumps of nickel atoms. A layer of adsorbed water excites subthreshold formation of 10^{21} cm^{-3} interstitial Ba and Cu1 atoms and transitions of oxygen from O1 to O5, and *vice versa* in the volume of crystallites, and the migration of defects and accumulation of Ba atoms in the surface layer, which block diffusive jumps of Ni within the volume of the crystals. These effects are related to the excitation of collective, low-frequency weakly damped motion of heavy holes in the crystal volume when defects are formed on the surface by physically adsorbed H_2O molecules, which is accompanied by Coulomb repulsion of cations from intermediate layers into interstitials and the migration of defects in the field of the collective excitations. © *1999 American Institute of Physics.*
[S1063-7761(99)01608-X]

1. INTRODUCTION

The existence of low-frequency, collective excitations of heavy (h) carriers, the so called acoustic plasmons, is a basic condition for the plasmon mechanism of high-temperature superconductivity.[1–6] The distinctive features of acoustic plasmons are their phonon vibration frequencies and the possibility of their dissipationless propagation through a crystal without Landau damping on light (l) and heavy carriers. The Landau damping is weakened if the phase velocity of the collective excitations of the h-carriers is such that

$$\mathbf{u}=\Omega_q/\mathbf{q}\gg\mathbf{v}_{Fh},\mathbf{v}_{Fl} \tag{1}$$

or

$$\mathbf{v}_{Fh}\ll\Omega_q/\mathbf{q}\ll\mathbf{v}_{Fl}, \tag{2}$$

where v_{Fh} and v_{Fl} are the Fermi velocities of the h- and l-carriers, $\mathbf{q}$ is the wave number, $\Omega=(4\pi n_h e^2/m_h^*)^{1/2}$ is the plasma frequency, and n_h and m_h^* are the concentration and effective mass of the h-carriers. If the inequality is not satisfied under normal conditions, then it is possible to amplify and excite acoustic plasmons which move without dissipation through a crystal, creating perturbations in the charge density, especially by introducing defects. In that case, weakly damped acoustic plasmons result from the interaction of electrons with perturbations in the charge density $\delta\rho_i$ in the neighborhood of the defects (i is the number of defects), whose Hamiltonian is

$$\delta H=\sum_i \delta\rho_i(\mathbf{q})V_{il}/\varepsilon(\mathbf{q},\omega)$$

and the matrix element[6] has the form

$$g_{pl}(\mathbf{k},\mathbf{q})=\int d\mathbf{r}\sum_{l,i}\Psi^*_{\mathbf{k}-\mathbf{q},l}(\mathbf{r})\frac{\mathbf{q}^2\varphi_{\mathbf{q}}}{4\pi e\varepsilon(\mathbf{q},\omega)}V_{il}\Psi_{\mathbf{k},l}(\mathbf{r}), \tag{3}$$

where $\Psi^*_{\mathbf{k}-\mathbf{q},l}$ and $\Psi_{\mathbf{k},l}=u_n(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}$ are the Bloch wave functions of an electron in the lth band (n is the number of the site), $\varphi_{\mathbf{q}}$ is the Fourier transform of the potential, V_{il} is the Coulomb part of the electron–plasmon interaction, $\varepsilon(\mathbf{q},\omega)$ is the permittivity, and ω is the frequency. Here the excitation of weakly damped acoustic plasmons can apparently occur with a small number of defects or local density perturbations. If so, then the formation of defects on the surface of high-temperature superconducting (HTSC) crystals allows weakly damped collective motion of the h-carriers to be excited throughout the crystal.

A weakly damped displacement of collective excitations of degenerate h-carriers with a periodically modulated density can be accompanied by low-energy subthreshold defect formation if the carriers are strongly localized and the excitation frequencies are low when

$$\Omega_{\mathbf{q}}=\mathbf{qu}\leqslant\omega_D, \tag{4}$$

where ω_D is the Debye frequency. Since the time for an atom to escape a lattice site satisfies $\tau_d\geqslant\omega_d^{-1}$ during subthreshold defect formation,[7] when $\Omega_{\mathbf{q}}\leqslant\omega_D$ holds the localization time for the antinodes of the charge of the acoustic plasmons on the lattice ions is sufficient to push them into interstices and to form a defect. When the motion is collective, the entire subsystem of h-holes participates in the subthreshold defect formation, so that dissipationless propagation of acoustic plasmons can be accompanied by the formation of an anomalously large number n_d of defects, satisfying $n_d\gg N_d^i$ (N_d^i is the number of defects on the surface) and can approach n_h.

 © 1999 American Institute of Physics

In addition, dissipationless collective motion of h-carriers can be accompanied by capture of n_d generated and intrinsic defects by the electric fields of the acoustic plasmons and by mass transport of defects through the crystal to scattering sites, such as the surface, intercrystallite boundaries, and vacancy clusters. When acoustic plasmons scatter on the surface and when there is a large number of migrating defects, it is possible for them to accumulate in the surface region, the atomic composition of the surface can change, and internode diffusive jumps can be blocked.

Note that subthreshold defect formation requires that the screening of the Coulomb repulsion of the ions in the interstices of the lattice on the part of the l-carriers be weakened or entirely absent, so that subthreshold defect formation can take place in the layers of an HTSC lattice, where the l- and h-layers are spatially separated. In $YBa_2Cu_3O_7$ these layers are the intermediate Ba–O and Cu1–O layers, since[8–11] the l-carriers are localized in wide bands ($Z1$) formed by dp-orbitals of the Cu2, O2, and O3 atoms in the cuprate layers, and the h-layers, in a narrow band ($Z2$) formed by dp-orbitals of Cu1, O1, and O4 atoms of the intermediate layers and Cu2 of the CuO_2 layers. In addition, the band structure is characterized by a wide vacancy band ($Z3$) formed by dp-orbitals of Cu1, O1 and O4 atoms, and a narrow band with weak overlapping of p_x- and p_y-orbitals of O2 and O3 atoms ($Z4$ band) can also form.[12] However, in the cuprate layers, subthreshold defect formation is improbable because of screening of the Coulomb repulsion forces by the l-carriers.

Local perturbations in the electron density at the surface of HTSC crystals can be created with the aid of physically adsorbed water molecules. Thus, physically adsorbed H_2O dipoles pull atoms, especially Ba,[13,14] out of $YBa_2Cu_3O_7$ and form vacancies. The screening of the perturbations in the charge density $\delta\rho_i$ in the neighborhood of the surface defects can excite weakly damped acoustic plasmons throughout the volume of the $YBa_2Cu_3O_7$.

Note that when $YBa_2Cu_3O_7$ interacts with water, it degrades.[15–19] The degradation depends on the interface (HTSC-liquid, HTSC-vapor) and the density of the ceramics and impurities in the HTSC or in the medium. Thus, under normal conditions, ceramics with a density higher than $4.5\ g/cm^3$ and single crystals of $YBa_2Cu_3O_7$ interact equally weakly with water vapor (the effect of intergrain boundaries is insignificant, since[20] they occupy 2.1–4.2% of the area), while the less dense samples decompose, with faster degradation if carbon dioxide or halogen impurities are present.

In an atmosphere of water vapor, degradation and implantation of H_2O molecules into the lattice take place in an adsorption process following chemical binding of the molecules with atoms on the HTSC surface and probably are characterized by different rates, since decomposition is determined by the rate of reactive diffusion and implantation is determined by the rate of diffusion into the lattice. Both processes take place after formation of a physically bound polylayer of definite thickness on the surface.[21] Differences in the surface morphology of the samples affect the thickness of the layer of adsorbate, since for equal vapor pressures p and adsorption times t, the thickness satisfies

$$h(p,t)\simeq m(p,t)(\mu_s s)^{-1}\omega_{H_2O}N_A^{2/3}M^{-2/3}\rho^{-1/3} \qquad (5)$$

(where m, μ_s, and M are the masses of the adsorbed water, sample, and a gram mole of water, s is the specific surface area of the adsorbent, ω_{H_2O} is the landing area of an H_2O molecule, ρ is the density of water, and N_A is the Avogadro number) is less on HTSC samples with a developed surface s. However, by varying p or t, on samples with different areas s it is possible to obtain the same value of h and a similar interaction process with the H_2O molecules. In addition, by varying h, it is possible to resolve the excitation of acoustic plasmons by a physically bound layer of water prior to the onset of chemical adsorption and the implantation of H_2O molecules into the lattice, and by varying t, to resolve the combined effect of initially physically bound water and then of a different number of H_2O molecules implanted into the lattice.

Note that the defect formation on an HTSC surface by physically bound H_2O molecules is possible if the thickness of the polylayer permits formation of a solvation coat of molecules whose interaction with the atoms in the lattice is sufficient to pull them off into the polylayer. For finite dimensions of the coat, defect formation is characterized by a threshold dependence on the polylayer thickness. In addition, in the volume of dense polycrystalline samples, where adsorption does not occur or is small and the sizes of the pores are small for formation of a solvation coat, defect formation and collective motion of the h-carriers can be excited under the influence of the adsorbate in the crystallites at the surface. However, as they move without dissipation in the volume of the crystal, acoustic plasmons propagate from crystal to crystal within the sample volume and this may stimulate subthreshold defect formation and accumulation of defects on the surface of crystallites that are not in contact with water.

Therefore, physically adsorbed water molecules can stimulate low-frequency collective excitations of h-carriers, whose dissipationless motion causes subthreshold formation of a large number of defects in the intermediate Ba–O and Cu1–O layers, the migration of defects in the volume of crystals, and their transport to the surface. Defect formation is characterized by a threshold dependence on the water layer thickness, accumulation of defects at the surface of crystals, and changes in the atomic composition of the surface.

This paper is a study of defect formation stimulated by water molecules adsorbed on the surface of $YBa_2Cu_3O_7$, and of the spectrum of the defects and their distribution. Defect formation in intermediate layers of $YBa_2Cu_3O_7$, where positrons are annihilated,[22] and the defect migration are studied by positron temporal spectroscopy. X-ray structural analysis is used to measure the lattice parameters a, b, and c and the parameter $\eta=(C_{O1}-C_{O5})/(C_{O1}+C_{O5})$ (where C_{O1} and C_{O5} are the concentrations of oxygen atoms at O1 and O5 sites), which characterizes the filling of O1 and O5 sites by oxygen.[23–26] The accumulation of defects in the surface layer of crystals was determined using atomic Auger analysis and by measuring the diffusion coefficient of nickel atoms.

2. EXPERIMENTAL SAMPLES AND TECHNIQUES

The objects of study were x-ray single-phase polycrystalline samples of $YBa_2Cu_3O_7$ with a density of 5.5 g/cm^3 (porosity $\approx$14%) and an average grain size of 5–15 μm, obtained by solid phase synthesis form a mixture of ground powders of Y_2O_3, $BaCO_3$, and CuO. Synthesis was at $\approx$900 °C in air for $\approx$20 h, followed by grinding of the resulting tablets, mixing and compressing at 950 °C in air for $\approx$24 h, and a final annealing in oxygen with a smooth reduction in temperature at 500–420 °C over $\approx$64 h. Three polycrystalline blocks were obtained, from which three types of samples were prepared with different amounts of intrinsic defects. Dispersed samples with particles sizes of 17–20 μm and a specific surface of 1 m^2/g were obtained by dispersing the polycrystalline samples.

Nonstoichiometric compounds $YBa_2Cu_3O_{7-\delta}$ with $\delta>0$ were obtained by annealing samples with $\delta=0$ in vacuum. The amount of oxygen was determined using x-ray structural analysis and a Q-1500 oxygen thermodesorption drift indicator.

Water molecules were adsorbed on the $YBa_2Cu_3O_7$ surface at room temperature and vapor pressures $0<p<19$ Torr after annealing of the samples in a vacuum of 10^{-3} Torr at temperatures below 180 °C for $\approx$2.5 h. The amount of adsorbate was determined using MacBain balances with a sensitivity of 2×10^{-5} g. Thermogravimetric measurements and differential thermogravimetric analysis (DTA) were performed on the dispersed samples using the Q-1500 drift indicator with a heating rate of 5° C/min.

The samples were placed once in an water-vapor atmosphere during the measurements employing temporal positron spectroscopy, x-ray phase analysis, and tracer atoms and repeatedly when determining the atomic composition of the surface. After exposure was completed, the polylayer of adsorbed water was removed from the crystallite surface.

The positron annihilation spectra were measured at room temperature on an instrument from ORTEC whose half-maximum resolution function is 220 ps with a resolution of 1 ps. A ^{22}Na positron source with an activity of 20 μCi was used, mounted between two samples of size $10\times10\times1.5$ mm^3. The positron annihilation rate λ_f and capture rate ν were determined from the expressions[27]

$$\lambda_f=I_1/\tau_1+I_2/\tau_2, \quad \nu=I_2(1/\tau_1-1/\tau_2), \tag{6}$$

where τ_1, τ_2 and I_1, I_2 are the lifetimes of the positrons in the quasi-free and bound states and the corresponding intensities.

X-ray structural analysis was done on DRON-2 and ADP-1 diffractometers. Dispersed samples of $YBa_2Cu_3O_7$ were used for measuring the parameters a, b, c, and η. The parameter η was determined from the ratio of the structural amplitude of the 012 and 102 reflections.[23–26]

The atomic composition of the surface was determined, using an 09-IOS-10-991 Auger spectrometer with an electron beam diameter of $\approx$50 μm and a current of 10^{-6} A, from an average of 4–5 measurements of the composition at different points on the sample surface. Prior to the measurements, the samples were kept in a vacuum of 10^{-3} Torr for about 2 days. During the measurements, the pressure was below 10^{-9} Torr. No degradation was observed in the sample composition in the vacuum owing to loss of weakly bound oxygen[28] or the action of the electron beam.[29]

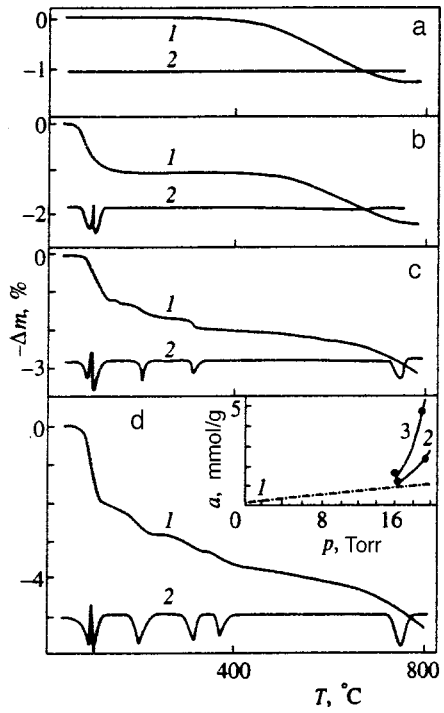

FIG. 1. Temperature variations in the mass loss according to thermogravimetric analysis (*1*) and the DTA signal (*2*) of samples of $YBa_2Cu_3O_7$ after adsorption of water: initial samples (a), 0.5 (b), 1.2 (c), and 2.2 mmol/g (d). The inset shows the adsorption isotherms ($T=21$ °C) for exposures of 15 (curve *1*), 120 (*2*) and 360 min (*3*).

The diffusion coefficient of nickel atoms was measured by the layer removal method[30] using radioactive tracers (^{63}Ni). The diffusion profiles developed in an air environment and were determined with a step size of 3–5 μm to a depth of 150–250 μm.[31]

3. EXPERIMENTAL RESULTS AND DISCUSSION

The formation of a physically bound polylayer and the evolution of the chemosorbed states of the H_2O molecules can be followed using data from a thermogravimetric analysis of the $YBa_2Cu_3O_7$ samples after adsorption of water at $p=18.7$ Torr (Fig. 1). After heating in vacuum to 180 °C and retention under normal conditions, the initial samples were characterized by a mass loss at $T\geqslant412$ K owing to desorption of oxygen O1 from the crystal lattice (Fig. 1a). For an exposure $t\leqslant120$ min and adsorption of $a<1$ mmol/g, the samples are characterized by a mass loss and a DTA signal implying absorption of heat at temperatures of 55–105 °C (Fig. 1b), which indicates the evaporation of water from the surface. The amount of absorbed heat, $Q\simeq25-30$ kJ/mol, is typical for desorption of physically bound water. Against a background of heat absorption, at $\approx$82 °C, there is a narrow exothermic DTA peak caused by the release of heat and indicating a phase transition within the polylayer that can be related to a transition from an orientation-ordered state of the H_2O dipoles to a disordered state in the near-surface water

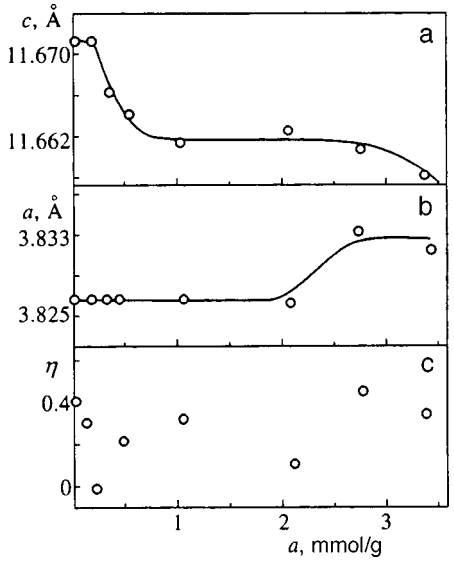

FIG. 2. The lattice parameters c (a), a (b), and η (c) as functions of the amount of adsorbed water.

layer during heating.[21,32] With increasing adsorption times $t>120$ min and amounts of adsorbed water up to $a \simeq 2$ mmol/g, in the thermogravimetric and DTA curves there are initially three and then four bound states of the H_2O molecule which correspond to a reduction in mass and the absorption of heat at temperatures $T_d \approx 208$, 330, 370, and 778 °C with heats of desorption $Q_d \simeq 38$, 99, 72, and 68 kJ/mol (Fig. 1c and 1d). The values of T_d and Q_d are typical of H_2O molecules bound by a hydrogen bond to atomic surroundings in a crystalline lattice.[33] The evolution of the polylayer and the bound states reflects the adsorption isotherm at $T=21$ °C (inset to Fig. 1d). For $p \leqslant 16$ Torr and $t \geqslant 15$ min, a increases with pressure and does not depend on exposure, i.e., a layer of physically bound water is formed on the surface and H_2O molecules are not implanted in the lattice. For $p>16$ Torr an increase in the absorption a with time t is observed owing to an increase in the amount of physically bound water and the number of molecules in the lattice.

Note that the presence of physically bound H_2O molecules on an $YBa_2Cu_3O_7$ surface causes a drop in the lattice parameter c with fixed a and random changes in η (Fig. 2). The parameter η varies over $0.02 \leqslant \eta \leqslant 0.4$, which indicates a migration of $\approx 1\times 10^{21}$ cm^{-3} oxygen atoms between O1 and O5 sites. Since the energy of O1↔O5 migration is $\approx 1.7-2.03$ eV,[34] while the collisional displacement energy of O1 and O5 atoms is at least 4.5–10 eV,[35,36] the mechanism for ejection of O1 and O5 atoms from lattice nodes is subthreshold and the heavy carriers which eject the O1 and O5 atoms are localized in intermediate layers and therefore occupy the narrow band $Z2$.

The reduction in c by physically adsorbed H_2O molecules can be related to compression of the lattice when the Coulomb repulsive force between Y, Ba, and Cu cations weakens during formation of cation defects and some of them escape. (The development of oxygen vacancies causes c to increase.)

When water molecules are implanted in the lattice, the drop in c is lessened and for $1.1<a<2.5$ mmol/g, c does not change. As the number of H_2O molecules in the lattice increases, when $a \geqslant 2.5$ mol/g holds, the volume of the lattice increases because of a sharp rise in a (Figs. 2a and 2b). The oscillations in η for $a>1.1$ mol/g are apparently caused by physically adsorbed water molecules (Fig. 2c).

Note that b does not change as a result of adsorption.

Thus, the effects of a physically bound water polylayer predominate at early times and for low levels of adsorption and it causes defects to develop, apparently in the cation sublattice, and leads to the migration of oxygen in the basis plane. As the adsorption time increases, this process is accompanied by the implantation of H_2O molecules in the lattice, where they form four nonequivalent states.

The lifetimes of positrons for different amounts of adsorbed water on three types of samples are listed in Table I. The samples of the first type were subjected predominantly

TABLE I. Annihilation parameters and characteristics of vacancy clusters before and after adsorption of water molecules

a, mmol/g	τ_1, ps	I_1, %	τ_2, ps	I_2, %	r_+, Å	N_+, 10^{-16} cm^{-3}
$0^{(1)}$	162±6	72±5	301±15	28±5	3.2	2.0
$0^{(2)}$	170±5	84±5	331±37	16±5	3.5	1.1
$0^{(3)}$	149±13	61±5	248±25	38±5	3.0	2.7
$0.08^{(1)}$	168±6	75±4	318±19	25±5	3.3	1.8
$0.31^{(1)}$	159±10	63±9	281±18	36±8	3.1	2.6
$0.34^{(2)}$	188±5	80±4	394±28	20±4	3.6	1.2
$0.45^{(1)}$	173±7	66±6	312±14	34±6	3.1	2.3
$0.61^{(1)}$	184±6	74±4	351±17	25±4	3.3	1.6
$0.63^{(2)}$	180±9	69±9	321±28	31±9	3.2	1.9
$0.83^{(1)}$	205±5	79±4	380±23	21±4	3.4	1.1
$1.62^{(2)}$	180±7	70±6	347±26	30±6	3.3	1.9
$1.83^{(3)}$	184±10	73±5	332±30	30±5	3.2	1.8
$2.71^{(3)}$	182±9	68±7	336±21	32±7	3.2	2.0
$2.78^{(3)}$	188±1	71±7	347±21	29±8	3.2	1.7
$2.84^{(3)}$	198±8	77±6	361±24	23±6	3.3	1.3
$2.92^{(3)}$	201±7	79±5	370±25	20±5	3.3	1.1

Note: The superscripts in the first column correspond to the three types of samples.

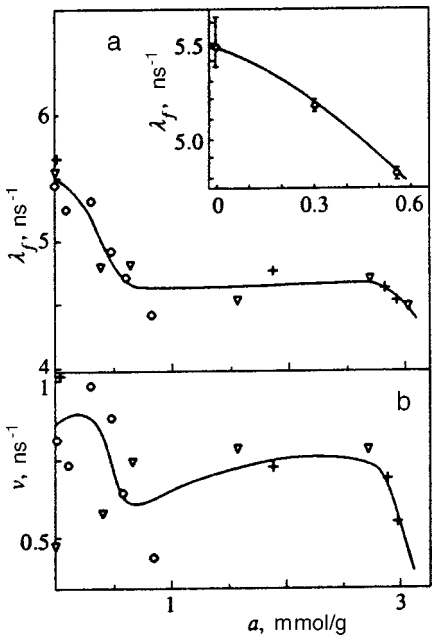

FIG. 3. The rates of annihilation (a) and capture (b) of positrons as functions of the amount of adsorbed water (the symbols ○, +, and ▽ correspond to samples of type 1, 2, and 3). The inset shows the rate of annihilation in $YBa_2Cu_3O_{7-\delta}$ as a function of the oxygen content.

to the action of physically adsorbed water molecules, while the second and third types were subject to the implantation of H_2O molecules into the crystal lattice, as well as to the water polylayer. The rates of positron annihilation and capture are plotted in Fig. 3 as functions of the amount of adsorbed water. Three regions of different behavior stand out in the $\lambda_f(a)$ curve: a region where λ_f drops smoothly, which corresponds to the effect of a physically adsorbed water layer; a region where λ_f is independent of a for $1.1<1<2.5$ mmol/g, in which H_2O molecules are implanted into the lattice after the polylayer has acted; and a region where λ_f decreases for $a>2.5$ mmol/g, which corresponds to an increase in the volume of the lattice as the number of molecules in the lattice increases. Here the reduction in λ_f owing to the physically bound H_2O dipoles correlates with the drop in the lattice parameter c, while implantation of H_2O molecules in the lattice for $a\leqslant 2.5$ mmol/g has no effect on λ_f and c (Figs. 2 and 3a).

The annihilation rate is related to the charge density of the electrons, $n^-(r)$ and positrons $n^+(r)$ by[37]

$$\lambda_f=\frac{\pi r_0^2 c}{e^2}\int d^3r n^+(\mathbf{r})n^-(\mathbf{r})\varepsilon[n^-(\mathbf{r})], \tag{7}$$

where

$$n^-(\mathbf{r})=e\sum_{i,l}\Psi^*_{i,l}(\mathbf{r})\Psi_{i,l}(\mathbf{r}),\quad n^+(\mathbf{r})=e\Psi^*_+(\mathbf{r})\Psi_+(\mathbf{r}),$$

$\Psi_{i,l}(\mathbf{r})$ and $\Psi_+(\mathbf{r})$ are the electron and positron wave functions, r_0 and e are the classical radius and charge of the electron, c is the speed of light, and $\varepsilon[n^-(\mathbf{r})]$ is the polarization factor. Since the positrons are annihilated in the intermediate layers,[22] whose electron density ρ includes the density ρ_{core} of the core electrons of Ba, Cu1, O1, O5, and O4 and of the electrons in the $Z2$ (ρ_{32}) and $Z3$ (ρ_{33}) bands, when the carriers are strongly localized in $Z2$ and the $Z3$ band is a vacancy band, we can write

$$\lambda_f=\pi r_0^2 c\varepsilon\rho\simeq\pi r_0^2 c\varepsilon(\rho_{\text{core}}+\rho_{32}+\rho_{33}), \tag{8}$$

where the polarization factor[38] ε is given by

$$\varepsilon(r_s)=1+0.1512r_s+2.414r_s^{3/2}-2.01r_s^2+0.4466r_s^{5/2}+0.1667r_s^3, \tag{9}$$

and $\varepsilon=2.3$ for $r_s=(3/4\pi\rho)^{1/3}=1$ ($\rho=1.6\times10^{24}\,\text{cm}^{-3}$).

The change in λ_f owing to a water polylayer for $a=0.9$ mmol/g is $\Delta\lambda_f=\lambda_f(0)-\lambda_f(a)=0.75\,\text{ns}^{-1}$, which corresponds to a density change $\Delta\rho=4.3\times10^{22}\,\text{cm}^{-3}$. Here $\Delta\rho$ is roughly an order of magnitude greater than the density of free carriers, which is given by[39] $\rho_{31}+\rho_{32}+\rho_{33}=5\times10^{21}\,\text{cm}^{-3}$ (ρ_{31} is the carrier density in the $Z1$ band), i.e., $\Delta\rho>\rho_{31}+\rho_{32}+\rho_{33}$ and $\Delta\rho>\rho_{32}+\rho_{33}$. Since $\rho_{32}\gg\rho_{33}$, we have $\Delta\rho\simeq\rho_{\text{core}}$ and

$$\Delta\lambda_f=\pi r_0^2 c\varepsilon\Delta\rho\simeq\pi r_0^2 c\varepsilon\Delta\rho_{\text{core}}. \tag{10}$$

Therefore, the reduction in λ_f is caused by the emergence of core electrons from the annihilation process during defect formation. In the Ba–O and Cu1–O layers, these defects are evidently Ba and Cu1 cations. When they fall into interstices, their positive charge repels a positron, reducing the overlap of $\Psi_{i,l}(\mathbf{r})$ and $\Psi_+(\mathbf{r})$; this prevents annihilation of the core electrons and reduces ρ.

Note that the formation of interstitial O4 and O1 defects is improbable. The emergence of atoms into the interstices can change the overlap of the wave functions $\Psi_{i,l}(\mathbf{r})$ and $\Psi_+(\mathbf{r})$ and the magnitude of λ_f, but, when the defect has a negative charge, the noninvolvement of the core electrons in annihilation becomes unlikely. When $Z3$ is a vacancy band, the redistribution of oxygen from O1 into O5 also apparently has no effect on λ_f, since the O1 atoms do not escape from the annihilation volume.

The appearance of Ba and Cu1 vacancies in the lattice is consistent with a reduction in c, since the loss of cations reduces the Coulomb repulsion among the Y, Ba, and Cu ions in the different layers and can make the layers come closer in the lattice.

The number of defects formed in the intermediate layers, $n_d\simeq\Delta\rho/n_e$, can be estimated assuming that the average number of core electrons of the interstitial atoms knocked out from annihilation is $n_e=45$ (Ba, O4, O1, Cu1 atoms). Then n_d equals $9.6\times10^{20}\,\text{cm}^{-3}$. Here n_d is reduced because of the large n_e, since the deep core electrons are not involved in annihilation, while the number of O1 and O4 atoms in the annihilation volume evidently does not change. However, given that the escape energy for Ba and Cu1 atoms from a lattice site is 20–25 eV,[35,36,40] the number n_d of defects formed in the volume of the crystals as a result of physically adsorbed water on the surface indicates a subthreshold mechanism and a collective defect formation process. Since subthreshold defect formation takes place in the Ba–O and Cu1–O layers and is caused by highly localized carriers, these carriers are h-carriers in the $Z2$ band.

Implantation of H_2O molecules in an $YBa_2Cu_3O_7$ lattice at $0.9 \leqslant a \leqslant 2.5$ mmol/g does not change λ_f, which may explain the absence of overlap in the wave functions of the electrons in the H_2O molecules and the positron when the molecules are localized in the interstices of the Ba–O and Cu1–O layers, where repulsion of a positron by the lattice cations and the protons of the molecule prevent annihilation, and in the layers of the lattice, to which positrons do not penetrate. The coordination bonding of the interstitial H_2O molecules with barium and oxygen ions does not affect λ_f.[41] The sharp drop in λ_f for $a \geqslant 2.5$ mmol/g is probably caused by distension of the lattice as the amount of H_2O in the volume of the $YBa_2Cu_3O_7$ increases and ρ decreases.

Three regions of different behavior can be identified in the $\nu(a)$ curve (Fig 3b). There is a region where ν changes rapidly because of the layer of physically adsorbed water for $a < 0.9$ mmol/g (the $\nu(a)$ curve is drawn through the points for a sample of the first type), where the behavior of ν reflects changes in the concentration N_+ and radius r_+ of the intrinsic defects in the samples that capture the positrons which serve as probes for subthreshold processes accompanying defect formation in the volume of the crystals. The others are regions where ν varies weakly as H_2O molecules are implanted into the lattice, with $0.9 \leqslant a \leqslant 2.5$ mmol/g and where it falls sharply as the volume of the lattice increases, with $a > 2.5$ mmol/g.

The capture rate is related to the concentration and radius of the defects (that capture positrons) by

$$\nu = 4\pi r_+ N_+ D_+, \tag{11}$$

where D_+ is the positron diffusion coefficient. The defect radius is[42]

$$r_+ = [x^2(1+y)^2/(2mU/\hbar^2)]^{1/2}, \tag{12}$$

where m is the mass of an electron, $\hbar$ is Planck's constant, U is the depth of the potential well of the defects, $x = (1/y) \times [\lambda_f \tau_2/(1+y^2) - 1]$, $x = \pi - \arctan(1/y)$. In the initial samples with $U = 2$ eV, $D_+ = 0.1$ cm^2/s,[43] $r_+ = 3$–3.5 Å, and $N_+ = (1.1$–$2.7) \times 10^{16}$ cm^{-3} (see the Table), i.e., the defects are vacancy clusters consisting of 9–13 point vacancies. As the Table shows, the behavior of $\nu(a)$ is caused by the changes in N_+ in the course of weak oscillations in r_+, so both an increase and a decrease in the number of vacancy clusters in the volume are observed owing to the physically adsorbed H_2O molecules on the surface. This is evidence of (the stimulation of) association and dissociation reactions of point defects and vacancy clusters and of the recombination of vacancy clusters on the surface during adsorption (which are possible when defects migrate in the crystals). Here the spread in ν may be related to the random predominance of association, dissociation, or recombination in the conversion of the defects in different exposures.

Note that when implantation into the lattice takes place, the water molecules do not fill vacancy clusters, since positron annihilation does not occur in water.[44] The drop in ν for $a > 2.5$ mmol/g is related to a reduction in N_+ for constant r_+, probably caused by compression of the vacancy clusters as the lattice distends.

The drop in λ_f caused by a physically adsorbed water polylayer for $a < 0.9$ mmol/g is analogous to the drop in λ_f as 4.2×10^{21} cm^{-3} O1 atoms are removed from Cu1–O1 layers when δ increases in $YBa_2Cu_3O_7$ from zero to 0.6 (inset to Fig. 3a). The escape of such a large number of oxygen atoms would lead to a growth in the lattice parameters and to suppression of superconductivity, while the parameters a and b remain constant, c decreases, and the temperature dependence of the dc resistance and T_c do not change. This shows that the amount of oxygen in the intermediate layers does not change and the drop in λ_f may be related to the nonparticipation of the core electrons of Ba and Cu1 in positron annihilation after they have been pushed out into the interstices. The constancy of a and b may be related to the fact that, with physical adsorption, subthreshold defect formation takes place in the intermediate layers, while defects do not develop in the yttrium layers and in the cuprate layers responsible for high-temperature superconductivity. In the Cu2–O layers, the energy spectrum of the l- and h-carriers may not change and the temperature T_c may not shift. Note that, for a constant oxygen content, transitions of 1.1×10^{21} cm^{-3} atoms from O1 to O5 sites do not affect T_c.[45] On the other hand, the drop in ρ_{32} associated with Ba and Cu1 defect formation in the intermediate layers and O1→O5 transitions can increase the number of holes in the $Z2$ and $Z1$ bands in the Cu2–O layers. Then the plasma frequency Ω and the average frequency $\tilde{\Omega}$ of the acoustic plasmons increase, but the critical temperature[5,6]

$$T_c = \tilde{\Omega} \exp\left[-\frac{1+\lambda}{\lambda - \mu_c^*(1+\lambda)} \right] \tag{13}$$

may not vary if the increase in $\tilde{\Omega}$ is compensated by a simultaneous enhancement in the Coulomb repulsion as the pseudopotential μ_c^* increases. Here

$$\mu_c^* = \mu_c^0 [1 + \mu_c^0 \ln(E_F/\tilde{\Omega})]^{-1}$$

is the Coulomb pseudopotential,

$$\lambda = \frac{2}{\pi} N(0) \int_0^\infty \frac{d\omega}{\omega} \langle \tilde{V}_c(\mathbf{q}) \operatorname{Im} \varepsilon^{-1}(\mathbf{q},\omega) \rangle = \mu_c^\infty - \mu_c^0,$$

is the coupling constant,

$$\mu_c^\infty = N(0) \langle 4\pi e^2/\mathbf{q}^2 \rangle, \quad \tilde{V}_c(\mathbf{q}) = 4\pi e^2/\varepsilon(\mathbf{q},\omega)\mathbf{q}^2,$$

E_F is the Fermi energy, and $N(0)$ is the density of l-carriers at the Fermi level.

The differences in defect formation by a physically adsorbed water polylayer and when H_2O molecules are implanted into an $YBa_2Cu_3O_7$ lattice show up in the temperature dependence of the surface resistance R_s measured at a frequency of 10 GHz by the cavity resonator method in the TE_{011} mode (Fig. 4). For $T > T_c$ the resistance R_s decreases under the influence of physically adsorbed water and increases when H_2O molecules are implanted into the crystal lattice, while for $T < T_c$, R_s increases in both cases. Therefore, in the normal state, an increase in the number of defects with subthreshold defect formation reduces R_s, while an increase in the number of H_2O molecules in the interstices increases R_s. In the superconducting state, an increase in the

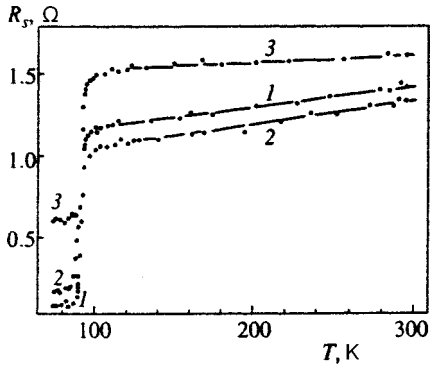

FIG. 4. Temperature variations in the surface resistance of $YBa_2Cy_3O_7$ after exposure for 0 (*1*), 90 (*2*), and 240 min (*3*) in water vapor at 18.7 Torr.

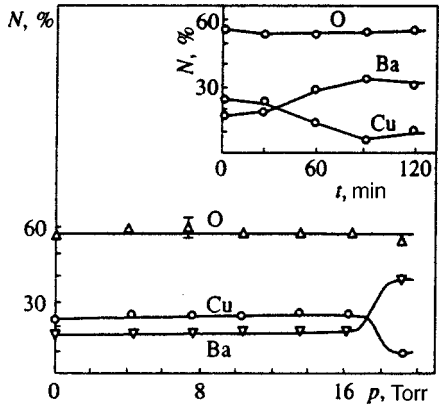

FIG. 5. Elemental composition of the surface as a function of water vapor pressure for an adsorption time of 120 min. The inset shows the atomic composition as a function of exposure time for a pressure of 18.7 Torr.

number of defects through subthreshold defect formation and implantation of H_2O molecules increases R_s. This behavior of R_s can be explained by noting that for $T>T_c$ the resistance is $R_s=(\sigma\delta)^{-1}$ (where σ is the resistivity and δ is the skin depth) and is determined by scattering of carriers in a skin layer with a depth $\delta\simeq 20\,\mu$m, while for $T<T_c$, the resistance is $R_s=-\mu_0\omega\,\mathrm{Im}\,\chi$ (where ω is the frequency, μ_0 is the magnetic permeability, and χ is the complex penetration depth[46]) and is determined by dissipation of carrier energy in a layer with a depth roughly equal to the penetration depth $\lambda\simeq 600-2000$ Å.[47] Since $\delta\gg\lambda$, the behavior of R_s is related to the different distributions of defects in the layers λ and $\delta-\lambda$. Assuming that physically bound water stimulates a nonuniform distribution in which defects accumulate within a thin surface layer $l<\lambda$, while H_2O molecules in the lattice are distributed uniformly within a layer of thickness δ, the growth in R_s for $T<T_c$ can be explained by an increase in λ as the number of defects increases, and for $T>T_c$ the growth in R_s is caused by enhanced scattering of l-carriers on H_2O molecules and the drop in R_s by weaker scattering on defects in a layer of thickness $\delta-l$.

Therefore, H_2O dipoles physically adsorbed on an $YBa_2Cu_3O_7$ surface stimulate subthreshold formation of $\sim 10^{21}\,\mathrm{cm}^{-3}$ Ba and Cu1 defects, O1↔O5 migration in the intermediate layers, and the recombination of more than $10^{15}\,\mathrm{cm}^{-3}$ vacancy clusters when there is no defect formation in the cuprate layers. Here the defects apparently accumulate in a thin surface layer of the crystallites. The H_2O molecules, as they are implanted into the crystal lattice, appear to be localized at interstices and form coordination bonds with the atomic surroundings.

The accumulation of an anomalously large number of defects owing to a physically adsorbed water polylayer in a thin surface layer of crystallites shows up through a change in the atomic composition of the surface and the blocking of diffusion transitions within the volume. The dependence of the surface composition on the water vapor pressure for a fixed adsorption time of $t=120$ min is shown in Fig. 5. This figure shows that the concentration of Ba and Cu atoms changes suddenly at $p>17$ Torr, while the amount of oxygen is essentially constant. Increasing the adsorption time from 30 to 120 min with a constant pressure of 18.7 Torr and an adsorption of $0.1\leqslant a\leqslant 0.3$ mmol/g leads to a smooth change in the Ba and Cu concentrations with a constant O content in the surface layer (inset to Fig. 5). Note that the number of yttrium atoms in the surface layer also did not change.

Therefore, the effect of physically bound water molecules on an $YBa_2Cu_3O_7$ surface has a threshold character: it begins at $p>17$ Torr and is characterized by an enrichment of the surface layer in barium atoms, a depletion in copper atoms, and constant content of oxygen and yttrium. The change in the number of Ba and Cu atoms on the surface indicates that the defects formed within the volume are interstitial atoms and vacancies of Ba and Cu. This is in good agreement with the reduction in $c(a)$ and $\lambda_f(a)$ as a rises to 1.1 mmol/g (Figs. 2 and 3a). In addition, the change in the amount of Ba and Cu by a factor of two indicates forced diffusion of the atoms under the influence of the physically adsorbed water. The diffusion field can be estimated from the excess concentration of Ba, $\Delta n_d=n(a)-n(0)\simeq 5\times 10^{14}\,\mathrm{cm}^{-2}$, for a Ba charge state of $Z=2+$ and $\varepsilon=4$:[48] $E=Ze\Delta n_d/\varepsilon_0\varepsilon\simeq 4.5\times 10^8$ V/cm. This magnitude of E is possible if the Ba and Cu atoms migrate through the crystal, trapped in the field of the weakly damped plasmons.

It may be assumed that the change in the atomic composition is caused by the effect of the electric field $\mathbf{E}_\mu$ created by n adsorbed H_2O molecules on the crystal surfaces:

$$\mathbf{E}_\mu=\sum_{i=1}^{n}\frac{3\mathbf{R}(\mu\mathbf{R})-\mathbf{R}^2\mu}{4\pi\varepsilon_0\varepsilon\mathbf{R}^5}, \tag{14}$$

where μ is the dipole moment. However, it is unlikely that the field will have a threshold effect and such a selective action on Ba, Cu, and Y cations and, specifically, on the diffusion of Ba to the surface and Cu within the volume, without affecting Y and oxygen, especially O1 whose energy to escape from a node is substantially lower than for Ba or Cu. With increasing a, as a polylayer of water develops on the surface, the effect of $\mathbf{E}_\mu$ becomes weaker because of the increase in ε. The thickness of the physically bound water layer increases as p rises, so that the adsorption[49] is given by

$$a=\Gamma p/62.4d(1-\varepsilon)T, \tag{15}$$

where Γ is the Henry constant, d is the specific mass of the sorbent, and ε is the porosity. As the atomic composition of

$YBa_2Cu_3O_7$ changes, estimates of h from Eq. (5) for dispersed samples with $0.2<a<0.6$ mmol/g, $\omega_{H_2O}=10.2$ Å,[50] and $\rho=1$ g/cm^3 give $40<h<100$ Å. For these values of h, the permittivity of the water layer is $\varepsilon\simeq 58-111$,[51] and $\mathbf{E}_\mu$ is attenuated by a factor of ε. In addition, with adsorption and a transition from a monolayer to a polylayer, the Coulomb barrier to chemosorption of the H_2O molecules decreases. Thus, the interaction energy of atomic species α in the HTSC in a charge state q_α^0 with atoms in the H_2O molecules is

$$U=\frac{1}{8\pi\varepsilon_0\varepsilon}\sum_{i,\alpha} q_\alpha^0 V_\alpha^i, \tag{16}$$

where

$$V_\alpha^i=\sum_{j,\beta}\frac{q_\beta^j}{|\mathbf{r}_{i\alpha}-\mathbf{r}_{j\beta}|}$$

is the electrostatic potential created by the j atoms of the β-th H_2O molecule at the i-th atom, reduced by a factor ε as a polylayer develops, which may explain the onset of the implantation of H_2O molecules into the $YBa_2Cu_3O_7$ crystal lattice after formation of a polylayer with $h>40$ Å. This thickness permits formation of a solvation coat which is sufficient to draw Ba into the polylayer, induce perturbations $\delta\rho_i$ at the surface, and induce shielding oscillations of the h-carriers in the volume. Thus, the change in the atomic composition of the surface is associated with volume processes of subthreshold defect formation and defect migration stimulated by the perturbations $\delta\rho_i$ owing to physically adsorbed H_2O molecules. It may be assumed that defects captured by the field of collective excitations of h-carriers that move without dissipation migrate through the crystals to places where they are scattered, in particular to the intergrain boundaries—surfaces where the Be atoms accumulate—while the Cu atoms are trapped in the volume. Part of the defects participate in association or dissociation reactions, including with vacancy clusters.

The buildup of a high concentration of defects in the surface layer and at the boundaries of $YBa_2Cu_3O_7$ crystallites serves to block diffusion jumps of nickel atoms in the volume and at the surface. Thus, in the original samples, diffusion is characterized by a surface component

$$D_s=3.16\times 10^{-10}\exp(-0.17/kT) \tag{17}$$

at temperatures[31] of 200–412 °C and by a volume component

$$D_v=1.0\times 10^{-2}\exp(-1.3/kT) \tag{18}$$

at temperatures of 412–500 °C (Fig. 6). Keeping the samples in water vapor for 120 min suppresses the volume component D_v and stimulates the surface slow component D_s^s and fast component D_s^r, which are given (in cm^2/s) by

$$D_{s1}^s=3.2\times 10^{-9}\exp(-0.27/kT),$$

$$D_{s1}^r=2.5\times 10^{-6}\exp(-0.37/kT) \tag{19}$$

(curves *1* and *1′* in Fig. 6). Increasing the time spent by the $YBa_2Cu_3O_7$ in water vapor to 360 min leads to a reduction in the preexponential factor in D_s^r and an increase in that for D_s^s without changing the activation energy for either:

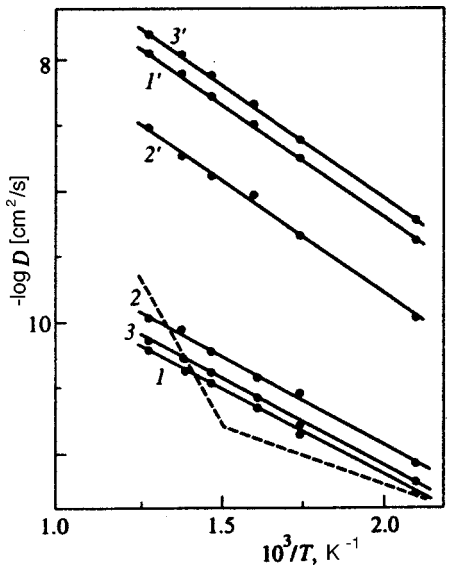

FIG. 6. Diffusion coefficient of Ni atoms as a function of temperature before (dashed curve) and after adsorption of water molecules for 120 min (*1*, *1′*), 360 min (*2*, *2′*), and 360 min with subsequent heating to 400 °C (*3*, *3′*).

$$D_{s2}^s=6.31\times 10^{-9}\exp(-0.28/kT),$$

$$D_{s2}^r=6.3\times 10^{-6}\exp(-0.37/kT) \tag{20}$$

(curves *2* and *2′* in Fig. 6).

Heating samples kept for 360 min in water vapor at a rate of 5 °C/min to 400 °C removes the H_2O molecules from the crystal lattice[52] and essentially restores both components:

$$D_{s3}^s=4.0\times 10^{-9}\exp(-0.27/kT),$$

$$D_{s3}^r=3.6\times 10^{-6}\exp(-0.37/kT) \tag{21}$$

(curves *3* and *3′* in Fig. 6). This heating does not reconstitute the components D_s and D_v of the original samples.

Note that the diffusion profiles in the volume of the samples were determined by removal of layers, while in the dense ceramics with a layer of adsorbed water, contact was made with grains lying mostly on the sample surface. Diffusion transitions, however, are blocked in crystals lying in the bulk of the ceramic that do not come into contact with the water layer. Defects can accumulate on the surface of crystallites lying within the volume when collective motion of the h-carriers is stimulated by perturbations in the charge density at the intergrain boundaries, perturbations created by acoustic plasmons that move without dissipation, excited within the crystallites and found on the sample surface.

Thus, physically bound water suppresses volume diffusion of Ni atoms and stimulates the two surface diffusion components, D_s^s and D_s^r. Implanting water molecules into the crystal lattice during chemosorption acts in opposite ways on the preexponential factors in D_s^s and D_s^r, without changing the activation energy. Annealing the H_2O molecules restores both components.

Suppressing volume diffusion transitions of Ni atoms can be evidence of the accumulation of a large number of

interstitial atoms in the surface region. Occupation of interstices by atoms increases the factor $q=m_d/M$, where m_d is the number of occupied interstices, M is the overall number of interstices, and for $q\to 1$ the diffusion coefficient takes the form[53]

$$D_{s,v}=(1-q)D_0\exp(-E/kT), \quad D_v\to 0, \tag{22}$$

i.e., volume transitions are blocked and diffusion is suppressed. In addition to this, interstitial atoms stimulate diffusion transitions along different nonequivalent interstices in the surface layer, which leads to the fast and slow components of D_s with different activation energies.

Implanting water molecules into the interstices of the lattice without changing the lattice parameters and electron density does not affect the activation energies for D_s^s and D_s^r. However, with localization in nonequivalent interstices the H_2O molecules can either increase or reduce the potential energy of nickel atoms at different interstices, such as u_O and u_T. In the special case of diffusion along two types of interstices,[53] we have

$$D=\alpha l^2\omega\frac{\lambda+q\mu-K}{q\mu^2}\exp\left(-\frac{\Delta}{kT}\right), \tag{23}$$

where α is a geometric factor, l is the jump length, ω is the vibration frequency of an atom at an interstice, $\mu=1-\varepsilon$, $K=\sqrt{(\lambda+3q\mu)^2-12q\mu}$, $\lambda=1+2\varepsilon$, and $\varepsilon=\exp[(u_O-u_T)/kT]$, the opposite behavior of D_s^s and D_s^r with increasing q when H_2O molecules are implanted into the interstices and Δ remains constant can be explained by an increase or a decrease in the difference u_O-u_T.

In the case of diffusion transitions from a node at an interstice, when[54]

$$D=\frac{1}{6}l^2(g\tau_g\tau_z)^{-1/2}\exp\left(-\frac{\Delta}{kT}\right), \tag{24}$$

where g is the number of vacant nodes and τ_g and τ_z are the residence times at a node and at an interstice, the behavior of D_s^s and D_{s2}^r during implantation of H_2O molecules into the lattice can be related to a drop in g with opposite changes in τ_g and τ_z.

Note that the behavior of D_{s2}^s and D_{s2}^r after heating indicates different types of interstitial defects which block the diffusion transitions. The recovery of D_s^s and D_s^r is caused by the removal of H_2O molecules which have been implanted into interstices during chemosorption, while the irreversible suppression of volume diffusion D_v is caused by the lack of effect of heating on the interstitial defects, in particular on Ba atoms that accumulate in the surface layer because of the physically adsorbed H_2O molecules.

Therefore, the blocking of diffusion transitions by Ni atoms shows that physically adsorbed H_2O molecules cause the surface layer of $YBa_2Cu_3O_7$ crystallites to be enriched by a large number of interstitial atoms. Here the blocking effect can be evidence of migration stimulated by the adsorbed molecules or of defect mass transport in the volume of the crystals.

To summarize, H_2O molecules adsorbed on $YBa_2Cu_3O_7$ surfaces stimulate subthreshold formation of anomalously high (comparable to that of the free carriers) concentrations of defects in intermediate layers (vacancies and interstitial Ba and Cu1 atoms), O1↔O5 transitions, migration and conversion reactions of defects in the volume of the crystals, and accumulation of defects, apparently in a thin layer at the surface and at the intercrystallite boundaries.

Here the subthreshold defect formation and the number and type of defects indicate that there is a subsystem of strongly localized holes which lie in the $Z2$ band, that the motion of the holes is excited by low energies and is collective, and that carriers move through the crystal at low frequencies ($\sim\omega_D$) and without dissipation. Subthreshold defect formation evidently takes place as a result of the Coulomb ejection of Ba and Cu1 cations into interstices by moving holes, which are predominantly localized at oxygen atoms, whose p-orbitals make the major contribution to the hybridization of the pd-orbitals of the band, and as a result of the migration of interstitial defects in the field of the collective excitations toward the boundaries and surfaces of the crystallites, where they accumulate, perhaps, during scattering of acoustic plasmons. The transitions of oxygen from O1 sites to O5 and back, as well as their random character, are probably caused by the mutual repulsion of holes and the neighboring atoms.

In conclusion, the author thanks V. T. Adonkin, G. N. Kashin, I. B. Kevdina, V. S. Mel'nikov, D. V. Morozovska, V. M. Pashkov, and G. M. Shalyapina for help in the measurements and for useful discussions.

*)E-mail: user@surfchen.freenet.kiev.ua

[1] J. Ruvalds, Phys. Rev. B **35**, 8869 (1987).
[2] A. Griffin and J. Pinder, Phys. Rev. B **39**, 11503 (1989).
[3] V. Z. Kresin, Phys. Rev. B **35**, 8716 (1987).
[4] J. Ashkenazi, C. J. Kuper, and P. Tyk, Solid State Commun. **63**, 1145 (1987).
[5] É. A. Pashitskiĭ, Fiz. Nizk. Temp. **21**, 995 (1995) [Low Temp. Phys. **21**, 763 (1995)].
[6] É. A. Pashitskiĭ, Fiz. Nizk. Temp. **21**, 1091 (1995) [Low Temp. Phys. **21**, 837 (1995)].
[7] M. M. Klinger, Ch. V. Lushchik, T. V. Mashovets *et al.*, Usp. Fiz. Nauk **147**, 523 (1985) [Sov. Phys. Usp. **28**, 994 (1985)].
[8] H. Krakauer and W. P. Pickett, Phys. Rev. B **37**, 7252 (1987).
[9] A. J. Freeman and J. Yu, Helv. Phys. Acta **61**, 401 (1988).
[10] M. Hirao, T. Uda, and Y. Murayama, Physica C **195**, 230 (1992).
[11] Rong Liu, B. W. Veal, A. P. Paulioas *et al.*, Phys. Rev. B **45**, 5614 (1992).
[12] Yu. A. Izyumov, N. M. Plakida, and Yu. N. Skryabin, Usp. Fiz. Nauk **153**, 621 (1989) [Sov. Phys. Usp. **32**, 1060 (1989)].
[13] V. A. Arbuzov, O. M. Bakunin, V. B. Vykhodets *et al.*, Sverkhprovodimost: Fiz., Khim., Tekh. **4**, 2410 (1991).
[14] V. I. Nefedov, A. N. Sokolov, M. A. Tyzikhov *et al.*, Poverkhnost', , No. 9, 22 (1989).
[15] V. M. Nefedov and A. N. Sokolov, ZhNKh **36**, 2723 (1989).
[16] E. A. Eremina, N. N. Oleĭnikov, V. I. Nefedov, and A. N. Sokolov, ZhVKhO **34**, 528 (1989).
[17] M. V. Pulyaeva, N. S. Granova, Yu. N. Velikhov, and V. V. Usenkova, Preprint IMK-92-13, Kharkov (1992).
[18] A. Barkatt, H. Hujaji, R. W. Vasantha *et al.*, MRS Bull. , September, 45 (1993).
[19] M. U. Kalanov, M. Karimov, M. S. Paĭzulakhanov *et al.*, Uzb. Fiz. Zhurn., No. 2, 56 (1993).
[20] S. Wen, X. Song, and J. Feng, Mater. Lett. **6**, 385 (1988).

[21] B. M. Gorelov, V. V. Dyakin, G. N. Kashin *et al.*, J. Electron Spectrosc. Relat. Phenom. **70**, 161 (1994).
[22] P. Turelici, A. L. Wach, R. K. Welzler *et al.*, J. Phys.: Condens. Matter **2**, 1635 (1990).
[23] V. G. Bar'yakhtar, A. K. Zalko-Titarenko, V. S. Melnikov *et al.*, Int. J. Mod. Phys. B **1**, 1259 (1988).
[24] K. Ikeda, M. Nagata, M. Ishika *et al.*, Jpn. J. Appl. Phys., Part 2 **27**, L202 (1988).
[25] A. K. Zhalko-Titarenko, V. S. Mikhalenkov, A. E. Morozovskiĭ *et al.*, Sverkhprovodimost: Fiz., Khim., Tekh. **3**, 948 (1990).
[26] V. N. Lysenko, V. T. Adonkin, V. V. Dyakin *et al.*, Sverkhprovodimost: Fiz., Khim., Tekh. **5**, 344 (1992).
[27] D. C. Connors and R. N. West, Phys. Rev. Lett. A **30**, 24 (1969).
[28] A. G. Merzhanov, A. V. Makarov, G. V. Romanov *et al.*, Pis'ma Zh. Tekh. Fiz. **15** (11), 4 (1989) [Sov. Tech. Phys. Lett. **15**, 414 (1989)].
[29] Y. Gao, T. J. Wagener, J. H. Weaver *et al.*, Phys. Rev. B **36**, 3971 (1987).
[30] B. I. Boltaks, *Diffusion in Semiconductors*, Infosearch, London (1963) [Russian orig. Fizmatgiz, Moscow (1961), 167 pp.].
[31] P. P. Gorbik, V. V. Dyakin, V. V. Zaitov *et al.*, Sverkhprovodimost: Fiz., Khim., Tekh. **3**, 1654 (1990).
[32] T. Marimoto and T. Ywaki, J. Chem. Soc., Faraday Trans. 2 **83**, 943 (1987).
[33] G. N. Zatsepina, *Physical Properties and Structure of Water*, Izd-vo. MGU, Moscow (1987) [in Russian].
[34] R. C. Baezoed, Phys. Rev. B **42**, 56 (1990).
[35] N. N. Degtyarenko, V. F. Elesin, and V. L. Mel'nikov, Sverkhprovodimost: Fiz., Khim., Tekh. **3**, 2516 (1990).
[36] V. V. Kirsanov, N. N. Musin, and E. I. Shamarina, Sverkhprovodimost: Fiz., Khim., Tekh. **7**, 427 (1994).
[37] P. A. Sterne, J. H. Kaiser, J. C. O'Brien, and R. H. Howell, Mater. Sci. Forum **105-110**, 469 (1992).
[38] E. Boronski and R. N. Nieminen, Phys. Rev. B **34**, 3820 (1986).
[39] D. M. Ginzberg, ed., *Physical Properties of High Temperature Superconductors*, World Scientific, Singapore (1989) [Russian transl. Mir, Moscow (1990)].
[40] R. C. Baezoed, Phys. Rev. B **38**, 11304 (1988).
[41] V. T. Adonkin, B. M. Gorelov, D. V. Morozovska, and V. M. Ogenko, Colloids Surf., A **101**, 233 (1995).
[42] P. A. Bulenkov, Yu. V. Zherdev, V. V. Palin *et al.*, Khim. Fiz. No. 1, 126 (1983).
[43] W. Swiatkowski, Acta Universitatis, Matematyka, Fizyka, Astronomia, No. 39, 3 (1982).
[44] V. T. Adonkin, V. V. Dyakin, V. M. Ogenko, and A. A. Chuiko, React. Kinet. Catal. Lett. **50**, 227 (1993).
[45] V. T. Adonkin, B. V. Alekseenko, A. P. Galushka *et al.*, Fiz. Tverd. Tela **35**, 1427 (1993) [Phys. Solid State **35**, 720 (1993)].
[46] N. H. Tec, M. B. Salamon, I. Datta *et al.*, Phys. Rev. B **45**, 5628 (1992).
[47] J. Buan, B. P. Stoikovic, N. Israeloff *et al.*, Phys. Rev. B **54**, 7462 (1996).
[48] A. G. Grachev, Sverkhprovod. Fiz. Khim. Tekh. **3**, 2516 (1990).
[49] D. P. Timofeev, *Adsorption Kinetics* [in Russian], Izd. AN SSSR, Moscow (1962).
[50] R. N. Meals and F. M. Lemis, *Silicones*, Reinhold Publ. Corp., New York (1963).
[51] M. M. Dubinin and V. V. Serpinskaya, eds., *Adsorption in Micropores* [in Russian], Nauka, Moscow (1983).
[52] P. P. Gorbik, B. M. Gorelov, V. V. Dyakin *et al.*, Ukr. Fiz. Zh. **38**, 84 (1993).
[53] A. A. Smirnov, *Theory of Diffusion in Implantation Alloys* [in Russian], Naukova Dumka, Kiev (1982).
[54] W. Seith, *Diffusion in Metallen*, Springer Verlag, Berlin (1939).

Translated by D. H. McNeill

Phase transitions and magnetic-transport phenomena in the system $La_{2/3}Ba_{1/3}(Mn_{1-x}Co_x)O_3$

I. O. Troyanchuk,*) L. S. Lobanovskiĭ, D. D. Khalyavin, V. P. Yarunichev, and N. V. Pushkarev

Institute of Solid-State and Semiconductor Physics, National Academy of Sciences of Belarus, 220072 Minsk, Belarus

H. Szymczak

Institute of Physics, Polish Academy of Sciences, 02-668 Warsaw, Poland

(Submitted 26 January 1999)
Zh. Éksp. Teor. Fiz. **116**, 604–610 (August 1999)

A study is performed of the crystalline structure, magnetization, and magnetotransport properties of the system $La_{2/3}Ba_{1/3}(Mn_{1-x}Co_x)O_3$ with perovskite structure. It is shown that cubic solid solutions exist over the entire range of cobalt concentrations $0 \leqslant x \leqslant 1$. Compositions with $x \leqslant 0.2$ are ferromagnets with maximum resistance near T_C. Compositions with $0.2 < x < 0.4$ manifest properties of inhomogeneous ferromagnets. Measurements of magnetic properties indicate the absence of long-range magnetic order in compositions with $0.5 \leqslant x \leqslant 0.9$, which are probably spin glasses. The spontaneous magnetization of cobaltate ($2\mu_B$ per formula unit) corresponds to ferromagnetic ordering of the moments of the Co^{3+} and Co^{4+} ions found in the intermediate spin state. It is conjectured that the magnetoresistance consists of an extrinsic and an intrinsic contribution. The first arises as a result of intergrain transport of spin-polarized charge carriers, and the second, as a result of magnetic ordering near T_C. The magnetoresistance is essentially independent of the spontaneous magnetization and decreases abruptly as the cobalt concentration is increased with a corresponding transition from long-range to short-range magnetic order. © *1999 American Institute of Physics.* [S1063-7761(99)01708-4]

1. INTRODUCTION

The giant magnetoresistance effect in manganates has attracted considerable attention. This has been due, on the one hand, to a clarification of the nature of the phenomenon and, on the other, to the possibility of practical applications. At present there are several points of view regarding this phenomenon.[1–4] Nagaev[5,6] has analyzed different situations realized in different types of magnetic semiconductors, including oxides. Note that on the basis of theoretical arguments it is hard to predict the magnitude of the effect in different magnetically ordered media. For this reason, the search for new magnetoresistive materials continues.[7,8]

Among oxides with mixed valence of the $3d$ ions, manganates and cobaltates with perovskite structure possess similar magnetic properties. Replacement of lanthanum ions by an alkaline-earth ion leads to ferromagnetism and metallic conductivity in both types of oxides. Manganates and cobaltates of rare-earth ions manifest the giant magnetoresistance effect during metamagnetic phase transitions.[9,10] However, the magnetotransport phenomena in oxides of the type $La_{1-x}Sr_xMnO_3$ and $La_{1-x}Sr_xCoO_3$ are completely different. Cobaltates of the form $La_{0.5}Sr_{0.5}CoO_{3-x}$ manifest a small positive magnetoresistive effect, which gradually decreases as the temperature is increased,[11,12] whereas manganates manifest a large negative magnetoresistive effect near the Curie point. Cobaltates remain metallic during the transition from the magnetically ordered phase to the paramagnetic phase, whereas manganates undergo a metal–insulator transition near T_C. Therefore it is of unquestionable interest to trace out the changes in their properties as the concentration is shifted from manganate to cobaltate.

2. EXPERIMENTAL PROCEDURE

Samples with compositions $La_{2/3}Ba_{1/3}(Mn_{1-x}Co_x)O_3$ ($0 \leqslant x \leqslant 1$, $\Delta x = 0.1$) were prepared by solid-state reactions at high temperatures. Especially pure quantities of the starting materials La_2O_3, $BaCO_3$, Mn_2O_3, and Co_3O_4 were mixed in stoichiometric relations and pressed into pellets. A preliminary anneal was performed at 1000 °C; next the pellets were ground up and pressed again. The final anneal was carried out at temperatures from 1470°C (manganate) to 1200 °C (cobaltate). The synthesis temperature was lowered uniformly as manganese was replaced by cobalt. The samples were cooled slowly (at a rate of 80°C per hour) to room temperature. X-ray phase analysis did not reveal any traces of foreign phases. The oxygen content in some of the samples was determined thermogravimetrically. It was found that the oxygen content was close to stoichiometric. According to the x-ray structural analysis, the resulting compositions were characterized by a cubic unit cell. The lattice constant increased as a function of the cobalt content from $a = 7.873$ to $a = 7.913$ Å.

The magnetic measurements were performed on a commercial vibration magnetometer in magnetic fields up to

 © 1999 American Institute of Physics

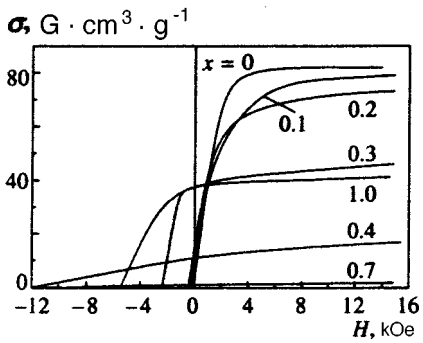

FIG. 1. Dependence of the magnetization on the field at 4.2 K. The measurements were made as the field was decreased after cooling in a field of 15 kOe.

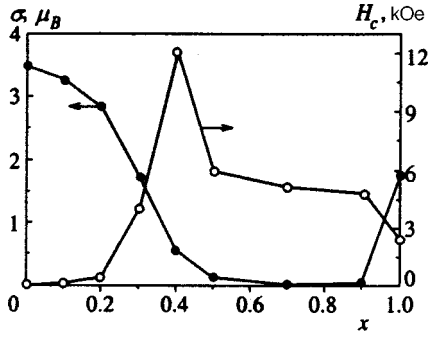

FIG. 2. Dependence of the spontaneous magnetization (left axis) and the coercive force (right axis) on the cobalt concentration.

16 kOe. The conductivity was measured on samples with dimensions $2\times2\times10$ mm. The contacts were formed by ultrasound indium soldering.

3. RESULTS AND DISCUSSION

The composition $La_{2/3}Ba_{1/3}MnO_3$, according to the magnetic measurements (Fig. 1), is a magnetically soft ferromagnet with magnetic moment $3.5\mu_B$ per formula unit at 5 K with Curie temperature 350 K.

As the cobalt content is increased, the ferromagnetic ordering is preserved up to a cobalt content of $x=0.2$. As the cobalt content increases further, the spontaneous magnetization begins to decrease steeply (Figs. 1 and 2). In compositions with $0.4\leqslant x\leqslant0.9$ the spontaneous magnetization does not exceed $0.1\mu_B$ per formula unit. The large paraprocess indicates that exchange interactions between ions of different nature are weakened (Fig. 1). As the cobalt content is increased, the coercive force grows steeply, reaching a maximum ~12 kOe at $x=0.4$ (Fig. 2). This is followed by a smooth fall of the coercive force to 3 kOe in pure cobaltate.

As the cobalt content is increased from $x=0.9$ to $x=1$, the spontaneous magnetization grows abruptly to $2\mu_B$ per formula unit. Such a spontaneous magnetic moment corresponds to parallel orientation of the cobalt magnetic moments, which are probably found in an intermediate spin state in pure cobaltate.

Results of a study of the temperature dependence of the magnetization are presented in Fig. 3. Compositions with cobalt concentration $x<0.4$ are characterized by a sharp transition from the magnetically ordered state to the paramagnetic state. In compositions with $0.4\leqslant x\leqslant0.9$ the transition is strongly washed out in temperature, which is evidence of an inhomogeneous clustered magnetic state of these solutions. The temperature of the ''magnetic order–disorder'' transition decreases smoothly as the cobalt concentration is decreased to $x=0.3$. For $0.5\leqslant x\leqslant0.9$ the freezing point temperature of the magnetic moments of the clusters depends weakly on the cobalt concentration. Blocking of the magnetic moments of the clusters takes place over a wide interval of temperatures.

Compositions with $0\leqslant x\leqslant0.2$ reveal a maximum in the temperature dependence of the resistivity. This maximum is located near the Curie temperature (Fig. 4). In the composition with $x=0.2$, the resistivity again decreases after the transition to the magnetically ordered state, and then grows again as the temperature is lowered, like a semiconductor. As the cobalt concentration increases further, the resistance increases abruptly, and no anomalies in the resistivity near the critical temperature are observed. The resistivity decreases abruptly in pure cobaltate. This compound is a metal both before and after the ferromagnet–paramagnet transition. Near T_C the resistivity decreases insignificantly in the transition to the magnetically ordered state.

The magnetoresistance $[R(H=0)-R(H)]/R(H=0)$ of compositions with $0<x<0.2$ is maximum near the Curie temperature and reaches 20% in a 9-kOe field.

Below the Curie temperature the magnetoresistance increases smoothly as the temperature is lowered, and near liquid-nitrogen temperatures becomes commensurate with the magnetoresistance near T_C. For $x=0.4$ no maximum in the magnetoresistance near T_C is observed. In this composition the magnetoresistance increases gradually as the tem-

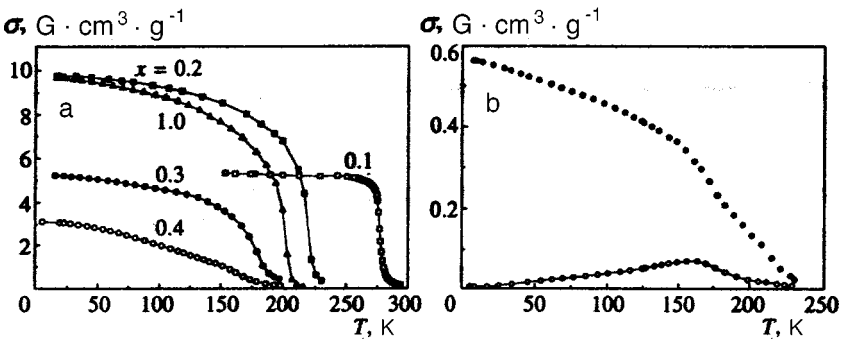

FIG. 3. a) Temperature dependence of the magnetization in a field of 100 Oe, measured after cooling in a field of 100 Oe. b) Analogous dependence for $x=0.5$ measured after cooling in a field of 100 Oe (filled circles) and in zero field (empty circles).

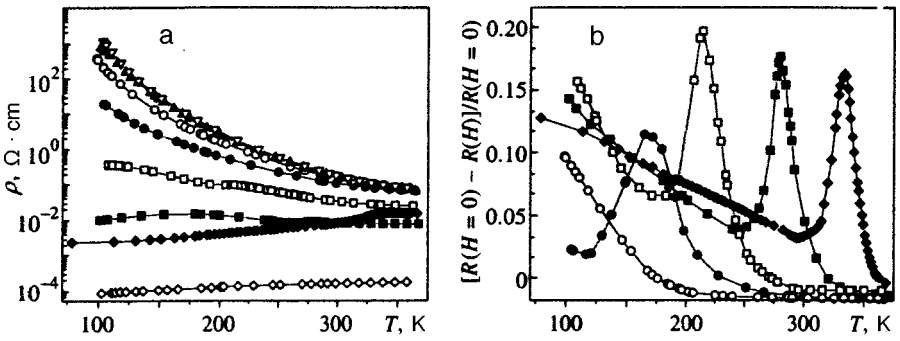

FIG. 4. Dependence of the resistivity (a) and magnetoresistance measured in a field of 9 kOe (b) on temperature for various cobalt concentrations: ◆ — $x=0$, ■ — $x=0.1$, □ — $x=0.2$, ● — $x=0.3$, ○ — $x=0.4$, ▲ — $x=0.5$, ▽ — $x=0.6$, ◇ — $x=1$.

perature is lowered. In all compositions with $0 \leqslant x \leqslant 0.3$ the low-temperature parts of the magnetoresistance near liquid-nitrogen temperature are similar in magnitude. For $x \geqslant 0.5$ the magnetoresistance is extremely small: it is less than 2% in a 9-kOe field at 120 K according to our data. In pure cobaltate the magnetoresistance is negative (around 3% at 120 K in a 9-kOe field). We did not detect a maximum in the magnetoresistance near T_C in $La_{2/3}Ba_{1/3}CoO_3$.

Usually the properties of perovskites are interpreted in terms of the concept of double exchange.[1,4] However, there are a number of factors contradicting this point of view.[6,13] We believe that the ferromagnetism of orthomanganates is due to the positive exchange interactions Mn^{3+}–O–Mn^{3+} and Mn^{3+}–O–Mn^{4+}. The cobalt ions in perovskites can exist in the oxidation states +2, +3, and +4. On the basis of our magnetic studies it is difficult to narrow down these possibilities. It has been reliably established that the cobalt ions in the perovskite phases $Ln(Co_{0.5}Mn_{0.5})O_3$ (Ln is a lanthanoid) exist in the divalent state, and manganese in the tetravalent state.[14,15]

From the results of x-ray structural studies of the compound $La_{0.5}Ca_{0.5}MnO_3$ doped with a small quantity of cobalt ions ($\sim$8% of the total number of manganese ions) we conclude that the cobalt ions are found in the divalent state.[16] Apparently, the configuration $Co^{2+}+Mn^{4+}$ in a perovskite matrix is more stable than the configuration $Co^{3+}+Mn^{3+}$. We therefore have grounds to assert that the cobalt ions in compositions with a large manganese content are found primarily in the divalent state. Co^{2+} ions, as a rule, should be surrounded by a larger number of Mn^{4+} ions in comparison with their average number. The Co^{2+}–O–Mn^{4+} exchange interaction in $La(Co_{0.5}Mn_{0.5})O_3$ is positive, whereas the superexchange Mn^{3+}–O–Co^{2+} is negative.[14] The superexchange Co^{2+}–O–Mn^{4+} is somewhat smaller in magnitude than the exchange interaction Mn^{3+}–O–Mn^{4+} since the Curie temperature for $La(Co_{0.5}Mn_{0.5})O_3$ is 240 K (Ref. 14). Therefore, when Co^{2+} ions are introduced the positive exchange interactions are weakened due to the appearance of additional bonds between cobalt and manganese. For $x=0.1$ the magnetic moments of manganese and cobalt are apparently aligned, whereas for $x=0.2$ nonferromagnetic regions are formed, which is manifested in a decrease of the magnetic moment to a value 10% below the calculated value (Fig. 2).

We believe that starting with this composition the solid solutions decay into microregions enriched with manganese or cobalt. In the manganese-rich microregions an antiferromagnetic structure is realized at low temperatures, while in the cobalt-rich regions a ferromagnetic structure is realized.

The ferromagnetism in these compositions is due to a positive interaction of the type Co^{2+}–O–Mn^{4+}. In accord with these assumptions, the coercive force grows abruptly (Fig. 2), since the compound $La(Co^{2+}_{0.5}Mn^{4+}_{0.5})O_3$ is a magnetically hard material.

As the cobalt content is increased above $x=0.3$ the cobalt ions begin to transition to the trivalent state, which leads to destruction of ferromagnetism based on the superexchange Co^{2+}–O–Mn^{4+}.

We think that the magnetic states in the solid solutions with $0.5 \leqslant x \leqslant 0.9$ have a clustered nature. In microregions containing an enhanced concentration of Co^{2+} ions, ferromagnetic clusters are formed, while in microregions enriched with Co^{3+} and Mn^{4+} ions antiferromagnetic or paramagnetic clusters are formed, depending on the nature of the magnetic state of Co^{3+}. From the available magnetic data it is hard to say in which magnetic state the Co^{3+} ions are found, whether it is the low-spin state or an intermediate spin state. The absence of long-range antiferromagnetic order is a factor testifying in favor of the low-spin state. In principle, different spin states of cobalt can coexist, depending on the immediate environment of both types of Co^{3+} ions. For $x=0.5$ a dependence of the magnetization on the prehistory of the sample appears at a quite high temperature $T=225$ K. This is due to the presence of clusters in which the positive superexchange Co^{2+}–O–Mn^{4+} dominates. In compositions

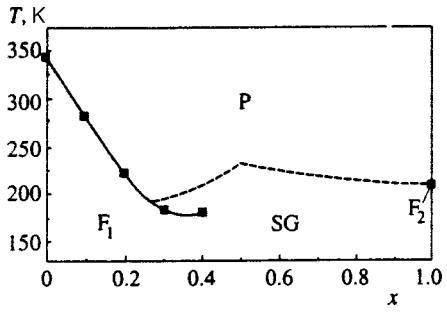

FIG. 5. Magnetic phase diagram of the system of solid solutions $La_{2/3}Ba_{1/3}(Mn_{1-x}Co_x)O_3$: F_1 — ferromagnet based on the exchange interactions Mn^{3+}–O–Mn^{3+}, Mn^{3+}–O–Mn^{4+}, and Co^{2+}–O–Mn^{4+}, P — paramagnet; SG — spin glass, and F_2 — ferromagnet based on the superexchange Co^{3+}–O–Co^{4+}. Below the dashed curve the magnetization depends on the magnetic prehistory.

close to $La_{2/3}Ba_{1/3}CoO_3$, Co^{4+} ions appear. The positive exchange interaction Co^{3+}–O–Co^{4+} leads to the appearance of ferromagnetic order in cobaltate.[17] The magnetic phase diagram is depicted in Fig. 5.

In the interpretation of the transport properties of orthomagnets it is very important to know which of the bands is responsible for the conductivity. In our view, the most justified point of view is the one according to which electron transport is realized via the $3d$ band formed by hybridization of e_g orbitals of manganese and $2p$ orbitals of oxygen.[18] At the Curie point T_C the compound $La_{2/3}Ba_{1/3}MnO_3$ undergoes a metal–insulator phase transition. This indicates that the width of the $3d$ band in this compound is close to the critical width at which the metallic state is realized. At the Curie point T_C the spin-split $3d$ bands coalesce, which gives the motion of the charge carriers a nonactivation character. The magnitude of the magnetoresistance in this case is determined by the magnitude of the conductivity jump in the region of the Curie point T_C and by how much T_C is shifted by the magnetic field toward higher temperatures, i.e., by the sensitivity of the magnetic order parameter to an external magnetic field. In pure cobaltate the connection between the conductivity and the type of magnetic state is much more tenuous than in manganate [Fig. 3(b)]. Therefore, the magnetoresistance peak, if it exists, is very small, and we did not record it. We think that the conduction band in cobaltate is much wider than in manganates as a consequence of the strong hybridization of the cobalt $3d$ orbitals and the manganese $2p$ orbitals, or the narrow $3d$ band overlaps with the wide valence p band.

When manganese is replaced by cobalt, some of the lattice sites occupied by cobalt ions become inaccessible for motion of charge carriers; this leads to suppression of the metallic state at low temperatures. However, the magnetoresistance peak in the compositions with $x=0.1$ and $x=0.2$ is not decreased in magnitude, which testifies to the strong influence of magnetic order on the conductivity. The magnetoresistance peak disappears at $x=0.4$, which corresponds to the boundary of the existence region of the ferromagnetic phase. However, in all of the compositions in the range $0 \leqslant x \leqslant 0.4$ there is one more component of the magnetoresistance, which gradually increases as the temperature is lowered and varies only weakly as manganese is replaced by cobalt.[19] This component can be attributed to intergrain transport of spin-polarized charge carriers. This transport depends strictly on the direction of the spins in the regions on both sides of the intergrain boundary, which can lead to a giant magnetoresistance effect.[19] This component of the magnetoresistance probably has only a weak dependence on the magnitude of the spontaneous magnetization (see Figs. 1 and 4).

This work was partially supported by the Foundation for Basic Research of the Republic of Belarus (Grant No. F97-097) and the Polish Committee for Scientific Research (Grant No. 2P03B 095 12).

*)E-mail: troyan@ifttp.bas-net.by

[1] A. J. Millis, P. B. Littlewood, and B. Shraiman, Phys. Rev. Lett. **74**, 5144 (1995).
[2] W. E. Pickett and D. J. Singh, Phys. Rev. B **55**, R8642 (1997).
[3] P. Majumdar and P. B. Littlewood, Nature (London) **395**, 479 (1998).
[4] D. I. Golosov, M. R. Norman, and K. Levin, Phys. Rev. B **58**, 8617 (1998).
[5] É. L. Nagaev, Usp. Fiz. Nauk **166**, 833 (1996).
[6] É. L. Nagaev, Usp. Fiz. Nauk **168**, 917 (1998).
[7] K. I. Kobayashi, T. Kimura, H. Sawada, K. Terakura, and Y. Tokura, Nature (London) **395**, 677 (1998).
[8] I. O. Troyanchuk, L. S. Lobanovsky, N. V. Kasper, M. Hervieu, A. Maignan, C. Michel, and H. Szymczak, Phys. Rev. B **58**, 14903 (1998).
[9] H. Kuwahara, Y. Tomioka, A. Asamitsu, Y. Moritomo, and Y. Tokura, Science **270**, 961 (1995).
[10] I. O. Troyanchuk, N. V. Kasper, D. D. Khalyavin, H. Szymczak, R. Szymczak, and M. Baran, Phys. Rev. Lett. **80**, 3380 (1998).
[11] G. Briceno, H. Chang, X. Sun, P. G. Schultz, and X. D. Xiang, Science **270**, 273 (1995).
[12] R. Mahendiran, A. K. Raychamdhuri, A. Chainani, and D. D. Sharma, J. Phys.: Condens. Matter **7**, L561 (1995).
[13] I. O. Troyanchuk, N. V. Kasper, D. D. Khalyavin, and E. F. Shapovalova, Phys. Rev. B **58**, 2422 (1998).
[14] N. Nishimori, K. Asai, and M. Mizoguchi, J. Phys. Soc. Jpn. **64**, 1326 (1995).
[15] I. O. Troyanchuk, N. V. Samsonenko, N. V. Kasper, and H. Szymczak, J. Phys.: Condens. Matter **9**, 8287 (1997).
[16] O. Toubemonde, F. Studer, A. Barnabe, A. Maignan, C. Martin, and B. Raveau, Eur. Phys. J. B **4**, 159 (1998).
[17] M. A. Senaris-Rodriguez and J. B. Goodenough, J. Solid State Chem. **118**, 323 (1995).
[18] W. Archibald, J. S. Zhou, and J. B. Goodenough, Phys. Rev. B **53**, 14445 (1996).
[19] Y. Qi, D. Y. Xing, and J. Dong, Phys. Rev. B **58**, 2783 (1998).

Translated by Paul F. Schippnick

Multicritical behavior of weakly disordered systems with two order parameters

V. V. Prudnikov,*) P. V. Prudnikov, and A. A. Fedorenko

Omsk State University, 644077 Omsk, Russia
(Submitted 13 February 1999)
Zh. Éksp. Teor. Fiz. **116**, 611–619 (August 1999)

We present a field-theoretic description of phase transitions in weakly disordered systems with two coupled order parameters. Using the two-loop approximation and the Padé–Borel summation technique, we carry out a renormalization-group analysis of the scaling functions for three-dimensional systems and identify the fixed points corresponding to stable multicritical behavior. We also study the effect of frozen point impurities on the nature of the phase diagrams.
© *1999 American Institute of Physics.* [S1063-7761(99)01808-9]

There is a broad class of systems in which the observed phase transitions cannot be described by a single order parameter transformed according to a single irreducible representation. This is especially evident in magnetic crystals where the magnetic structure is described by two or more irreducible representations (the antiferromagnets Cr_2TeO_6, $KCuF_3$, $GdAlO_3$, MnF_2, and the like). Structural phase transitions that require several order parameters for their description have been detected in $KMnF_3$, boracites, and other substances. The phase diagrams of such systems contain a singular multicritical point, which is either bicritical or tetracritical.[1,2] In the first case, two lines of second-order phase transitions intersect at this point, while in the second there are four lines of second-order phase transitions that intersect at this point. Near the multicritical point, the system exhibits specifically critical behavior characterized by competition of the different types of ordering. Here, if bicritical behavior is realized, one order parameter expels the other, while tetracritical behavior allows for a mixed phase in which both types of ordering coexist.

The model Hamiltonian of a homogeneous system with two coupled order parameters ϕ and ψ belonging to two different irreducible representations of dimensionalities n and m has the form

$$\mathscr{H}_0=\int d^dx\left\{\frac{1}{2}[r_1\phi^2+r_2\psi^2+(\nabla\phi)^2+(\nabla\psi)^2]+\frac{u_{10}}{4!}(\phi^2)^2+\frac{u_{20}}{4!}(\psi^2)^2+\frac{2u_{30}}{4!}\phi^2\psi^2\right\}, \tag{1}$$

$$\phi^2=\sum_{i=1}^{n}\phi_i^2, \quad \psi^2=\sum_{i=1}^{m}\psi_i^2, \quad (\nabla\phi)^2=\sum_{i=1}^{n}(\nabla\phi_i)^2,$$

$$(\nabla\psi)^2=\sum_{i=1}^{m}(\nabla\psi_i)^2.$$

An analysis of the problem of phase transitions in such a system was done by Lyuksyutov *et al.*[3] and (independently) by Kosterlitz *et al.*[4] in the one-loop approximation by the Wilson renormalization-group technique within the scope of the ε-expansion method, where $\varepsilon=4-d$ (d is the dimensionality of the system). Lyuksyutov *et al.*[3] attempted to follow the dependence of the multicritical behavior on the numbers n and m. However, numerous studies of systems with one order parameter done in recent years suggest that usually the one-loop approximation (and the more so when used within the ε-expansion method) yields predictions that differ substantially from real critical behavior. To establish the situation with multicritical phenomena and to find the dependence of the multicritical behavior on the structure of the order parameters, we used in Ref. 5 the two-loop approximation to provide a field-theoretic description of a system whose Hamiltonian is (1). We used the mass theory of Parisi,[6] which makes it possible to describe three-dimensional systems directly, without resorting to the ε-expansion. Studies of critical phenomena have shown[7] that such an approach provides the most meaningful description of critical behavior and its use together with methods of summation of asymptotically convergent series yields extremely accurate results. In Ref. 5 we carried out a renormalization-group analysis of the scaling functions in the two-loop approximation combined with the Padé–Borel summation technique and identified the fixed points corresponding to stable bicritical and tetracritical behavior.

For a system with two coupled order parameters there are three types of stable fixed points corresponding to different values of n and m. The corresponding domains of existence in the nm plane obtained in the one-loop approximation[3,4] are depicted in Fig. 1a and those obtained in the two-loop approximation[5] in Fig. 1b. The fact that the pattern changes so dramatically indicates that there is little resemblance between the one-loop representation and real multicritical behavior. Note that the first type corresponds to an isotropic fixed point, at which $u_1^*=u_2^*=u_3^*$ and the Hamiltonian (1) effectively coincides with the Hamiltonian of a system that has a single $(n+m)$-component order parameter and complete $SO(n+m)$ symmetry, which is higher than the $SO(n)\times SO(m)$ symmetry of the initial system (this is a manifestation of an asymptotic symmetry, due to fluctuations, at the multicritical point). For points of the second type, all three values u_i^* are finite and may not coincide. The lowest symmetry of the initial Hamiltonian, $SO(n)\times SO(m)$, corresponds to such points. Points of the third type correspond to decoupled order parameters, since at

 © 1999 American Institute of Physics

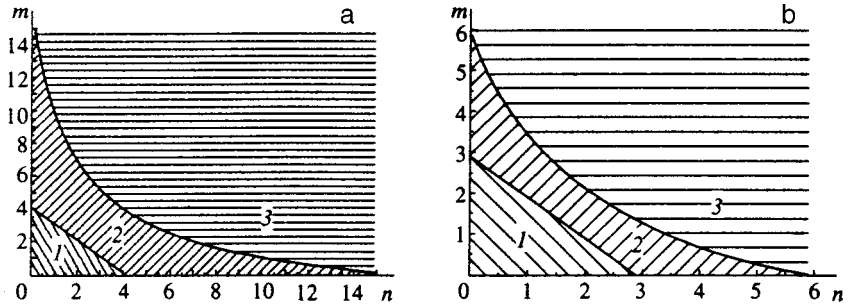

FIG. 1. The regions of stability of the fixed points determined in Refs. 3 and 4 in the first order of the ε-expansion (a) and in Ref. 5 via the field-theoretic approach in the two-loop approximation with $d=3$ (b).

these points $u_3^*=0$. Here the higher symmetry $SO(n)\oplus SO(m)$ corresponds to such points. According to Ref. 5, the edge of the stability region of an isotropic fixed point is the straight line $n+m=2.9088$, i.e., $SO(2)$ is the highest asymptotic symmetry of the system, and the domain of existence of points of the second type becomes so narrow that it contains only five physically interesting point. The appreciable change in the values of the fixed points and in the conditions for their stability, established in Ref. 5, is responsible for an appreciable change in the phase diagrams in the critical region and results in other types of the symmetry of the system at the multicritical point.

We study the effect of frozen point impurities on the multicritical behavior of a system with two coupled order parameters. As is well known,[8] the disorder in a system generated by the presence of frozen impurities manifests itself in the form of random perturbations of the local critical temperature or in the form of random fields. Because a random field breaks the symmetry of the system with respect to a change in the sign of the order parameter, the statistical properties of systems with this type of disorder differ substantially. Ferro- and antiferromagnetic systems containing nonmagnetic impurity atoms in the absence of an external magnetic field may serve as an example of disordered systems with a perturbation of the random-critical-temperature type, while in a uniform magnetic field the presence of nonmagnetic impurity atoms in anisotropic antiferromagnets manifests itself in the form of random fields.[9] In the present paper we study the multicritical behavior of system with a disorder of the random-temperature type. Such behavior can occur in disordered systems in which, as in MnAs (see Ref. 10), a sequence of phase transitions is described by introducing two coupled order parameters of different nature, parameters that correspond to the structural and ferromagnetic phase transitions, or in XY-like antiferromagnets of the type of Cr_2TeO_6, $KCuF_3$, and the like,[11] in which a multicritical point appears in a zero external magnetic field. In some cases the description of the multicritical behavior of disordered binary alloys consisting of atoms of two species with a mixed exchange interaction may correspond to the introduction of disorder of the random-critical-temperature type in a system with coupled order parameters.[12,13]

The effect of disorder of the random-temperature type on the multicritical behavior of the systems was studied by Izyumov *et al.*,[12] Laptev and Skryabin,[13] and Lisyanskiĭ and Filippov[14] by the ε-expansion method in the one-loop approximation. However, above we have clearly shown, using the example of a homogeneous system, that the predictions of the one-loop approximation do not agree with the real multicritical behavior. In disordered systems one can expect even larger discrepancies, a fact suggested by studies of disordered systems characterized by a single order parameter.[15,16] For disordered systems described by the Ising model in the one-loop approximation, accidental degeneracy occurs in the system of renormalization-group equations for the interaction vertices. This makes it impossible to use the given approximation in studies of the unique class of disordered systems in which the presence of an impurity has a real effect on the characteristics of the critical behavior of such systems. In the present paper we discuss the results of applying the field-theoretic approach in the two-loop approximation directly for three-dimensional systems.

The Hamiltonian of a system with two coupled order parameters that contains frozen impurities of the random-temperature type can be represented in the form

$$\mathscr{H}[\phi,\psi]=\mathscr{H}_0[\phi,\psi]+\mathscr{H}_{\rm imp}[\phi,\psi], \tag{2}$$

where $\mathscr{H}_0[\phi,\psi]$ is the Hamiltonian of the homogeneous system [Eq. (1)], and the term $\mathscr{H}_{\rm imp}[\phi,\psi]$, which specifies the interaction of the impurities and the order-parameter fluctuations, can be written

$$\mathscr{H}_{\rm imp}[\phi,\psi]=\frac{1}{2}\int d^dx[V_1(x)\phi^2+V_2(x)\psi^2]. \tag{3}$$

Here the $V_i(x)$ are the potentials of the random field of the impurities with a Gaussian distribution, whose correlators in the case of point impurities are given by the expressions

$$\begin{aligned}
&\langle\langle V_i(x)\rangle\rangle=0,\\
&\langle\langle V_1(x)V_1(x')\rangle\rangle=-u_{40}\delta(x-x'),\\
&\langle\langle V_2(x)V_2(x')\rangle\rangle=-u_{50}\delta(x-x'),\\
&\langle\langle V_1(x)V_2(x')\rangle\rangle=-u_{60}\delta(x-x').
\end{aligned} \tag{4}$$

Applying the replica method, we can easily average over the random distribution of impurities and reduce the problem of statistically describing a weakly disordered system to that of statistically describing a homogeneous system with the effective Hamiltonian

$$\begin{aligned}
\mathscr{H}_{\rm repl}[\phi,\psi]=\sum_{\alpha=1}^{k}\mathscr{H}_0[\phi_\alpha,\psi_\alpha]+\frac{1}{2}\sum_{\alpha=1}^{k}\sum_{\beta=1}^{k}[u_{40}\phi_\alpha^2\phi_\beta^2\\
+u_{50}\psi_\alpha^2\psi_\beta^2+2u_{60}\phi_\alpha^2\psi_\beta^2],
\end{aligned} \tag{5}$$

which contains k images (''replicas'') of the initial homogeneous component $\mathscr{H}_0$ and a number of additional terms with impurity vertices u_{40}, u_{50}, and u_{60}, which specify the effective interaction, via the impurity field, of the $(k\times n)$- and $(k\times m)$-component order parameters. This statistical model is thermodynamically equivalent to the initial disordered model in the limit $k\to 0$.

It is known that in the field-theoretic approach[17] the asymptotic critical behavior and the structure of the phase diagrams in the fluctuation region are determined by the Callan–Symanzik renormalization-group equation for the vertex parts of the irreducible Green's functions. To calculate the β-functions (the scaling functions) as functions of the renormalized interaction vertices u_i $(i=1,\ldots,6)$, which enter into the renormalization-group equation, we used a standard method based on the Feynman diagrammatic technique and the renormalization procedure.[7] As a result, in the two-loop approximation, we arrived at the following expressions for the β-functions:

$$\begin{aligned}\beta_1(u)=&-u_1+\frac{n+8}{6}u_1^2+\frac{m}{6}u_3^2+24u_1u_4-\frac{41n+190}{243}u_1^3\\&-\frac{2m}{27}u_3^3-\frac{23m}{243}u_1u_3^2-\frac{184m}{81}u_1u_3u_6\\&-\frac{16m}{9}u_3^2u_6-\frac{400n+2096}{81}u_1^2u_4\\&-\frac{5920}{27}u_1u_4^2-\frac{8m}{9}u_3^2u_4,\end{aligned}$$

$$\begin{aligned}\beta_2(u)=&-u_2+\frac{m+8}{6}u_2^2+\frac{n}{6}u_3^2+24u_2u_5\\&-\frac{41m+190}{243}u_2^3-\frac{2n}{27}u_3^3-\frac{23n}{243}u_2u_3^2\\&-\frac{184n}{81}u_2u_3u_6-\frac{16n}{9}u_3^2u_6\\&-\frac{400m+2096}{81}u_2^2u_5-\frac{5920}{27}u_2u_5^2-\frac{8n}{9}u_3^2u_5,\end{aligned}$$

$$\begin{aligned}\beta_3(u)=&-u_3+\frac{2}{3}u_3^2+\frac{n+2}{6}u_1u_3+\frac{m+2}{6}u_2u_3+4u_3u_4\\&+4u_3u_5+16u_3u_6-\frac{5(n+m)+72}{486}u_3^3\\&-\frac{23(n+2)}{486}u_1^2u_3-\frac{23(m+2)}{486}u_2^2u_3\\&-\frac{n+2}{9}u_1u_3^2-\frac{m+2}{9}u_2u_3^2\\&-\frac{20(n+m)+432}{81}u_3^2u_6-\frac{8(n+3)}{9}u_3^2u_4\\&-\frac{8(m+3)}{9}u_3^2u_5-\frac{368}{27}u_3u_4^2-\frac{368}{27}u_3u_5^2\\&-\frac{92(n+2)}{81}u_1u_3u_4-\frac{92(m+2)}{81}u_2u_3u_5\\&-\frac{8(n+2)}{3}u_1u_3u_6-\frac{8(m+2)}{3}u_2u_3u_6\\&-64u_3u_6^2-64u_3u_4u_6-64u_3u_5u_6,\end{aligned}\tag{6}$$

$$\begin{aligned}\beta_4(u)=&-u_4+16u_4^2+\frac{n+2}{3}u_1u_4+\frac{m}{3}u_3u_6-\frac{3040}{27}u_4^3\\&-\frac{2m}{27}u_3^2u_6-\frac{8m}{3}u_3u_6^2-\frac{400(n+2)}{81}u_1u_4^2\\&-\frac{23(n+2)}{243}u_1^2u_4-\frac{5m}{243}u_3^2u_4-\frac{184m}{81}u_3u_4u_6,\end{aligned}$$

$$\begin{aligned}\beta_5(u)=&-u_5+16u_5^2+\frac{m+2}{3}u_2u_5+\frac{n}{3}u_3u_6-\frac{3040}{27}u_5^3\\&-\frac{2n}{27}u_3^2u_6-\frac{8n}{3}u_3u_6^2-\frac{400(m+2)}{81}u_2u_5^2\\&-\frac{23(m+2)}{243}u_2^2u_5-\frac{5n}{243}u_3^2u_5-\frac{184n}{81}u_3u_5u_6,\end{aligned}$$

$$\begin{aligned}\beta_6=&-u_6+8u_6^2+\frac{n+2}{6}u_1u_6+\frac{m+2}{6}u_2u_6+\frac{n}{6}u_3u_4\\&+\frac{m}{6}u_3u_5+4u_4u_6+4u_5u_6-\frac{64}{3}u_6^3-\frac{4(n+2)}{3}u_1u_6^2\\&-\frac{4(m+2)}{3}u_2u_6^2-\frac{23(n+2)}{486}u_1^2u_6\\&-\frac{23(m+2)}{486}u_2^2u_6-\frac{368}{27}u_4^2u_6-\frac{368}{27}u_5^2u_6-32u_4u_6^2\\&-32u_5u_6^2-\frac{n}{27}u_3^2u_4-\frac{4n}{9}u_3u_4^2-\frac{m}{27}u_3^2u_5\\&-\frac{4m}{9}u_3u_5^2-\frac{5(n+m)}{486}u_3^2u_6-\frac{20(n+m)}{81}u_3u_6^2\\&-\frac{92(n+2)}{81}u_1u_4u_6-\frac{92(m+2)}{81}u_2u_5u_6\\&-\frac{16n}{9}u_3u_4u_6-\frac{16m}{9}u_3u_5u_6.\end{aligned}$$

Perturbation-theory series are known to be asymptotically convergent and the vertices of the interaction of the order-parameter fluctuations in the fluctuation region $r_1, r_2 \to 0$, are large enough so that Eqs. (6) can be used directly. Thus, to extract the necessary physical information from these expressions we employed the generalized Padé–Borel method used for the summation of asymptotically convergent series. Here the direct and inverse Borel transformations generalized to the six-parameter case and retaining the symmetry of the system have the form

TABLE I. Values of the fixed points of a disordered system and the eigenvalues of the stability matrix.

n	m	u_1^*	u_2^*	u_3^*	u_4^*	u_5^*	u_6^*	b_i $(i=1,\ldots,6)$
1	1	1.588 92	1.588 92	0	−0.034 48	−0.034 48	0	$0.4612\pm0.222i$, 0.0362, $0.4612\pm0.222i$, 0.0362
1	2	1.588 92	0.938 32	0	−0.034 48	−0.000 26	0	$0.4612\pm0.222i$, 0.0183, 0.0183,0.6671,0.0017
1	2	1.588 92	0.934 98	0	−0.034 48	0	0	$0.4612\pm0.222i$, 0.0172, 0.0172,0.6673,−0.0017
1	3	1.588 92	0.829 62	0	−0.034 48	0	0	$0.4612\pm0.222i$, 0.0834, 0.0834,0.1315,0.6814
1	3	1.588 92	1.283 57	0	−0.034 48	−0.070 98	0	$0.4612\pm0.222i$, 0.3266, 0.3266,5.9782,−3.1324
2	2	0.938 32	0.938 32	0	−0.000 26	−0.000 26	0	0.6671,0.0017,0.0017, 0.0005,0.0005,0.6671
2	2	0.934 98	0.934 98	0	0	0	0	0.6673,−0.0017,−0.0017, −0.0017,−0.0017,0.6673
2	3	0.938 32	0.829 62	0	−0.000 26	0	0	0.6671,0.0017,0.0659, 0.0659,0.1315,0.6814
2	3	0.934 98	0.829 62	0	0	0	0	0.6673,−0.0017,0.1315, 0.6814,0.0648,0.0648
3	3	0.829 62	0.829 62	0	0	0	0	0.6814,0.1315,0.1315, 0.6814,0.1315,0.1315

$$f(u_1,\ldots,u_6)=\sum_{i_1,\ldots,i_6} c_{i_1,\ldots,i_6} u_1^{i_1}u_2^{i_2}u_3^{i_3}u_4^{i_4}u_5^{i_5}u_6^{i_6} = \int_0^\infty e^{-t}F(u_1t,\ldots,u_6t)\,dt, \tag{7}$$

$$F(u_1,\ldots,u_6)=\sum_{i_1,\ldots,i_6}\frac{c_{i_1,\ldots,i_6}}{(i_1+\cdots+i_6)!}u_1^{i_1}u_2^{i_2}u_3^{i_3}u_4^{i_4}u_5^{i_5}u_6^{i_6}.$$

To analytically continue the Borel image of a function, we introduce a power series in an auxiliary parameter λ,

$$\tilde{F}(u_1,\ldots,u_6,\lambda)=\sum_{k=0}^{\infty}\lambda^k\sum_{i_1,\ldots,i_6}\frac{c_{i_1,\ldots,i_6}}{k!} \times u_1^{i_1}u_2^{i_2}u_3^{i_3}u_4^{i_4}u_5^{i_5}u_6^{i_6}\delta_{i_1+\cdots+i_6,k}, \tag{8}$$

to which we apply the Padé [L/M] approximation at the point $\lambda=1$. To calculate the β-functions in the two-loop approximation, we use the [2/1] approximant. Multicritical behavior is determined by the existence of a stable fixed point satisfying the system of equations

$$\beta_i(u_1^*,u_2^*,u_3^*,u_4^*,u_5^*,u_6^*)=0, \quad i=1,\ldots,6. \tag{9}$$

The requirement that a fixed point be stable reduces to the requirement that the eigenvalues b_i of the matrix

$$B_{i,j}=\frac{\partial\beta_i(u_1^*,u_2^*,u_3^*,u_4^*,u_5^*,u_6^*)}{\partial u_j} \tag{10}$$

lie in the right complex half-plane.

The resulting system of summed β-functions for each n and m contains a broad spectrum of fixed points. Table I lists stable fixed points of a system for the values of n and m of the uppermost interest to physics and a number of fixed points that are unstable in the two-loop approximation, which we will need in our further analysis. Table I also lists the eigenvalues of the stability matrix for the corresponding fixed points.

An analysis of the nature of the fixed points and their stability suggests the following: the presence of impurities in the system causes fluctuation decoupling of the order parameters and ensures only one type of stable multicritical behavior, the tetracritical behavior with the common symmetry of the system being $SO(n)\oplus SO(m)$. Here, for one-component order parameters $(n=m=1)$, the presence of impurities is important and leads to a critical behavior with exponents corresponding to those of the disordered Ising model.[16] As for the cases with $n=1$, $m=2$, and $n=2$, $m=2$, although calculations show that the fixed point with finite values of the impurity vertices u_4^* and u_5^* for both order parameters is stable, we believe that in the higher orders of the approximation the fixed point that is stable is the one at which, for the same general effect of decoupling of the order parameters, the values of the impurity vertices are finite only for one-component order parameters. A possible indication of this is, on the hand, the weak stability of fixed points of the first type accompanied by the weak instability of fixed points of the second type and, on the other, that a similar situation arises in the analysis of the effect of impurities on the critical behavior of systems with one order parameter in the two-loop approximation.[18] For $n,m\geqslant 3$ only the homogeneous fixed point that is stable coincides with the third-type point of a homogeneous system is tetracritical in nature. Thus, when the order parameters of the system are characterized by a number of components that is larger than, or equal to, two, the presence of impurities has no effect on the characteristics of their critical behavior and the multicritical behavior is tetracritical.

The phase diagrams for the Hamiltonian (1) of a homogeneous system in the mean-field approximation (i.e., without fluctuations) are well known (see, e.g., Ref. 3). For in-

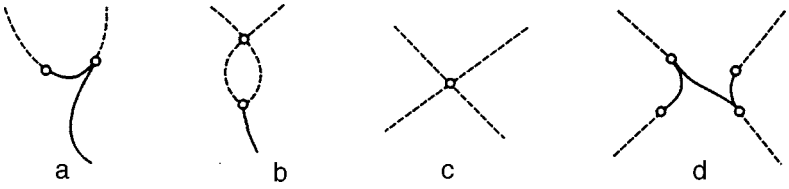

FIG. 2. Possible phase-diagram types. The solid lines correspond to curves representing first-order phase transitions, while the dashed lines correspond to curves representing second-order phase transitions.

stance, when $u_3^2 < u_1 u_2$ holds the tetracritical point is realized, with the result that there cam be a mixed phase with $\phi \neq 0$ and $\psi \neq 0$. In the opposite case, $u_3^2 \geqslant u_1 u_2$, the phase diagram has a bicritical point and there is no mixed phase. However, as shown in Refs. 3 and 5, allowance for fluctuations may dramatically change the phase diagram in the critical region. To establish whether this is the case, we must study the phase portrait of the system on the basis of the solution of the system of equations ($r = r_1 = r_2$)

$$r \frac{\partial u_i}{\partial r} = \beta_i(u_j), \tag{11}$$

which specifies the phase trajectories in the space of the vertices u_i. As $r \to 0$, depending on the bare values of the vertices, u_{i0}, either the phase trajectories leave the stability region of the Hamiltonian (1) and a first-order phase transition takes place, or they end up at a stable fixed point from the set of the points considered above with a definite symmetry of the system. In their motion the phase trajectories may cross regions in which the vertices meet the condition for tetra- or bicriticality. As a result, in the critical region on the phase diagrams corresponding to the mean-field theory there appear segments of curves representing first-order phase transitions.

Determining the regions for the values of the vertices u_i of the replica Hamiltonian (5) of a disordered systems in which the system is stable is complicated by the need to examine the limit $k \to 0$. However, allowing for the fact that the impurity vertices u_{40}, u_{50}, and u_{60} are proportional to $c(1-c)$, where c is the concentration of the impurity atoms, and limiting ourselves to weakly disordered systems, we assume that the former stability conditions ($u_{10} > 0$, $u_{20} > 0$, and $u_{30} > -(u_{10}u_{20})^{1/2}$) and the conditions for tetracritical or bicritical behavior are met, just as they are for a homogeneous system. The only additional condition imposed on the impurity vertices u_{40}, u_{50}, and u_{60} is that they be negative, a requirement that follows from the fact that the corresponding correlators in (4) are positive. In view of this, when we construct the phase portrait and the possible phase diagrams for weakly disordered systems and allow for fluctuations in the multicritical region in the solution of Eqs. (11) with the β-functions of (6) summed by the Padé–Borel method, it is sufficient to follow the changes in the values of u_1, u_2, and u_3 and assume that they depend on the vertices u_4, u_5, and u_6 parametrically.

Because the presence of impurities in systems with two order parameters substantially limits the possible types of stable fixed points, the number of possible types of phase diagrams changes dramatically in relation to that number in homogeneous systems. What is important in such changes is that for disordered systems there can be no phase diagram with a bicritical point. For disordered systems with interacting fields whose vertices have bare values that meet the bicriticality condition $u_{30}^2 \geqslant u_{10}u_{20}$, critical fluctuations and fluctuations of the local critical temperature destroy the stability of bicritical behavior and decouple the order parameters. As a result, the phase diagrams with bicritical behavior outside the critical region contain segments of curves representing first-order phase transitions, with the corresponding diagram being the one in Fig. 2a. The numerical solution of Eq. (11) shows that for all bare values of the vertices, u_{i0}, lying in the bicritical region there can be no phase trajectories that take these segments to a stable fixed tetracritical point. As a result, the phase diagram predicted by Laptev and Skryabin[13] and depicted in Fig. 2b is not realized. But if the bare values of the system's vertices satisfy the tetracriticality condition, the only possible diagrams are those depicted in Figs. 2c and 2d. Thus, of all the phase diagrams discussed in Refs. 3 and 5 and realized in homogeneous systems with two order parameters, only three types of phase diagram, those depicted in Figs. 2a, c, and d, are realized in disordered systems.

In conclusion we would like to express the hope that the differences in the multicritical behavior of homogeneous and disordered systems with competing order parameters established here will find their reflection in the design and analysis of experiments on the multicritical behavior of the corresponding systems.

This study was made possible by a Grant from the Russian Fund for Fundamental Research (Grant No. 97-02-16124).

*)E-mail: prudnikv@univer.omsk.su

[1] K. S. Aleksandrov, A. T. Anistratov, B. V. Beznosikov, and N. V. Fedoseeva, *Phase Transitions in Crystals of* ABX_3 *Halide Compounds* [in Russian], Nauka, Novosibirsk (1981).
[2] Y. Shapira, in R. Pynn and A. Skjeltorp (Eds.), *Multicritical Phenomena*, Plenum Press, New York (1984), p. 35.
[3] I. F. Lyuksyutov, V. L. Pokrovskiĭ, and D. E. Khmel'nitskiĭ, Zh. Éksp. Teor. Fiz. **69**, 1817 (1975) [Sov. Phys. JETP **42**, 923 (1975)].
[4] J. M. Kosterlitz, D. R. Nelson, and M. E. Fisher, Phys. Rev. B **13**, 412 (1976).
[5] V. V. Prudnikov, P. V. Prudnikov, and A. A. Fedorenko, JETP Lett. **68**, 950 (1998).
[6] G. Parisi, J. Stat. Phys. **23**, 49 (1980).
[7] E. Brezin, J. C. Le Guillou, and J. Zinn-Justin, in *Phase Transitions and Critical Phenomena*, Vol. 6, C. Domb and M. S. Green (Eds.), Academic Press, New York (1976), p. 125.
[8] V. S. Dotsenko, Usp. Fiz. Nauk **165**, 481 (1995) [Phys. Usp. **38**, 457 (1997)].
[9] S. Fishman and A. Aharony, J. Phys. C **12**, L729 (1979).
[10] V. E. Naĭsh, Yu. N. Skryabin, and V. N. Syromyatnikov, Fiz. Met. Metalloved. **52**, 1147 (1981).

[11] D. Mukamel, Phys. Rev. B **14**, 1303 (1976).
[12] Y. A. Izyumov, Y. N. Skryabin, and V. M. Laptev, Phys. Status Solidi B **87**, 441 (1978).
[13] V. M. Laptev and Yu. N. Skryabin, Fiz. Tverd. Tela (Leningrad) **22**, 2949 (1980) [Sov. Phys. Solid State **22**, 1722 (1980)].
[14] A. A. Lisyanskiĭ and A. É. Filippov, Ukr. Fiz. Zh. **32**, 626 (1987).
[15] G. Jug, Phys. Rev. B **27**, 609 (1983).
[16] I. O. Mayer, J. Phys. A **22**, 2815 (1989).
[17] D. J. Amit, *Field Theory: The Renormalization Group and Critical Phenomena*, McGraw-Hill, New York (1978).
[18] V. V. Prudnikov and A. N. Vakilov, Zh. Éksp. Teor. Fiz. **101**, 1853 (1992) [Sov. Phys. JETP **74**, 990 (1992)].

Translated by Eugene Yankovsky

Regularity criteria for linear chains

V. V. Lobzin*)

Institute of Terrestrial Magnetism, the Ionosphere, and Radio-Wave Propagation, Russian Academy of Sciences, 142092 Troitsk, Moscow Region, Russia

V. R. Chechetkin

Institute of Innovative and Thermonuclear Research, 142092 Troitsk, Moscow Region, Russia
(Submitted 26 March 1999)
Zh. Éksp. Teor. Fiz. **116**, 620–635 (August 1999)

The structural analysis of linear chains of arbitrary fixed shape is discussed in the context of a spectral approach. The shape of the chain is characterized by a set of scalar and pseudoscalar invariants, which remain constant under translations and rotations. The statistical properties of the set of invariants are compared with the analogous characteristics for a freely linked chain. The proposed criteria have the self-averaging property for relatively short ($\sim$100–300 links) chains and can be used to discern possible latent periodicities and symmetries in a system. As examples, two applications of the theory are considered: the structural analysis of chains generated by random walks on a cubic lattice and protein C_α backbones.
© *1999 American Institute of Physics.* [S1063-7761(99)01908-3]

1. INTRODUCTION

The conformations of long polymer molecules are characterized in a first approximation by the positions of the fixed atoms in repeating molecular groups.[1–3] The molecular backbone delineated in this way corresponds to a certain three-dimensional chain. For a whole series of polymer molecule—an extensive class of proteins at the head of the list—aggregation into a three-dimensional structure under specified physical conditions in solution is a strictly defined process (to within relatively insignificant thermal vibrations). The specificity of the aggregation of proteins imparts specificity to molecular identification and the selection of chemical reactions involving a particular protein. Existing databases house x-ray structural data on approximately 7500 protein structures. It is speculated that their number will exceed 12 000 in another two years.[4]

The physical analysis of fixed molecular backbones presents major difficulties because the structures in question simultaneously exhibit attributes of both ordered and random structures. Therefore, for example, the protein backbones (or C_α backbones when identified by the position of the distinguished carbon atom) are usually classified in internal terms on the basis of a mutual comparison of the structures.[5–7] Such global characteristics as the radius of gyration or the ratio of the principal axes of the inertia ellipsoid[1,8] are too crude and do not mirror the internal chirality of the molecules. On the other hand, typical proteins are comparatively short in length ($\sim 100-400$ links) and do not have a clearly defined self-similarity. For protein backbones, therefore, the scaling functions of various characteristics also fail to reflect detailed aggregation features (cf. Refs. 9 and 10).

The introduction of a quantitative measure of regularity for three-dimensional configurations of linear chains has an important information-bearing aspect, in that any given measure of complexity capable of reflecting the specificity of molecular interactions or evolutionary features can be linked to the specificity of aggregation.[11–14] The direct extension of classical information theory[15] to this case produces a ''combinatorial explosion.'' Suppose, for example, that a link of a linear chain is described by the polar angles $0 \leqslant \varphi \leqslant 2\pi$ and $0 \leqslant \theta \leqslant \pi$ (Ref. 1). If the space (φ, θ) is partitioned into M segments, n-link fragments can then assume M^n distinct conformations. Consequently, even for a coarse partition into quadrants ($M=6$) the possible analysis is already limited to three-link fragments, because the number M^n of distinct combinations must not exceed the total length L ($\sim 100-400$ for proteins). On the other hand, the adopted regularity criteria must reflect not only the local characteristics, but also the global characteristics of the chain structure. Finally, they must possess the self-averaging property for possibly shorter chains to enable the analysis of individual structures with $L \sim 100-400$.

In this paper we show that fairly simple criteria suitable for practical applications can be obtained by the spectral approach. A similar approach has been used previously to analyze symbolic sequences and genomic sequences of DNA.[16] This approach yields effective criteria for the assessment of global regularity[17] and can be used to discern latent periodicities in a system[18] or to analyze the character of correlations.[19] It will be shown below that despite the difference in the statements of the problem, many of the cited results[16–19] can be extended to the case of three-dimensional linear chains.

The paper is organized as follows. A general statement of the problem is set forth in Sec. 2, and a system of invariants characterizing the shape of the linear chains is constructed. Such invariants provide an effective means for discerning possible latent periodicities and symmetry in the system (Sec. 3). The statistical characteristics of the system

 © 1999 American Institute of Physics

of scalar and pseudoscalar invariants introduced in Sec. 2 are investigated in Sec. 4 for random, freely linked chains. A series of distinct regularity criteria can be introduced by means of comparison with the characteristics for random, freely linked chains (Sec. 5). In Sec. 6 the general theory is illustrated in specific examples of chains generated by random walks on a cubic lattice and the C_α backbones of protein structures. Finally, Sec. 7 closes with remarks on possible generalizations of the results.

2. SPECTRAL REPRESENTATION AND SYSTEM OF INVARIANTS FOR A LINEAR CHAIN

Let the Cartesian coordinates of the nodes of a chain (or fixed atoms in the polymer molecule) be given: $\{\mathbf{r}_m\}$, $m=1,\ldots,L$. Since the shape of a fixed chain configuration remains constant in the presence of translations and rotations of the system as a whole, the shape-related internal structural characteristics of the chain must be described in terms of a system of invariants that do not change in the presence of translations and rotations. In principle, any system of invariants can be used to describe shape, provided only that the number of independent invariants is sufficiently large (comparable with the number of internal degrees of freedom $3L-6$) and permits the shape to be characterized in detail. The system of invariants introduced in this section has the additional advantage that standard methods of spectral analysis are applicable.[20] It will be shown below that such a system of invariants admits a reasonably complete statistical analysis and gives suitable quantitative criteria of the regularity of a chain.

We introduce the vector Fourier harmonics

$$\boldsymbol{\rho}(q_n)=(L-1)^{-1/2}\sum_{m=1}^{L-1}\Delta\mathbf{r}_m\exp(-iq_nm),$$

$$q_n=2\pi n/(L-1),\quad n=0,1,\ldots,L-2, \tag{2.1}$$

$$\Delta\mathbf{r}_m=\mathbf{r}_{m+1}-\mathbf{r}_m. \tag{2.2}$$

Forming all possible scalar and mixed products from harmonics $\boldsymbol{\rho}(q_n)$ with different q_n, we obtain a system of scalar and pseudoscalar invariants that remain constant in translations and rotations.

For such a system of invariants to characterize the internal regularity of a chain, one other condition must be imposed on the system. We demonstrate the sense of the condition in a specific example. We distinguish bonds from 1 to Δm and perform parallel translation of the chain fragment between 1 and Δm from beginning to end. The vector Fourier harmonics for a chain with these fragments, $\boldsymbol{\rho}^{\mathrm{tr}}(q_n)$, and for the initial chain, $\boldsymbol{\rho}^{\mathrm{in}}(q_n)$, are interrelated by the equation

$$\boldsymbol{\rho}^{\mathrm{tr}}(q_n)=\exp(-iq_n\Delta m)\boldsymbol{\rho}^{\mathrm{in}}(q_n). \tag{2.3}$$

Since the internal regularity of both chains is identical, it follows from Eq. (2.3) that the sum of the wave vectors for the harmonics $\boldsymbol{\rho}(q_n)$ in the scalar and mixed products must be equal to $2\pi k$, where k is an integer. Finally, for the system of invariants we obtain

$$I(q_{n_1},\ldots,q_{n_r})=\mathrm{Inv}\{\rho_{\alpha_1}(q_{n_1})\ldots\rho_{\alpha_r}(q_{n_r})\},$$

$$q_{n_1}+\ldots+q_{n_r}=0 \bmod 2\pi. \tag{2.4}$$

In Eq. (2.4) the Greek-letter subscript for $\rho_\alpha(q_n)$ corresponds to projection in cartesian coordinates, $\alpha\in\{x,y,z\}$. We assume that all possible convolutions are eliminated from different components $\rho_\alpha(q_n)$ by means of Kronecker deltas $\delta_{\alpha\beta}$ and Levi Civita symbols $\varepsilon_{\alpha\beta\gamma}$. Convolutions with an odd number $\varepsilon_{\alpha\beta\gamma}$ give pseudoscalar invariants that change sign in mirror reflections, and convolutions with an even number of symbols $\varepsilon_{\alpha\beta\gamma}$ give scalar invariants that do not change in reflections.

Making use of the fact that the real-valuedness of $\Delta\mathbf{r}_m$ leads to the relation

$$\boldsymbol{\rho}(q_n)=\boldsymbol{\rho}^*(2\pi-q_n) \tag{2.5}$$

(the asterisk everywhere signifies the complex conjugate), we obtain the elementary scalar and pseudoscalar invariants

$$F(q_n)=\boldsymbol{\rho}(q_n)\cdot\boldsymbol{\rho}^*(q_n), \tag{2.6}$$

$$H(q_n)=i[\boldsymbol{\rho}(q_n)\times\boldsymbol{\rho}^*(q_n)]\cdot\boldsymbol{\rho}(0). \tag{2.7}$$

From now on we shall refer to the invariants $F(q_n)$ as structure factors. An exact sum rule can be obtained directly from the definition (2.1):

$$\sum_{q_{n_1}+\ldots+q_{n_r}=0 \bmod 2\pi}\rho_{\alpha_1}(q_{n_1})\ldots\rho_{\alpha_r}(q_{n_r})$$

$$=(L-1)^{(r-2)/2}\sum_{m=1}^{L-1}\Delta r_{m,\alpha_1}\ldots\Delta r_{m,\alpha_r}. \tag{2.8}$$

It follows at once from Eq. (2.8) that all pseudoscalars have zero mean spectral density, because the convolution of any of the symbols $\varepsilon_{\alpha\beta\gamma}$ with the right-hand side yields zero. For the structure factors $F(q_n)$ the harmonic with $q_n=0$ is proportional to the square of the span (distance between the ends) of the chain, and the average value of harmonics with $q_n\neq 0$ is

$$\overline{F}=\frac{1}{L-2}\sum_{n=1}^{L-2}F(q_n)$$

$$=\frac{1}{L-2}\sum_{m=1}^{L-1}(\mathbf{r}_{m+1}-\mathbf{r}_m)^2-\frac{(\mathbf{r}_L-\mathbf{r}_1)^2}{(L-1)(L-2)}. \tag{2.9}$$

Since the bond lengths in the chain are usually approximately equal, $\Delta\mathbf{r}_m^2\approx\mathrm{const}$, a given average spectral height $\overline{F}$ corresponds to a fixed span.

Based on the Wiener–Khinchin relation, the structure factors $F(q_n)$ can be set in correspondence with the correlation function

$$K(m_0)=\frac{1}{L-1}\sum_{n=0}^{L-2}F(q_n)\exp(-iq_nm_0),$$

$$m_0=0,\ldots,L-2. \tag{2.10}$$

The correlation function $K(m_0)$ can also be written in the form

$$K(m_0)=\frac{1}{L-1}\sum_{m=1}^{L-1}\Delta\mathbf{r}_m^c\cdot\Delta\mathbf{r}_{m+m_0}^c, \tag{2.11}$$

where $\Delta\mathbf{r}_m^c$ denotes cyclically continued differences in the coordinates:

$$\Delta\mathbf{r}_m^c=\begin{cases}\Delta\mathbf{r}_m, & 1\leqslant m\leqslant L-1,\\ \Delta\mathbf{r}_{m-L+1}, & L\leqslant m\leqslant 2L-2.\end{cases} \tag{2.12}$$

The definition (2.10), in turn, also produces various sum rules, the most important of which is[16]

$$\sum_{m_0=1}^{L-2}(K(m_0)-\overline{K})^2=\frac{1}{L-1}\sum_{n=1}^{L-2}(F(q_n)-\overline{F})^2, \tag{2.13}$$

where $\overline{K}$ is the average value of $K(m_0)$ with $m_0\neq 0$ [cf. Eq. (2.9)].

Condition (2.5) and definitions (2.6) and (2.10) impart symmetry to the corresponding spectra:

$$F(q_n)=F(2\pi-q_n),\quad K(m_0)=K(L-1-m_0). \tag{2.14}$$

Consequently, for $F(q_n)$ and $K(m_0)$ the spectra can be limited to their left halves only, $1\leqslant n\leqslant N$ and $1\leqslant m_0\leqslant N$, where

$$N=[(L-1)/2]. \tag{2.15}$$

In this expression the brackets signify the integral part of a number.

Analogous relations are readily obtained for the pseudoscalars $H(q_n)$ and other higher invariants as well.

3. STRUCTURE FACTORS, LATENT PERIODICITIES, AND LATENT SYMMETRIES

We first inquire how physical information about structure can be acquired directly from the spectra of the constructed sets of invariants. The common application of the Fourier transform to discern latent periodicities against a background of random contributions is well known.[20] Thus, we consider a chain composed of repeating l-link fragments. The vector harmonics $\boldsymbol{\rho}(q_n)$ then have the form

$$\boldsymbol{\rho}(q_n)=\mathbf{b}(q_n)+\exp(-ilq_n)\mathbf{b}(q_n)+\ \ldots\ +\exp(-i(M-1)lq_n)\mathbf{b}(q_n), \tag{3.1}$$

where $Ml=L-1$, and the Fourier transform for $\mathbf{b}(q_n)$ is obtained only by summation over the iteration length:

$$\mathbf{b}(q_n)=(L-1)^{-1/2}\sum_{m=1}^{l}(\mathbf{r}_{m+1}-\mathbf{r}_m)\exp(-iq_nm),\qquad q_n=2\pi n/(L-1),\quad n=0,\ldots,L-2. \tag{3.2}$$

The corresponding scalar structure factors are

$$F(q_n)=\mathbf{b}(q_n)\cdot\mathbf{b}^*(q_n)\frac{\sin^2(Mlq_n/2)}{\sin^2(lq_n/2)}, \tag{3.3}$$

and their spectrum consists of equally spaced peaks at $lq_n=2\pi k$, $k=0,\ldots,l-1$. Quasi-random variations of the length l can induce splitting of peaks and the suppression of peaks with $k\geqslant 2$. In general, therefore, latent periodicities can be discerned either by estimating the statistical significance for the sum of the equidistant harmonics or from the individual peaks, or by a combination of these two criteria.[18]

It is significant that the spatial Fourier harmonics disclose latent symmetries as well as latent periodicities. We demonstrate this statement in the example of the elementary repeated Mth-order symmetry corresponding to the group C_M (Refs. 21 and 22) (M-star configurations). The vector Fourier harmonics for this case have the form

$$\boldsymbol{\rho}(q_n)=\mathbf{b}(q_n)+\exp(-ilq_n)\hat{R}\mathbf{b}(q_n)+\ \ldots\ +\exp(-i(M-1)lq_n)\hat{R}^{M-1}\mathbf{b}(q_n). \tag{3.4}$$

Here $\mathbf{b}(q_n)$ are again given by Eq. (3.2), $Ml=L-1$, and the 3×3 rotation matrix $\hat{R}$ satisfies the condition

$$\hat{R}^M=\underbrace{\hat{R}\ \ldots\ \hat{R}}_{M}=\hat{I}, \tag{3.5}$$

where $\hat{I}$ is the unit matrix. The matrix $\hat{R}$ describes rotation about a certain axis $\mathbf{n}$ through the angle φ, $\hat{R}^2$ describes rotation through 2φ, etc. Condition (3.5) then gives $\varphi=2\pi/M$.

Let $\mathbf{n}$ be directed along the $\mathbf{z}$ axis of a Cartesian coordinate system. Introducing the combinations

$$\mathbf{b}_\pm=\mathbf{b}_x\pm i\mathbf{b}_y. \tag{3.6}$$

we can show that

$$\hat{R}\,\mathbf{b}_z=\mathbf{b}_z,\quad \hat{R}\,\mathbf{b}_\pm=\exp(\pm 2\pi i/M)\mathbf{b}_\pm. \tag{3.7}$$

Bearing in mind that

$$\mathbf{b}=\mathbf{b}_x+\mathbf{b}_y+\mathbf{b}_z=\frac{1}{2}(1-i)\mathbf{b}_+ +\frac{1}{2}(1+i)\mathbf{b}_-+\mathbf{b}_z, \tag{3.8}$$

and taking Eqs. (3.4), (3.7), and (3.8) into account, for $F(q_n)$ we obtain a series of equally spaced peaks at $lq_n=2\pi k$, $k=0,\ldots,l-1$, corresponding to $\mathbf{b}_z$, a series at $lq_n=2\pi k+2\pi/M$, $k=0,\ldots,l-1$, corresponding to $\mathbf{b}_+$, and a series at $lq_n=2\pi k-2\pi/M$, $k=1,\ldots,l-1$, corresponding to $\mathbf{b}_-$. Since we have $(L-1)/l=M$ and $q_n=2\pi n/(L-1)$, Mth-order rotational symmetry corresponds to peak harmonics labeled $n=Mk, Mk\pm 1$ (here k is an integer). If $\mathbf{b}_z=0$ holds, only split peaks with $n=Mk\pm 1$ are left. Analogous relations are easily derived for other symmetry subgroups with the identification of appropriate irreducible representations.[21,22]

If the repeating elements correspond, for example, to spiral structures, right- and left-handed helixes can be distinguished by the analysis of pseudoscalars. We also note that a system of pseudoscalars associated with the topological characteristics of the backbone of the chain may prove to be more useful in a number of applications.[23]

4. STATISTICAL CHARACTERISTICS FOR RANDOM, FREELY LINKED CHAINS

Since the transformation (2.1) is reversible, it can be regarded as a more suitable representation of the chain configuration. Comparing different structural characteristics for

fixed chain configurations with the average characteristics for a random, freely linked chain, we can obtain several quantitative regularity criteria for a chain having an arbitrary shape. In this section we discuss a general theory for analyzing the statistical characteristics of random chains.

For simplicity we shall assume below that all bonds are equal and have unit length, $\Delta\mathbf{r}_m \equiv \mathbf{n}_m$ (where $\mathbf{n}_m$ is the unit vector). The random analog for a chain with an arbitrary fixed shape is a random, freely linked chain with the same number of links and the same span $|\mathbf{r}_L - \mathbf{r}_1|$ [see Eq. (2.9)]. The statistical distributions for invariants of such a chain can be obtained by means of the characteristic function (cf. Refs. 1, 24, and 25)

$$Z = \frac{1}{\Omega}\int d\mathbf{n}\int d\mathbf{n}_1 \ldots \int d\mathbf{n}_{L-1}\,\delta(\mathbf{n}_1 + \ldots + \mathbf{n}_{L-1} - R\mathbf{n}) \times \exp\left(i\sum_{n=0}^{L-2} \mathbf{u}(q_n)\cdot\boldsymbol{\rho}(q_n)\right). \tag{4.1}$$

In the integration with the δ-function the quantity R is assumed to be fixed, the harmonics $\boldsymbol{\rho}(q_n)$ are given by Eq. (2.1) with $\Delta\mathbf{r}_m = \mathbf{n}_m$, and it is convenient to impose on the auxiliary vector variables $\mathbf{u}(q_n)$ a condition analogous to (2.5):

$$\mathbf{u}(q_n) = \mathbf{u}^*(2\pi - q_n). \tag{4.2}$$

The normalization factor Ω is equal to

$$\Omega = \int d\mathbf{n}\int d\mathbf{n}_1 \ldots \int d\mathbf{n}_{L-1}\,\delta(\mathbf{n}_1 + \ldots + \mathbf{n}_{L-1} - R\mathbf{n}). \tag{4.3}$$

Different averages can be obtained by differentiating Z:

$$\langle \rho_{\alpha_1}(q_{n_1})\rho_{\alpha_2}(q_{n_2})\ldots\rangle = \left.\frac{\partial^{\cdots} Z}{i\partial u_{\alpha_1}(q_{n_1}) i\partial u_{\alpha_2}(q_{n_2})\ldots}\right|_{\{\mathbf{u}(q_n)=0\}} \tag{4.4}$$

(angle brackets everywhere signify averaging over the ensemble of realizations).

Rewriting the exponential in (4.1) in the form

$$\exp\left(i\sum_{n=0}^{L-2} \mathbf{u}(q_n)\cdot\boldsymbol{\rho}(q_n)\right) = \exp\left(i\mathbf{g}_0\cdot\mathbf{n}R + i\sum_{m=1}^{L-1}\mathbf{n}_m\cdot\mathbf{g}_m\right), \tag{4.5}$$

where

$$\mathbf{g}_m = (L-1)^{-1/2}\sum_{n=1}^{L-2}\mathbf{u}(q_n)\exp(-iq_n m), \tag{4.6}$$

$$\mathbf{g}_0 = \mathbf{u}(0)/(L-1)^{1/2}, \tag{4.7}$$

we see at once that also for $m \neq 0$

$$\langle n_{m_1,\alpha_1} n_{m_2,\alpha_2}\ldots\rangle = \left.\frac{\partial^{\cdots} Z}{i\partial g_{m_1,\alpha_1} i\partial g_{m_2,\alpha_2}\ldots}\right|_{\{\mathbf{g}_m=0\}}. \tag{4.8}$$

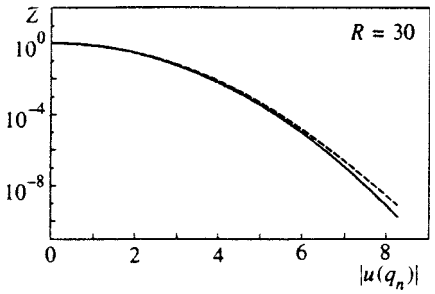

FIG. 1. Comparison of the exact expression (4.9) (solid curve) with the approximate expression (4.11) (dashed curve) for a chain having $L-1 = 100$ links and a fixed span R.

Invoking the integral Fourier transform for the δ-function in (4.1), we can reduce the expression for Z to a triple integral:

$$Z = \frac{1}{\Omega}\int d^3\mathbf{g}\,\frac{\sin(|\mathbf{g}-\mathbf{g}_0|R)}{|\mathbf{g}-\mathbf{g}_0|R}\prod_{m=1}^{L-1}\frac{\sin(|\mathbf{g}+\mathbf{g}_m|)}{|\mathbf{g}+\mathbf{g}_m|}. \tag{4.9}$$

From Eqs. (4.1)–(4.9), in particular, we obtain

$$\langle\boldsymbol{\rho}(q_n)\boldsymbol{\rho}^*(q_n)\rangle = \langle F(q_n)\rangle = \overline{F}, \tag{4.10}$$

where $\overline{F}$ is the spectral average defined by Eq. (2.9). Equation (4.10) reflects an important quasi-ergodic property: Averaging over the ensemble of random realizations is equivalent to averaging over the spectrum. In particular, all averages for pseudoscalars are equal to zero (see Sec. 2).

For $L \gg 1$, by analogy with Ref. 16, we can obtain an asymptotic cumulant expansion for Z using the exact sum rule (2.8). In the principal approximation with respect to $\sim L^{-1}$ we obtain

$$Z \approx \frac{\sin(|\mathbf{u}_0|R/(L-1)^{1/2})}{|\mathbf{u}_0|R/(L-1)^{1/2}}\exp\left(-\frac{\overline{F}}{3}\sum_{n=1}^{N}\mathbf{u}(q_n)\cdot\mathbf{u}^*(q_n)\right). \tag{4.11}$$

Condition (4.2) has been taken into account here, so that the summation in (4.11) is restricted to half the spectrum [see Eq. (2.15)].

In Fig. 1 the exact expression (4.9) is compared with the approximation (4.11) for the case when all the variables $\mathbf{u}(q_n)$ are equal to zero with the exception of one pair of complex conjugate variables for an arbitrarily selected value of q_n. The total number of links is $L-1=100$. It is evident from Fig. 1 that the exact expression almost coincides with the approximate expression in the range $Z \geqslant 10^{-6}$. Consequently, the majority of practical applications can be limited to the approximation (4.11). For large $|\mathbf{u}(q_n)|$, however, the asymptotic forms differ, and for $Z \leqslant 10^{-15}$ the discrepancy between the asymptotic forms is roughly an order of magnitude.

5. REGULARITY CRITERIA FOR LINEAR CHAINS

We now look at some specific inferences drawn from the general expression (4.11). The characteristic function Z is known to be associated with the multicomponent probability distribution function by the Fourier transform.[24,25] In the en-

suing discussion we are interested only in the amplitude distribution for harmonics with $q_n \neq 0$. In the principal approximation with respect to $\sim L^{-1}$ the corresponding multicomponent probability density function can be written in the form

$$p(F_1, \ldots, F_N) = p(F_1) \ldots p(F_N), \tag{5.1}$$

where the density function for the individual structure factors, $p(F_n)$, is given by the equation

$$p(F)dF = \frac{1}{2} f^2 \exp(-f) df, \tag{5.2}$$

$$f_n \equiv 3F(q_n)/\overline{F}. \tag{5.3}$$

The probability that the amplitude of an individual structure factor F_n will not be greater than F is equal to

$$P(F_n \leqslant F) = \int_0^F p(F')dF' = 1 - \left(1 + f + \frac{f^2}{2}\right) \exp(-f), \tag{5.4}$$

and, accordingly, the probability of exceeding F supplements (5.4):

$$P(F_n > F) = 1 - P(F_n \leqslant F). \tag{5.5}$$

The probability $P(F_n > F)$ dictates the fraction of structure factors in the half-spectrum, $N(F)/N$, with heights greater than the specified quantity F. Consequently, the proximity to a random distribution can be estimated by comparing the dependence of $N(F)/N$ on F for specific half-spectra with the dependence (5.5) for the random analogs. The statistical significance of deviations from the dependence (5.5) can be estimated by means of the Kolmogorov–Smirnov test.[24]

Equations (5.2) and (5.4) correspond to the probability distribution for the sum of three independent random variables with an identical Rayleigh distribution.[24] It is readily perceived that this result is a direct consequence of definition (2.6), because the scalar product $\boldsymbol{\rho}(q_n)\boldsymbol{\rho}^*(q_n)$ includes a sum over three components in Cartesian coordinates.

The statistical significance of the individual high peaks can be estimated by comparing them with the spikes in random half-spectra.[26] This problem has important bearing on the search for latent periodicities and symmetries in configurations of a general form (Sec. 3). The probability that the amplitudes of all structure factors in the half-spectrum will not exceed F is equal to

$$P(F_n \leqslant F; N) = [P(F_n \leqslant F)]^N, \tag{5.6}$$

where $P(F_n \leqslant F)$ is given by Eq. (5.4), and the probability that the amplitude of at least one of the N structure factors will exceed F is complementary to (5.6):

$$P(F_n > F: N) = 1 - [P(F_n \leqslant F)]^N. \tag{5.7}$$

The threshold values of the relative amplitudes (5.3) for different numbers N as determined by the conditions $P(F_n > F; N) = 0.1$ and 0.05 are given in Table I.

The global regularity of the chain configuration is conveniently evaluated by means of the spectral entropy (see the detailed discussion in Ref. 16):

TABLE I. Criteria for singular harmonics at various thresholds of significance in a half-spectrum of N harmonics.

N	f_{thr} $P = 0.1$	f_{thr} $P = 0.05$
50	10.33	11.20
100	11.17	12.02
150	11.65	12.50
200	11.99	12.84
250	12.25	13.10
300	12.47	13.31
350	12.65	13.49
400	12.80	13.64
450	12.94	13.78
500	13.06	13.90

$$S = -\sum_{n=1}^{L-2} \left(\frac{F(q_n)}{\overline{F}}\right) \ln\left(\frac{F(q_n)}{\overline{F}}\right). \tag{5.8}$$

Since the heights of the structure factors in the random half-spectra are distributed with relative uniformity, the quantity S for random configurations with a specified span of the chain (or with specified $\overline{F}$) attains a maximum (to within relatively small random deviations). Averaging (5.8) by means of the distribution function (5.1), we obtain

$$\langle S \rangle_{\text{random}} = -[\ln(1/3) + \Gamma'(4)/\Gamma(4)](L-2) \simeq -0.1575 \ldots (L-2), \tag{5.9}$$

where Γ and Γ' are the gamma function and its derivative. Analytical estimation for random deviations $\langle (\Delta S)^2 \rangle_{\text{random}}$ requires exceeding the limits of the approximation (5.1) and taking into account correlations between structure factors with different wave numbers. In the next section we obtain a numerical estimate for $\langle (\Delta S)^2 \rangle_{\text{random}}$ on the basis of numerical simulation results. Assuming that the deviations from $\langle S \rangle_{\text{random}}$ are Gaussian, from these data we can readily estimate the statistical significance of the deviation $(\langle S \rangle_{\text{random}} - S)/[2\langle (\Delta S)^2 \rangle_{\text{random}}]^{1/2}$ for the spectral entropy S corresponding to an arbitrary fixed configuration. Another useful quantity in applications is $\Delta S_{\text{rel}} = (\langle S \rangle_{\text{random}} - S)/|\langle S \rangle_{\text{random}}|$, where $|\langle S \rangle|$ is the absolute value of the average entropy for random configurations.[17] For random deviations it satsifies $\Delta S_{\text{rel}} \propto (L-2)^{-1/2}$ and is small for long random chains with $L \gg 1$, whereas for regular deviations we have $\langle (S \rangle_{\text{random}} - S) \propto (L-2)$. The ratio ΔS_{rel} can therefore be used to compare the total degree of regularity for chains with different numbers of links. The spectral entropy (5.8) also reflects certain information characteristics associated with the given chain configuration.

6. APPLICATIONS OF THE RESULTS

6.1. Chains generated by random walks on a cubic lattice

We illustrate applications of the results in two concrete examples. For the first we consider chains generated by a random walk on a cubic lattice. Figure 2 shows the half-spectra for a specific random realization with $L - 1 = 200$

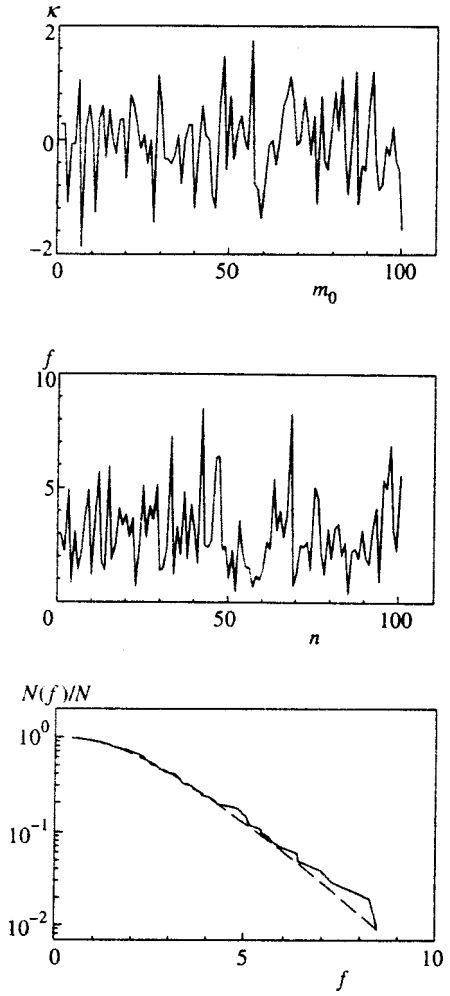

FIG. 2. Half-spectra for the correlation functions [in the variables (6.1)] and structure factors [in variables (5.3)] for a chain of $L-1=200$ links obtained from a random walk on a cubic lattice, as functions of f for the fraction of structure factors $N(f)/N$ exceeding a given level f. The dashed curve corresponds to Eq. (5.5).

chains. For the structure factors we use the dimensionless ratios (5.3), and for the correlation functions (2.10) we use the variables

$$\kappa(m_0)=(K(m_0)-\overline{K})/[2\langle(\Delta K)^2\rangle]^{1/2}, \tag{6.1}$$

$$\langle(\Delta K)^2\rangle=\overline{F^2}/3(L-1). \tag{6.2}$$

Here $\overline{K}$ is the average value of $K(m_0)$ with $m_0\neq 0$, and we have used the sum rule (2.13) and the distribution (5.2) to calculate the mean square deviation for random configurations $\langle(\Delta K)^2\rangle$. For random configurations the ratios $\kappa(m_0)$ with different m_0 can be regarded approximately as independent Gaussian variables. The lower part of Fig. 2 shows a graph of the fraction of structure factors $N(f)/N$ exceeding a given value f as a function of f. The dashed curve corresponds to Eq. (5.5). For the spectral entropy (5.8) we obtain $S/(L-2)=-0.155$ in accordance with the prediction (5.9).

Figure 3 shows the spectral entropy (5.8) and the mean square deviations obtained by averaging over 200 random realizations with $L-1=100, 150, \ldots, 400$ as functions of $(L-2)$. For the average entropy and the mean square deviations we obtain

$$\langle S\rangle=(-0.1567\pm 0.0003)(L-2), \tag{6.3}$$

$$\langle(\Delta S)^2\rangle=(0.092\pm 0.015)(L-2), \tag{6.4}$$

respectively, in good agreement with (5.9), $\langle S\rangle_{\text{random}}\approx -0.1575(L-2)$.

6.2. Structural characteristics for protein C_α backbones

Our second example is an analysis of the structural characteristics of protein C_α backbones (with respect to the position of a fixed carbon atom). We recall that the aggregation of the three-dimensional protein structure (or tertiary structure) evolves from so-called secondary structure elements.[1,3,27,28] Secondary structure elements are universal and comprise either fragments of right-handed α-helixes with a period $p\approx 3.6$ (in link-count units) or approximately plane fragments of β elements joined by connective fragments, which make up a separate group of elements.

Figure 4 shows data for representatives of two of the four principal structure classes of protein molecules.[27,28] For the molecular backbones we use a diagram representation,[29] in which α elements correspond to ribbon helixes, β elements correspond to ribbons with arrows, and connective elements correspond to fragments resembling pieces of wire. The half-spectra for the correlation functions and structure factors are given in the variables (6.1) or (5.3). The alphanumeric notation is consistent with the code used to label the given structure in the Brookhaven data base.[4]

Since both structures contain α-helixes, it is evident from Fig. 4 that they correspond to peaks in the spectra for the structure factors with periods $p=(L-1)/n\approx 3.6$, which correspond to latent periodicities. Structure 256BA has an approximate second-order symmetry, which quickly gener-

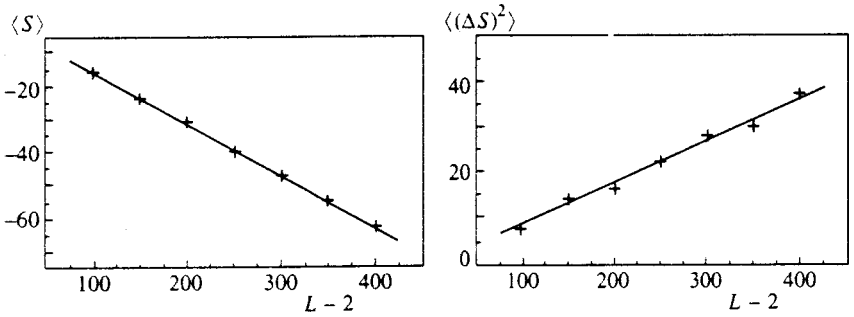

FIG. 3. Spectral entropy (5.8) and mean square deviations obtained by averaging over 200 random realizations for each value of L, as functions of $L-2$.

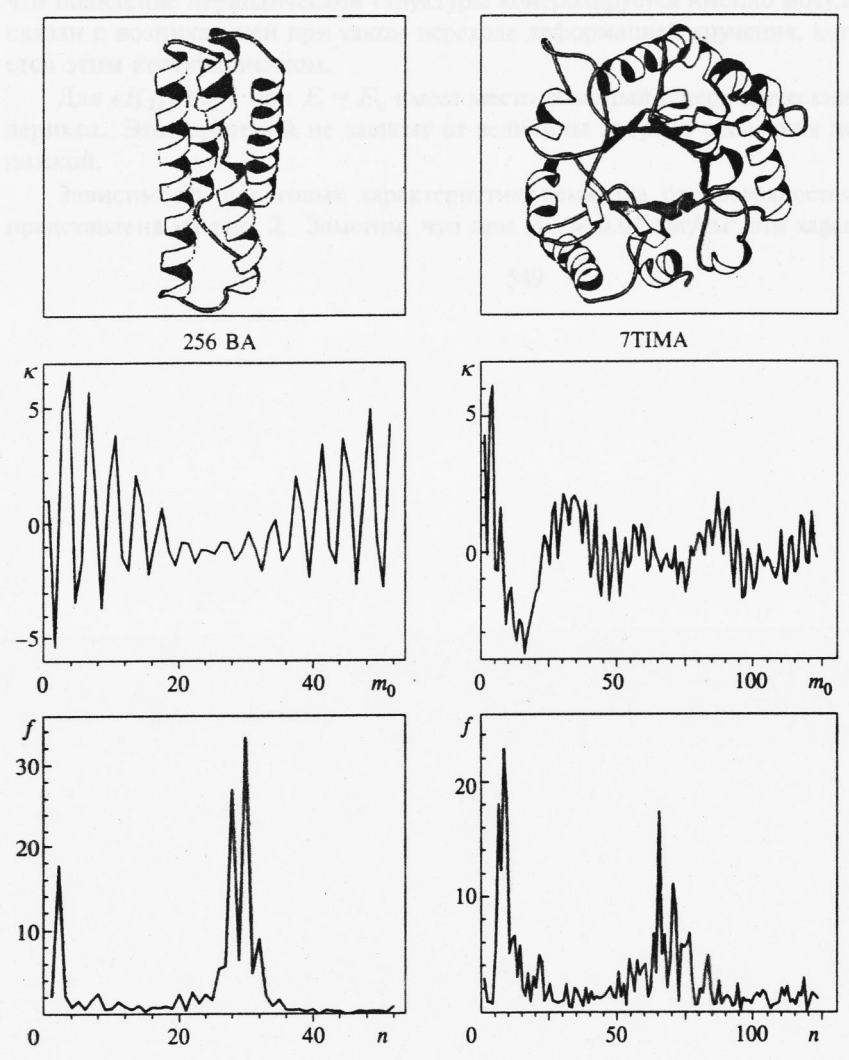

FIG. 4. Diagram representations of structures[29] and half-spectra for the normalized correlation functions κ and for the structure factors f for cytochrome b562 (structure 256BA) and triose phosphate isomerase (7TIMA).

ates a second peak at $n=2$ with amplitude $f=17.9$ (cf. Table I in Sec. 5). Structure 7TIMA, in turn, has an approximate eighth-order rotational symmetry, which generates characteristic split peaks: $n=7$, $f=12.3$; $n=8$, $f=22.9$; $n=9$, $f=17.6$.

The individual high peaks in the spectra for the structure factors do not always correspond exclusively to latent periodicities and symmetries. Some of the peaks with small wave numbers q_n can reflect a specific long-range order associated with the compaction of protein structures,[1,27,28] because their aggregation ties in with the imposition of a number of steric and energy constraints and has a cooperative character.

Figure 5 shows the fraction of the number of structure factors $N(f)/N$ exceeding a given value f as a function of f for structure 7TIMA (left graph). To separate the influence of latent periodicities and symmetries in this structure, we have discarded all high structure factors with $f \geqslant 10.0$ (it follows from Table I that such factors are in excess of typical random spikes), and for all other structure factors we have reevalu-

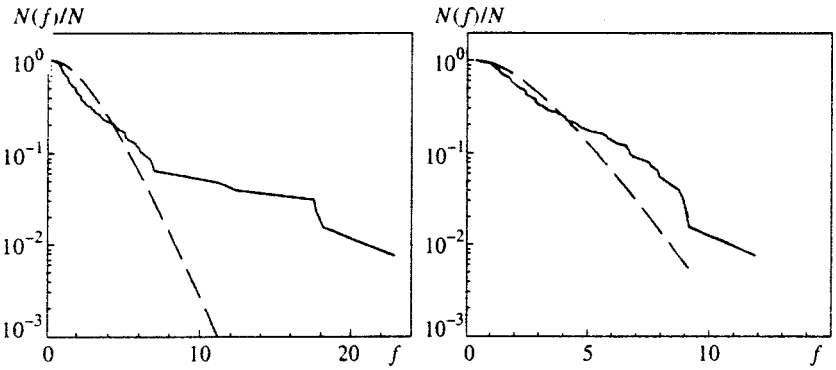

FIG. 5. Dependence on f of the fraction of structure factors $N(f)/N$ exceeding a given f for triose phosphate isomerase (7TIMA). The right graph is plotted after the rejection of harmonics with $f \geqslant 10$ and subsequent reevaluation of the average spectral values (see text). The dashed curves correspond to Eq. (5.5).

ated the spectral average. The resulting dependence is now described reasonably well by Eq. (5.5) (right graph in Fig. 5). Of course the agreement improves for more irregular structures. An analysis of structures lacking distinct symmetries and relatively low values of $\Delta S_{\rm rel}$ (see below) shows that for such structures the functional relation (5.5) is satisfied quite well even without the preliminary extraction of regularity effects.

Because of the high degree of regularity of the structures in Fig. 4, the corresponding values of $\Delta S_{\rm rel}=(\langle S\rangle_{\rm random}-S)/|\langle S\rangle_{\rm random}|$ are also large: 5.470 (256BA) and 1.996 (7TIMA). Clearly, the values of $\Delta S_{\rm rel}$ decreases as the order of symmetry and the complexity of the structure increase.

The formation of a tertiary protein structure from secondary structure elements corresponds to a characteristic decrease of the correlation functions at distances $m_0\sim 10-20$ (see Fig. 4). The same behavior is observed on the part of the correlations of other structural and physicochemical characteristics along the protein chain.[30,31]

7. CONCLUSION

It follows from the results of the study that the spectral approach provides a simple and effective means for investigating quantitatively the structural characteristics of linear chains having an arbitrary specified shape. The method can be used to discern possible latent periodicities and symmetries in a system, and beginning with relatively short chains, $L\geqslant 100$, it is already possible to estimate the total regularity of a chain in terms of the spectral entropy (5.8). Data assembled in existing data bases from x-ray structural examinations for polymer chains are simply incorporated into the scheme.

In the study we have used a freely linked chain as the initial random model. In the next approximation this model must take into account excluded volume effects.[1–3] Unfortunately, these effects are difficult to treat analytically, and lengthy numerical computations are involved. It is significant, however, that after the effects associated with latent periodicities and symmetries have been separated, the statistics (5.2)–(5.5) approximately describes the influence of the irregular structural background in real structures (see Fig. 5).

The formal unification of the digital representation of data and the proximity of the statistics for random analogs makes the method suitable for the investigation of possible correlations and the influence of physicochemical characteristics on the structure of the backbones of polymer molecules, whereupon the role of various factors in the aggregation of three-dimensional structures can be assessed in detail.

Both small-scale and large-scale structural characteristics are taken into account in the spectral approach, and the identification of long-range order effects is entirely natural.

Even for an elementary set of invariants such as the structure factors (2.6) the total number of independent invariants $(L-1)/2$ is found to be commensurate with the total number of degrees of freedom $3L-6$. They can therefore be used to identify essentially any fixed structure (to within the cyclic permutation of fragments). Sets of invariants for chains of unequal length can be compared by the standard padding-with-zeros technique.[20]

In summary, the method admits the sufficiently complete investigation of chains having an arbitrary fixed shape.

The authors are grateful to A. A. Vedenov and members of the seminar directed by M. A. Lifshits for a discussion of the results of the study and for helpful remarks. We are also indebted to Yu. P. Lysov for generous assistance in obtaining data on protein structures.

*)E-mail: lobzin@top.izmiran.troitsk.ru

[1] P. J. Flory, *Statistical Mechanics of Chain Molecules*, Wiley-Interscience, New York (1968).
[2] P. G. de Gennes, *Scaling Concepts in the Physics of Polymers*, Cornell Univ. Press, Ithaca, N.Y. (1979).
[3] A. Yu. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules*, AIP Press, New York (1994) [Russian orig. Nauka, Moscow (1989)].
[4] R. A. Laskowski, E. G. Hutchinson, A. D. Michie *et al.*, Trends Biochem. Sci. **22**, 488 (1997).
[5] C. A. Orengo, D. T. Jones, and J. M. Thornton, Nature (London) **372**, 631 (1994).
[6] L. Holm and C. Sander, Science **283**, 595 (1996).
[7] T. J. P. Hubbard, A. G. Murzin, S. E. Brenner, and C. Chothia, Nucleic Acids Res. **25**, 236 (1997).
[8] P. Biswas, A. Paramekanti, and B. J. Cherayil, J. Chem. Phys. **104**, 3360 (1996).
[9] X. Yi, Phys. Rev. E **49**, 5903 (1994).
[10] G. A. Arteca, Phys. Rev. E **54**, 3044 (1996).
[11] G. Jumarie, Physica A **184**, 499 (1992).
[12] T. G. Dewey, Phys. Rev. E **54**, R39 (1996).
[13] A. Fernandez and A. Belinky, J. Phys. A Math. Gen. **29**, L433 (1996).
[14] L. F. Luo, *Collected Works on Theoretical Biophysics*, Inner Mongolia Univ. Press, Hohhot (1997).
[15] C. Shannon, *Papers on Information Theory and Cybernetics* [Russian translation], IIL, Moscow (1963).
[16] A. Yu. Turygin and V. R. Chechetkin, Zh. Éksp. Teor. Fiz. **106**, 335 (1994) [JETP **79**, 186 (1994)]; V. R. Chechetkin and A. Yu. Turygin, J. Phys. A Math. Gen. **27**, 4875 (1994).
[17] V. R. Chechetkin, L. A. Knizhnikova, and A. Yu. Turygin, J. Biomol. Struct. Dyn. **12**, 271 (1994); V. R. Chechetkin and V. V. Lobzin, Phys. Lett. A **222**, 354 (1996).
[18] V. R. Chechetkin and A. Y. Turygin, J. Theor. Biol. **175**, 477 (1995); V. R. Chechetkin and V. V. Lobzin, J. Biomol. Struct. Dyn. **15**, 937 (1998).
[19] V. R. Chechetkin and A. Y. Turygin, J. Theor. Biol. **178**, 205 (1996); V. R. Chechetkin and V. V. Lobzin, J. Theor. Biol. **190**, 69 (1998).
[20] S. L. Marple Jr., *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, NJ (1987).
[21] M. Hamermesh, *Group Theory and Its Application to Physical Problems*, Addison-Wesley, Reading, Mass. (1962).
[22] R. L. Flurry, *Symmetry Groups: Theory and Chemical Applications*, Prentice-Hall, Englewood Cliffs, N.J. (1980).
[23] V. R. Chechetkin and V. V. Lobzin, Phys. Lett. A **250**, 443 (1998).
[24] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., Wiley, New York (1968) Vol. 2, 2nd ed., Wiley, New York (1971).
[25] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 2nd ed., North-Holland, Amsterdam-New York (1992).
[26] M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, Berlin-New York (1983).
[27] T. Creighton, *Proteins: Structures and Molecular Properties*, Freeman, New York (1993).
[28] V. M. Stepanov, *Structure and Functions of Proteins* [in Russian], Vysshaya Shkola, Moscow (1996).
[29] P. Kraulis, J. Appl. Crystallogr. **24**, 946 (1991).
[30] N. D. Socci, W. S. Bialek, and J. N. Onuchic, Phys. Rev. E **49**, 3440 (1994).
[31] O. Weiss and H. Herzel, Z. Phys. C **204**, 183 (1998).

Translated by James S. Wood

Screening of the static electric field in periodic Si–SiO_2 quantum wells: Effect on the electroinduced quadratic nonlinear-optical response

V. V. Savkin*)

M. V. Lomonosov Moscow State University, 119899 Moscow, Russia
(Submitted 31 August 1998)
Zh. Éksp. Teor. Fiz. **116**, 636–645 (August 1998)

The screening of a static electric field in periodic quantum wells is investigated theoretically. It is shown that the dependences of the potential on the quantum well number and the electric field from the applied voltage are oscillatory. The experimentally observed oscillatory dependences of the electroinduced quadratic nonlinear-optical response of periodic Si–SiO_2 quantum wells are explained assuming a specific morphology of the silicon layers.
© *1999 American Institute of Physics.* [S1063-7761(99)02008-9]

1. INTRODUCTION

Many effects occurring in solid-state microstructures are due to the presence of a static electric field. Semiconductor heterostructures (for example, metal–oxide–semiconductor structures) are of great interest. The distribution of the static field in such structures was first found in Ref. 1. Low-dimensional systems (for example, periodic quantum wells and superlattices), whose unique physical properties are related to quantum-well effects,[2] are structures of this class. Periodic quantum wells and superlattices consist of alternating thin (thickness ~ 10 Å) layers of different substances (for example, Si–SiO_2, GaAs–AlGaAs, InGaAs–AlAs, ZnSe–ZnCdSe, and so on). Spatial confinement of electron motion strongly influences the linear–optical and the nonlinear-optical properties of such systems, for example, second harmonic generation.[3] If the system of interest is in an external field, the nonlinear-optical effects depend strongly on the magnitude of the external perturbation and on the spatial properties of the systems under study.[4]

Experiments studying the azimuthal anisotropic dependences of the second harmonic and the influence of a static electric field on the SHG in periodic quantum wells have been described in Refs. 5 and 6. Periodic Si–SiO_2 quantum wells were prepared as follows. Silicon and silicon dioxide layers were alternately deposited on a 0.1-mm-thick silicon substrate in a manner so that 40 periods were obtained in the structure. The thicknesses of the layers and the parameters of the radiation used in the experiment are presented in Fig. 1. The result of the experiments described in Ref. 5 were the azimuthal anisotropic dependences of the second-harmonic intensity in transmission geometry for parallel polarization of the pump and second-harmonic waves. These dependences were found to be completely anisotropic and to possess two-fold symmetry.

In the experiments described in Ref. 6 the second-harmonic intensity at the maximum of the azimuthal curve was measured as a function of the constant voltage applied to the periodic quantum wells in the same geometry and with the same polarization of the pump and second-harmonic waves. Oscillatory dependences of the intensity of the electroinduced second harmonic on the voltage applied to the periodic quantum wells were obtained in these experiments (see Fig. 2). Thus far, no models explaining the nonmonotonic nonlinear-optical response in the systems studied have been proposed.

In the present paper we report the results of a theoretical study of the dependence of the quadratic optical response of periodic quantum wells in an external static electric field. In the calculations, attention is focused mainly on the analysis of the screening of the static electric field in a system of periodic, two-dimensional quantum wells and the morphology of the silicon layers of the structure. It is shown that the nonmonotonic dependence of the electroinduced second harmonic on the voltage applied to the structure is due to the essentially quantum screening of the electrostatic field in the periodic quantum wells. To find the nonlinear-optical response of the structure it is necessary to know, first, the nonlinear properties of the structure and, second, the distribution of the static electric field in periodic Si–SiO_2 quantum wells. Accordingly, we shall divide this problem into two parts. In Part 1 we shall consider the distribution of the static field in the structure (the electrostatic part), and in Part 2 we shall solve the nonlinear-optical problem (the nonlinear-optical part). In the electrostatic part of the problem attention will be focused mainly on calculations of the screening of the field in a system of periodic quantum wells and in a single quantum well, while in the nonlinear-optical part of the problem attention will be focused mainly on the symmetry properties of the dipole nonlinear susceptibility tensor $\chi_{ijkl}^{(3),D}$ of the silicon layers of the periodic Si–SiO_2 quantum wells.

2. ELECTROSTATIC PROBLEM

In the electrostatic part of this work we shall calculate the screening of the external static electric field in periodic Si–SiO_2 quantum wells. The field distribution in the structure as a whole will be found first (disregarding the detailed screening in each individual layer), after which the local field in a single quantum well will be calculated.

 © 1999 American Institute of Physics

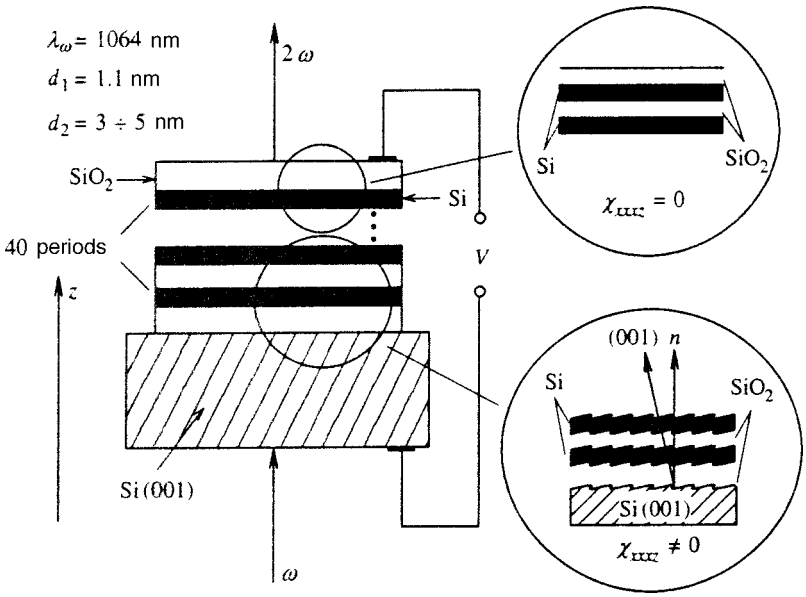

FIG. 1. The geometry of an experiment for studying the influence of a static electric field on optical second-harmonic generation in periodic $Si–SiO_2$ quantum wells and the morphology of the silicon layers. Insets: Different layers of the structure: flat and terrace-type. For these layers one component of the dipole nonlinear susceptibility tensor (χ_{xxxz}), which contributes to the second-harmonic generation from the terrace-type layers ($\chi_{xxxz} \neq 0$) and which does not contribute from the flat layers ($\chi_{xxxz} = 0$), is indicated.

The electrostatic problem can be solved by assuming that the electric current in the structure is zero. The periodic $Si–SiO_2$ quantum wells are treated as a collection of independent two-dimensional quantum wells (it is assumed that the electron wave functions do not overlap), but infinitesimal tunneling is assumed to exist. This makes it possible to introduce a single chemical potential for all quantum wells and thereby make thermodynamic equilibrium possible in the system. It is also assumed that the Fermi level of the system in the first silicon layer (closest to the substrate) lies at the center of the band gap. When an external electrostatic field is applied, screening of this field by the intrinsic charge carriers (holes in the valence band and electrons in the conduction band) occurs in the structure of the periodic quantum wells. It is assumed that there are no dopant charges in the periodic quantum wells. Electric neutrality of the system is ensured by the presence of charge on the outer electrodes and by the fact that the signs of the electron and hole charges are different. Further, assuming that the structure consists of 40 interacting charged layers and that $d_1 \ll d_2$ (where d_1 is the thickness of the silicon layer, and d_2 is the thickness of the silicon oxide layer), it becomes possible to switch from the continuous z coordinate (the z axis is directed perpendicular to the layers of the structure; see Fig. 1) to a discrete coordinate N — the quantum-well number. For definiteness, the potential $\varphi(N)$ and the field $E(N)$ are assumed to be their values at the center of the Nth quantum well.

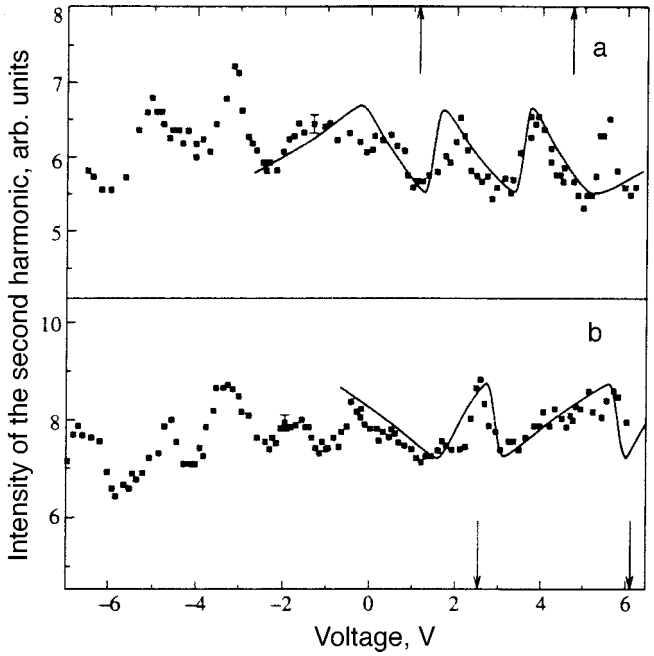

FIG. 2. Second-harmonic intensity as a function of the voltage applied to the periodic $Si–SiO_2$ quantum wells for samples with the following parameters: a — $d_1 = 11$ Å, $d_2 = 30$ Å and b — $d_1 = 11$ Å, $d_2 = 50$ Å; dots — experimental data, solid curve — theoretical calculation (d_1 and d_2 — thicknesses of the silicon and silicon oxide layers, respectively). The arrows indicate the voltage range corresponding to filling of the first quantum subband.

In Part 1 of the calculation of the screening of the electrostatic field in the structure, the distribution $\varphi(N)$ of the potential of the screened external field is found as a function of the quantum-well number. A computational method to the Thomas–Fermi method,[7] specifically, its discrete analog for the two-dimensional case, is used to solve this problem.

To solve the electrostatic part of the problem, we found the dependence of the charge-carrier density on the electrostatic potential in the two-dimensional layer at finite temperature. The formula was obtained assuming that only the first subband in the potential well is filled. The calculation was performed similarly to the calculation performed in Ref. 8 for the three-dimensional case. The final expression for this dependence is

$$n(\varphi,T) = -\frac{mkTe}{\pi\hbar^2}\ln\left(1+\exp\left[\frac{e\varphi+\mu}{kT}\right]\right) + \frac{mkTe}{\pi\hbar^2}\ln\left(1+\exp\left[\frac{-(e\varphi-\mu+\Delta)}{kT}\right]\right), \qquad (1)$$

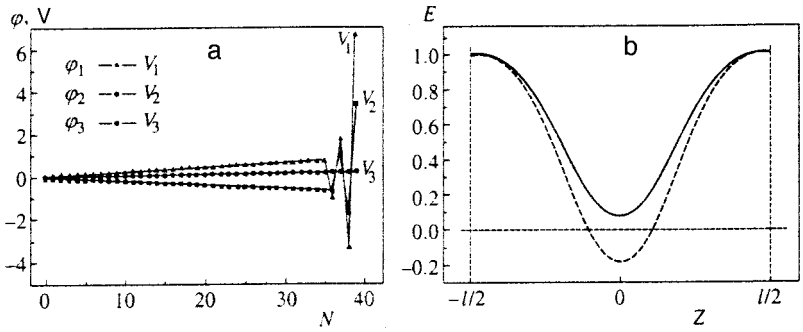

FIG. 3. Screening of the electric field in periodic quantum wells: a — potential φ as a function of the quantum-well number N ($d_1=11$ Å and $d_2=50$ Å; d_1 and d_2 — thicknesses of, respectively, the silicon and silicon oxide layers) for three differenr values of the applied voltage $V_1>V_2>V_3$ (the alternating sign of the derivative of the potential in the initial wells for different voltages V_1, V_2, and V_3 explains the appearance of the oscillatory dependence of the nonlinear-optical response) and b — intensity of the screened electrostatic field as a function of the coordinate of a single quantum well (the well width is 11 Å ; the external field is 1): dashed line — non-self-consistent calculation, solid line — self-consistent calculation (Hartree approximation).

where φ is the electric potential, T is the absolute temperature, Δ is the band gap, μ is the chemical potential, and m and $e>0$ are, respectively, the electron mass and charge (assumed to be the same for the electron and the hole). In this formula the first term corresponds to the electron density, and the second term corresponds to the hole density. The zero point of the potential is measured with respect to the first quantum level of the conduction band, so that the band gap Δ appeared only in the second term.

Assuming that the current flowing through the structure is zero, we make use of the Poisson electrostatic equation $\Delta\varphi=-4\pi\rho$. Switching to a description of the screened field as a function of the discrete variable N, we represent the three-dimensional charge density ρ as follows:

$$\rho(\mathbf{r})=\sum_{N=1}^{40}\delta[z-(d_1+d_2)N]n(\varphi(N)],$$

where $n[\varphi(N)]$ is the two-dimensional charge density determined from Eq. (1). If the Poisson equation is integrated once as a function of z from the center of the $(N-1)$st quantum well to the center of the Nth quantum well, then an equation relating the electric field $E(N)$ with the electron density $n(N)=n(\varphi(N))$ and therefore with the potential $\varphi(N)$ according to Eq. (1) is obtained:

$$-E(N)+E(N-1)=-2\pi[n(N)+n(N-1)]/\varepsilon_1, \qquad (2)$$

where ε_1 is the static permittivity of silicon. Taking account of the dielectric layer in the model will not change the form of the equation, since in this layer there are no free charges and therefore the electric field is constant. As a second equation we shall employ the formula for finite increments that expresses the potential difference in neighboring wells in terms of the mean field with allowance for the different static permittivities of silicon and the silicon oxide:

$$-\varphi(N)+\varphi(N-1)=\frac{1}{2}\left(d_1+d_2\frac{\varepsilon_2}{\varepsilon_1}\right)[E(N)+E(N-1)], \qquad (3)$$

where ε_2 is the static permittivity of silicon oxide. The equations (1), (2), and (3) written out above with the boundary conditions $\varphi(0)=0$ and $\varphi(40)=U_0$ form a system of equations whose solution determines the dependence of the potential $\varphi(N)$ on the quantum-well number.

The behavior of the solutions of the system of equations can be analyzed qualitatively by solving certain particular cases analytically.

First, in the case $n(\varphi,T)=0$ and $T=0$ (the case of a homogeneous insulator) this system possesses a trivial solution — the electrostatic potential is proportional to the quantum-well number. Second, if $\Delta=0$ and $T=0$ (the case of metallic layers at zero temperature), the system of equations becomes linear. A search for a solution in the form $\varphi(N)=Aa^N$ (where A and a are constants that depend on the properties of the structure and the boundary conditions) gives

$$\varphi(N)=C_1a^N+C_2a^{-N}. \qquad (4)$$

In the expression (4)

$$a=\left[1-\sqrt{\frac{d_1+d_2\varepsilon_2/\varepsilon_1}{\varepsilon_1a_0}}\right]\left[1+\sqrt{\frac{d_1+d_2\varepsilon_2/\varepsilon_1}{\varepsilon_1a_0}}\right]^{-1},$$

where $a_0=\hbar^2/me^2$ is the Bohr radius, and the constants C_1 and C_2 are determined from the boundary conditions. Hence it is evident that for $d_1+d_2\varepsilon_2/\varepsilon_1<\varepsilon_1a_0$ the solution is a monotonic function, while for $d_1+d_2\varepsilon_2/\varepsilon_1>\varepsilon_1a_0$ the solution assumes an oscillatory form.

Depending on the parameters of the samples, the monotonic or oscillatory behavior of the potential as a function of the quantum-well number also occurs when $T\neq0$ and $\Delta\neq0$.

Numerical solutions were obtained for periodic Si–SiO_2 quantum wells. Figure 3a shows the dependence of the potential on the quantum-well number for three different boundary conditions V_1, V_2, and V_3 (for a sample with $d_1=11$ Å and $d_2=50$ Å). The curves presented make it possible to trace the appearance of the oscillatory dependence of the electric field in the structure on the applied voltage. Figure 3a shows that as the voltage on the structure increases monotonically ($V_1>V_2>V_3$), the derivative of the potential in all quantum wells, except the last few, is sign-alternating (the potential either decreases or increases). This signifies the

presence of an oscillatory dependence of the applied voltage on the electrostatic field in the structure. This field determines the intensity of the electroinduced second-harmonic intensity.

The second part of the electrostatic problem is the calculation of the screening of the electrostatic field in a single quantum well. The need to study this problem arises because in the last few quantum wells of the structure, where the potentials are large compared with the initial wells, the screening in the quantum wells can strongly influence the electroinduced nonlinear-optical response. The local potential distribution in each quantum well has been studied in the Hartree approximation.[9] It was assumed here that each quantum well is a two-dimensional potential well with infinitely high walls, because for the parameters of the quantum wells employed in the experiment the Born parameter B is

$$B = U_0 m l^2/\hbar^2 \sim 10^2 \gg 1,$$

where l is the width and U_0 is the depth of the potential well. The calculation was performed with only the first quantum level filled. The electron density in the first subband, for which the second level is still unoccupied, was obtained from the expression for the two-dimensional density of states. This quantity then determines the voltage range of applicability of our calculations. The following expression was obtained as the Hartree equation (taking into account two terms in the perturbation theory series):

$$\delta\Psi = \sum_{m=1}^{2} \langle \Psi_1 | \hat{V} | \psi_{2m} \rangle \Psi_{2m}/(E_1 - E_{2m}), \tag{5}$$

where the perturbation operator

$$\hat{V} = e\left[E_z z - 4\pi e n \int_{-l/2}^{l/2} \Psi_1(z')\,\delta\Psi(z')|z'-z|dz' \right], \tag{6}$$

n is the electron or hole density in the first quantum level, and E_z is the external field. The function Ψ_{2m} is the wave function of electrons located in the level with number $2m$, and E_{2m} is the energy of this level.

The equation (5) is an integral equation with a degenerate kernel, and therefore it possesses an analytic solution. Multiplying the left- and right-hand sides of Eq. (5) by the corresponding expressions and integrating the equation so obtained, we obtain an algebraic system of equations, solving which gives the final answer. The unknown quantity in this equation is the correction $\delta\Psi$ to the electron wave function. Finding $\delta\Psi$ from Eq. (5), we can determine the coordinate dependence of the potential in the quantum well:

$$\Phi(z') = -4\pi e n \int_{-l/2}^{l/2} \Psi_1(z)\,\delta\Psi(z)|z-z'|dz. \tag{7}$$

The final result is not presented here because the formulas are too complicated. The dependence of the screened electrostatic field on the coordinate for the quantum well with $l = 11$ Å and only the first quantum level being filled is presented in Fig. 3b.

Accordingly, we have calculated the screening of the electrostatic field in periodic $Si-SiO_2$ quantum wells. In the calculation of the screening of the field, it was shown first that an oscillatory dependence of the field in the structure on the constant voltage applied to the periodic $Si-SiO_2$ quantum wells arises.

3. NONLINEAR–OPTICAL PROBLEM

The screening of the static electric field can strongly influence the nonlinear-optical properties of the structures, specifically, the generation of an electroinduced second harmonic. Since in centrosymmetric systems the generation of the second optical harmonic is forbidden in the dipole approximation, this effect will be described by the dipole cubic and quadrupole quadratic nonlinear susceptibilities:

$$P_i^{2\omega}(\mathbf{r}) = \chi_{ijkl}^{(3),D} E_j^{\omega}(\mathbf{r}) E_k^{\omega}(\mathbf{r}) \mathbf{E}_l^0(\mathbf{r}) + \chi_{ijkl}^{(2),Q} E_j^{\omega}(\mathbf{r})\,\partial E_k^{\omega}(\mathbf{r})/\partial x_l, \tag{8}$$

where $\chi_{ijkl}^{(3),D}$ and $\chi_{ijkl}^{(2),Q}$ are, respectively, the dipole and quadrupole nonlinear susceptibility tensors of the medium, $\mathbf{E}^{\omega}$ is the pump field, and $\mathbf{E}^0$ is the external static electric field. If it is assumed that the silicon layers of the structure, which are the source of the second-harmonic generation, are flat (∞/m symmetry; the allowed components are written out below), then for the indicated geometry of the experiment (Fig. 1) such layers will not be the source of the second harmonic, since these components of the susceptibility tensors are zero. Therefore it can be inferred that the silicon layers are not flat. In Ref. 5 it is indicated that there exist anisotropic dependences of the second-harmonic intensity in periodic quantum wells. This signifies the absence of symmetry for layers in a plane perpendicular to the z axis (Fig. 1). For this reason, to solve the nonlinear-optical part of the problem the morphology of the silicon layers of the structure is assumed to be as follows. Because the correct cut of the silicon substrate deviates from the exact $\langle 001 \rangle$ face, the inner silicon layers of the periodic quantum wells will possess a specific structure, reminiscent of terraces (in what follows, inner layers are layers close to the silicon substrate and outer layers are layers close to the interface with air). The height of a single step of the terraces equals the lattice constant of the crystal structure. However, when subsequent layers are deposited, this specific structure of the layers is lost, and the outer layers of the periodic quantum wells will be flat (see Fig. 1). The contribution of the inner layers to second-harmonic generation is due to the fact that these layers (terrace-type structure) of the periodic quantum wells possess the specific symmetry $2/m$. For the outer layers (flat and ∞/m symmetry) the $\chi_{ijkl}^{(3),D}$ tensor possesses 15 independent nonzero components with the indices i, j, k, and l: *xxxx*, *xxyy*=*xyxy*, *xxzz*=*xzxz*, *xyyx*, *xzzx*, *yxxy*, *yxyx*=*yyxx*, *yyyy*, *yyzz*=*yzyz*, *yzzy*, *zxxz*, *zxzx*=*zzxx*, *zyyz*, *zyzy*=*zzyy*, *zzzz*. For the inner layers with $2/m$ symmetry these and the following components of the tensor $\chi_{ijkl}^{(3),D}$ are different from zero: *xxxz*, *xxzx*=*xzxx*, *xyyz*, *xyzy*=*xzyy*, *xzzz*, *yxyz*=*yyxz*, *yxzy*=*yzxy*, *yyzx*=*yzyx*, *zxxx*, *zxyy*=*zyxy*, *zxzz*=*zzxz*, *zyyx*, *zzzx* (28 independent nonzero components in all). Next, by virtue of the geometry of the experiment (see Fig. 1: the constant field is directed along the z axis, and the pump and second-harmonic waves possess only x and y components) only the components $\chi_{xxxz}^{(3),D}$, $\chi_{xyyz}^{(3),D}$, $\chi_{yxyz}^{(3),D} = \chi_{yyxz}^{(3),D}$ contribute to second-

harmonic generation in the dipole approximation for layers with terrace-type structure, while for layers with a flat structure none of the components written out above contributes. To observe anisotropic dependences of the second harmonic[5] it is sufficient that for layers with terrace-type structure the angle between the normal to the substrate and the normal to the exact $\langle 001 \rangle$ face be of the order of a tenth of a degree. A detailed analysis of the effect of the magnitude of this angle on second-harmonic generation and on the existence of anisotropic dependences of the second harmonic can be found in Refs. 10 and 11. The number of layers with terrace-type structure is not important for us, since this number will ultimately influence only the intensity of the second harmonic, which for us is a free parameter. However, it is important to note that the terrace-type layers contributing to second-harmonic generation should emphasize the oscillatory behavior of the derivative of the potential as a function of the applied voltage (see Fig. 3), as is observed in the inner layers of the structure. In the calculations it was assumed that the 20 inner layers of the periodic quantum wells possess terrace-type structure (they contribute to second-harmonic generation), and the outer 20 layers have a flat structure (they do not contribute to second-harmonic generation).

In calculating the propagation factors (Green's functions, which are solutions to Maxwell's equations for a source in the form of a flat current screen), the reflections of the pump and second-harmonic waves only from the substrate and vacuum were taken into account, and for periodic $Si-SiO_2$ quantum wells an effective-medium model with permittivity $\varepsilon = 1.5$ was used.

A comparison with the experimental data can be made using the dependences obtained above for the screening of the field in periodic $Si-SiO_2$ quantum wells and the assumptions made about the nonlinear-optical properties of the medium.

4. RESULTS AND DISCUSSION

As a result, theoretical curves (solid curves in Fig. 2) of the second-harmonic intensity versus the voltage applied to the periodic $Si-SiO_2$ quantum wells were obtained. The theoretical curves were extended by 15% along the horizontal axis. The arrows in the figures show the range where the calculations are applicable. Thus it can be concluded that in the range of applicability of our model the theoretical calculation agrees qualitatively well with the experimental data. The discrepancy between the theoretical calculations and experimental data at voltages less than 1 V is probably due to the approximations used to calculate screening of the static field, specifically, the assumptions that only the first subband is filled and that there is no electric current through the structure. The nonmonotonic nonlinear-optical response is mainly due to the characteristic features of the screening of the electric field in the structure. The appearance of a potential that is an oscillatory function of the quantum-well number is due to the essentially quantum mechanical (Thomas–Fermi) description of the interaction of the silicon layers in the structure that excludes self-action.

It can be shown that the structure under study does not possess symmetry relative to the sign of the applied voltage. In the calculations it was implicitly assumed that tunneling of charge carriers is forbidden in the system of periodic $Si-SiO_2$ quantum wells at the right-hand end of the structure (where there is no substrate), which terminates with an oxide layer. At the same time, tunneling is allowed on the left-hand side (where a substrate is present), and the substrate and the periodic quantum wells possess a common Fermi level. For this reason, a jump is present in the chemical potential μ on the right-hand side of the system, between the last oxide layer and the electrode. Therefore the system is insensitive to a change in sign of the voltage. In any case, the oscillatory part of the derivative of the potential is located in the inner layers of the structure, i.e., the layers close to the substrate.

In summary, we have calculated the screening of the static electric field in periodic $Si-SiO_2$ quantum wells. A characteristic feature of the screening of the static field in a system of two-dimensional quantum wells and the existence of a specific $2/m$ symmetry of the inner silicon layers of the structure are responsible for the nonmonotonic dependence of the second-harmonic intensity on the voltage applied to the periodic quantum wells.

I wish to thank A. N. Rubtsov for an invaluable contribution to this work. I also thank O. A. Aktsipetrov, A. A. Nikulin, and A. A. Fedyanin for valuable remarks. This work was supported by INTAS-93 [Grant No. 370 (ext)], the Russian Fund for Fundamental Research (project 97-02-17919, 97-02-17923, DFG 98-02-04092), programs supporting the leading science schools (Project 96-15-96420), the Federal Integrated Program ''Educational–Scientific Center for Fundamental Optics and Spectroscopy,'' the programs ''Physics of Solid-State Nanostructures'' and ''Fundamental Metrology'' of the Ministry of Science and Technology of the Russian Federation, and the program of the Russian Ministry of Education ''Universities of Russia'' and the Soros Educational Program ''Soros Students.''

*)E-mail: volodya@astral.ilc.msu.su

[1] C. G. B. Garrett and W. H. Brattain, Phys. Rev. **99**, 376 (1955).
[2] A. I. Ekimov, F. Hache, M. C. Schanne-Klein *et al.*, J. Opt. Soc. Am. B **10**, 100 (1993).
[3] O. A. Aktsipetrov, A. I. Ekimov, and A. A. Nikulin, JETP Lett. **55**, 435 (1992).
[4] M. M. Fejer, S. J. B. Yoo, B. L. Byer *et al.*, Phys. Rev. Lett. **62**, 1041 (1989).
[5] O. A. Aktsipetrov, V. N. Golovkina, A. I. Zayats *et al.*, Dokl. Ross. Akad. Nauk **340**, 171 (1995) [Phys. Dokl. **40**, 12 (1995)].
[6] O. A. Aktsipetrov, P. V. Elyutin, A. A. Fedyanin *et al.*, Surf. Sci. **325**, 343 (1995).
[7] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (Pergamon Press, New York) [Russian original, Nauka, Moscow, 1989, p. 312].
[8] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon Press, New York) [Russian original, Nauka, Moscow, 1995, Part 1, p. 194].
[9] A. S. Davydov, *Quantum Mechanics*, 2nd edition (Pergamon Press, Oxford, 1976) [Russian original, Nauka, Moscow, 1973, p. 347].
[10] H. M. van Driel, Appl. Phys. A: Solids Surf. **59**, 545 (1994).
[11] G. Lupke, D. J. Bottomley, and H. M. van Driel, Phys. Rev. B **47**, 10389 (1993).

Translated by M. E. Alferieff
Edited by S. J. Amoretty

Effect of plasmon–phonon excitations on the coefficient of reflection from the surface of hexagonal silicon carbide

A. V. Mel'nichuk

Institute of Semiconductor Physics, Ukrainian National Academy of Sciences, 252650 Kiev, Ukraine
(Submitted 22 December 1998)
Zh. Éksp. Teor. Fiz. **116**, 646–654 (August 1999)

The coefficient $R(\nu)$ of reflection from the surface of 6H–SiC single crystals is studied in the case in which the long-wavelength optical vibrations of the lattice are coupled with the electron plasma. It is shown for the first time that the anisotropy of the properties of electrons and phonons in 6H–SiC gives rise to special features in the spectrum of the coupled vibrations and the transparency regions. It is found, in particular, that if the axis of the crystal lies in the polarization plane of the incident radiation ($0<\theta<90°$), for $30\,\mathrm{cm}^{-1}\leqslant\nu_{p\perp}<320\,\mathrm{cm}^{-1}$ the spectrum of $R(\nu)$ acquires three regions of transparency and opacity, and for $\nu_{p\perp}\geqslant 320\,\mathrm{cm}^{-1}$ four such regions, which are absent in an isotropic medium. The width of these regions is found to depend on the electron concentration in the conduction band and on the angle θ. © *1999 American Institute of Physics.* [S1063-7761(99)02108-3]

The reflection of electromagnetic waves from a superconducting surface is widely used in science and technology to manufacture various optical devices and to study the features of the energy and crystal structure of matter.[1] Gurevich and Tarkhanyan[2] found that for a uniaxial semiconductor the reflection spectrum has regions of transparency and opacity and that the number of these regions depends on the electron concentration and the orientation of the optical axis **C** with respect to the surface of the crystal. Other researchers (see Ref. 3) used IR-reflection spectroscopy to do a comprehensive study of isotropic structures in the event of formation of coupled plasmon–phonon excitations.

Single crystals of hexagonal silicon carbide (6H–SiC) belong to the space group C_{6V}^4 ($P6_3mc$) and, in contrast to ZnO, its plasmon subsystem is highly anisotropic.[4] The experimental spectra of reflection from the surface of 6H–SiC with a high electron concentration were first obtained by Il'in *et al.*[5] The method of modified faulted total internal reflection was used in Ref. 6 in the region of excitation of surface plasmon–phonon polaritons to obtain a reflection surface of heavily doped 6H–SiC. The effect of the anisotropy of the crystal lattice and effective electron masses and the damping coefficients for phonons (γ_f) and plasmons (γ_p) on the coefficient of reflection from the surface of 6H–SiC single crystals in the IR part of the spectrum was studied in Refs. 7 and 8. However, the literature contains no data on the effect of the anisotropy of the plasmon–phonon subsystem in 6H–SiC on the reflection coefficient for the ordinary and extraordinary waves.

In the present paper we will study the frequency dependence of the coefficient $R(\nu)$ of reflection from the surface of heavily doped single crystals of hexagonal silicon carbide (the 6H polytype) when there is a connection between long-wavelength optical vibrations and plasma vibrations ν_p of the free carriers for different orientations of the optical axis **C** of the crystal with respect to the crystal surface xy.

The experimental IR-reflection spectra in the 200 to $1400\,\mathrm{cm}^{-1}$ frequency range were measured with a SPECORD M-80 spectrometer and an attachment for reflection with a standard mirror. The reflection spectra were recorded by a polarizer with a degree of polarization equal to 0.98. All measurements were done at room temperature. The $5\times5\times0.5$ mm hexagonal singe crystals of 6H–SiC used in the experiments had a natural surface etched in hydrofluoric acid for 15 minutes. The electron concentration in the conduction band were determined by measuring the transparency of the sample at $\mathbf{E}\perp\mathbf{C}$ with $\lambda=0.628\,\mu$m. The data were found to be in good agreement with the results of measurements of the Hall effect for these samples.

If we allow for damping, the frequency dependence of the dielectric constant, $\varepsilon(\nu)$, in the plasmon–phonon interaction region can be written as[4]

$$\varepsilon_{\perp,\parallel}(\nu)=\varepsilon_{\infty\perp,\parallel}+\frac{\varepsilon_{\infty\perp,\parallel}(\nu_{L\perp,\parallel}^2-\nu_{T\perp,\parallel}^2)}{\nu_{T\perp,\parallel}^2-\nu^2+i\nu\gamma_{f\perp,\parallel}}-\frac{\varepsilon_{\infty\perp,\parallel}\nu_{p\perp,\parallel}^2}{\nu(\nu+i\nu\gamma_{p\perp,\parallel})},\quad(1)$$

where $\varepsilon_{\infty\perp,\parallel}$ are the components of the dielectric constant tensor perpendicular to the **C** axis of the crystal and parallel to that axis in the limit $\nu\to\infty$, and $\nu_{T\perp,\parallel}$ and $\nu_{L\perp,\parallel}$ are the frequencies of the transverse and longitudinal optical vibrations of the lattice polarized perpendicular to the axis and along the axis. Calculations of the coefficient of reflection from the flat surface of a ''semi-infinite'' uniaxial polar semiconductor in the IR region of the spectrum were done, with the absorption coefficient (damping) ignored, by the formulas given in Ref. 1:

$$R(\nu)=\left|\frac{1-n(\nu)}{1+n(\nu)}\right|^2,\quad(2)$$

where n is the refractive index, which is known to be related to the dielectric constant by the formula $n(\nu)=\sqrt{\varepsilon(\nu)}$. The refractive index of 6H–SiC has frequency intervals in which $n(\nu)$ is purely imaginary. In these intervals the crystal is not

 © 1999 American Institute of Physics

transparent at the given wavelength (total reflection, $R(\nu)=1$). The frequencies corresponding to the singularities of this equation can be found from the condition $\varepsilon(\nu)=0$.

When the electric field vector **E** is perpendicular or parallel to the plane determined by the wave vector **K** and the optical axis **C** of the crystal, transverse and longitudinal–transverse waves with refractive indices $n_\perp$ and $n_\parallel$, respectively, propagate in 6H–SiC single crystals.

For $\mathbf{E}\perp\mathbf{C}$, $\theta=0$ and 90° (case 1), the ordinary wave is transverse, θ is the angle between the crystal axis and the direction of propagation of the electromagnetic wave, and the x and y axes are on the surface of the sample. For the transverse wave the refractive index is

$$n_\perp(\nu)=\sqrt{\varepsilon_\perp(\nu)}. \tag{3}$$

Note that the refractive index of the transverse wave is independent of the direction of propagation. The vector **E** in the transverse wave is directed along the y axis. If we ignore the damping of plasmon and phonon vibrations, we can determine the frequencies of the zeros of the coefficient of reflection of the transverse wave and the so-called cut-off frequencies (the frequencies of longitudinal plasmon–phonon excitations), $\Omega_{L\pm}$ ($\theta=0$) or $\nu_{L\pm}$ ($\theta=90°$), by Eqs. (1)–(3) with $R(\nu)=0$ and $n_\perp^2=0$, respectively:

$$\nu_{01,2}=\frac{1}{\sqrt{2\alpha_{\infty\perp}}}\{\alpha_{0\perp}\nu_{L\perp}^2\pm[(\alpha_{0\perp}\nu_{L\perp}^2+\nu_{p\perp}^2)^2 - 4\alpha_{\infty\perp}\nu_{T\perp}^2\nu_{p\perp}^2]^{1/2}\}^{1/2}, \tag{4}$$

where $\alpha_{0\perp}=1-1/\varepsilon_{0\perp}$, with $\varepsilon_{0\perp}$ the value of dc dielectric constant tensor of the lattice, and $\alpha_{\infty\perp}=1-1/\varepsilon_{\infty\perp}$; and

$$\Omega_{L\pm}=\frac{1}{\sqrt{2}}\{\nu_{L\perp}^2+\nu_{p\perp}^2\pm[(\nu_{L\perp}^2+\nu_{p\perp}^2)^2 - 4\nu_{T\perp}^2\nu_{p\perp}^2]^{1/2}\}^{1/2} \quad \text{at } \theta=0. \tag{5}$$

An expression for $\nu_{L\pm}$ can be obtained if in Eqs. (3) and, respectively, (5) we replace $\perp$ by $\parallel$ everywhere. At $n_\perp=\infty$ we have resonance, $\nu_R=\nu_{T\perp}$. The incident wave undergoes total reflection in two frequency ranges:

$$\begin{aligned}&\nu<\Omega_{L-}, \quad \nu_{T\perp}<\nu<\Omega_{L+} \quad \text{at } \theta=0,\\ &\nu<\nu_{L-}, \quad \nu_{T\parallel}<\nu<\nu_{L+} \quad \text{at } \theta=90°,\end{aligned} \tag{6}$$

and between these ranges the crystal is transparent to the given wave. The coefficient $R(\nu)$ rapidly decreases from unity to zero and then again increases as the frequency grows.

The anisotropy in 6H–SiC (according to Ref. 9, for 6H–SiC $\nu_{p\perp}=2.682\nu_{p\parallel}$) leads to new nontrivial features in the reflection spectra. Thus, for $\mathbf{E}\parallel\mathbf{C}$, $0<\theta<90°$ (case 2), an extraordinary wave propagates in 6H–SiC. The extraordinary wave is a mixed wave, i.e., neither purely transverse nor purely longitudinal, and the group and phase velocities are not parallel and largely depend on the direction of propagation of the wave. The reflection coefficient is determined by Eq. (2), as it is in case 1. However, as shown by Gurevich and Tarkhanyan,[2] the refractive index of the longitudinal–transverse wave depends on the connection between the electromagnetic field and the phonon and plasmon subsystems of the uniaxial semiconductor and that with the angle θ:

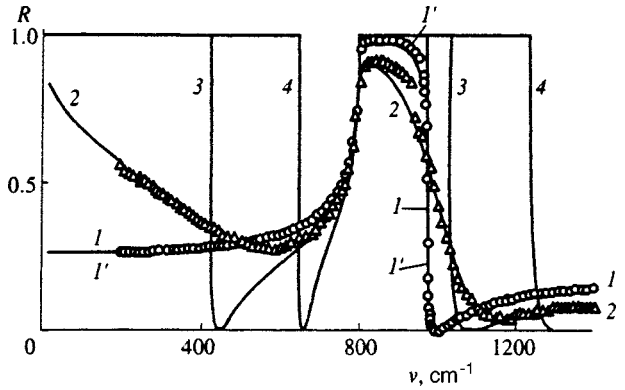

FIG. 1. The frequency dependence of the coefficient $R(\nu)$ of reflection of IR radiation from 6H-SiC single crystals at $\mathbf{E}\perp\mathbf{C}$ and $\theta=0$. Experimental results: ○—sample SN-1, △—sample SL-4. Results of calculations: *1′*—$\nu_p=15\ \text{cm}^{-1}$, $\gamma_p=18\ \text{cm}^{-1}$, and $\gamma_f=3\ \text{cm}^{-1}$; *2*—$\nu_p=740\ \text{cm}^{-1}$, $\gamma_p=830\ \text{cm}^{-1}$, and $\gamma_f=12\ \text{cm}^{-1}$.

$$\begin{aligned}&n_\parallel^2=\frac{a}{\nu^2(b+c)},\\ &a=(\nu^2-\Omega_{L+}^2)(\nu^2-\Omega_{L-}^2)(\nu^2-\nu_{L+}^2)(\nu^2-\nu_{L-}^2),\\ &b=(\nu^2-\Omega_{L+}^2)(\nu^2-\Omega_{L-}^2)(\nu^2-\nu_{T\parallel}^2)\frac{\sin^2\theta}{\varepsilon_{\infty\parallel}},\\ &c=(\nu^2-\nu_{L+}^2)(\nu^2-\nu_{L-}^2)(\nu^2-\nu_{T\perp}^2)\frac{\cos^2\theta}{\varepsilon_{\infty\perp}}.\end{aligned} \tag{7}$$

When the extraordinary wave propagates along the crystal axis ($\theta=0$ and $n_\parallel^2=\varepsilon_\perp$) or perpendicular to that axis ($\theta=90°$ and $n_\parallel^2=\varepsilon_\parallel$), it splits into a longitudinal and a transverse wave. The refractive index coincides with (3) for the transverse wave with its wave vector directed along the axis, and in the case $\theta=90°$ differs only in that $\perp$ is replaced by $\parallel$. For other directions $0<\theta<90°$ in the crystal there exists a longitudinal–transverse wave, which in contrast to such a wave in the isotropic case has new regions of transparency and opacity. The reason is that for $0<\theta<90°$ the connection between electromagnetic, plasma, and optical vibrations gives rise to coupled plasmon–phonon excitations, which are absent in the case mentioned earlier. The number and domain of existence of these excitations depend on the electron concentration in the conduction band and on the relative position of the frequencies $\nu_{T\perp,\parallel}$, $\nu_{L\perp,\parallel}$, $\nu_{L\pm}$, and $\Omega_{L\pm}$ (see Ref. 10).

Figure 1 depicts the frequency dependence of the experimental and calculated coefficients of reflection from the 6H-SiC surface at $\mathbf{E}\perp\mathbf{C}$ and $\theta=0$. The circles (○) and triangles (△) indicate the experimental reflection spectra of undoped ($n_0=5\times10^{15}\,\text{cm}^{-3}$, sample SN-1) and heavily doped ($n_0=1\times10^{19}\,\text{cm}^{-3}$, sample SL-4) silicon carbide (the 6H polytype). The calculated spectra of IR radiation reflection, $R(\nu)$ (curves *1*, *3*, and *4*) were obtained by Eqs. (1)–(3) in the range from zero to $1400\,\text{cm}^{-1}$ for $\nu_{p\perp}=15$, 550, and $1000\,\text{cm}^{-1}$ in the absence of damping in the phonon and plasmon subsystems. The zeros of the reflection coefficient

for the transverse wave can be found by formula (4) and are, respectively, at $\nu_{01}=0.86$ (curve *1*), 442 (curve *2*), and 655 cm^{-1} (curve *3*) for the low-frequency minimum and $\nu_{02}=998$, 1076, and 1322 cm^{-1} for the high-frequency minimum. The position of the high-frequency minimum can be determined more precisely compared to that of the low-frequency minimum, since the fact that damping is neglected begins to have an effect earlier at lower frequencies of IR radiation. The curves *1′* and *2* were calculated by equations that allow for the contribution of damping in the phonon and plasmon subsystems of 6H–SiC single crystals.[4,7] A comparison of the calculated spectra with the experimental spectra at $\mathbf{E}\perp\mathbf{C}$ and $\theta=0$ made it possible to determine ν_p, γ_p, and γ_f (see the caption to Fig. 1). Figure 1 shows that the small values of the plasmon (γ_p) and the phonon (γ_f) damping in the 0–800 and 990–1400 cm^{-1} frequency ranges have almost no effect on the spectrum of $R(\nu)$ (curves *1′* and *1*). The results of calculations disagree with the experimental points in the 800–940 cm^{-1} frequency range. An analysis shows that the discrepancy between the calculated IR-reflection spectra and the experimental data on $R(\nu)$ in the 800–940 cm^{-1} frequency range is due to the appearance of a thin (<0.05 μm) faulted layer. In this range, as shown in Ref. 11, IR radiation interacts very actively with the surface layer and the ''depth'' of this interaction dis at most 1 μm. The small discrepancy in $R(\nu)$ for sample SN-1 in the 900–1000 cm^{-1} frequency range is due to the effect on the reflection coefficient of vibrations in the plasmon subsystem. The analysis of the reflection spectrum for heavily doped 6H–SiC becomes much more complicated when we account for plasmon and phonon damping. The value of γ_p in 6H–SiC is comparable to the plasma frequency and can even by larger. The spectra were compared by the method of least squares while the square deviation (averaged over all the points) of the calculated reflection coefficient from the experimental reflection coefficient was being recorded. For sample SL-4, the plasmon damping coefficient was obtained by analyzing the variance in the reflection spectra and was found to be 830 cm^{-1}. Figure 1 shows ($\triangle$ and curve *2*) that the reflection spectrum changes significantly in the entire IR spectrum accessible to measurements. A computer experiment shows that as the damping coefficient for the plasmon subsystem increases the broad band of the reflection coefficient narrows in such a way that the high-frequency edge of the band smooths out and shifts to the low-frequency region. For $\gamma_{p\perp}\geqslant\nu_{p\perp}\approx\nu_{T\perp}$, the reflection band in the spectra of $R(\nu)$ is separated from the metallic reflection region at $\gamma_p=0$. The spectrum of $R(\nu)$ becomes a narrow band with a sharp peak in the region of ''residual rays'' at the frequency $\nu_{T\perp}$, with the reflection coefficient for 6H–SiC in this band amounting to 0.98. The fact that there is damping in the phonon subsystem reduces the reflection coefficient near the residual-ray peak. For sample SL-4, the optical phonon damping coefficient was accounted for by the method described in Ref. 7. With $\gamma_f=12$ cm^{-1}, satisfactory agreement between the calculated spectrum (curve *2*) and the experimental spectrum ($\triangle$) was achieved at $R_{\max}(\nu)=0.82$. The method employed earlier[4] for ZnO was used to determine, to within the value of the effective mass,[7,12] the mobility and

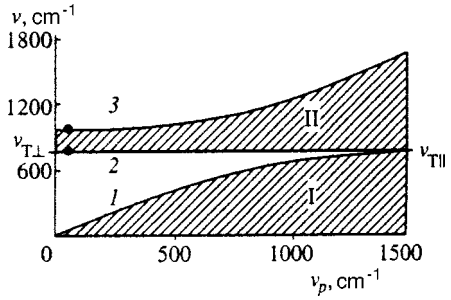

FIG. 2. Transparency and opacity regions in 6H–SiC for $\mathbf{E}\perp\mathbf{C}$ and $\theta=0$: curve *1*—Ω_{L-}, line *2*—$\nu_{T\perp}=797$ cm^{-1}, and curve *3*—Ω_{L+}; $\mathbf{E}\perp\mathbf{C}$ and $\theta=90°$: curve *1*—ν_{L-}, line *2*—$\nu_{T\parallel}=788$ cm^{-1}, and curve *3*—ν_{L+}. The two black dots represent the experimental values of the width of an opacity region for sample SN-1, $\theta=0$.

conductivity of 6H–SiC samples SL-4, which were found to be $\mu=27$ $cm^2\,V^{-1}\,s^{-1}$ and $\sigma=250$ $Ohm^{-1}\,cm^{-1}$, respectively. These data on μ and σ agree with the values obtained for 6H–SiC by other researchers.[12,13]

Figure 2 depicts the regions of transparency and opacity (dashed areas I and II) at $\theta=0$ and 90° for single crystals of 6H–SiC with different degrees of doping. The electron concentration in silicon carbide was varied from 10^{15} to 10^{20} cm^{-3}. An increase in the concentration of free charge carriers in this range for 6H–SiC enlarges the low- and high-frequency opacity regions, respectively, from zero to $\Omega_{L-}\rightarrow\nu_{T\perp}$ (I) and from $\nu_{T\perp}$ to Ω_{L+} (II). Here the frequency ν_0 shifts by 680 cm^{-1} for the low-frequency region and by 490 cm^{-1} for the high-frequency region. At electron concentrations $n_0>5\times10^{19}$ cm^{-3} single crystals of 6H–SiC, as Fig. 2 clearly shows, are opaque almost throughout the IR spectrum. The two black dots represent the experimental values of the width of an opacity region for sample SN-1 ($\Omega_{L+}-\nu_{T\perp}=173.1$ cm^{-1}). The results of calculations are in good agreement with the experimental results.

The shape of the spectrum of $R(\nu)$ depends on $n_0(\nu_p)$. The anisotropy of the effective electron masses and of the crystal lattice in 6H–SiC at $\mathbf{E}\parallel\mathbf{C}$ for $0<\theta<90°$ leads to an increase in the IR-reflection spectra of the number of transparency and opacity regions and to a dependence of this number on the electron concentration in the conduction band. For $\nu_{p\perp}<30$ cm^{-1} there are two transparency regions and two opacity regions. As the electron concentration is increased ($30\text{ cm}^{-1}\leqslant\nu_{p\perp}<320\text{ cm}^{-1}$) another pair of opacity and transparency regions begin to manifest themselves due to the fact that resonance and cut-off frequencies appear in the first transparency region (from the low-frequency side).

Figure 3 depicts the spectra of $R(\nu)$ for nitrogen-doped silicon carbide (the 6H polytype) at $\mathbf{E}\parallel\mathbf{C}$ for $\theta=30°$ (curve *1*) and $\theta=60°$ (curve *2*). The calculation of the frequency dependence of the reflection coefficient was done by formula (2) with allowance for (7) and the use of the data on sample SL-2 with $n_0=1.4\times10^{18}$ cm^{-3}. The cut-off and resonance frequencies for SL-2 are listed in Tables I and II. As θ increases, the resonance frequency ν_{R1} of the longitudinal–transverse wave increases from 134.9 cm^{-1} at $\theta=30°$ to 200.6 cm^{-1} at $\theta=60°$. The zeros of the reflection coefficient were calculated on a computer while solving Eq. (4). At

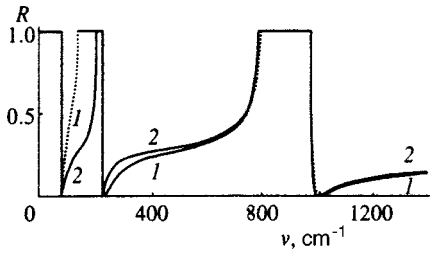

FIG. 3. Spectra of $R(\nu)$ in 6H–SiC (sample SL-2) at $\mathbf{E}\|\mathbf{C}$ for $\nu_{p\perp}=280\,\mathrm{cm}^{-1}$. Curve *1* corresponds to $\theta=30°$ and curve *2* to $\theta=60°$.

$\theta=30°$ we have $R_{\min}(\nu=0)$ at the frequencies $\nu_{01}=86\,\mathrm{cm}^{-1}$, $\nu_{02}=236\,\mathrm{cm}^{-1}$, and $\nu_{03}=1012\,\mathrm{cm}^{-1}$ (curve *1*), which change their values as θ increases to 60° and become $\nu_{01}=89\,\mathrm{cm}^{-1}$, $\nu_{02}=230\,\mathrm{cm}^{-1}$, and $\nu_{03}=1003\,\mathrm{cm}^{-1}$ (curve *2*). Three opacity regions in the IR-reflection spectra manifest themselves in the ranges $0-\nu_{L-}$, $\nu_{R1}-\Omega_{L-}$, and $\nu_{R2}-\nu_{L+}$, respectively. The width of the transparency regions depends on the direction in which the electromagnetic wave propagates, i.e., the angle θ, due to the angular dependence of the resonance frequencies at constant cut-off frequencies. For 6H–SiC, $\varepsilon_{\infty\perp}/\varepsilon_{\infty\|}<m_\|^*/m_\perp^*$ (here $m_\perp^*$ and $m_\|^*$ are the components of the electron effective mass tensor perpendicular to the crystal axis and parallel to the axis), with the result that the width of the transparency regions increases with the angle θ between the crystal axis and the wave vector. The region *3* for sample SN-2 begins at the frequency $\nu_{R1}=134.9\,\mathrm{cm}^{-1}$ for $\theta=30°$ and $200.6\,\mathrm{cm}^{-1}$ for $\theta=60°$ and ends at the frequency $\Omega_{L-}=226.8\,\mathrm{cm}^{-1}$.

Figure 4 depicts the calculated IR-reflection spectra for heavily doped 6H–SiC with the use of the parameters of sample SL-4 with $\nu_{p\perp}=740\,\mathrm{cm}^{-1}$ at $\theta=30°$ (Fig. 4a) and 60° (Fig. 4b). For $\nu_{p\perp}\geqslant 320\,\mathrm{cm}^{-1}$, another opacity and transparency region appears in the reflection spectra of 6H-SiC in the second transparency region (from the high-frequency side). Thus, the maximum number of opacity regions in the IR-reflection spectra for 6H–SiC is four ($i=1-4$); these regions are separated by transparency regions. The high-frequency ends of the first three transparency regions from the low-frequency side ($n_\|\to\infty$, $R(\nu)=1$) coincide with the resonance frequencies of the longitudinal–transverse wave (7) ($\nu=\nu_{R1,2,3}$), which can be found by solving the bicubic equation

TABLE I. Dependence of the cut-off frequency of the longitudinal-transverse wave on ν_p in 6H–SiC.

Frequency	SN-1		SL-2	SL-4		SL-8
$\nu_{p\perp}$, cm^{-1}	15	100	280	740	1000	1950
ν_{L-}, cm^{-1}	3.0	30.5	85.2	222.3	296.7	527.4
Ω_{L-}, cm^{-1}	8.2	82	226.8	538.7	645.5	761.7
ν_{L+}, cm^{-1}	964.1	964.4	966.0	977.9	990.2	1086.3
Ω_{L+}, cm^{-1}	970.1	971.8	983.9	1094.8	1234.7	2040.5

TABLE II. Dependence of the resonance frequency of the longitudinal-transverse wave on ν_p and θ in 6H–SiC.

	SL-2		SL-4		
θ, °	ν_{R1}, cm^{-1}	ν_{R2}, cm^{-1}	ν_{R1}, cm^{-1}	ν_{R2}, cm^{-1}	ν_{R3}, cm^{-1}
10	92.6	796.7	240.3	796.7	980.6
30	134.9	794.8	345.0	794.8	1002.3
60	200.6	790.3	489.5	790.3	1060.9
85	226.0	788.1	537.4	788.1	1093.7

$$(\nu^2-\Omega_{L+}^2)(\nu^2-\Omega_{L-}^2)(\nu^2-\nu_{T\|}^2)\frac{\sin^2\theta}{\varepsilon_{\infty\|}}$$

$$+(\nu^2-\nu_{L+}^2)(\nu^2-\nu_{L-}^2)(\nu^2-\nu_{T\perp}^2)\frac{\cos^2\theta}{\varepsilon_{\infty\perp}}=0. \qquad (8)$$

The frequencies ν_{R1} and ν_{R3} correspond to low-frequency and high-frequency coupled longitudinal–transverse plasmon–phonon excitations and depend on the electron concentration in the conduction band and on the angle θ. The frequency ν_{R2} varies from $\nu_{T\perp}$ as $\theta\to 0$ to $\nu_{T\|}$ as $\theta\to 90°$.

The above results suggest that in 6H–SiC there is degeneracy of the transparency and opacity regions from two regions when $\theta=0$ or 90° to four regions when $0<\theta<90°$. Figure 4 shows that the reflection spectrum for sample SL-4 has four cutoffs at the frequencies of the longitudinal plasmon–phonon excitations and three resonances at the frequencies of longitudinal–transverse plasmon–phonon excitations (Tables I & II). The zeros of the reflection coefficient were calculated numerically and are, respectively $\nu_{01}=225\,\mathrm{cm}^{-1}$, $\nu_{02}=551\,\mathrm{cm}^{-1}$, $\nu_{03}=984\,\mathrm{cm}^{-1}$, and $\nu_{04}=1143\,\mathrm{cm}^{-1}$ at $\theta=30°$ (Fig. 4a), and $\nu_{01}=231\,\mathrm{cm}^{-1}$, $\nu_{02}=543\,\mathrm{cm}^{-1}$, $\nu_{03}=998\,\mathrm{cm}^{-1}$, and $\nu_{04}=1113\,\mathrm{cm}^{-1}$ at $\theta=60°$ (Fig. 4b).

Figure 5 depicts the transparency and opacity regions for the heavily doped sample SL-4 of silicon carbide. For a given ν_p the cut-off frequencies $\Omega_{L\pm}$ and $\nu_{L\pm}$ are constants that do not depend on θ, i.e., the opacity regions 1 and 2 remain practically unchanged. As for the opacity regions 3 and 4 (and the transparency regions between them), they

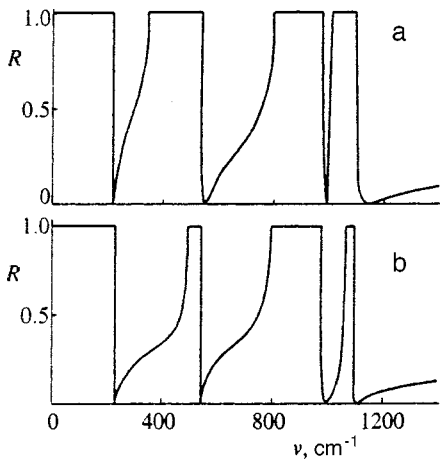

FIG. 4. Spectra of $R(\nu)$ in heavily doped 6H–SiC (sample SL-4) at $\mathbf{E}\|\mathbf{C}$: a—$\theta=30°$, and b—$\theta=60°$.

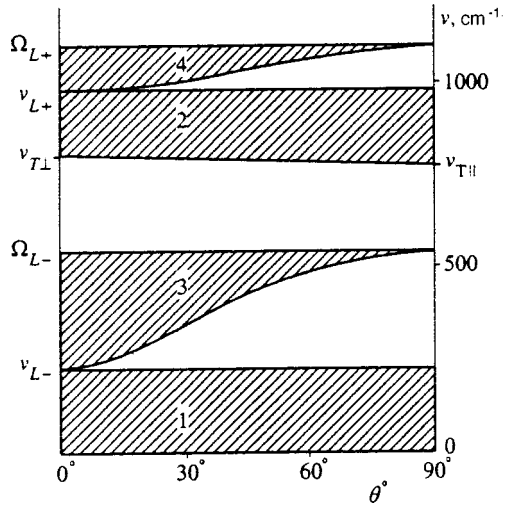

FIG. 5. Transparency and opacity regions in heavily doped 6H–SiC (sample SL-4).

narrow (and the transparency regions broaden) as θ increases, since their beginning coincides with the resonance frequency, which depends on θ. As θ increases, the second opacity region increases in size from $|\nu_{T\perp}-\nu_{L+}|=180.9\,\mathrm{cm}^{-1}$ as $\theta\to 0$ to $|\nu_{T\parallel}-\nu_{L+}|=189.9\,\mathrm{cm}^{-1}$ as $\theta\to 90°$.

The degeneracy of the transparency and opacity regions in doped single crystals of 6H–SiC has yet to be observed in experiments. The difficulty here is the large plasmon damping ($\gamma_p>\gamma_T$) in the samples. However, a computer experiment showed that as the damping in the plasmon subsystem gets weaker ($\gamma_p<0.5\nu_T$), the degeneracy of the transparency and opacity regions can be experimentally detected in heavily doped single crystals of 6H–SiC ($n_0>2\times10^{18}\,\mathrm{cm}^{-3}$). In Ref. 9 the method of modified faulted total internal reflection was used for the first time with 6H–SiC to obtain the experimental spectrum of surface plasmon–phonon polaritons of a new type existing in the third opacity region (Fig. 5).

Thus, we have studied the coefficients of reflection from the surface of a 6H–SiC single crystal with different degrees of doping. In the presence of a connection between long-wavelength optical vibrations and the electron plasma, the splitting of the transparency regions was detected for the first time in heavily doped 6H–SiC ($\mathbf{E}\|\mathbf{C}$, $0<\theta<90°$). For $30\,\mathrm{cm}^{-1}\leqslant\nu_{p\perp}<320\,\mathrm{cm}^{-1}$ three minima appear in the reflection spectra, and for $\nu_{p\perp}\geqslant320\,\mathrm{cm}^{-1}$ there are four minima and the same number of transparency and opacity regions (Figs. 3 and 4), which is due to the appearance of new coupled plasmon–phonon excitations in anisotropic silicon carbide (the 6H polytype). Figure 5 depicts two new, i.e., not previously studied, opacity regions (3 and 4) in 6H–SiC with $n_0=10^{19}\,\mathrm{cm}^{-3}$, with the first being in the interval $\nu_{L-}<\nu<\Omega_{L-}$ and caused by the presence of free electrons, and the second being in the interval $\nu_{L+}<\nu<\Omega_{L+}$ and caused by the screening of the field of the electromagnetic waves by the optical vibrations of the crystal lattice (it lies above the region of ''residual rays''). An increase in the electron concentration in the conduction band in 6H–SiC causes broadening of the opacity regions and to a narrowing of the transparency regions. When we have $\mathbf{E}\|\mathbf{C}$ and $\theta=0$ or 90°, the reflection spectra contain only two transparency and opacity regions, which agrees with the isotropic case $\mathbf{E}\perp\mathbf{C}$.

[1] Yu. I. Ukhanov, *Optical Properties of Semiconductors* [in Russian], Nauka, Moscow (1977).
[2] L. É. Gurevich and R. G. Tarkhanyan, Fiz. Tekh. Poluprovodn. **6**, 1895 (1972) [Sov. Phys. Semicond. **6**, 1631 (1973)].
[3] V. A. Kizel', *Reflection of Light* [in Russian], Nauka, Moscow (1973).
[4] E. F. Venger, A. V. Melnichuk, L. Yu. Melnichuk, and Yu. A. Pasechnik, Phys. Status Solidi B **188**, 823 (1995).
[5] M. A. Il'in, A. A. Kukharskiĭ, E. P. Rashevskaya, and V. K. Subashiev, Fiz. Tverd. Tela (Leningrad) **13**, 2478 (1971) [Sov. Phys. Solid State **13**, 2078 (1973)].
[6] A. V. Mel'nichuk, Poverkhnost' **7**, 76 (1998).
[7] A. V. Mel'nichuk and Yu. A. Pasechnik, Fiz. Tverd. Tela (Leningrad) **34**, 423 (1992) [Sov. Phys. Solid State **34**, 227 (1992)].
[8] F. Engelbrecht and R. Helbig, Phys. Rev. B **48**, 15 698 (1993).
[9] A. V. Mel'nichuk and Yu. A. Pasechnik, Fiz. Tverd. Tela (St. Petersburg) **40**, 636 (1998) [Phys. Solid State **40**, 582 (1998)].
[10] A. V. Mel'nichuk, L. Yu. Mel'nichuk, and Yu. A. Pasechnik, Zh. Tekh. Fiz. **68**(1), 58 (1998) [Tech. Phys. **43**, 52 (1998)].
[11] A. V. Goncharenko, A. V. Mel'nichuk, Yu. A. Pasechnik, and A. A. Shilov, manuscript deposited at the Ukrainian Scientific Research Institute of Scientific and Technical Information, Vol. 834, 18 (1991).
[12] H. Harima, S. Nakachima, and T. Uemura, J. Appl. Phys. **78**, 1996 (1995).
[13] *Silicon Carbide* (Proc. Int. Conf. on Silicon Carbide, University Park, Pennsylvania, October 20–23, 1968), H. Henisch and R. Roy (Eds.), Pergamon Press, New York (1969).

Translated by Eugene Yankovsky

Superconductivity of strongly correlated electrons in copper and ruthenium oxides within the $t-J-I$ model

E. V. Kuz'min

Krasnoyarsk State University, 660075 Krasnoyarsk, Russia

S. G. Ovchinnikov*) and I. O. Baklanov

L. V. Kirenskiĭ Institute of Physics, Siberian Branch of the Russian Academy of Sciences, 660036 Krasnoyarsk, Russia

(Submitted 15 January 1999)
Zh. Éksp. Teor. Fiz. **116**, 655–670 (August 1999)

We propose a $t-J-I$ model with direct ferromagnetic exchange I to explain the superconductivity of copper oxides and the ruthenate Sr_2RuO_4 on the basis of the analysis of the electronic structure of these substances. We analyze the possible p- and d-type superconducting solutions. Solutions of the s type with singlet pairings are impossible in the strong-electron-correlations regime, and p-type solutions correspond to triplet superconductivity and is formed near the ferromagnetic instability threshold in ruthenates. The solution with the $d_{x^2-y^2}$ symmetry near the antiferromagnetic instability threshold corresponds to copper oxides. We also discuss the reason for the high values of the superconducting transition temperature ($T_c \sim 100$ K) in copper oxides and the low values ($T_c \sim 1$ K) in ruthenates. © *1999 American Institute of Physics.* [S1063-7761(99)02208-8]

1. INTRODUCTION

Even in the earliest research on superconductivity in Sr_2RuO_4 (Ref. 1) attempts were made to compare the results with those for copper oxides, the reason being that Sr_2RuO_4 and La_2CuO_4 have similar structures. At the same time, the values of the superconducting transition temperature T_c differ dramatically: $T_c \sim 1$ K for ruthenates and $T_c \sim 100$ K for cuprates. Latter it was found that magnetic and superconducting properties differ, too. While in cuprates superconductivity occurs against the background of strong antiferromagnetic fluctuations, ruthenates are close to the ferromagnetic instability threshold.[2] The Cooper pairs in the superconducting state in Sr_2RuO_4 are of the p type,[3] in contrast to the d-type pairing in cuprates. For Sr_2RuO_4, solutions of the p type were obtained by Mazin and Singh[4] as a result of band-structure calculations. The literature on superconductivity in the strong-electron-correlation regime within the Hubbard and $t-J$ models is vast, and this is reflected in Refs. 5 and 6.

In the present paper we propose a $t-J-I$ model that takes into account the antiferromagnetic (J) and ferromagnetic (I) exchange interactions simultaneously. Three possible types of superconducting state are studied by this model: single states of the s and d types and triple states of the p type. In addition to the ordinary self-consistency equations for the gap and the chemical potential, in the strong-electron-correlation regime there is an additional self-consistency condition (constraint) that excludes two-particle states (''doubles'') at a single site. Solutions of the s type to not meet this condition, but solutions of the $d_{x^2-y^2}$ and p-types do. The coupling constant of d-type solutions is determined primarily by antiferromagnetic exchange interaction, while the coupling constant of p-type solutions is determined primarily by ferromagnetic exchange interaction. The special features of the momentum-dependence of the gap in the equations for T_c lead to $T_c \sim 100$ K for d-type pairings and to $T_c \sim 1$ K for p-type pairings, with the coupling constants being equal.

2. THE $t-J-I$-MODEL HAMILTONIAN

The Hamiltonian of the $t-J-I$ model can be written

$$H=\sum_{\mathbf{f}\sigma}(\varepsilon-\mu)X_{\mathbf{f}}^{\sigma\sigma}-t\sum_{\mathbf{f}\boldsymbol{\delta}\boldsymbol{\sigma}}X_{\mathbf{f}}^{\sigma 0}X_{\mathbf{f}+\boldsymbol{\delta}}^{0\sigma}$$

$$+J\sum_{\mathbf{f}\boldsymbol{\delta}}K_{\mathbf{f},\mathbf{f}+\boldsymbol{\delta}}^{(-)}-I\sum_{\mathbf{f}\boldsymbol{\delta}}K_{\mathbf{f},\mathbf{f}+\boldsymbol{\delta}}^{(+)}, \tag{1}$$

$$K_{\mathbf{fm}}^{(\pm)}=\mathbf{S_f}\cdot\mathbf{S_m}\pm\frac{1}{4}n_{\mathbf{f}}n_{\mathbf{m}}, \quad X_{\mathbf{f}}^{\uparrow\uparrow}+X_{\mathbf{f}}^{\downarrow\downarrow}+X_{\mathbf{f}}^{00}=1. \tag{2}$$

The Hamiltonian (1) is given on a lattice of N sites ($\mathbf{f}$ and $\mathbf{m}$ are the lattice sites) with periodic boundary conditions and with z nearest neighbors, and $\boldsymbol{\delta}$ is the vector connecting the nearest neighbors. The Hamiltonian describes a system of N_e electrons in the subspace of local states $|0\rangle$ (holes or vacant lattice sites) and $|\sigma\rangle$ (one-electron states with a spin projection $\sigma=\uparrow$ or $\sigma=\downarrow$), so that $0\leqslant N_e\leqslant N$. In this basis the states and the transitions between them are described by the Hubbard X-operators $X_{\mathbf{f}}^{pq}=|p\rangle\langle q|$ acting on the states $|0\rangle$ and $|\sigma\rangle$ (doubles are excluded automatically), and $\mathbf{S_f}$ and $n_{\mathbf{f}}$ are the operators of the spin and number of particles at site $\mathbf{f}$. The signs in (1) are chosen so that all the parameters, t, J, and I, are positive. The energy ϵ of the one-electron level will be assumed to be zero.

 © 1999 American Institute of Physics

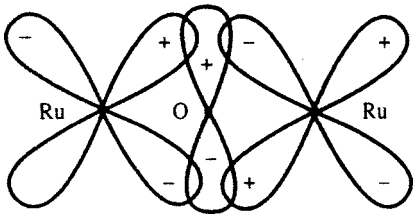

FIG. 1. The $(d_{xy}-p)-\pi$ bond in Sr_2RuO_4.

As in the $t-J$ model, the antiferromagnetic exchange is an indirect cation–anion superexchange. The ferromagnetic exchange I has the form common to this model and is generated by direct overlap of cation d-orbitals of neighboring sites. In cuprates, where the electronic states near the Fermi level are formed primarily by the $d_{x^2-y^2}-p-\sigma$ bond, direct overlap can be ignored, $J \gg I$. A characteristic feature of the electronic structure of Sr_2RuO_4 is that near the Fermi surface the states are formed by the $(t_{2g}-p)-\pi$ bond. In this case there is indirect 180-degree cation–anion–cation interaction J and direct overlap of t_{2g}-orbitals of neighboring cations (Fig. 1). According to Goodenough,[7] for Ru^{4+} the antiferromagnetic exchange is small, so that the model with $I>J$ corresponds to Sr_2RuO_4. The importance of strong electron correlations for cuprates is well know, and for ruthenates the need to allow for such correlations stems from the large effective mass of the electrons belonging to the γ-band formed by d_{xy}-orbitals, $m \approx 12m_e$, a fact corroborated by experiments involving quantum oscillations.[8] According to Riseman *et al.*,[9] it is the γ-band that forms the superconducting state. The other d-orbitals of ruthenium form the hole α-band and the electron β-band, which in our model act as a reservoir for electrons. According to the band calculations of Singh,[10] the number of electrons in the γ-band, n_γ, is 1.28. After the γ-band splits into two Hubbard subbands due to strong electron correlations, the lower subband becomes completely filled by one electron per atom, while the upper subband is partially filled by ''additional'' electrons (i.e., doubles) with a concentration $n_2=0.28$ (Fig. 2). In view of the hole–double symmetry, it is convenient to go over to the hole representation: the upper Hubbard subband with a concentration n_2 of DOUBLES is equivalent to the lower subband with the same concentration of holes, $n_0=n_2$, or the electron concentration $n=1-n_0$.

The Hamiltonian (1) describes states in the lower Hubbard band and makes it possible to study two very different scenarios of the system's behavior.

1. If $J>I$, as is the case for cuprates, then at $n=1$ $(n_0=0)$ the ground state has long-range antiferromagnetic (AF) order. As is known, the AF state is destroyed at low hole concentrations, $(n_0)_{AF} \sim 0.05$. In the region where $n_0 > (n_0)_{AF}$, superconductivity may emerge due to antiferromagnetic (J) exchange interactions.

2. If $I>J$, which is the case for Sr_2RuO_4, then at $n=1$ the ground insulator state is ferromagnetic (F). In the presence of holes there is competition between the saturated F state and the normal (nonmagnetic) N state, whose chemical potential is lower. As a result of this competition, for electron concentrations $n<n_F$, which is equivalent to hole concentrations $n_0>(n_0)_F$, the system passes to the N state, and in this region superconductivity may form due to ferromagnetic (I) exchange interactions.

It is convenient to normalize the Hamiltonian (1) to the halfwidth of the initial electron band, $w=zt$. If we introduce the Fourier transforms of the Hubbard operators,

$$X_{\mathbf{k}\sigma}=\frac{1}{\sqrt{N}}\sum_{\mathbf{f}} e^{i\mathbf{k}\cdot\mathbf{f}}X_{\mathbf{f}}^{0\sigma}, \quad X_{\mathbf{q}}^{\sigma\sigma'}=\frac{1}{\sqrt{N}}\sum_{\mathbf{f}} e^{i\mathbf{q}\cdot\mathbf{f}}X_{\mathbf{f}}^{\sigma\sigma'}, \quad (3)$$

where the vectors $\mathbf{k}$ and $\mathbf{q}$ belong to the first Brillouin zone, we arrive at the model Hamiltonian in the form

$$\frac{H}{zt}\equiv h=h_{\rm kin}+h_{\rm int}, \quad (4)$$

where

$$h_{\rm kin}=\sum_{\mathbf{k}\sigma}(\omega_{\mathbf{k}}-\tilde{\mu})X_{\mathbf{k}\sigma}^{\dagger}X_{\mathbf{k}\sigma},$$

$$\omega_{\mathbf{k}}=-\frac{1}{z}\sum_{\boldsymbol{\delta}} e^{i\mathbf{k}\cdot\boldsymbol{\delta}}=-\gamma_{\mathbf{k}}, \quad (5)$$

$$h_{\rm int}=\frac{1}{2}\sum_{\mathbf{q}\sigma}\gamma_{\mathbf{q}}\{g(X_{\mathbf{q}}^{\sigma\bar{\sigma}}X_{-\mathbf{q}}^{\bar{\sigma}\sigma}-X_{\mathbf{q}}^{\sigma\sigma}X_{-\mathbf{q}}^{\bar{\sigma}\bar{\sigma}}) -\lambda(X_{\mathbf{q}}^{\sigma\bar{\sigma}}X_{-\mathbf{q}}^{\bar{\sigma}\sigma}+X_{\mathbf{q}}^{\sigma\sigma}X_{-\mathbf{q}}^{\bar{\sigma}\bar{\sigma}})\}, \quad (6)$$

with $\bar{\sigma}=-\sigma$, $g=J/t$, $\lambda=I/t$, and $\mu/zt=\tilde{\mu}$ the dimensionless chemical potential. The Hamiltonian $h_{\rm kin}$ [Eq. (4)] describes the kinetic energy of the electrons and at $U=\infty$ is the Hamiltonian of the Hubbard model.

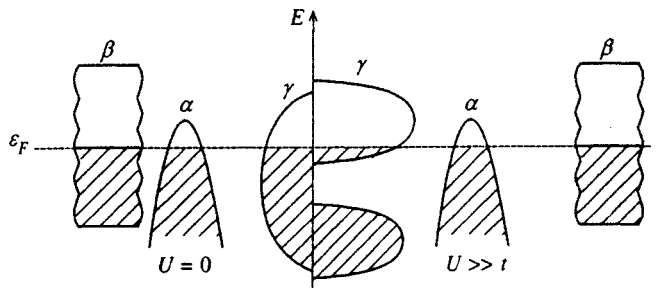

FIG. 2. Band structure of Sr_2RuO_4 in the limit of free electrons $(U=0)$ and in the strong correlation regime $(U \gg t)$.

3. EQUATIONS OF MOTION

Using the algebra of X-operators, we arrive at the equation of motion for a quasi-Fermi operator $(\hbar=1)$:

$$i\dot{X}_{\mathbf{k}\sigma}=[K_{\mathbf{k}\sigma},h]=(\omega_{\mathbf{k}}-\mu)X_{\mathbf{k}\sigma}+L_{\mathbf{k}\sigma},$$

$$L_{\mathbf{k}\sigma}=L_{\mathbf{k}\sigma}^{(\rm kin)}+L_{\mathbf{k}\sigma}^{(\rm int)}, \quad (7)$$

$$L_{\mathbf{k}\sigma}^{(\rm kin)}=\frac{1}{\sqrt{N}}\sum_{\mathbf{p}}\omega_{\mathbf{p}}(X_{\mathbf{k}-\mathbf{p}}^{\bar{\sigma}\sigma}X_{\mathbf{p}\bar{\sigma}}-X_{\mathbf{k}-\mathbf{p}}^{\bar{\sigma}\bar{\sigma}}X_{\mathbf{p}\sigma}), \quad (8)$$

$$L_{\mathbf{k}\sigma}^{(\mathrm{int})}=\frac{1}{\sqrt{N}}\sum_{\mathbf{p}}\ \omega_{\mathbf{k}-\mathbf{p}}\{(\lambda-g)X_{\mathbf{k}-\mathbf{p}}^{\bar{\sigma}\sigma}X_{\mathbf{p}\bar{\sigma}}+gX_{\mathbf{k}-\mathbf{p}}^{\sigma\bar{\sigma}}X_{\mathbf{p}\sigma}+\lambda X_{\mathbf{k}-\mathbf{p}}^{\sigma\sigma}X_{\mathbf{p}\sigma}\},\tag{9}$$

where the nonlinear operator $L_{\mathbf{k}\sigma}$ describes the correlations of electrons with spin projections in the opposite directions and in the same direction.

We introduce the irreducible operator (see Ref. 11)

$$\overline{L}_{\mathbf{k}\sigma}=L_{\mathbf{k}\sigma}-\frac{\langle\{L_{\mathbf{k}\sigma},X_{\mathbf{k}\sigma}^{\dagger}\}\rangle}{\langle\{X_{\mathbf{k}\sigma},X_{\mathbf{k}\sigma}^{\dagger}\}\rangle}X_{\mathbf{k}\sigma}-\frac{\langle\{L_{\mathbf{k}\sigma},X_{-\mathbf{k}\bar{\sigma}}\}\rangle}{\langle\{X_{-\mathbf{k}\bar{\sigma}}^{\dagger},X_{-\mathbf{k}\bar{\sigma}}\}\rangle}X_{-\mathbf{k}\bar{\sigma}}^{\dagger},\tag{10}$$

which has the property of being ''mean-orthogonal'': $\langle\{\overline{L}_{\mathbf{k}\sigma},X_{\mathbf{k}\sigma}^{\dagger}\}\rangle=\langle\{\overline{L}_{\mathbf{k}\sigma},X_{-\mathbf{k}\bar{\sigma}}\}\rangle=0$. Then Eq. (7) becomes

$$i\dot{X}_{\mathbf{k}\sigma}=\left(\omega_{\mathbf{k}}-\tilde{\mu}+\frac{C_{\mathbf{k}\sigma}}{1-n_{\bar{\sigma}}}\right)X_{\mathbf{k}\sigma}+\frac{\Delta_{\mathbf{k}\sigma}}{1-n_{\sigma}}X_{-\mathbf{k}\bar{\sigma}}^{\dagger}+\overline{L}_{\mathbf{k}\sigma},\tag{11}$$

where $X_{\mathbf{k}\sigma}=\langle\{L_{\mathbf{k}\sigma},X_{\mathbf{k}\sigma}^{\dagger}\}\rangle$ and $\Delta_{\mathbf{k}\sigma}=\langle\{L_{\mathbf{k}\sigma},X_{-\mathbf{k}\bar{\sigma}}\}\rangle$.

The generalized Hartree–Fock approximation, or the mean-field approximation, corresponds to the linear part of Eq. (11), i.e., we ignore the irreducible operator $\overline{L}_{\mathbf{k}\sigma}$. It is in this approximation that we will study the possibility of superconductivity manifesting itself. In Eq. (11), $C_{\mathbf{k}\sigma}/(1-n_{\bar{\sigma}})$ describes the renormalization of the spectrum, and $\Delta_{\mathbf{k}\sigma}$ is the possible superconducting gap. Spectrum renormalization can be calculated in general form, but it is sufficient to limit ourselves to an approximation of the Hubbard I type:

$$\xi_{\mathbf{k}\sigma}=\omega_{\mathbf{k}}-\tilde{\mu}+\frac{C_{\mathbf{k}\sigma}}{1-n_{\bar{\sigma}}}\approx(1-n_{\bar{\sigma}})\omega_{\mathbf{k}}-gn_{\bar{\sigma}}-\lambda n_{\sigma}-\tilde{\mu}.$$

In the nonmagnetic ground state, $n_{\uparrow}=n_{\downarrow}=n/2$, the dependence on the spin projection disappears and the modified spectrum can be written

$$\xi_{\mathbf{k}}=c(n)(\omega_{\mathbf{k}}-m),$$

$$m=\frac{(g+\lambda)n/2+\tilde{\mu}}{c(n)},\quad c(n)=1-\frac{n}{2},\tag{12}$$

where m is the effective chemical potential.

The expression for the gap $\Delta_{\mathbf{k}\sigma}$ has the form

$$\Delta_{\mathbf{k}\sigma}=\frac{1}{N}\sum_{\mathbf{p}}\ \omega_{\mathbf{p}}(\langle X_{-\mathbf{p}\sigma}X_{\mathbf{p}\bar{\sigma}}\rangle-\langle X_{-\mathbf{p}\bar{\sigma}}X_{\mathbf{p}\sigma}\rangle)+\frac{1}{N}\sum_{\mathbf{p}}\ \omega_{\mathbf{k}-\mathbf{p}}\{(\lambda-g)\langle X_{-\mathbf{p}\sigma}X_{\mathbf{p}\bar{\sigma}}\rangle+g\langle X_{-\mathbf{p}\bar{\sigma}}X_{\mathbf{p}\sigma}\rangle\}.\tag{13}$$

We introduce the anomalous means

$$B_{\mathbf{p}}=\langle X_{-\mathbf{p}\downarrow}X_{\mathbf{p}\uparrow}\rangle.\tag{14}$$

Then, using the symmetry property $\omega_{\mathbf{p}}=\omega_{-\mathbf{p}}$, we find that

$$\Delta_{-\mathbf{k}\downarrow}=-\Delta_{\mathbf{k}\uparrow}=\Delta_{\mathbf{k}}=\frac{1}{N}\sum_{\mathbf{p}}\ [2\omega_{\mathbf{p}}-g(\omega_{\mathbf{k}+\mathbf{p}}+\omega_{\mathbf{k}-\mathbf{p}})+\lambda\omega_{\mathbf{k}+\mathbf{p}}]B_{\mathbf{p}}.\tag{15}$$

The first term on the right-hand side of Eq. (15) reflects the presence of kinematic electron correlations and originates in the kinetic term in the Hamiltonian (this is known as kinematic pairing[12]), and the other terms are the consequence of exchange interactions.[1)]

Using Eq. (11) (with the irreducible operator $\overline{L}_{\mathbf{k}\sigma}$ discarded) and the relationships (12) and (15), we arrive in the mean-field approximation at the following system of equations:

$$i\dot{X}_{\mathbf{k}\uparrow}=\xi_{\mathbf{k}}X_{\mathbf{k}\uparrow}=\frac{\Delta_{\mathbf{k}}}{c(n)}X_{-\mathbf{k}\downarrow}^{\dagger},$$

$$i\dot{X}_{-\mathbf{k}\downarrow}^{\dagger}=-\xi_{\mathbf{k}}X_{-\mathbf{k}\downarrow}^{\dagger}-\frac{\Delta_{\mathbf{k}}^{*}}{c(n)}X_{\mathbf{k}\uparrow}.\tag{16}$$

4. MEANS AND SELF-CONSISTENCY EQUATIONS

Using the system of equations (16), we arrive at a system of equations for the two-time retarded anticommuting Green's functions[11] and its solution:

$$\langle\langle X_{\mathbf{k}\uparrow}|X_{\mathbf{k}\uparrow}^{\dagger}\rangle\rangle_{E}=c(n)\frac{E+\xi_{\mathbf{k}}}{E^{2}-E_{\mathbf{k}}^{2}},$$

$$\langle\langle X_{-\mathbf{k}\downarrow}^{\dagger}|X_{\mathbf{k}\uparrow}^{\dagger}\rangle\rangle_{E}=-\frac{\Delta_{\mathbf{k}}^{*}}{E^{2}-E_{\mathbf{k}}^{2}},\tag{17}$$

where

$$E_{\mathbf{k}}^{2}=\xi_{\mathbf{k}}^{2}+\frac{|\Delta_{\mathbf{k}}|^{2}}{c^{2}(n)}.\tag{18}$$

The spectral theorem[11] yields the following expressions for the means:

$$n_{\mathbf{k}}=\langle X_{\mathbf{k}\uparrow}^{\dagger}X_{\mathbf{k}\uparrow}\rangle=\langle X_{\mathbf{k}\downarrow}^{\dagger}X_{\mathbf{k}\downarrow}\rangle=\frac{c(n)}{2}\left(1-\frac{\xi_{\mathbf{k}}}{E_{\mathbf{k}}}\tanh\frac{E_{\mathbf{k}}}{2\tau}\right)\equiv c(n)f_{\mathbf{k}},\tag{19}$$

$$B_{\mathbf{k}}^{*}=\langle X_{\mathbf{k}\uparrow}^{\dagger}X_{-\mathbf{k}\downarrow}^{\dagger}\rangle=\frac{\Delta_{\mathbf{k}}^{*}}{2E_{\mathbf{k}}}\tanh\frac{E_{\mathbf{k}}}{2\tau},\tag{20}$$

where $E_{\mathbf{k}}>0$, and $\tau=k_{B}T/zt$ is the dimensionless temperature.

In the superconducting phase we have three self-consistency equations.

1. An equation that links the electron concentration n with the effective chemical potential m:

$$n=\frac{1}{N}\sum_{\mathbf{k}\sigma}\ n_{\mathbf{k}\sigma}=\frac{2c(n)}{N}\sum_{\mathbf{k}}\ f_{\mathbf{k}},\quad f_{\mathbf{k}}=\frac{1}{2}\left(1-\frac{\xi_{\mathbf{k}}}{E_{\mathbf{k}}}\right),\tag{21}$$

where $f_{\mathbf{k}}$ is the distribution function (19) at $T=0$.

2. The constraint condition, or the sum rule for anomalous means (the exclusion of doubles as a consequence of the algebra of X-operators):

$$\frac{1}{N}\sum_{\mathbf{k}} B_{\mathbf{k}}=\frac{1}{N}\sum_{\mathbf{k}} B_{\mathbf{k}}^{*}=0. \tag{22}$$

3. The equation for the energy gap $\Delta_{\mathbf{k}}$ (see below), which has meaning only if condition (22) is met.

But first we must examine the possible nonsuperconducting states.

N state. We begin with the normal (nonmagnetic) phase, or the N state ($\Delta_{\mathbf{k}}\equiv 0$). The distribution function $f_{\mathbf{k}}$ becomes the Fermi step $f_{\mathbf{k}}^{0}=\theta(m-\omega_{\mathbf{k}})$ and Eq. (21) becomes

$$\frac{n}{2-n}=\frac{1}{N}\sum_{\mathbf{k}} \theta(m-\omega_{\mathbf{k}})=\int_{-1}^{m}\rho(\omega)d\omega\equiv g(m), \tag{23}$$

where $\rho(\omega)=\rho(-\omega)$ is the density of states corresponding to the dispersion law for $\omega_{\mathbf{k}}$. The system energy (per lattice site) is

$$\epsilon_0=\frac{1}{N}\sum_{\mathbf{k}\sigma}\omega_{\mathbf{k}}\left(1-\frac{n}{2}\right)f_k^0-\frac{1}{4}(g+\lambda)n^2$$

$$=(2-n)\int_{-1}^{m}\omega\rho(\omega)d\omega-\frac{1}{4}(g+\lambda)n^2. \tag{24}$$

F state. In the model with $U=\infty$ ($J=0$), long-range ferromagnetic (F) order sets in in the region of high concentrations, $n>n_F(\lambda)$. This critical concentration can easily be found by comparing the energies of the saturated ferromagnetic state $[\epsilon_F(n,\lambda)]$ and the normal state $[\epsilon_0(n,\lambda)]$. The energy of the F state (per lattice site) is

$$\epsilon_F(n,\lambda)=\nu(m_F)-\frac{1}{2}\lambda n^2,$$

$$n=g(m_F)=\int_{-1}^{m_F}\rho(\omega)d\omega,\quad \nu(m_F)=\int_{-1}^{m_F}\omega\,\rho(\omega)d\omega, \tag{25}$$

where m_F is the chemical potential in the F state, and n is the electron concentration. By comparing the energies of the N and F states we can find $n_F(\lambda)$. For instance, for a square lattice, $n_F\approx 0.91$ at $\lambda=0.3$, and the domain of existence of the ground F state grows as λ increases: at $\lambda=1$ we have $n_F\approx 0.6$. A similar situation occurs in the three-dimensional case.

AF state. When $J\gg I$ and $n\to 1$, the system exhibits long-range antiferromagnetic (AF) order. As noted in Sec. 2, the AF state is destroyed at low hole concentrations, $n_0\sim 0.05$ (see Refs. 5 and 6). We do not discuss this state in detail in this paper.

5. SYMMETRY PROPERTIES OF ANOMALOUS MEANS AND SOLUTIONS FOR THE GAP

We analyze the structure and symmetry properties of the anomalous averages $B_{\mathbf{k}}$ and the gap $\Delta_{\mathbf{k}}$. We represent the anomalous averages $B_{\mathbf{p}}\equiv\langle X_{-\mathbf{p}\downarrow}X_{\mathbf{p}\uparrow}\rangle$ as

$$B_{\mathbf{k}}=B_{\mathbf{k}}^{(s)}+B_{\mathbf{k}}^{(a)},\quad B_{\mathbf{k}}^{(s)}=\frac{1}{2}(B_{\mathbf{k}}+B_{-\mathbf{k}})=B_{-\mathbf{k}}^{(s)},$$

$$B_{\mathbf{k}}^{(a)}=\frac{1}{2}(B_{\mathbf{k}}-B_{-\mathbf{k}})=-B_{-\mathbf{k}}^{(a)}, \tag{26}$$

i.e., as the sum of the symmetric (s) and antisymmetric (a) parts. We immediately note that the sum rule (22) for the antisymmetric part $B_{\mathbf{k}}^{(a)}$ is satisfied automatically. Performing an inverse Fourier transformation, we obtain an expression for the symmetric part:

$$(B_{\mathbf{k}}^{(s)})^{*}=\frac{1}{2}\langle X_{\mathbf{k}\uparrow}^{\dagger}X_{-\mathbf{k}\downarrow}^{\dagger}+X_{-\mathbf{k}\uparrow}^{\dagger}X_{\mathbf{k}\downarrow}^{\dagger}\rangle$$

$$=\frac{1}{\sqrt{2}}\sum_{\mathbf{r}} e^{i\mathbf{k}\cdot\mathbf{r}}\frac{1}{N}\sum_{\mathbf{f}}\langle Z_{\mathbf{f},\mathbf{f}+\mathbf{r}}^{\dagger}\rangle, \tag{27}$$

$$Z_{\mathbf{fm}}^{\dagger}=\frac{1}{\sqrt{2}}(X_{\mathbf{f}}^{\uparrow 0}X_{\mathbf{m}}^{\downarrow 0}+X_{\mathbf{m}}^{\uparrow 0}X_{\mathbf{f}}^{\downarrow 0})=Z_{\mathbf{mf}}^{\dagger}, \tag{28}$$

where $Z_{\mathbf{fm}}^{\dagger}$ is the operator of creation of a singlet pair at an arbitrary pair of sites **f** and **m**. Thus, the symmetric part of an anomalous mean corresponds to singlet pairings.

Similarly, for the antisymmetric part we have

$$(B_{\mathbf{k}}^{(s)})^{*}=\frac{1}{2}\langle X_{\mathbf{k}\uparrow}^{\dagger}X_{-\mathbf{k}\downarrow}^{\dagger}-X_{-\mathbf{k}\uparrow}^{\dagger}X_{\mathbf{k}\downarrow}^{\dagger}\rangle$$

$$=\frac{1}{\sqrt{2}}\sum_{\mathbf{r}} e^{i\mathbf{k}\cdot\mathbf{r}}\frac{1}{N}\sum_{\mathbf{f}}\langle T_{\mathbf{f},\mathbf{f}+\mathbf{r}}^{\dagger}\rangle, \tag{29}$$

$$T_{\mathbf{fm}}^{\dagger}=\frac{1}{\sqrt{2}}(X_{\mathbf{f}}^{\uparrow 0}X_{\mathbf{m}}^{\downarrow 0}-X_{\mathbf{m}}^{\uparrow 0}X_{\mathbf{f}}^{\downarrow 0})=-T_{\mathbf{mf}}^{\dagger}, \tag{30}$$

where $T_{\mathbf{fm}}^{\dagger}$ is the operator of creation of a triplet pair with $S^{z}=0$ at an arbitrary pair of sites **f** and **m**. Thus, the antisymmetric part of an anomalous mean corresponds to triplet pairings.

Let us examine two alternative lattices: a square lattice ($d=2$) and a simple cubic lattice ($d=3$). For these lattices we have

$$\gamma_{\mathbf{k}}=\frac{1}{d}\sum_{j}^{d}\cos k_j,\quad \omega_{\mathbf{k}}=-\gamma_{\mathbf{k}} \tag{31}$$

(the lattice constant $a=1$). We introduce two functions,

$$C_j=\frac{1}{N}\sum_{\mathbf{p}}\cos p_j B_{\mathbf{p}}^{(s)},\quad S_j=\frac{1}{N}\sum_{\mathbf{p}}\sin p_j B_{\mathbf{p}}^{(a)}. \tag{32}$$

Since

$$\gamma_{\mathbf{k}\pm\mathbf{p}}=\frac{1}{d}\sum_{j}^{d}(\cos k_j\cos p_j\mp\sin k_j\sin p_j),$$

the gap (15) can be written

$$\Delta_{\mathbf{k}}=\Delta_{\mathbf{k}}^{(s)}+\Delta_{\mathbf{k}}^{(a)}, \quad \Delta_{\mathbf{k}}^{(s)}=2\Delta_0+\frac{2g-\lambda}{d}\sum_j \cos k_j C_j,$$

$$\Delta_0=\frac{1}{N}\sum_{\mathbf{p}} \omega_{\mathbf{p}} B_{\mathbf{p}}^{(s)}, \quad \Delta_{\mathbf{k}}^{(a)}=\lambda\frac{1}{d}\sum_j \sin k_j S_j \tag{33}$$

and contains a symmetric (s) part in the momenta and an antisymmetric (a) part. The expression (18) for the spectrum $E_{\mathbf{k}}$ also contains $|\Delta_{\mathbf{k}}|^2$ with mixed symmetry. From the general physical requirement imposed on the spectrum, $E_{\mathbf{k}}=E_{-\mathbf{k}}$, it follows that $|\Delta_{\mathbf{k}}|^2=|\Delta_{-\mathbf{k}}|^2$, which leads either to the class of symmetric solutions $\Delta_{\mathbf{k}}^{(s)}=\Delta_{-\mathbf{k}}^{(s)}$ (singlet pairings) or to the class of antisymmetric solutions $\Delta_{\mathbf{k}}^{(a)}=-\Delta_{-\mathbf{k}}^{(a)}$ (triplet pairings).

Actually, there can be several solutions within one class, with each solution corresponding to a linear combination of cosines (for symmetric solutions) or sines (for antisymmetric solutions). In the general case we number the solutions by the label l, denote the gap of the lth type by $\Delta_{\mathbf{k}l}$, and the spectrum (18) with such a gap by $E_{\mathbf{k}l}$. The expression for the anomalous means becomes

$$B_{\mathbf{k}l}=\frac{\Delta_{\mathbf{k}l}}{2E_{\mathbf{k}l}}\tanh\frac{E_{\mathbf{k}l}}{2\tau}, \quad E_{\mathbf{k}l}=\sqrt{\xi_{\mathbf{p}}^2+\frac{|\Delta_{\mathbf{k}l}|^2}{c^2(n)}}. \tag{34}$$

Denoting the corresponding linear combination of trigonometric functions by $\psi_l(\mathbf{k})$ and the dimensionless coupling constant of the interaction that forms a gap of the lth type by α_l, we arrive at the following types of solution.

Symmetric solutions of the s type (singlet pairings), $l=0$:

$$\Delta_{\mathbf{k}0}=(2+\alpha\omega_{\mathbf{k}})\Delta_0, \quad \Delta_0=\frac{1}{N}\sum_{\mathbf{p}} \omega_{\mathbf{p}} B_{\mathbf{p}}^{(s)},$$

$$\psi_0(\mathbf{k})=\omega_{\mathbf{k}}, \quad \alpha=2g-\lambda. \tag{35}$$

The constraint condition (22) for the anomalous means $B_{\mathbf{k}0}$ with the gap $\Delta_{\mathbf{k}0}$ is not met, which implies that there can be no solutions of the s type. Note that many equations for the gap of the s type have already been proposed (see, e.g., Ref. 5), but the constraint condition was not taken into account.

Antisymmetric solutions of the p type ($l=1$) and symmetric solutions of the d type ($l=2$) can be written in a unique form:

$$\Delta_{\mathbf{k}l}=\alpha_l\psi_l(\mathbf{k})\Delta_l, \quad \Delta_l=\frac{1}{N}\sum_{\mathbf{p}} \psi_l(\mathbf{p})B_{\mathbf{p}l}. \tag{36}$$

Combining (34) and (36), we arrive at the gap equation

$$\frac{1}{\alpha_l}=\frac{1}{N}\sum_{\mathbf{p}}\frac{\psi_l^2(\mathbf{p})}{2E_{\mathbf{p}l}}\tanh\frac{E_{\mathbf{p}l}}{2\tau}. \tag{37}$$

The solutions of this equation are meaningful only if the sum rule (22) holds, and at $T=0$ this sum rule can be written

$$\frac{1}{N}\sum_{\mathbf{p}}\frac{\psi_l(\mathbf{p})}{E_{\mathbf{p}l}}=0. \tag{38}$$

In explicit form, we have the following types of solution.

1. *Antisymmetric solutions of the p-type (triplet pairings), $l=1$:*

$$\psi_p(\mathbf{k})=\frac{1}{d}\sum_j \sin k_j, \quad \alpha_p=\lambda. \tag{39}$$

2. *Symmetric solutions of the d-type (singlet pairings), $l=2$:*

$$\psi_d(\mathbf{k})=\frac{\cos k_x-\cos k_y}{2}, \quad \alpha_d\equiv\alpha=2g-\lambda \tag{40}$$

for the square lattice. For p-type solutions the sum rule is satisfied automatically, while for d-type solutions the validity of (40) follows from the symmetry properties.

In conclusion of this section we examine the symmetry properties of the solutions from a general position. The system Hamiltonian (1) is written in terms of the exchange operators (2). The operator $K_{\mathbf{fm}}^{(-)}$ has an eigenvalue equal to -1 when it acts on a singlet pair and a zero eigenvalue when it acts on a triplet pair. Antiferromagnetic exchange ($J>0$ and $J/t=g>0$) ensures attraction between the electrons in a singlet pair and ''ignores'' triplet pairs. For this reason, J-exchange takes no part in the formation of p-type triplet superconductivity. This fact is reflected by the presence in the expression (15) for the gap of the momentum-symmetric contribution $g(\gamma_{\mathbf{k}+\mathbf{p}}+\gamma_{\mathbf{k}-\mathbf{p}})$, which yields only symmetric solutions corresponding to singlet pairings. On contrast to J-exchange, ferromagnetic direct exchange ($I>0$ and $I/t=\lambda>0$) acts on any pairs: the operator $K_{\mathbf{fm}}^{(+)}$ has an eigenvalue equal to $+1/2$ when it acts on a triplet pair and an eigenvalue equal to $-1/2$ when it acts on a singlet pair. With allowance for the sign in the Hamiltonian, I-exchange leads to attraction between the electrons in a triplet pair and the only term responsible for the formation of p-type triplet superconductivity. On the other hand, since in the singlet state of a pair this type of exchange leads to repulsion, competition between the exchange interactions emerges in this case, and this is reflected by the coupling constant $\alpha=2g-\lambda$.

Our equations for the gap and T_c in the mean-field approximation coincide in structure with similar equations obtained by the diagrammatic technique for X-operators in the $t-J$ model[5,6] when Cooper instability of the normal phase is examined.

6. COMPARISON OF *p*- AND *d*-TYPE SOLUTIONS

We write the spectrum $E_{\mathbf{k}l}$ in the form ($l=p,d$)

$$E_{\mathbf{k}l}=c(n)\sqrt{(\omega_{\mathbf{k}}-m)^2+\psi_l^2(\mathbf{k})D_l^2}, \quad D_l^2=\frac{\alpha_l^2|\Delta_l|^2}{c^4(n)}. \tag{41}$$

Then from (37) we obtain in explicit form the equations for the effective gap D_l as a function of concentration (chemical potential) and the dimensionless temperature τ,

$$\frac{2c(n)}{\alpha_l}=\frac{1}{N}\sum_{\mathbf{p}}\frac{\psi_l^2(\mathbf{p})}{\sqrt{(\omega_{\mathbf{p}}-m)^2+\psi_l^2(\mathbf{p})D_l^2}} \times\tanh\frac{c(n)\sqrt{(\omega_{\mathbf{p}}-m)^2+\psi_l^2(\mathbf{p})D_l^2}}{2\tau}, \tag{42}$$

and an equation for the transition temperature $\tau_c^{(l)}$ ($D_l^2\to 0$ as $\tau\to\tau_c^{(l)}$),

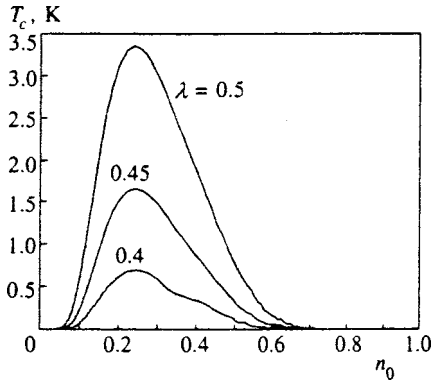

FIG. 3. Concentration dependence of the transition temperature T_c in Sr_2RuO_4 with triplet pairing of the p type; $\lambda = I/t$, with I the ferromagnetic exchange.

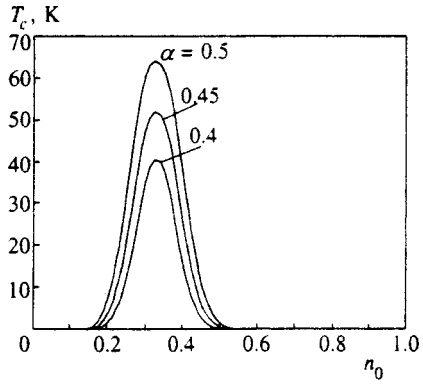

FIG. 4. Concentration dependence of the transition temperature T_c in copper oxides with singlet pairing of the d type; $\alpha = 2g - \lambda$ and $g = J/t$, with J the antiferromagnetic exchange ($J > I$).

$$\frac{2c(n)}{\alpha_l} = \frac{1}{N}\sum_{\mathbf{p}} \frac{\psi_l^2(\mathbf{p})}{|\omega_{\mathbf{p}} - m|} \tanh \frac{c(n)|\omega_{\mathbf{p}} - m|}{2\tau_c^{(l)}}. \tag{43}$$

When stating the problem, we can use these equations with $l = d$ to describe cuprates, which at a certain concentration of holes (or electrons) pass from the AF state to the superconducting (SC) state with singlet pairing of the d type, and at $l = p$ they should describe superconductivity with triplet pairings of the p type in Sr_2RuO_4, which is above the ferromagnetic stability threshold. Although the dimensionless coupling constants $\alpha_d = \alpha$ and $\alpha_p = \lambda$ may differ in value, when they are equal, the values of the transition temperatures $\tau_c^{(d)}$ and $\tau_c^{(p)}$ depend significantly on the properties of the functions $\psi_d^2(\mathbf{k})$ [Eq. (40)] and $\psi_p^2(\mathbf{k})$ [Eq. (39)].

6.1. Results of calculations

The equations for the effective gap [Eqs. (42)] at $\tau = 0$ and for the transition temperature [Eqs. (43)] were solved numerically by summing over the momenta $\mathbf{p}$ in the first Brillouin zone of a square lattice (it contained 10^6 values of $\mathbf{p}$ and about 100 values of m in the interval from -1 to $+1$). We used the relationship that links the electron concentration n with the effective chemical potential m in the normal phase [Eq. (23)], i.e.,

$$n = \frac{2g(m)}{1 + g(m)}, \quad g(m) = \int_{-1}^{m} \rho(\omega)\, d\omega, \tag{44}$$

where $\rho(\omega)$ is the density of states. For a square lattice we have

$$\rho(\omega) = \frac{2}{\pi^2} K(\sqrt{1-\omega^2}) \approx \frac{1}{\pi} - \left(\frac{1}{2} - \frac{1}{\pi}\right) \ln|\omega|,$$

$$\int_{-1}^{+1} \rho(\omega) d\omega = 1, \tag{45}$$

where K is the complete elliptic integral of the first kind. The results of calculating $T_c = zt\tau_c$, which is considered a function of the hole concentration, are depicted in Figs. 3 and 4 for typical values of the coupling constants λ and α. The fitting in the curves was to the experimental data on Sr_2RuO_4. At low concentrations of doubles in the upper Hubbard subband, $n_2 = 0.28$ (this is equivalent to holes in the lower Hubbard subband), the dispersion law is described fairly well by the quadratic expression $\varepsilon(\mathbf{k}) = -ztc(n)\omega_{\mathbf{k}} \approx \varepsilon(0) + p^2/2m^*$, where $1/m^* = (4/3)ta^2/\hbar^2$ for $n_2 \sim 1/3$. Plugging in the values of the effective mass $m^* = 12m_e$ and of the atomic separation $a = 1.93$ Å, we find that $t \sim 0.1$ eV and $zt \sim 0.4$ eV. Within our theory, the maximum value of $\tau_c(n_0)$ corresponds to $n_0 \sim 1/3$ and the effective chemical potential $m = 0$. Note that in this sense ruthenates are ''self-doped,'' since the concentration $n_2 = n_0 = 0.28$ is close to the optimal concentration with respect to the maximum in T_c. In cuprates the value of T_c depends on the degree of doping to which the hole concentration n_0 corresponds. What is indicative is the fact that even if all the parameters are equal (including $\alpha = \lambda$), the transition temperatures differ substantially: $T_c^{(d)} \gg T_c^{(p)}$. Let us analyze the reason for this difference.

6.2. Gap anisotropy

The functions $\psi_p^2(\mathbf{k})$ and $\psi_d^2(\mathbf{k})$ describe the anisotropy of the superconducting gap in p- and d-type pairings, respectively. Figure 5a depicts the Brillouin zone of a square lattice and the constant-energy surface $\omega_{\mathbf{k}} = 0$ (the square $ABCD$). If the constant-energy surface corresponds to the chemical potential $m = 0$, the states inside it in the nonsuperconducting phase at $T = 0$ are filled, which is denoted by hatching. Note that for free electrons such filling corresponds to the concentration $n = 1$, for the saturate F state $n = 1/2$, and for the case of strongly correlated electrons $n = 2/3$, in accordance with (44) and (45). The behavior of the functions $\psi_p^2(\mathbf{k})$ and $\psi_d^2(\mathbf{k})$ at $\omega_{\mathbf{k}} = 0$ is depicted in Figs. 5b and 5c. The gap ''collapses'' (vanishes) not only at isolated points of the constant-energy surface (d type) but also on its fragments BC and DA (p type). Reasoning in a similar manner, we can study gap anisotropy for an arbitrary constant-energy surface.

6.3. Mean values of *p*- and *d*-type functions on constant-energy surfaces

We now replace summation over momenta with integration over constant-energy surfaces corresponding to the law

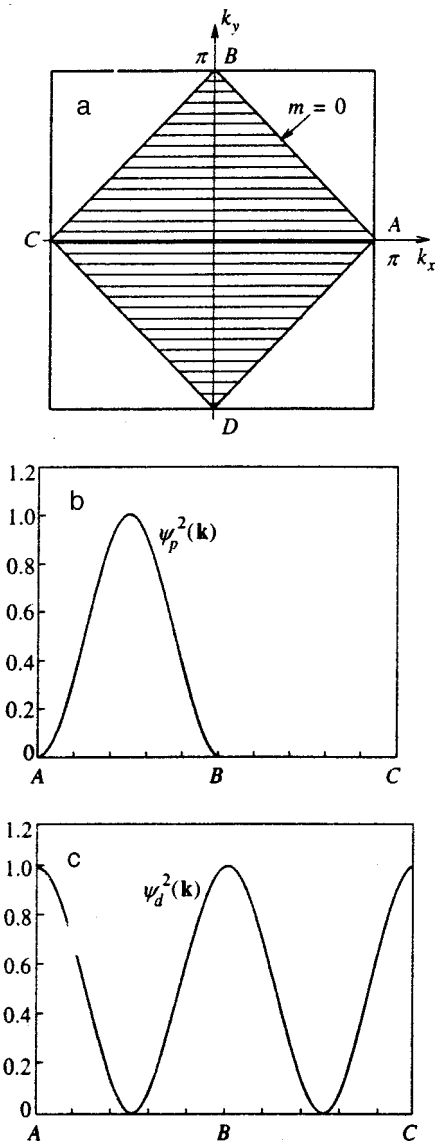

FIG. 5. Gap anisotropy on the constant-energy surface $\omega_{\mathbf{k}}=0$ (the square $ABCD$): (a) the Brillouin zone of a square lattice and the filled states with the chemical potential $m=\omega_{\mathbf{k}}=0$; (b) anisotropy of the p type; and (c) anisotropy of the d type (in both cases the behavior of the functions along the lines AB, CD, BC, and DA is the same).

of dispersion for $\omega_{\mathbf{k}}$. Let $\omega_{\mathbf{k}}=\omega$ by the equation of the constant-energy surface $(-1\leqslant\omega_{\mathbf{k}}\leqslant+1)$. Then for the function $A_{\mathbf{k}}$, which can be expressed explicitly in terms of the dispersion law for $\omega_{\mathbf{k}}$, i.e., $A_{\mathbf{k}}=A(\omega_{\mathbf{k}})$, we have

$$\frac{1}{N}\sum_{\mathbf{k}} A(\omega_{\mathbf{k}})=\int_{-1}^{+1}\rho(\omega)A(\omega)d\omega, \tag{46}$$

$$\rho(\omega)=\frac{1}{N}\sum_{\mathbf{k}}\delta(\omega-\omega_{\mathbf{k}})=\frac{1}{(2\pi)^d}\oint_{(\sigma_\omega)}\frac{d\sigma_\omega}{|\nabla_{\mathbf{k}}\omega_{\mathbf{k}}|}$$

$$=\frac{d}{d\omega}\left(\frac{1}{(2\pi)^d}\int_{v_\omega}d^d\mathbf{k}\right). \tag{47}$$

Here $\rho(\omega)$ is the density of states corresponding to the dispersion law for $\omega_{\mathbf{k}}$, σ_ω is the surface area of the $(d-1)$-dimensional constant-energy surface, and v_ω is the d-dimensional volume of the $\mathbf{k}$-space encompassed by the constant-energy surface ω (the volume occupied by the unit cell is $a^d=1$). In the two-dimensional case (the square lattice) considered here,

$$\omega_{\mathbf{k}}=-\frac{\cos k_x+\cos k_y}{2},\quad |\nabla_{\mathbf{k}}\omega_{\mathbf{k}}|=\frac{\sqrt{\sin^2 k_x+\sin^2 k_y}}{2}, \tag{48}$$

and the density of states (45) has a logarithmic singularity. Unfortunately, the functions $\psi_l^2(\mathbf{k})$ present in the theory cannot be expressed explicitly in terms of $\omega_{\mathbf{k}}$, and for this reason we introduce the mean values of these functions over the constant-energy surface $\omega_{\mathbf{k}}=\omega$:

$$\langle\psi_l^2(\mathbf{k})\rangle_{\omega_{\mathbf{k}}=\omega}\equiv\psi_l^2(\omega)=\frac{1}{(2\pi)^2}\oint_{(\sigma_\omega)}\frac{\psi_l^2(\mathbf{k})}{|\nabla_{\mathbf{k}}\omega_{\mathbf{k}}|}d\sigma_\omega. \tag{49}$$

Note that direct summation of the functions $\psi_l^2(\mathbf{k})$ over the Brillouin zone yields the obvious and same result $1/2d=1/4$, so that these functions must satisfy the integral condition

$$\frac{1}{N}\sum_{\mathbf{k}}\psi_l^2(\mathbf{k})=\int_{-1}^{+1}\psi_l^2(\omega)d\omega=\frac{1}{4}. \tag{50}$$

We analyze the properties of these functions. The integrand for a p-type function in (49),

$$\frac{\psi_p^2(\mathbf{k})}{|\nabla_{\mathbf{k}}\omega_{\mathbf{k}}|}=|\nabla_{\mathbf{k}}\omega_{\mathbf{k}}|+\frac{\sin k_x\sin k_y}{\sqrt{\sin^2 k_x+\sin^2 k_y}}, \tag{51}$$

has no singularity, with the result that

$$\psi_p^2(\omega)=\frac{1}{(2\pi)^2}\oint_{(\sigma_\omega)}|\nabla_{\mathbf{k}}\omega_{\mathbf{k}}|d\sigma_\omega$$

$$=\frac{2}{\pi^2}(E(\sqrt{1-\omega^2})-\omega^2K(\sqrt{1-\omega^2})), \tag{52}$$

where E is the complete elliptic integral of the second kind. Here we used the symmetry properties, in view of which the second term in (51) yields zero when it is integrated. The function (52) is smooth and is approximated very well by the expression

$$\psi_p^2(\omega)=\frac{2}{\pi^2}(1-|\omega|^\nu),\quad \nu=\frac{\pi^2}{16-\pi^2}\approx 1.61. \tag{53}$$

The integrand for a d-type function in (49),

$$\frac{\psi_d^2(\mathbf{k})}{|\nabla_{\mathbf{k}}\omega_{\mathbf{k}}|}=\frac{1}{2}\frac{(\cos k_x-\cos k_y)^2}{\sqrt{\sin^2 k_x+\sin^2 k_y}}, \tag{54}$$

has singularities of the same type as the density of states (45). Calculations yield

$$\psi_d^2(\omega)=(1-\omega^2)\rho(\omega)-2\psi_p^2(\omega). \tag{55}$$

For the sake of comparison, we depict the functions $\psi_p^2(\omega)$ and $\psi_d^2(\omega)$ in Fig. 6, with each normalized to 1/4 by (50).

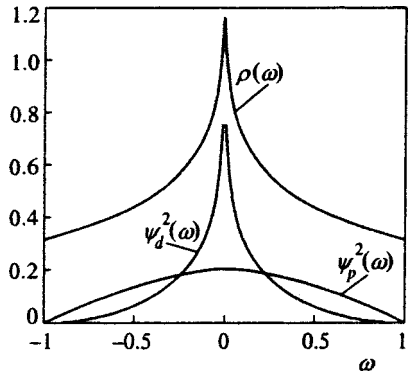

FIG. 6. Density of electronic states in a square lattice, $\rho(\omega)$, and the mean value of p- and d-type functions on constant-energy surfaces ω.

6.4. Transition temperature

Let us now turn to Eqs. (43) for the transition temperature. We replace summation over the Brillouin zone with integration over constant-energy surfaces. It is convenient to divide the integral into two parts: over a narrow layer of width 2δ near the chemical potential m, and over the remain part of the zone. The value of δ can always be chosen so that the argument of the hyperbolic tangent is larger than 2.0 and the tangent is, to a high accuracy, equal to unity, i.e., $c(n)\delta/2\tau_c \geqslant 2$. Thus, we have the integrals

$$I(m,\tau_c^{(l)}) \approx \int_{-1}^{+1} \frac{d\omega \psi_l^2(\omega)}{|\omega-m|} \tanh \frac{c(n)|\omega-m|}{2\tau_c^{(l)}}$$
$$= \int_{m-\delta}^{m+\delta} \frac{d\omega \psi_l^2(\omega)}{|\omega-m|} \tanh \frac{c(n)|\omega-m|}{2\tau_c^{(l)}}$$
$$+ \int_{-1}^{m-\delta} \frac{d\omega \psi_l^2(\omega)}{|\omega-m|} + \int_{m+\delta}^{+1} \frac{d\omega \psi_l^2(\omega)}{|\omega-m|}$$
$$\equiv I_S^{(l)} + I_B^{(l)}, \tag{56}$$

which are divided into an integral over the thin lawyer (S) and an integral over the part of the zone outside the layer (B).

Triplet pairings of the p type. Since $\psi_p^2(\omega)$ is a smooth and slowly varying function, the integral over the layer in the logarithmic approximation is

$$I_S^{(p)} = 2\psi_p^2(m) \ln \frac{1.14 c(n)\delta}{\tau_c^{(p)}}, \tag{57}$$

where $2\gamma/\pi = 1.14$, with γ the Euler constant. Calculating the integral $I_B^{(p)}$, we can write the solution of Eq. (43):

$$\tau_c^{(p)} \approx 1.14 c(n)\delta \exp\left\{ -\frac{c(n)/\lambda - I_B^{(p)}/2}{\psi_p^2(m)} \right\}$$
$$= 1.14 c(n)\delta \exp\left\{ -\frac{c(n)}{(\lambda+\lambda_B)\psi_p^2(m)} \right\} \tag{58}$$

with $\delta \gtrsim 2\tau_c^{(p)}/c(n)$. In the second part of (58) we have introduced the notation

$$\frac{c(n)}{\lambda} - \frac{I_B^{(p)}}{2} \equiv \frac{c(n)}{\lambda+\lambda_B}, \quad \lambda_B = \lambda^2 \frac{I_B^{(p)}}{2c(n) - \lambda I_B^{(p)}}, \tag{59}$$

where the function λ_B describes the effect of ''band enhancement'' of the parameter λ. Since $\psi_p^2(m) \leqslant 0.2$, the exponential factor proves to be small, which leads to small values of $\tau_c^{(p)}$ for triplet superconductivity.

Singlet pairings of the d type. From Eq. (55) and Fig. 6 it follows that $\psi_d^2(\omega)$ changes rapidly in the vicinity of $\omega = 0$. Hence in this region the integral over the layer must be calculated with allowance for the logarithmic singularity in $\rho(\omega)$ as $\omega \to 0$. Below we estimate the integral over the layer by replacing the density of states $\rho(\omega)$ with its mean value $\rho(\omega,\delta)$ in the interval $[\omega-\delta, \omega+\delta]$. Then $\psi_d^2(\omega) \to \langle \psi_d^2(\omega,\delta)\rangle$ and, as in the previous case, we find that

$$\tau_c^{(d)} \approx 1.14 c(n)\delta \exp\left\{ -\frac{c(n)}{(\alpha+\alpha_B)\langle \psi_d^2(m,\delta)\rangle} \right\}, \tag{60}$$

with $\delta \gtrsim 2\tau_c^{(d)}/c(n)$. Here we have also introduced the function α_B according to the definition

$$\frac{c(n)}{\alpha} - \frac{I_B^{(d)}}{2} \equiv \frac{c(n)}{\alpha+\alpha_B}, \quad \alpha_B = \alpha^2 \frac{I_B^{(d)}}{2c(n) - \alpha I_B^{(d)}}; \tag{61}$$

α_B describes the effect of band enhancement of the parameter α.

Representing the solutions in the form (58) or (60) is convenient when we wish to do a comparison with the BCS theory, in which $T_c \propto \exp\{-1/N(\varepsilon_F)V\}$, where $N(\varepsilon_F)$ is the density of states at the Fermi level, and V is the effective attraction. The dimensionless parameter $N(\varepsilon_F)V$ is similar to the expressions $(\alpha+\alpha_B)\langle \psi_d^2(m,\delta)\rangle$ and $(\lambda+\lambda_B)\psi_p^2(m)$ used in our theory. In contrast to the case of triplet superconductivity, the function $\langle \psi_d^2(m,\delta)\rangle$ in the exponent of the exponential function is large compared to $\psi_p^2(m)$ in the vicinity of $m=0$, which ensures, other things being equal, substantially larger values of $\tau_c^{(d)}$.

7. CONCLUSION

The proposed $t-J-I$ model makes it possible to study the superconductivity of strongly correlated electrons with different symmetries (s, p, or d) of the order parameters, facilitates comparison of the high-T_c superconductivity of copper oxides and the superconductivity of the ruthenate Sr_2RuO_4 with low T_c within a unified approach. The above comparison answers the question of why there is high-T_c superconductivity ($T_c \sim 100$ K) in layered copper oxides. In light of our results, the features that copper oxides exhibit are as follows: in two-dimensional CuO_2 layers, owing to the $(d_{x^2-y^2}-p)-\sigma$ bond and strong electron correlations, there form quasiparticles with particle-to-particle hops occurring against the background of strong antiferromagnetic fluctuations. These fluctuations lead to $d_{x^2-y^2}$-type pairing, with the gap anisotropy being such that the Van Hove singularity, which increases the ordinary logarithmic contribution in the equations for T_c and the order parameter, manifests itself in full.

The authors are grateful to N. M. Plakida for discussing the results. This work was made possible by grants from the Federal Program of Support for Integration of Basic Research and Higher Education (Grant No. 69) and the Krasnoyarsk Regional Scientific Fund (Grant 8F0032).

*)E-mail: sgo@post.krascience.rssi.ru

1)Note that in their monograph,[5] Izyumov *et al.* used the diagrammatic technique in the $t-J$ model to derive a similar expression for the gap, where, however, the term $-g\omega_{\mathbf{k}-\mathbf{p}}$ (in our notation) replaces the **p**-symmetric combination $-g(\omega_{\mathbf{k}+\mathbf{p}}+\omega_{\mathbf{k}-\mathbf{p}})$ in (15). This discrepancy has certain consequences, which we leave for Sec. 6.

[1] Y. Maeno, H. Hasimoto, K. Yoshida, S. Nishizaki, T. Fujita, and F. Lichtenberg, Nature (London) **372**, 532 (1994).
[2] T. Oguchi, Phys. Rev. B **51**, 1385 (1995).
[3] T. M. Rice and H. Sigrist, J. Phys.: Condens. Matter **7**, L643 (1995).
[4] I. I. Mazin and D. J. Singh, Phys. Rev. Lett. **79**, 733 (1997).
[5] Yu. A. Izyumov, M. I. Katsnel'son, and Yu. N. Skryabin, *The Magnetism of Collectivized Electrons* [in Russian], Nauka, Moscow (1994).
[6] Yu. A. Izyumov, Usp. Fiz. Nauk **167**, 465 (1997) [Phys. Usp. **40**, 445 (1997)].
[7] J. B. Goodenough, *Magnetism and the Chemical Bond*, Wiley, New York (1963).
[8] A. P. Mackenzie, S. R. Julian, A. J. Diver, G. J. McMullan, M. P. Ray, G. G. Lonzarich, Y. Maeno, S. Nishizaki, and T. Fujita, Phys. Rev. Lett. **76**, 3786 (1996).
[9] T. M. Riseman, P. G. Kealey, E. M. Forgan, A. P. Mackenzie, L. M. Galvin, A. W. Tyger, S. L. Lee, C. Ages, D. McK. Paul, C. M. Aegerter, R. Cubitt, Z. O. Hao, T. Akima, and Y. Maeno, Nature (London) **396**, 242 (1998).
[10] D. J. Singh, Phys. Rev. B **52**, 1358 (1995).
[11] S. V. Tyablikov, *Methods in the Quantum Theory of Magnetism* [in Russian], Nauka, Moscow (1975) [English translation of an earlier Russian edition: Plenum Press, New York (1967)].
[12] R. O. Zaĭtsev and V. A. Ivanov, Fiz. Tverd. Tela (Leningrad) **29**, 2554 (1987) [Sov. Phys. Solid State **29**, 1475 (1987)].

Translated by Eugene Yankovsky

Magnetoresistance and Hall effect in $La_{0.8}Sr_{0.2}MnO_3$

A. E. Kar'kin*)

Institute of Metal Physics, Ural Branch of Russian Academy of Sciences, 620219 Ekaterinburg, Russia

D. A. Shulyatev and A. A. Arsenov

Moscow Institute of Steel and Alloys, 107071 Moscow, Russia

V. A. Cherepanov and E. A. Filonova

Ural State University, 620083 Ekaterinburg, Russia

(Submitted 15 February 1999)
Zh. Éksp. Teor. Fiz. **116**, 671–683 (August 1999)

A comparative study of the longitudinal ρ_{xx} and transverse ρ_{xy} resistivities and magnetic susceptibility χ_{ac} of $La_{0.8}Sr_{0.2}MnO_3$ single crystals and ceramic samples has been conducted in a wide range of temperatures $T=1.7-370$ K and magnetic fields, $H=0-13.6$ T. It turned out that the relation $\rho_{xy}\sim\rho_{xx}$, which is expected to hold in the case of carrier scattering by magnetic fluctuations, applies to the single crystals. In polycrystals, an additional H-dependent contribution to the resistivity tentatively attributed to plane (near grain boundaries) and bulk ''defects'' of the magnetic sublattice has been detected. The scattering of carriers by these defects does not make a notable contribution to the anomalous Hall effect and magnetic susceptibility χ_{ac}. As a result, the curve of ρ_{xy} versus ρ_{xx} seems to be steeper than a linear dependence. Under the assumption that the materials under investigation are metals with constant carrier concentrations, the conductivity $\sigma=1/\rho_{xx}$ due to the critical magnetic scattering calculated in the molecular field approximation reproduces the main features of experimental data, namely, the drop in the amplitude and shift of the resistivity peak near the Curie point with increasing magnetic field H and also a relatively slow change in the derivative $d\sigma/dH$ with increasing temperature in the region $T\leqslant T_C$. The large hole concentration of about two per unit cell derived from Hall measurements indicates that carriers of opposite signs can coexist in these materials. © *1999 American Institute of Physics.* [S1063-7761(99)02308-2]

1. INTRODUCTION

The negative magnetoresistance detected in perovskites containing manganese, like $La_{1-x}Sr_xMnO_3$, over a wide range of doping levels x in both ferromagnetic metals and ferromagnetic insulators is related in one way or another to changes in the long-range or short-range magnetic order, because it was observed only in the ferromagnetic region or at temperatures slightly higher than the Curie point T_C, where ferromagnetic fluctuations are notable. The most common interpretation of this effect is associated with delocalization of charge carriers, which is assumed to take place either when magnetic moments are ordered below T_C or under applied magnetic field.[1–8] The relatively high resistivity ρ near T_C (5–30 m$\Omega\cdot$cm for samples with $x\geqslant 0.2$) and its derivative with respect to temperature ($d\rho/dT<0$), at $T>T_C$ which is typical of semiconductors, provide evidence in favor of this interpretation.[1,4,5,7] The parameter characterizing the region of a metal–insulator transition (MIT), namely, Mott's minimal conductivity

$$\sigma_M\approx 0.03(e^2/\hbar)n^{1/3} \tag{1}$$

is approximately $0.1\,(\text{m}\Omega\cdot\text{cm})^{-1}$ for the carrier concentration $n\sim 3\times 10^{21}\,\text{cm}^{-3}$ (0.2 hole per unit cell). According to this estimate, compounds of this class with doping levels $x>x_c$ (here the critical doping level $x_c\approx 0.15$ for $La_{1-x}Sr_xMnO_3$ or $La_{1-x}Ba_xMnO_3$) should be insulators above T_C and metals at low temperatures, so the concept of MIT induced by magnetic disorder seems quite feasible.

There are, however, several observations that can hardly be interpreted in terms of this MIT model. In many cases where the conductivity is smaller than or on the order of σ_M estimated above, it nonetheless exhibits a clearly metallic behavior. This concerns both the region $T\geqslant T_C$, where one often observes $d\rho/dT>0$ (Ref. 2) and the region of low temperatures, where $\sigma<\sigma_M$, but the conductivity remains constant as $T\to 0$ (Refs. 4 and 5). Further, the conductivity growth induced by magnetic field H is interpreted by the MIT model in terms of decrease (closing) of the gap or pseudogap ΔE. It is plausible that at $T\leqslant T_C$ the value ΔE should drop as a linear function of H: $\Delta E\approx\Delta E_0-a_H H$ (Ref. 9), so that the carrier concentration and conductivity should follow the function

$$n\propto\sigma\propto\exp[(a_H H-\Delta E_0)/T] \quad \text{for} \quad \Delta E>0,$$

$$n\propto\sigma\sim\text{const} \quad \text{for} \quad \Delta E=0,$$

where a_H is a constant independent of both the field and temperature. The conductivity should increase rapidly (exponentially) with H in the range of relatively low fields and

 © 1999 American Institute of Physics

saturate at higher magnetic fields, whereas $d\sigma/dH$ should rapidly (approximately following the function $\exp[\Delta E/T]$) increase with T. These predictions of the $\sigma(H)$ behavior conspicuously contradict experimental data, because the conductivity grows with H linearly at $T \leqslant T_C$ without any sign of saturation, and $d\sigma/dH$ is a relatively slow function of temperature.[3,8,10] According to the model, the carrier concentration n should depend strongly (exponentially) on the field intensity and temperature. The carrier concentrations in $Nd_{0.5}Sr_{0.5}MnO_3$ (Ref. 4) and $Pr_{0.5}Sr_{0.5}MnO_3$ (Ref. 5) derived from measurements of the Hall constant, $n_H = 1/(R_H e)$, at $T < T_C$ are, however, almost constant with both the field and temperature.

Another feasible model describing properties of these materials is based on the assumption that at all temperatures (both above and below T_C) and doping levels $x > x_c$ the conductivity is of metallic nature, the carrier concentration is independent of the magnetic field and temperature, and the effect of negative magnetoresistance is fully determined by the drop in the amplitude of scattering by magnetic impurities under magnetic field. This model, however, should account for large values of ρ at $T \geqslant T_C$. If we set the free path in the conventional Drude formula to the Fermi length λ_F, we obtain the conductivity $\sigma \approx 0.6(e^2/\langle\hbar\rangle)n^{1/3}$, which yields an estimate of the carrier concentration $n \sim 10^{-3}-0.1$ hole per unit cell. This value is considerably smaller than both the concentration derived from Hall measurements, $n_H = 0.5-1.0$ hole per unit cell,[4,5,11] and the estimate $n_H = 0.2$ derived from simple chemical considerations. Thus, the model of metallic conductivity in these materials and in many other systems with strong electronic correlations also runs into difficulties.

Another feature of materials of this type is the notable difference between the $\rho(T,H)$ functions for single crystals and polycrystalline samples in the region of low temperatures, which was graphically demonstrated[2] for $La_{2/3}Sr_{1/3}MnO_3$. This difference was attributed to the scattering in regions near grain boundaries in polycrystals. In the range of low magnetic fields, $H < 0.5$ T, the measurement data for $\rho(T,H)$ were interpreted in terms of spin polaron tunneling through grain boundaries, but no mechanism of magnetoresistance has been suggested for the range of stronger magnetic fields.[2]

This very simplified description of the state of the problem demonstrates that our understanding of the transport in these materials is far from clear. Neither the nature of conductivity in the vicinity of T_C and at higher temperatures (whether the material is a metal or an insulator), or the cause of such a sharp increase in the conductivity induced by magnetic field has been clarified.

In this paper we report the results of a detailed comparative study of the magnetoresistance and Hall effect as functions of the temperature and magnetic field in single crystals and polycrystalline samples with the nominal composition $La_{0.8}Sr_{0.2}MnO_3$. Earlier publications reporting on samples with similar chemical compositions, $La_{2/3}Sr_{1/3}MnO_3$ did not contain, unfortunately, measurements of the Hall effect and magentoresistance near the Curie point. Previously the Hall effect was investigated largely in epitaxial films of various compositions,[4,5] but their defect structure from the viewpoint of resistivity versus magnetic field and temperature is intermediate between single crystals and ceramic samples, and their properties are closer to those of ceramics. In one recent paper[11] measurements of the Hall effect in $La_{1-x}Sr_xMnO_3$ single crystals were reported. Measurements of these magnetic materials yield more information because of the anomalous Hall effect, which occurs as a result of asymmetric carrier scattering due to magnetic disorder (magnetic impurities or defects in the spin lattice). Since the scattering of this type affects both the longitudinal ρ_{xx} and transverse ρ_{xy} components of the resistivity tensor, their field and temperature dependences yield interesting data concerning the mechanism of negative magnetoresistance in these materials.

2. EXPERIMENTAL

$La_{0.8}Sr_{0.2}MnO_3$ single crystals were grown using the floating-zone technique with radiative heating without a crucible.[6] Polycrystalline samples of this material were manufactured by the conventional technique of ceramic synthesis from starting materials La_2O_3, $SrCO_3$, and Mn_2O_3. The weighted quantities of the starting materials were ground and annealed in several stages in a temperature interval 850–1100 °C. The annealing time on the last stage was 70 h. The powders manufactured by this method were compressed and annealed for 20 h at 1300 °C. The close values of T_C and ρ at $T > T_C$ measured in our single crystals and ceramics indicate that their stoichiometric compositions are close.

The resistivity $\rho_{xx} = \rho$ and susceptibility χ_{ac} were measured using samples with sizes of $1 \times 1 \times 6$ mm, the Hall effect parameters ρ_{xy} and ρ were measured in square plane samples with sizes of 1.5×1.5 mm and thicknesses of 120 μm (ceramics) and 80 μm (single crystals) with contacts fabricated by ultrasonic soldering of indium and set symmetrically at the corners. Measurements of ρ_{xx} and ρ_{xy} were conducted using the four-terminal technique at a direct current of 1 mA. In measurements of single crystals, the (110) axis was aligned with the current density and directed perpendicular to the magnetic field. In order to minimize errors, the current and magnetic field directions in Hall measurements were altered, and the current and voltage contacts were interchanged. The resistivity of long samples was measured in the field range $H = 0-13.6$ T as the temperature was varied between 1.7 and 370 K, whereas the parameters of plane samples (ρ_{xy} and ρ) were measured at fixed temperatures from 4.2 to 360 K switched with a step of 20 K by scanning the magnetic field. Since $\rho_{xy} \ll \rho_{xx}$, the typical Hall voltage was on the order of 10 nV against the background of an asymmetrical signal with an amplitude on the order of 10 μV; therefore, the uncertainty in ρ_{xy} was 10–30%. The susceptibility χ_{ac} was measured by the conventional two-coil technique in an ac field on the order of 10^{-4} T oscillating at a frequency of 1 kHz and aligned with the dc magnetic field $H = 0-0.1$ T.

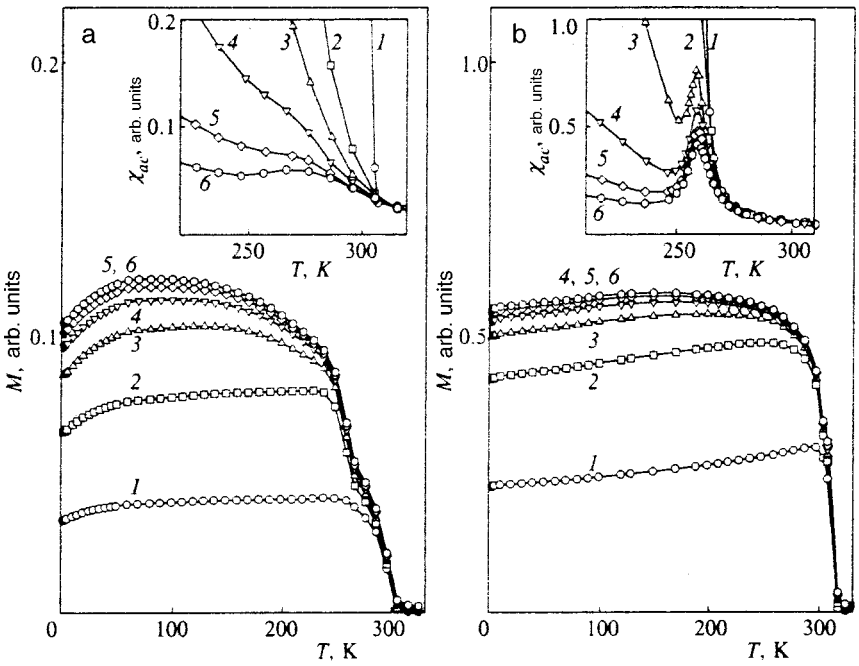

FIG. 1. Calculated of magnetization M as a function of temperature based on measurement data for susceptibility χ_{ac} in (a) single crystals and (b) polycrystalline samples of $La_{0.8}Sr_{0.2}MnO_3$. Curves *1–6* correspond to magnetic field ranging between 0 and 0.1 T. The insets show χ_{ac} near the Curie point.

3. RESULTS AND DISCUSSION

Figure 1 shows magnetizations $M(H,T)$ of single-crystal and ceramic samples derived from measurements of the susceptibility $\chi_{ac}(H,T)$ under magnetic fields of up to 0.1 T. The magnetization almost saturates in the field range $H \geqslant 0.06$ T. At $H = 0.1$ T the ceramic sample displays a peak in the susceptibility at $T = T_C$, which is typical of ferromagnets, whereas in the single crystal the feature near T_C has a smaller amplitude.

The observed behavior of $\rho(T,H)$ is in fair agreement with earlier measurements[2,12] of samples with similar compositions. The value of ρ near T_C in ceramics is a factor of 1.5 higher than in single crystals, and about half of this difference can be attributed to the ceramic porosity, since the absolute value of resistivity at high temperatures is only 30% higher than in single crystals. The curves of ρ versus magnetic field and temperature for ceramics and single crystals near T_C are also very similar.

The differences in $\rho(T,H)$ in the low-temperature region are considerably larger (Fig. 2). At $T < 50$ K the ceramic resistivity can be described by a polynomial $\rho = \rho_0 - a_1 T + a_3 T^3$, where the parameters a_1 and ρ_0 strongly depend on H. The resistivity of single crystals in this temperature interval can be described by a similar formula (see also Fig. 6), but the parameters ρ_0, a_1, and a_3 are more than an order of magnitude lower and nearly independent of the magnetic

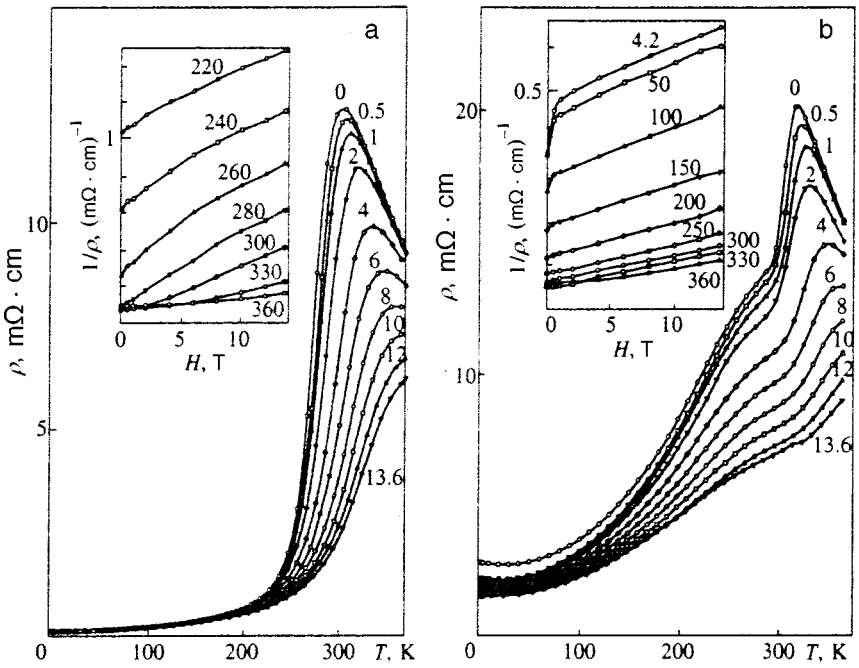

FIG. 2. Resistivity of (a) single crystals and (b) polycrystalline samples of $La_{0.8}Sr_{0.2}MnO_3$ versus temperature. The numbers near the curves show the magnetic field H in tesla. The insets show the conductivity $1/\rho$ as a function of magnetic field, the numbers show the temperature in kelvins.

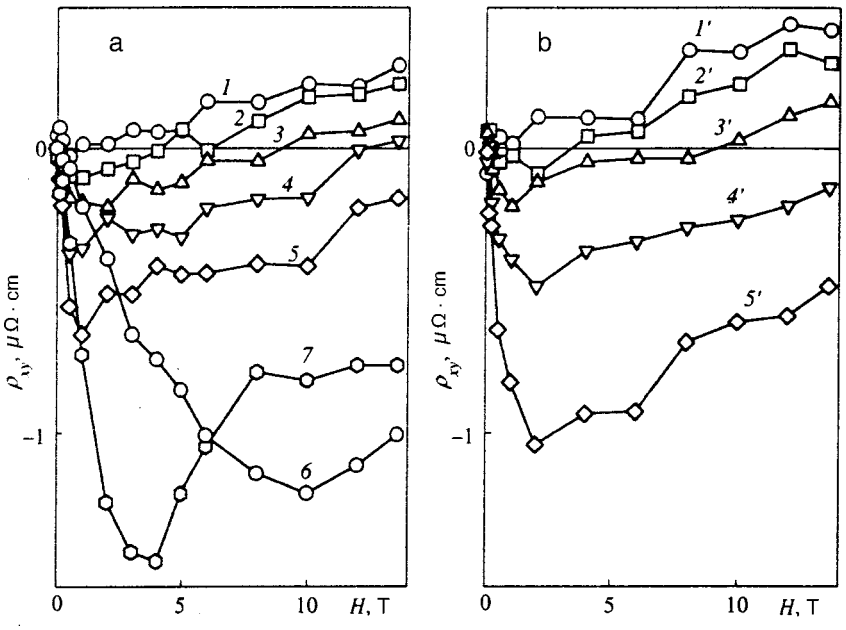

FIG. 3. Transverse resistivity ρ_{xy} versus magnetic field H for (a) single crystals and (b) polycrystalline samples of $La_{0.8}Sr_{0.2}MnO_3$ at different temperatures: (*1*) 4.2 K, (*2*) 60 K; (*3*) 160 K; (*4*) 200 K; (*5*) 260 K; (*6*) 300 K; (*7*) 340 K; (*1′*) 4.2 K; (*2′*) 120 K; (*3′*) 200 K; (*4′*) 260 K; (*5′*) 320 K.

field. In the region of intermediate temperatures $50<T<150$ K the resistivity of single crystals (Fig. 6) and ceramics is closer to functions like $\rho=a_2T^2$. At $T<150$ K the conductivity $\sigma=1/\rho$ of ceramics rapidly grows in the range of low magnetic fields $H\leqslant 0.5$ T, in the range of higher fields it grows almost linearly with the field, and at $T\geqslant 150$ K the dependence is linear throughout the studied field range (Fig. 2). A remarkable feature of the conductivity $\sigma(H)$ of single crystals and ceramics is that $d\sigma/dH$ is almost constant over the entire region of the ferromagnetic state, $T<T_C$ (constant magnetoresistance), although the conductivity σ varies considerably in this region.

Figure 3 shows $\rho_{xy}(H)$ at several temperatures. At T fairly close to T_C, the dominant contribution is due to the anomalous Hall effect of the electronic nature, whereas the normal Hall effect due to holes makes a notable contribution only at low temperatures. In ceramics the normal contribution at low temperatures is a factor of approximately 1.5 larger than in single crystals, which can also be attributed to the porosity of ceramic samples. The value of R_H is relatively low, an order of magnitude lower than that expected at the carrier concentration of 0.2 hole per unit cell. The amplitude of the anomalous Hall effect in ceramics is a factor of almost 1.5 lower than in single crystals (with allowance for porosity, by a factor of 2), and its changes with the magnetic field and temperature are very similar (Fig. 3).

The total Hall signal is given by the formula[13]

$$\rho_{xy}=R_H(T)B+\mu_0R_A(T)M(B,T), \tag{2}$$

where R_H and R_A are the normal and anomalous Hall constants, M is the magnetization. In the limit of strong magnetic field, when the magnetization $M(B,T)$ saturates at M_S, we have $\rho_{xy}=R_H(T)B+\mu_0R_A(T)M_S$. The linear extrapolation of $\rho_{xy}(B,T)$ curves from the region of strong magnetic field allows one to separate the two contributions. The Hall carrier concentration, $n_H=1/(R_He)$, and R_A were determined in earlier studies.[4] As a matter of fact, this technique can be used only in the region of sufficiently low temperatures. The anomalous Hall constant R_A is determined by the same parameters (spin correlators) as the resistivity ρ (Refs. 13–18). In the case of scattering by spin fluctuations, $R_A\propto\rho$ (Ref. 14), even though there are more exotic models which predict the dependence $R_A\propto\rho^2$ (Ref. 18). For ''conventional'' ferromagnets, ρ, just as R_A, is almost constant with B [see Eq. (2)] and the linear extrapolation procedure is quite legitimate. Nonetheless, if ρ and R_A are determined by the same scattering processes whose amplitudes are functions of field, then R_A is also a function of B, i.e, $R_A\propto\rho(B)$, so that Eq. (2) can be rewritten as

$$\rho_{xy}=R_H(T)B+c_H\rho(T,B)M(T,B),$$

where c_H is a constant. Thus, ρ_{xy} is a nonmonotonic and fairly complex function of the magnetic field B, so the conventional procedure for separation of the normal and anomalous contributions to the Hall effect can be used only in the temperature range where ρ can be treated as a field-independent parameter. In the case of a single crystals, this is the range of low temperatures where the resistivity ρ is small and the anomalous contribution to ρ_{xy} is negligible.

In order to proceed further, we assume that R_H is independent of T, as in the case of a conventional metal. After subtracting the normal contribution ($R_H=0.2\cdot10^{-9}$ m^3/C), in the limit of strong magnetic field, where the demagnetization factor and difference between B and H can be ignored (since $\mu_0M_S\approx 0.6$ T for $La_{0.8}Sr_{0.2}MnO_3$, this corresponds to $H\geqslant 1$ T), we obtain

$$\rho_A=\rho_{xy}-R_HH\sim\rho(T,H)M(T,H). \tag{3}$$

In order to check out this formula, we calculated the magnetization in the molecular field approximation using the relation $m=B_2((m+h)/t)$, where $B_2(x)$ is the Brillouin function for spin $j=2$, $m=M/M_S$, $t=T/T_C$, $h=H/H_{mol}$, and the molecular field $H_{mol}=k_BT_C/(2\mu_B)\approx 200$ T (Ref. 13). Figure 4 shows the anomalous contribution ρ_A and prod-

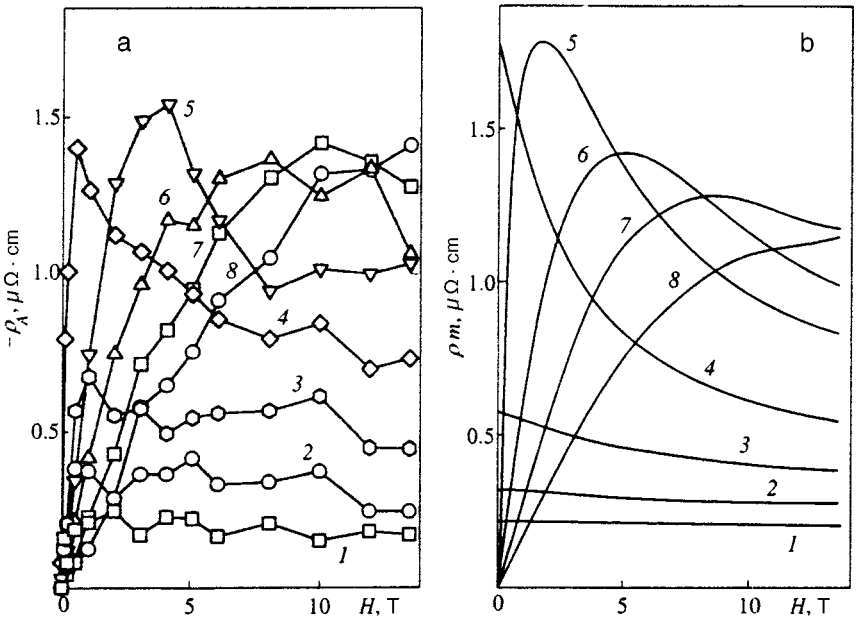

FIG. 4. (a) Anomalous contribution to the transverse resistivity, $\rho_{xy}-R_H H$, versus magnetic field and (b) resistivity ρ times magnetization m calculated in the molecular field approximation for the $La_{0.8}Sr_{0.2}MnO_3$ single crystal at different temperatures: (*1*) 160 K; (*2*) 200 K; (*3*) 240 K; (*4*) 280 K; (*5*) 300 K; (*6*) 320 K; (*7*) 340 K; (*8*) 360 K.

uct ρm as functions of magnetic field H, data for ρ_{xy} and ρ were obtained from the same sample. There is fairly good agreement between the curves describing these two parameters, although some details of these curves (such as peak positions) are notably different. Note that in the region of strong magnetic fields ρm is almost a linear function of H, which is determined by the field dependence of both resistivity ρ and magnetization m. For this reason, the procedure of linear extrapolation from the range of high magnetic field cannot be used for separating the normal and anomalous components of the Hall effect at least at $T \geqslant 180$ K ($0.6T_C$), so the apparent increase in the Hall constant R_H when the temperature approaches T_C (Ref. 4) may be caused by the strong field dependence of the anomalous Hall constant $R_A = \rho_A/M$. In order to check out Eq. (3), we plotted ρ_A/m against $\rho-\rho_0$ using measurements for $H \geqslant 1$ T and $T = 160-360$ K, where ρ_A is markedly different from zero (Fig. 5). For simplicity, we set the residual resistivity ρ_0 due to scattering by nonmagnetic impurities to 0.11 mΩ·cm, which equals the experimental value of ρ at $T = 4.2$ K and $H = 13.6$ T. The processing of these data by the formula $\rho_A/m = \rho^k$ yields $k = 0.95 \pm 0.05$. Thus, the agreement between the measurement data for the single crystals and Eq. (3) is fairly good.

The curves of $\rho_{xy}(H,T)$ for single crystals and ceramics are similar (Fig. 3), whereas the curves of $\rho(H,T)$ are different in the region of low temperatures (Fig. 2). Only near the Curie point, at $T = (0.9-1.15)T_C$, the curves of ρ_A/m and ρ are correlated, whereas at lower temperatures there is no such correlation. This means that ''additional'' scattering mechanisms, which do not occur in single crystals, contribute little to the anomalous Hall effect. The cause of this difference is not quite clear since, even if additional scatterers are concentrated in thin layers near grain boundaries, they should contribute to both the resistivity and Hall effect.

The only mechanism capable of accounting for the $\rho(T)$ curve like that observed in the single-crystal sample near the Curie point (Fig. 2) in a metallic ferromagnet is critical scattering of charge carriers by spin fluctuations.[13,19–21] Critical scattering of a high intensity and the effect of negative magnetoresistance were observed in experiments more than thirty years ago in systems like $Eu_{1-x}Gd_xS$ with relatively low carrier concentrations of $10^{18}-10^{21}$ cm^{-3} (Ref. 21). The critical scattering cross section Σ can be estimated by the formula[20]

$$d\Sigma/d\Omega \propto \chi(q)/\chi_0, \tag{4}$$

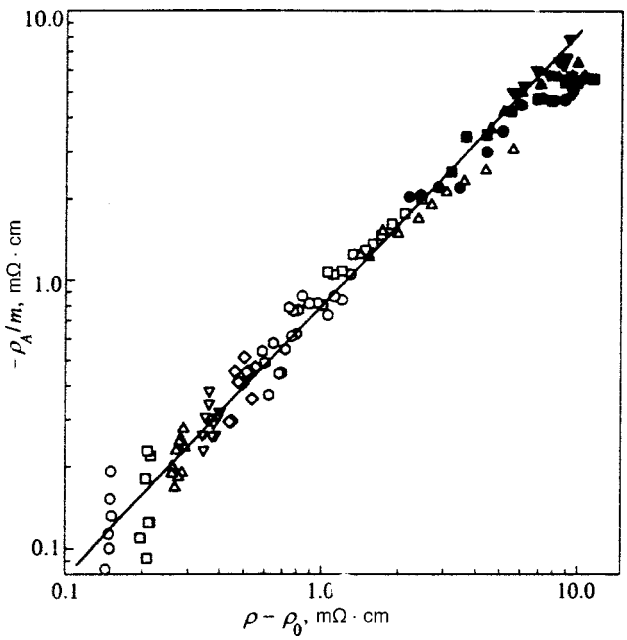

FIG. 5. Ratio between the anomalous contribution ρ_A to the Hall effect and calculated magnetization m as a function of resistivity $\rho-\rho_0$ (see text) for a $La_{0.8}Sr_{0.2}MnO_3$ single crystal in the temperature range 160 to 360 K. The open symbols show measurements at $T < T_C = 300$ K, the filled symbols show measurements at $T \geqslant T_C$.

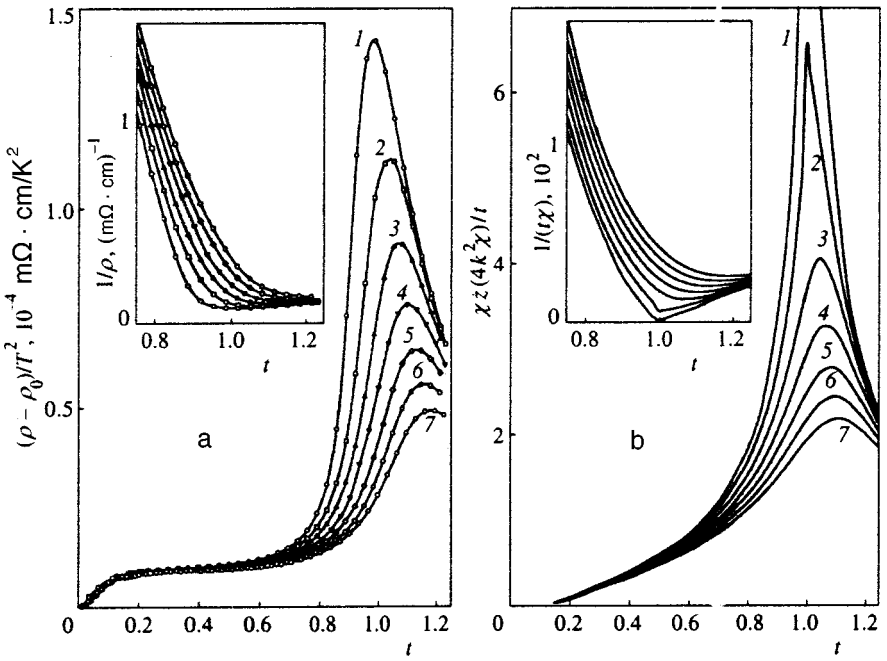

FIG. 6. (a) Measurements of $(\rho-\rho_0)/T^2$ for a $La_{0.8}Sr_{0.2}MnO_3$ single crystal and (b) calculations of $\chi z/t$ versus reduced temperature $t=T/T_C$ (see text). The insets show measurements of conductivity $1/\rho$ and $1/(t\chi)$. Curves *1–7* correspond to magnetic fields 0, 2, 4, 6, 8, 10, and 12 T, respectively.

where q is the scattering vector, Ω is the solid angle, χ_0 is the paramagnetic susceptibility, $\chi(q)$ is the generalized susceptibility expressed at small q by the formula[20]

$$\chi(q)\sim[1/\chi(H,T)+(qR)^2]^{-1},$$

$\chi(H,T)=dm/dH$ is the magnetic susceptibility, and R is the phenomenological parameter on the order of several interatomic distances. In order to calculate ρ, Eq. (4) should be integrated over q with weight factor q^2. Since $\chi_0\propto T_C/T=1/t$, we write the result in the form[20,21]

$$\rho/\rho_\infty=t\chi(H,T)z(y),$$

$$z(y)=2[y-\ln(1+y)]/y^2,$$

$$y=4(k_FR)^2\chi(H,T), \tag{5}$$

$k_F=2\pi/\lambda_F$ is the Fermi wave vector, and ρ_∞ is the resistivity in the high-temperature limit, which corresponds to the paramagnetic scattering. When parameter y is small, which corresponds to temperatures not too close to T_C (small χ) and large wavelengths λ_F, the function $z(y)$ is close to unity, so that

$$\rho/\rho_\infty=t\chi(H,T). \tag{6}$$

This is a rapidly growing function which has a narrow peak near T_C and which tends to a constant in the high-temperature limit. At small λ_F typical of conventional metals like iron, the effect of $z(y)$ is flattening of the peak near T_C; therefore the resistivity ρ monotonically increases in the temperature interval $T<T_C$ and is constant in the range $T>T_C$.

At finite y the shape of function $\rho(T,H)$ is determined by unknown parameters k_F and R. However, the experimental curves of $\rho(T,H)$ can be compared with calculations by Eq. (5) with sufficient accuracy using approximation (6) since in the region $y\leqslant 0.5$ the function $1/z(y)$ can be approximated by the linear function $1+(8/3)y$ with an error of several percent, so the conductivity is described by the equation

$$\frac{\sigma}{\sigma_\infty}=\frac{1}{t\chi(H,T)}+\frac{8}{3}\frac{(k_FR)^2}{t}. \tag{7}$$

Since t is a relatively slow function near the Curie point, a decrease in λ_F simply translates along the abscissa axis the curve of $1/[t_\chi(H,T)]$, which is described by a linear function $\propto|t-1|$ near the Curie point and whose slope on the left is twice as large.

In calculating $\rho(H,T)$ one can use either experimental data for $\chi(H,T)$ or cross sections $d\Sigma(H,T)/d\Omega$ derived from measurements of small-angle neutron scattering. Such data are available for $La_{2/3}Ca_{1/3}MnO_3$ (Ref. 3) and really show that these three characteristics vary concurrently near T_C. Here we use only the susceptibilities $\chi(H,T)$ calculated in the molecular field approximation. Calculations of $1/[t_\chi(H,T)]$ and measurements of $\sigma(H,T)$ at $T=(0.75-1.25)T_C$ and $H=0-13.6$ T are plotted in Fig. 6. A rather good qualitative agreement between the behavior of $\rho(H,T)$ and $1/[t\chi(H,t)]$ is clearly seen. In the range $T<T_C$ the calculated function $1/[t_\chi(H,T)]$ reproduces two basic features of $\sigma(H,T)$, namely, its linear decrease near T_C and the shift of the curve proportional to H. In the region $T>T_C$ the calculations also reproduce a drop in the slope of the $\sigma(H,T)$ curve and a slow nonlinear increase in σ with H, but there is no quantitative agreement between calculated and experimental curves. Note that the calculation does not use any fitting parameters, and the only parameter in the susceptibility calculations is spin $j=2$.

Since the molecular field approximation takes into account only the long-range magnetic order, it fails to describe the magnetization M and susceptibility χ (spin waves) at low temperatures, and the predicted critical indices are incorrect:

$M \propto (1-t)^{\alpha}$, $1/\chi \propto (t-1)^{\beta}$, $\alpha = 1/2$, and $\beta = 1$. More realistic critical indices $\alpha = 0.4$ and $\beta = 4/3$ (Ref. 3), in particular, lead to a smaller slope of $1/\chi$ at $T > T_C$, which is in better agreement with experimental data on $\sigma(T)$. On the other hand, χ has no singularity at $T = T_C$ even in high-quality samples. The peak in the susceptibility at $T = T_C$ (Fig. 1) is sharper in ceramic samples than in single crystals, in agreement with the behavior of ρ in these samples (Fig. 2).

It follows from Eq. (5) that the feature in the resistivity and the negative magnetoresistance, should almost vanish at $k_F R \sim 1$. At $R = 5-10$ Å the carrier concentration at which magnetoresistance can be observed should be smaller than $3 \cdot 10^{19} - 3 \cdot 10^{20}\,\mathrm{cm}^{-3}$ or 0.05–0.5 hole per unit cell for $La_{0.8}Sr_{0.2}MnO_3$. Although this estimate has a poor accuracy, one can, probably, encounter such a situation when the magnetoresistance, diminishes with the doping level.

In order to compare experimental curves of $\rho(T,H)$ with the calculated curves from Eq. (5) over a wider temperature range, we plotted $(\rho - \rho_0)/T^2$ and $\chi z/t$ versus temperature in Fig. 6 using the fitting parameter $k_F R = 0.2$. In the interval $0.5 \leqslant t \leqslant 1.25$ the calculations of $\chi z/t$ adequately reproduce the decrease in the resistivity with the increasing magnetic field, but the shift of the peak due to temperature is a factor of 1.5 slower than on experimental curves. In the range $t < 0.5$ the molecular field approximation does not apply, therefore, we cannot clearly identify the cause of the resistivity behavior $\rho \propto t^3$ at $t < 0.1$ and $\rho \propto t^2$ at $0.1 \leqslant t \leqslant 0.5$ (Fig. 6)—whether it is due to scattering by localized magnetic moments (spin waves) or some other processes.

In ceramic samples, the peak in $\rho(T)$ near T_C (Fig. 2) is markedly lower than a similar peak in single crystals. Since the anomalous Hall effect in ceramics is also two times as small, the resistivity can be expressed in this case in the form of the sum of two components. The first is about half the total resistivity, is controlled by the critical scattering, and has a noticeable magnitude only near T_C. The second, additional component ρ_{add} is related in some way to distortions in the crystal and/or magnetic lattice near grain boundaries and can be described by the formula

$$\rho_{\mathrm{add}} = \rho_1(T) M + \rho_2 H. \tag{8}$$

The first term on the right, linear in the magnetization M, is naturally associated with scattering processes in regions close to grain boundaries, where the magnetization can change its direction on a length of several lattice constants, whereas the second is due to deviations from the magnetic order in a much larger volume of the sample. As the temperature approaches T_C, the magnetization drops and becomes an almost linear function of the field (Fig. 1), which corresponds to the drop in the nonlinear conductivity component (Fig. 2). The ''bulk'' component $\rho_2 H$ in Eq. (8), however, does not contribute to the anomalous Hall effect, nor is it detected in measurements of parameters determined by processes in the bulk, such as magnetization.[2] On the other hand, specific heat measurements of ceramic samples ($0 < x < 0.3$) demonstrate a considerable magnetic contribution in the low-temperature region approximately proportional to T (Refs. 22 and 23). This behavior, which is more typical of spin glasses than conventional ferromagnets, indicates that there can be additional magnetic excitations, which can also lead to magnetic scattering of charge carriers.

The Hall concentration of carriers of about two holes per unit cell seems grossly exaggerated. There should therefore be a mechanism that leads to a smaller Hall constant, such as coexistence of two types of carriers of opposite signs. Probably, the reversal of the thermo-emf experimentally detected in such compounds provides evidence in favor of this assumption.[24] This effect should be conspicuous if the concentrations of carriers of opposite signs (or k_F) are notably different. Near the Curie point, the conductivity σ is proportional to k_F^2, so the thermo-emf $S = (S_h \sigma_h - S_e \sigma_e)/(\sigma_h + \sigma_e)$, where the subscripts e and h label electrons and holes, should have a feature with the shape of a peak because the shapes of σ_h and σ_e versus temperature are notably different in this temperature range. Similar features near T_C may turn up on curves of the Hall constant.

Note that the critical scattering model applies, apparently, only to materials like $La_{1-x}Sr_xMnO_3$ or $La_{1-x}Ba_xMnO_3$, which (for $x \geqslant x_c$) do not demonstrate effects that could lead to a notable restructuring of the electron spectrum and thereby cause a metal–insulator transition. Such effects can induce structural or electronic (charge patterning) phase transitions, or a ferromagnetic transition in cases when it leads to a considerable change in the unit cell volume. These are, probably, the cases when the magnetic field causes ''giant'' changes in the conductivity of several orders of magnitude.[25–29]

Thus, we have shown that, in $La_{0.8}Sr_{0.2}MnO_3$ single crystals, the main cause of scattering of charge carriers, alongside the effect of negative magnetoresistance, is associated with magnetic critical scattering, whereas interaction with defects of other types is relatively weak and noticeable only at low temperatures. The calculations performed in the molecular field approximation adequately reproduce basic features of the conductivity σ and Hall effect as functions of temperature T and magnetic field H, in particular, the approximately linear increasing of σ with H, and the weak temperature dependence of σ at $T \leqslant T_C$ (a constant magnetoresistance). In the region where the critical scattering dominates, there is a linear dependence between the longitudinal ρ and anomalous transverse ρ_A resistivities. In ceramic samples, there is, in addition to the critical scattering, a contribution to the resistivity which is controlled by both the magnetization M and applied magnetic field H. This component caused by scattering near grain boundaries and within them contributes little to the anomalous Hall effect, which leads to an apparent violation of the relation $\rho_A \sim \rho$. Our analysis indicates that $La_{0.8}Sr_{0.2}MnO_3$ is a metal with a low carrier concentration in the entire range of temperatures. The incredibly high carrier concentration of about two holes per unit cell indicates that there should be carries of the opposite sign in these materials.

We acknowledge helpful discussions of our results with Yu. N. Skryabin, M. V. Sadovskii, and É. Z. Valiev. This work was supported by the Russian Fund for Fundamental Research (Project 97-02-17315), MNTTs (Project 636), and the Program of State Support for Leading Scientific Schools of Russian Federation (Project 96-15-96515).

*)e-mail: karkin@orar.e-burg.su

[1] H. L. Ju, C. Know, Qi Li *et al.*, Appl. Phys. Lett. **65**, 2108 (1994).
[2] H. Y. Hwang, S.-W. Cheong, N. P. Ong, and B. Batlog, Phys. Rev. Lett. **77**, 2041 (1996).
[3] J. M. De Teresa, M. R. Ibarra, P. A. Algarabel *et al.*, Nature **386**, 256 (1997).
[4] P. Wagner, D. Mazilu, L. Trappeniers *et al.*, Phys. Rev. B **55**, R14721 (1997).
[5] P. Wagner, I. Gordon, A. Vantomme *et al.*, Europhys. Lett. **41**, 49 (1998).
[6] A. M. Balbashov, S. G. Karabashev, Y. M. Mukovskii, and S. N. Zver'kov, J. Cryst. Growth **167**, 365 (1996).
[7] G. J. Snyder, M. R. Beasley, and T. H. Geballe, Appl. Phys. Lett. **69**, 4254 (1996).
[8] A. J. Millis, B. I. Shraiman, and R. Mueller, Phys. Rev. Lett. **77**, 175 (1996).
[9] N. G. Bebenin and V. V. Ustiniv, J. Phys.: Condens. Matter **10**, 6301 (1998).
[10] J. P. Mitchel, D. N. Argiriou, C. D. Porter *et al.*, Phys. Rev. B **54**, 6172 (1996).
[11] A. Asamitsu and Y. Tokura, Phys. Rev. B **58**, 47 (1998).
[12] V. E. Arkhipov, V. P. Dyakina, S. G. Karabashev *et al.*, Fiz. Met. Metalloved. **84**, 632 (1997).
[13] S. V. Vonsovskii, *Magnetism* [in Russian], Nauka, Moscow (1971).
[14] Yu. Kagan and L. A. Maksimov, Fiz. Tverd. Tela **7**, 536 (1965) [*sic*].
[15] J. Kondo, Prog. Theor. Phys. **27**, 772 (1962).
[16] J. Kondo, Prog. Theor. Phys. **28**, 846 (1963).
[17] E. A. Stern, Phys. Rev. Lett. **15**, 62 (1963).
[18] L. Berger, Phys. Rev. B **2**, 4559 (1970).
[19] P. G. De Gennes and J. Friedel, Phys. Chem. Solids **4**, 71 (1958).
[20] Yu. A. Izyumov and R. P. Ozerov, *Magnetic Neutronography* [in Russian], Nauka, Moscow (1966).
[21] S. Methfessel and D. Mattis, *Magnetic Semiconductors*, Springer, Heidelberg (1968).
[22] B. F. Woodfield, M. L. Wilson, and J. M. Byers, Phys. Rev. Lett. **78**, 3201 (1997).
[23] A. N. Petrov, V. A. Cherepanov, and E. A. Filonova, submitted to Fizika Tverdogo Tela.
[24] M. A. Senaris-Rodriguez and J. B. Goodenough, J. Solid State Chem. **118**, 323 (1995).
[25] A. Asamitsu, Y. Moritomo, and Y. Tokura, Phys. Rev. B **53**, R2952 (1996).
[26] M. McCormack, S. Jin, T. H. Tiefel *et al.*, Appl. Phys. Lett. **64**, 3045 (1994).
[27] A. Asamitsu, Y. Morimoto, Y. Tomloka *et al.*, Lett. Nature **373**, 407 (1995).
[28] S. J. L. Billinge, R. G. DiFrancesco, G. H. Kwei *et al.*, Phys. Rev. Lett. **77**, 715 (1996).
[29] V. Caignaert, A. Maignan, and B. Raeau, Solid State Commun. **95**, 357 (1995).

Translation was provided by the Russian Editorial office.
Edited by S.J. Amoretty

Effect of ordering of mobile oxygen on spectra of two-magnon and electronic Raman scattering of light in $YBa_2Cu_3O_{6+x}$ crystals with different doping levels

A. A. Maksimov,*) D. A. Pronin, S. V. Zaĭtsev, and I. I. Tartakovskiĭ†)

Institute of Solid State Physics, Russian Academy of Sciences, 142432 Chernogolovka, Moscow Region, Russia

M. V. Klein

Science and Technology Center for Superconductivity and Department of Physics, University of Illinois at Urbana-Champaign, 104 S. Goodwin Avenue, Urbana, Illinois 61801, USA

B. W. Veal

Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA

(Submitted 1 March 1999)
Zh. Éksp. Teor. Fiz. **116**, 684–703 (August 1999)

The effect of ordering of mobile chain-site oxygen in $YBa_2Cu_3O_{6+x}$ crystals at different doping levels x on the kinetics of the intensity change of the two-magnon line and the extended structureless electronic continuum in optical Raman spectra and on the superconducting transition temperature T_c, has been studied in detail. An increase in the chain-site oxygen content x leads to a higher contribution of free carriers to the electronic continuum in Raman spectra. The kinetics of the electronic continuum becomes slower with x, whereas the relaxation rate of the two-magnon scattering is a nonmonotonic function of the stoichiometric index. Computer simulations of the relaxation of nonequilibrium states using the Monte Carlo technique qualitatively describe the kinetics observed in experiments. Our results lead us to a conclusion about local inhomogeneities in the electronic and spin systems in CuO_2 planes with scales of several lattice constants. © *1999 American Institute of Physics.*
[S1063-7761(99)02408-7]

1. INTRODUCTION

It is well known that physical properties of the $YBa_2Cu_3O_{6+x}$ single crystal, which is a metal-oxide compound, are determined by the content x of mobile oxygen in chains. It has been established that not only the oxygen content x but also the degree of its ordering plays an important role. This has been verified in numerous experiments using the technique of short-time annealing of superconducting $YBa_2Cu_3O_{6+x}$ crystals with intermediate x at a temperature $T \approx 150\,^\circ$C (Ref. 1). At this temperature, the oxygen mobility in chains is so high that positions of oxygen atoms are almost fully disordered. This procedure of $YBa_2Cu_3O_{6+x}$ sample processing at temperatures no higher than $T \leqslant 200\,^\circ$C does not deplete them of oxygen, which has been checked out in numerous testing experiments.[2] After the annealing, samples are rapidly quenched in liquid nitrogen, and then the relaxation of the disordered state of chain oxygen is measured at the room temperature. Thus changes in the superconducting transition temperature T_c (Refs. 2 and 3), structural lattice parameters,[4] optical properties,[5] and spectra of two-magnon and electronic Raman scattering of light[6] have been investigated. These experiments have unambiguously confirmed the decisive effect of mobile oxygen ordering on physical properties of the $YBa_2Cu_3O_{6+x}$ compound, especially on its superconductivity.

Direct measurements of the relaxation kinetics of Raman spectra and optically induced conductivity of $YBa_2Cu_3O_{6.4}$ after the annealing, however, were only conducted in our earlier work.[7] It turned out that different physical quantities have different relaxation rates, and they can be tentatively separated into two groups: slow and fast parameters. Such parameters as the intensities of the two-magnon scattering, extended electronic continuum, and the charge-transfer band in optical conductivity relax rapidly to their equilibrium values within several tens of minutes. The superconducting transition temperature T_c and low-frequency optical conductivity relax in several hours. Full relaxation of various physical parameters to their initial equilibrium values requires several days.

In the reported work, we have studied the kinetics of Raman spectra changes associated with oxygen ordering in chain planes of $YBa_2Cu_3O_{6+x}$ samples which have $T_c = 18-74$ K with the oxygen content ranging between $x \approx 0.4$ and 0.7, respectively. The main purpose of our study was a detailed investigation in $YBa_2Cu_3O_{6.4}$ crystals of the ''fast'' kinetics of the intensities of the two-magnon scattering and electronic continuum bands in samples with different stoichiometric values.

In order to describe the kinetics observed in experiments and account for the differences among characteristic relaxation times, we have conducted computer simulations of diffusion of mobile oxygen by the Monte Carlo technique on the base of the model of pair interactions suggested earlier.[8,9]

 © 1999 American Institute of Physics

For the quantitative description of oxygen ordering in chains, we have introduced two order parameters describing correlation between a real structure and idealized OI and OII structures at a given time moment over a correlation length r. This paper presents a detailed description of the model and calculations of order parameters of systems relaxing from disordered states as functions of time till the moment when they achieve their equilibrium values. Our computer simulations indicate that different correlation lengths correspond to different relaxation times, which allows us to ascribe the differences between relaxation rates of various physical parameters observed in experiments to differences between characteristic lengths that determine spatial scales over which these parameters change. The decisive factor for rapidly relaxing quantities is local ordering of oxygen, whereas other parameters are affected by formation of longer oxygen chains.

Thus, the kinetic investigation of various physical quantities yield information about characteristic spatial scales of their changes. An extraordinary and remarkable trait of electronic Raman scattering spectra in all cuprate high-T_c superconducting (HTSC) crystals is the presence of a structureless continuum, whose origin has not been elucidated so far. In this connection, we have studied the kinetics of Raman scattering spectra in the process of chain oxygen ordering in crystals with different oxygen contents, which has allowed us, in particular, to clarify the origin of the extended structureless electronic continuum in $YBa_2Cu_3O_{6+x}$ spectra.

2. EXPERIMENT

2.1. Samples

$YBa_2Cu_3O_{6+x}$ single crystals used in our experiments were fabricated from the melt of the mixture of Y_2O_3, $BaCuO_3$, and CuO oxides.[1,2] The technological cycle included the annealing of samples at a temperature of $\sim 520\,^\circ$C for several days in the O_2–N_2 atmosphere with a partial oxygen pressure corresponding to the required stoichiometry. This time was sufficient to achieve a predetermined equilibrium oxygen stoichiometry throughout the crystal volume with the oxygen content x controlled to within 0.01 (Ref. 1).

We studied in our experiments $YBa_2Cu_3O_{6+x}$ twin crystals with relatively low temperatures of the superconducting transition ranging between 18 and 74 K with oxygen contents x between 0.4 and 0.7, respectively. The sample sizes were $1\times1\times0.2$ mm, they had fairly narrow superconducting transition widths, $\Delta T_c \lesssim 1$ K, and mirror surfaces. The small widths of the superconducting transitions and the dedicated measurements of Raman spectra with a high spacial resolution ($\lesssim 1.5\,\mu$m) from different surface areas, supported the assumption of the high homogeneity of the oxygen content in our samples.

In order to obtain a high degree of oxygen disorder, a tested crystal was placed in a specially designed, vacuum-tight, thin-wall copper container with a small inside volume and annealed in a furnace at a temperature $\sim 150\,^\circ$C for 15–20 min. After that the container with the sample was quickly (in less than 0.1 s) immersed in liquid nitrogen. The container was then warmed to room temperature, the sample was extracted and either used in Raman measurements at the room temperature or quickly placed in a low-temperature cryostat for measurements of T_c. The typical time of the sample exposure to the room temperature before the first measurement was between two and five minutes. The fact that the sample was insulated by the vacuum-tight container when its temperature changed abruptly was important for protection of its surface, which is especially important for optical measurements.

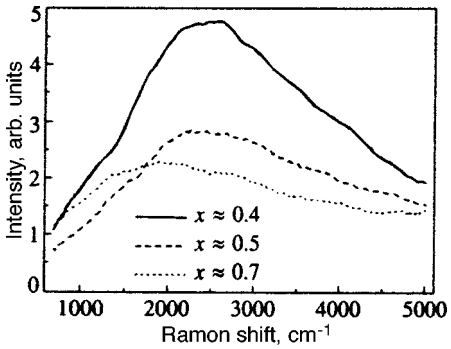

FIG. 1. Optical Raman spectra of $YBa_2Cu_3O_{6+x}$ crystals in the $x'y'$ polarization recorded at room temperature at different values of x.

2.2. Experimental techniques

Raman scattering spectra were recorded by a Dilor XY-500 triple grating monochromator equipped with a multichannel light detector. Samples were excited by the Ar^+ laser line with $\lambda = 4880$ Å, and the power density in the laser spot was $P < 4$ W/cm^2. In kinetic studies, Raman spectra were recorded in real time with a time step of 20–60 s in the pseudo-back-scattering configuration from a fixed section of the **ab**-plane on the crystal surface at room temperature. The laser light intensity during the recording was controlled to an accuracy within 1%.

In measurements of the superconducting transition temperature, the dynamic magnetic susceptibility of crystals was measured as a function of temperature. T_c was fixed at the middle level of the magnetic screening signal.

The good reproducibility of Raman spectra and measured T_c in multiple cycles of thermal processing of one sample indicates that the total oxygen content in the sample remained constant during short-time annealings at temperatures less than 200 °C (Ref. 2).

2.3. Experimental results

Figure 1 shows optical Raman spectra of $YBa_2Cu_3O_{6+x}$ crystals with various oxygen contents. The plots contain high-frequency sections of Raman spectra in the $x'y'$ polarization, where the two-magnon Raman scattering band is usually observed in dielectric phases of HTSC materials.[10,11] We have also detected the two-magnon band in our experiments with superconducting samples with the mobile oxygen content of up to $x \approx 0.7$ and the critical temperature $T_c \leqslant 75$ K. The superconducting transition at $T < T_c$ has little effect on short-range antiferromagnetic correlations in $YBa_2Cu_3O_{6+x}$. Further doping of $YBa_2Cu_3O_{6+x}$ to $x \geqslant 0.7$

destroys antiferromagnetic correlations, which is observed experimentally, first, as softening and then almost complete elimination of the two-magnon scattering line, i.e., superconductivity and antiferromagnetism coexist far beyond the insulator–superconductor transition in $YBa_2Cu_3O_{6+x}$.

Immediately after quenching the samples annealed at 150 °C for 15 to 20 min in liquid nitrogen, we observed a drop in the superconducting transition temperature T_c, whose rate increased with decreasing oxygen content x in the test sample. The measured critical temperature T_c largely recovers to its equilibrium value in a sample held at room temperature for about 10 h. Moreover, the intensities of the two-magnon scattering, I_{2ML}, and extended continuum, I_{EC}, in Raman spectra recorded at room temperature increase after quenching, and the growth in the continuum intensity increases more rapidly than the intensity of the two-magnon scattering. These two parameters change most rapidly during the first 30–100 min at room temperature. Full recovery of all physical parameters to their initial (equilibrium) values takes several days.

We represented our experimental results in the form

$$\hat{\mathbf{f}}(\mathscr{A}(t)) \equiv \ln\left|\frac{\mathscr{A}(t)-\mathscr{A}(t_\infty)}{\mathscr{A}(t_0)-\mathscr{A}(t_\infty)}\right|, \tag{1}$$

where $\mathscr{A}(t_\infty)$ is the equilibrium value of the measured physical quantity, and $\mathscr{A}(t_0)$ is the value measured immediately after the annealing. In this representation, if the parameter relaxes to its equilibrium value following a simple exponential law with a time constant τ, where $\mathscr{A}(t)-\mathscr{A}(t_\infty) = [\mathscr{A}(t_0)-\mathscr{A}(t_\infty)]\exp(-t/\tau)$, the function $\hat{\mathbf{f}}(\mathscr{A})$ is linear. The electronic continuum intensity I_{EC} was defined as the mean intensity in the frequency interval $\omega \approx 400-800\,\mathrm{cm}^{-1}$, and that of the two-magnon scattering, I_{2ML}, was determined by subtracting the continuum intensity from the peak amplitude.

Figure 2 shows relative changes in T_c of the annealed $YBa_2Cu_3O_{6+x}$ samples as a function of the relaxation time at the room temperature expressed in terms of $\hat{\mathbf{f}}(T_c)$ [Eq. (1)]. Immediately after quenching, the critical temperature of the sample with $x \approx 0.4$ dropped from 18 to 2.9 K, and in the sample with $x \approx 0.5$ it dropped from 52.5 to 47.4 K. The superconducting transition temperature is a slowly changing parameter. Characteristic relaxation times τ for T_c derived from experimental data (straight lines in Fig. 2) range between 400 and 600 min. In the intervals of time longer than those plotted in Fig. 2 (greater than 900 min), the relaxation rate of the critical temperature to its equilibrium value is lower.

Measurements of the electronic continuum intensity I_{EC} as a function of exposure time at room temperature after the annealing and quenching of $YBa_2Cu_3O_{6+x}$ crystals with various oxygen content are plotted in the similar form in Fig. 3. It is clear that in the initial, relatively fast stage of relaxation, these measurements can be fairly accurately approximated by exponential functions (straight lines in Fig. 3) with time constants of several tens of minutes. Note that the relaxation time grows notably with the oxygen content. After the initial stage, the relaxation process at longer times is much slower.

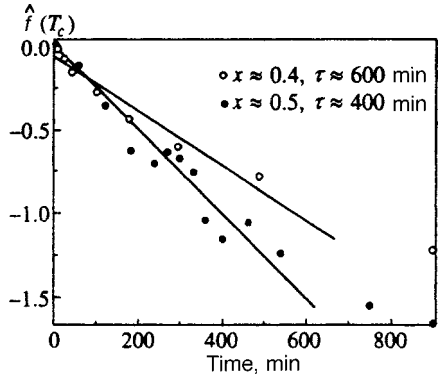

FIG. 2. Superconducting transition temperature T_c transformed using function $\hat{f}(T_c)$ [Eq. (1)] versus time for $YBa_2Cu_3O_{6+x}$ samples with $x \approx 0.4$ and 0.5. The straight lines are approximations of the initial sections of experimental curves by simple exponential functions of time.

While the kinetics of the electronic continuum intensity I_{EC} becomes slower at higher concentrations of chain oxygen, the two-magnon scattering intensity demonstrates a more complex, nonmonotonic dependence on the stoichiometry. Figure 4a, which shows the experimental curves of the two-magnon scattering intensity versus time in the form of $\hat{\mathbf{f}}(I_{2ML})$ [Eq. (1)], clearly demonstrates that their relaxation to equilibrium values is faster than in the case of the electronic continuum (Fig. 3). The relaxation rate of this process gradually changes from larger to smaller values so that the initial sections of the $I_{2ML}(t)$ curves cannot be described by single exponential functions. Nonetheless, model calculations plotted in Fig. 4b are in reasonable agreement with the experimental data. The next section describes in detail the model used in our computer simulations and their results.

3. NUMERICAL CALCULATIONS

3.1. Model description

As was demonstrated by Morris *et al.*,[12] the oxygen mobility is higher in the plane of CuO_x chains ($x<1$), and at

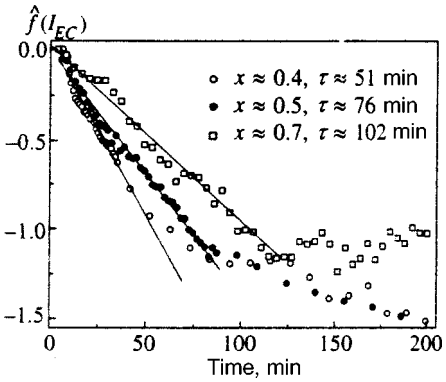

FIG. 3. Intensity of electronic continuum in Raman spectra versus time expressed in terms of function $\hat{f}(I_{EC})$ [Eq. (1)] measured in $YBa_2Cu_3O_{6+x}$ samples with $x \approx 0.4$, 0.5, and 0.7. The straight lines are approximations of the initial sections of experimental curves by simple exponential functions of time.

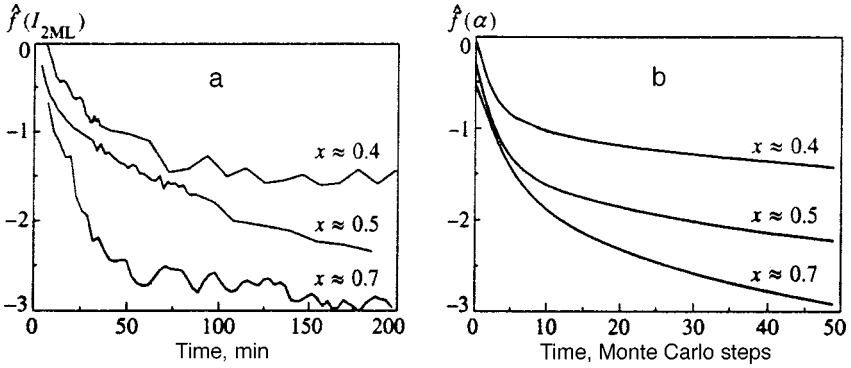

FIG. 4. (a) Intensities of two-magnon Raman scattering as functions of time after the annealing and quenching of $YBa_2Cu_3O_{6+x}$ crystals transformed using function $\hat{f}(I_{2ML})$ [Eq. (1)] for samples with $x \approx 0.4$, 0.5, and 0.7. The curve for $x \approx 0.5$ is shifted by -0.25 along the ordinate axis, the curve for $x \approx 0.7$ is shifted by -0.5. (b) Calculations of the order parameter $\hat{f}(\alpha)$ [Eq. (3)] calculated using a cluster with dimensions of 3×3 lattice constants ($\Sigma_{r=1}^{3} \alpha_1^r$) for samples with various oxygen contents $x = 0.4$, 0.5, and 0.7. The curves for $x = 0.5$ and 0.7 are also shifted along the ordinate axis by -0.25 and -0.5, respectively.

room temperature oxygen atoms diffuse largely in this plane. In modeling the oxygen diffusion, we used a square lattice of possible oxygen positions O1 and O5 with a lattice constant $a = a'/\sqrt{2}$, where a' is the lattice constant of $YBa_2Cu_3O_{6+x}$ crystals in the CuO_x plane. Figure 5 shows the CuO_x plane. All possible oxygen sites form a lattice with the constant $a'/\sqrt{2}$ turned through an angle of 45° with respect to the **a** and **b** crystal axes. Oxygen sites labeled as O1 and O5 are arranged in the chess-board pattern.

We take into consideration only oxygen transitions to eight nearest sites. These transitions can be classified with three groups. Transitions of the first group, NN (nearest neighbors) are transitions to the four neighboring sites through distance a (henceforth $a = a'/\sqrt{2}$). Transitions of the second and third type, NNN and NNNCu (next NN), are diagonal transitions through distance $\sqrt{2}a$, and the difference between them is in the positions of copper sites.

In calculating the probabilities of hops and describing the diffusion-induced oxygen ordering we have used the ASYNNNI (asymmetric next-nearest neighbors Ising) model, which was suggested for describing this system previously.[8,9] In this model, the full system energy is calculated as a function of oxygen configuration by summing all energies of pair interactions. Three constants, V_1, V_2, and V_3, of pair interactions are considered (Fig. 5): V_1 is the energy of repulsion between two neighboring oxygens at lattice sites O1 and O5; V_2 is the energy of attraction between two oxygens located at like lattice sites at a distance $\sqrt{2}a$ and separated by a copper atom (intrachain attraction); V_3 is the energy of repulsion between two atoms at like lattice sites at a distance $\sqrt{2}a$ (interchain repulsion). In our calculations we have assumed $V_2 = -0.3478V_1$ and $V_3 = 0.1594V_1$, where $V_1 = 367$ meV (Refs. 8 and 9). The full system energy is expressed as a sum of all energies of interactions between neighboring oxygen atoms:

$$E = V_1 \sum^{\mathrm{NN}} n_i n_j + V_2 \sum^{\mathrm{NNNCu}} n_i n_j + V_3 \sum^{\mathrm{NNN}} n_i n_j, \qquad (2)$$

where n_i and n_j are occupancies of oxygen sites: $n = 0$ if a site is vacant and $n = 1$ if a site is occupied. In order to calculate the probability of a transition to a neighboring vacant site, one should find the energy change ΔE when a probing atom is transferred to the vacant site. If $\Delta E < 0$ (i.e., the transition of the probing atom to the vacant site leads to a drop in the total energy), we assume that the absolute transition probability $P_{\Delta E<0} = \exp(-Q/T)$, where $Q \approx 1$ eV is the energy barrier height. Otherwise ($\Delta E > 0$), we have $P_{\Delta E>0} = \exp[-(\Delta E + Q)/T]$. If the neighboring site is occupied, $P = 0$. Note that $\exp(-Q/T)$ is a common factor and simply determines the time scale of the process. Therefore, without loss of generality, at a constant temperature we can set $Q = 0$ in the calculations. The probability of remaining at the same site was equated to unity in our calculations. The sum of probabilities of all possible transitions (with $P > 0$) was normalized in such a way that the sum of relative probabilities p was equal to unity, $\Sigma p = 1$. The probabilities p calculated for each oxygen atom are input parameters for the generator of random numbers which determines further the behavior of the atom: whether it goes to a vacant site or remains at the same site, with the probability of each event is proportional to p.

In calculating oxygen diffusion, we used a 128×128 square lattice with cyclic boundary conditions. The Monte Carlo technique is based on sequential repetition of calculation steps. On each step, we considered each oxygen atom and calculated relative probabilities of its behavior. In accordance with the output of the random number generator, the atom either remains at its site or moves to a vacant neighboring site. The order of atoms processed at each step is random and changes from step to step. After each step of the Monte Carlo process, the configuration of atoms was fixed for the calculation on the next step. Usually our calculations included 5000 steps. Note that the unit time in this process is one Monte Carlo step, which corresponds in the real world to several minutes at room temperature, given the definition of relative probabilities described above.

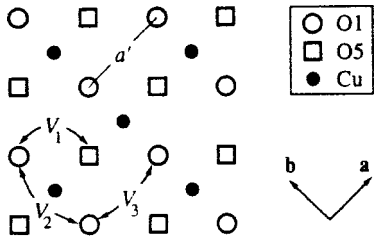

FIG. 5. Configuration of oxygen (O1 and O5) and copper (Cu) atoms in the chain plane of $YBa_2Cu_3O_{6+x}$ crystals and parameters of pair interaction V_1, V_2, and V_3 between nearest neighbors used in computer simulations. The vectors **a** and **b** are aligned with the axes of the crystal lattice.

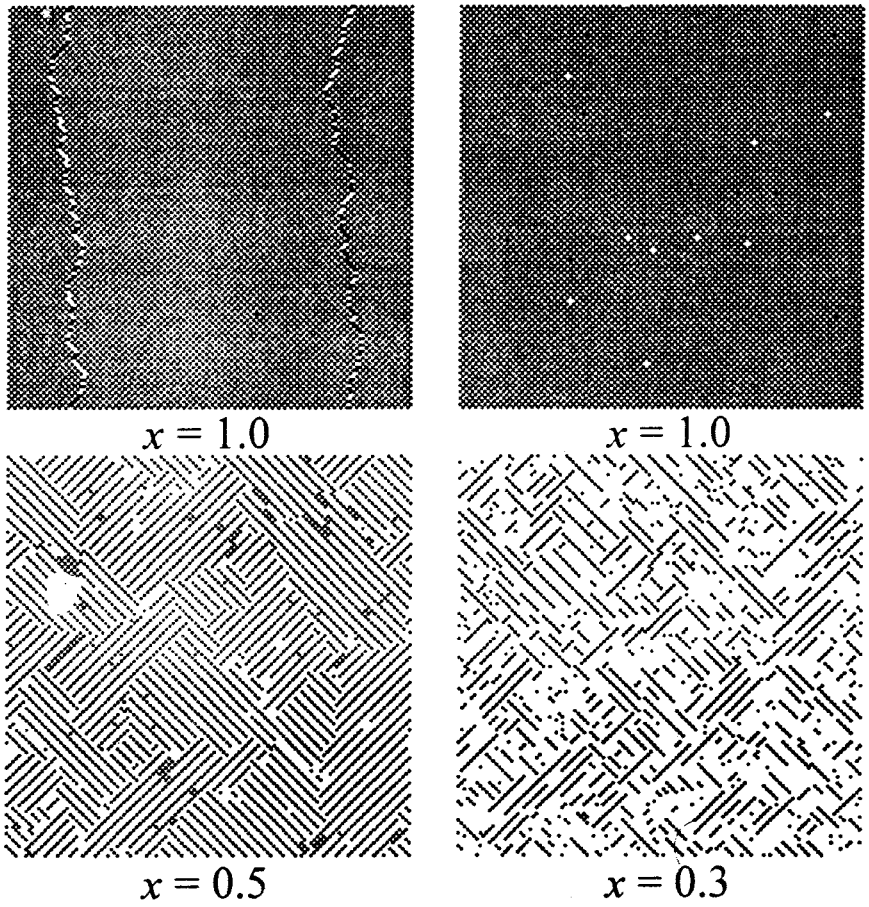

FIG. 6. Calculated positions of oxygen atoms in the CuO_x plane of $YBa_2Cu_3O_{6+x}$ crystals after 5000 Monte Carlo steps at different x. The full circles show atoms at O5 sites, the open circles show atoms at O1 sites.

3.2. Order parameters

In studying ordering of oxygen atoms in the square lattice determined by pair interaction between them, we analyzed the initial configuration of atoms corresponding to their random distribution. The system evolution at the room temperature $T=0.07V_1$ ($T\approx 300$ K) was then analyzed over 5000 Monte Carlo steps. During this time interval, the system transfers to a state which can be assumed to be at equilibrium when $T=300$ K.

Visually, the oxygen configuration after 5000 Monte Carlo steps is fairly ordered (Fig. 6) and corresponds to the orthorhombic OI structure at the oxygen content $x=1$ (all O1 sites are occupied and O5 sites are vacant, or vice versa). At $x=0.5$ the resulting structure is similar to OII, when sites of one sort, for example, O5, are vacant, and other sites, O1, are half occupied. In this case, oxygen atoms form fairly long chains alternating with chains of vacancies. At oxygen concentrations different from $x=0.5$ and $x=1$, the pattern does not decompose into OI and OII phases of large sizes, but forms as a configuration with a uniform density of oxygen atoms in an intermediate phase. Note that if the matrix for numerical calculations is sufficiently large, 128×128, formation of twin regions (with oxygen distributions at O1 and O5 sites reversed) with dimensions of about $50a$ can be detected (Fig. 6). The twins are fairly stable and can persist without changes in their patterns up to 25000 Monte Carlo steps.

If it is not sufficient to visualize ordering of the atomic pattern obtained in our calculations, a numerical characteristic (order parameter) should be introduced for studying the kinetics of this process and its effect on various physical parameters such as T_c. Different groups of researchers used different physical quantities as order parameters. So, Poulsen *et al.*[9] for example, calculated the total system energy as a function of time, which was defined as a sum of all pair interactions [Eq. (2)]. In the process of relaxation, the system energy drops. McCormack *et al.*[13] defined twelve sorts of elementary clusters with oxygen atoms at their centers. Depending on the cluster sort, it was assumed to inject into the cuprate plane a certain charge h_i. The values h_i for each cluster sort were determined on the basis of band calculations[14] for ordered structures. The authors then calculated the fraction of clusters of each sort in the structure and determined the average charge density h over the lattice. Considering T_c to be proportional to the average charge density in the cuprate plane, they obtained on the qualitative level T_c as a function of the oxygen concentration. That publication lists only the values calculated by averaging over large numbers of steps for the equilibrium state at room temperature, but does not characterize kinetic properties of the ordering process. The resulting curve of T_c versus x has a shape similar to the experimental curve;[15] in particular, it has the notorious feature at $x\approx 0.5$. Note, however, that both the total system energy and charge densities are parameters averaged over the entire cuprate plane. For this reason, their utilization as order parameters in analyzing various physical

processes associated with ordering of the oxygen distribution in $YBa_2Cu_3O_{6+x}$ crystals is limited because these parameters take into account formation of an ordered structure only indirectly.

Ceder *et al.*[16] calculated the numbers of copper atoms without neighboring oxygens and of those with one or two oxygens at neighboring lattice sites. This scheme is based on a simple chemical mechanism of charge transfer (see for details Ref. 17) and cannot yield T_c as a function of x. These authors, however, investigated the kinetics of the number of copper atoms without neighboring oxygen atoms and detected two time constants of about 100 and 2000 Monte Carlo steps.

Poulsen *et al.*[18] calculated numbers of oxygens belonging to OI and OII clusters. The minimal structures, which were dubbed clusters, had dimensions of 4×4 and 8×8, respectively. They then assumed that oxygen atoms belonging to OII clusters provided $T_c=60$ K, and atoms belonging to OI clusters yielded $T_c=90$ K, so that the resulting T_c was calculated as a weighted average over all oxygen atoms. The calculations for room-temperature equilibrium configurations at different oxygen concentrations x were in good agreement with experimental data. Unfortunately, this technique cannot be applied to weakly order structures which contain few clusters of the dimensions given above. Consequently, this method cannot be used in studies of the ordering kinetics of oxygen configurations, when the initial state is highly disordered.

In this paper, we suggest using as an order parameter a factor of structural correlations in a system which is free from all drawbacks listed above. For each structure we calculated on each step parameters α_1 and α_2' corresponding to the correlation between the oxygen atom environment and that expected in phases OI and OII, respectively. The calculation was performed as follows. In calculating an order parameter over a distance scale r, occupations of all O1 and O5 sites on the boundary of the square of width $2r$ centered at the probe oxygen site were analyzed. The scales were taken to be integer numbers $r=1,2,3,\cdots,6$, where the unit length is the lattice constant a (the separation between nearest O1 and O5 sites). If the occupancy at the square boundary was equal to that expected for the ideal structure (irrespective of whether it was vacant or occupied), the site was equipped with number 1, and if the occupancy was different, the site was ascribed -1. These alternating numbers were added for all sites on the square boundary and the sum divided by the number of sites under investigation. The resulting correlator between the actual structure and configurations OI and OII at distance r was calculated for all atoms in the lattice and averaged over these atoms.

Thus, one can calculate for any distribution of oxygens in the lattice, i.e., on any step of the Monte Carlo calculation and at all concentrations x, two order parameters α_1 and α_2', which characterize the correlation between a specific structure and phases OI or OII. At $x=1$ for the ideal OI structure $\alpha_1=1$, and at $x=0.5$ for the ideal OII configuration $\alpha_1=0.5$. In real calculations, after a sufficient number of iteration steps (5000 Monte Carlo steps) $\alpha_1\lesssim x$. If the order parameter $\alpha_1=0$, it corresponds to full disorder in the distribution of oxygen atoms, irrespective of the concentration x. In the case of random distribution of oxygen atoms $\alpha_2'=1/2-x/2$. At $x=1$ we have for the ideal OI structure $\alpha_2'=0.5$, and at $x=0.5$ we have for the OII structure $\alpha_2'=1$. In real calculations, the values of the order parameter are slightly smaller. In order to bring the order parameter to zero in the case of random distribution, we introduced the order parameter defined as $\alpha_2=\alpha_2'-1/2+x/2$. The order parameters defined in this section are universal and suitable for characterizing both ordered and disordered structures.

3.3. Final order parameters

In the reported work, we have analyzed the order parameters α_1 and α_2 calculated for correlation distances $r=1,2,3,...,6$ at oxygen concentrations $x=0.1,0.2,0.3,...,1$ and at temperature $T=0.07V_1$ (approximately 300 K). We have also investigated variations in the ordering kinetics caused by widening the set of allowed transitions. Three sets of allowed transitions were used in the model: (1) oxygen atoms can transfer only to four nearest neighboring sites over distance a (O1–O5 or O5–O1); (2) the same transitions plus two transitions to sites at distance $\sqrt{2}a$ without a jump over a copper atom (O1–O1 or O5–O5); (3) all eight possible transitions. Our analysis has demonstrated that introduction of additional transitions leads to a faster system relaxation and shortens the time constant, but does not fundamentally change the result. All the results discussed in the following sections have been obtained using the second set of allowed transitions.

In this section, we will discuss only the final values of order parameters (after 5000 Monte Carlo steps), which are assumed to be equilibrium values for the specific temperature, without analyzing their evolution on the 5000 Monte Carlo steps. Figure 7 plots equilibrium values of the order parameters α_1 and α_2 at room temperature for different correlation lengths $r=1,2,3,...,6$. The parameter α_1 gradually increases with x and is close to unity at $x=1$, whereas α_2 has a peak at $x=0.5$. The order parameter decreases with increasing correlation length r, which indicates that the ordering is not ideal (in an ideal structure, the order parameter does not depend on r). Note that the order parameter calculated at $r=1$ is markedly different from all others, because the model takes into account only interactions between nearest neighbors in calculating the energy (constants V_1, V_2, and V_3). If the temperature in the calculations is increased, the degree of correlation with the OI structure (parameter α_1) changes little, whereas the degree of correlation with the OII configuration (parameter α_2) changes noticeably, in particular, the peak at $x\approx0.5$ flattens rapidly.

Of great interest is the behavior of the sum $\alpha_1+\alpha_2$ as a function of the oxygen concentration x. Figure 8 shows the order parameter α_1 calculated at $r=1$, 2, and 3 plus the order parameter α_2 at $r=2$, 4, and 6, respectively. In other words, these sums have been calculated at equal lattice constants of OI and OII sublattices (the unit cell size in the OII phase is twice as large as in the OI phase). This sum of the order parameters shows an unexpectedly good correlation with the critical temperature T_c of $YBa_2Cu_3O_{6+x}$ samples annealed at

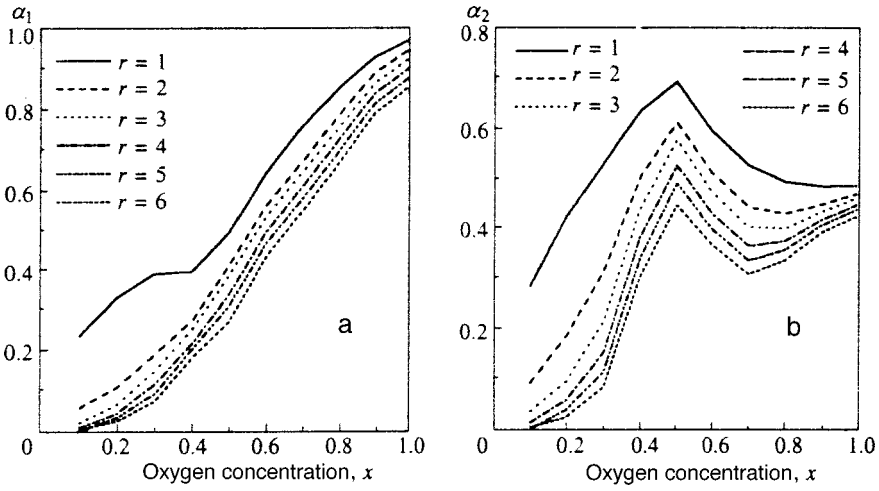

FIG. 7. Order parameters (a) α_1 and (b) α_2 after 5000 Monte Carlo steps as functions of oxygen concentration x for different correlation lengths r; $T=0.07V_1\approx 300$ K.

room temperature, plotted against the oxygen concentration. The curve has the characteristic feature in the shape of a flat section in the region $x=0.5-0.7$, just as the critical temperature curve,[15] moreover, the ratio between the order parameter sums at $x=0.5$ and $x=1$ is about 2/3 ($T_c^{x=0.5}/T_c^{x=1.0}=60\,\mathrm{K}/90\,\mathrm{K}=2/3$).

The results listed above clearly indicate that structural order parameters α_1 and α_2 are directly related to the superconducting properties of $YBa_2Cu_3O_{6+x}$ crystals, so there is good reason to utilize them in studying the effect of the kinetics of oxygen ordering on the physical properties of these crystals. The increase in T_c caused by aging of an annealed sample is associated with the specific ordering of atoms and vacancies (phase OII), which takes place at the room temperature but is absent at higher temperatures.

3.4. Ordering kinetics

This section describes investigations of kinetics of oxygen ordering based on the model discussed in the previous sections. The initial distribution is a lattice with randomly located atoms. The concentration of oxygen atoms corresponds to the stoichiometric index x varied between 0.1 and 1 with a step of 0.1.

In the initial time interval within 100 Monte Carlo steps, regions of small dimensions ($R\approx 10a$) are formed in such a way that the atoms inside them are all at equivalent crystallographic sites, either O1 or O5 (twins). Atomic transitions inside twins form site O1 to site O5, on the contrary, have a very low probability. In the process of further relaxation till the 5000th Monte Carlo step, twin boundaries move spontaneously, as a result, the twins grow, and it is possible to obtain in a lattice with dimensions of 128×128 after 5000 steps either a single twin occupying almost the entire area and having an almost ideal structure (with less than ten defects) or two twins with almost ideal internal structures and occupying approximately equal areas (Fig. 6). As a result of numerous computer simulations with 5000 Monte Carlo steps at $x=1$, we found that the probabilities of these two realizations are approximately equal. There is every reason to assume that both these configurations are stable (up to 25000 Monte Carlo steps). In order to study the evolution of twins over longer times, one should take a larger matrix for calculations.

In our study attention was focused on the initial stage of the relaxation at times $t<100$ Monte Carlo steps. Introduction of order parameters, which depend on the correlation length, has allowed us to trace the emergence of both the short-range order (at $r=1$ the ordering of nearest neighbors) and ordering over longer ranges (up to $r=6$). On each step, six order parameters α_1^r were calculated. At the initial moment, all order parameters at all r and oxygen concentrations x are zero, which corresponds to a random distribution. After 5000 Monte Carlo steps, the order parameters approach x, and their values diminish with the correlation length r (Fig. 7). Note that chance plays an important role in the problem under discussion. The differences between the order parameters calculated at initial steps of the Monte Carlo process in two independent simulations are comparable to the order parameter itself. Note also that the order parameters calculated in one simulation are somewhat ''noisy'' (the noise amplitude may be as high as 10%). This circumstance makes it impossible to compare order parameters obtained in single computer simulations. In our analysis, we have used the parameters obtained by averaging over 100 simulations as functions of time.

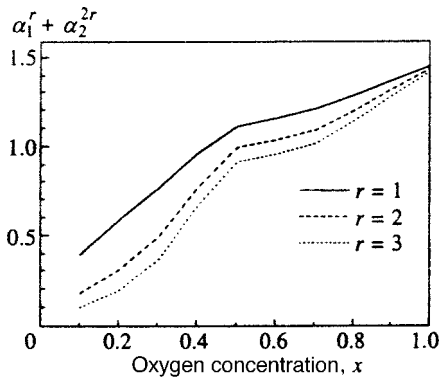

FIG. 8. Sum of calculated equilibrium order parameters $\alpha_1^r+\alpha_2^{2r}$ versus oxygen concentration x for correlation lengths $r=1$, 2, and 3.

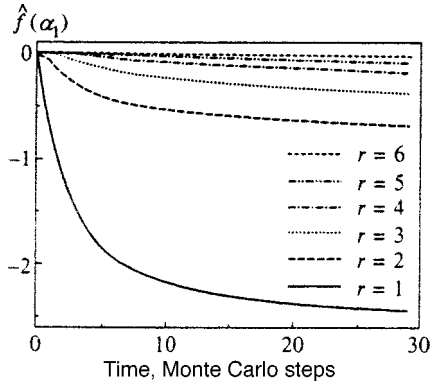

FIG. 9. Calculated order parameter α_1 in the $YBa_2Cu_3O_{6.4}$ crystal as a function of the number of Monte Carlo steps expressed in terms of the function $\hat{f}(\alpha_1)$ [Eq. (3)] for different correlation lengths.

It is interesting to represent the resulting functions of time in the form similar to that selected for experimental data [Eq. (1)]:

$$\hat{f}(\alpha(t)) \equiv \ln\left|\frac{\alpha(t)-\alpha(t_\infty)}{\alpha(t_\infty)}\right|, \tag{3}$$

where $\alpha(t)$ is the order parameter versus time. The equilibrium values $\alpha(t_\infty)$ were determined at the time corresponding to the 5000th Monte Carlo step, and the initial values $\alpha(t_0)$ were set to zero in our calculations. The transformation of the order parameters described by Eq. (3) allowed us to separate an exponential component and easily determine the time constant of the relaxation process. Figure 9 shows calculations of the order parameter α_1 expressed using the function $\hat{f}(\alpha_1)$ for different correlation lengths at $x=0.4$. One can see that the kinetics is slow for $r>3$, whereas for shorter correlation lengths there are two notably different stages: the faster initial stage and considerably slower changes at longer times. The explanation of this behavior at small r is that only local ordering takes place at the initial stage, and many small twins with sizes of several lattice constants are generated. This process occurs on the first five to ten Monte Carlo steps and leads to short-range ordering. The stage of the slower relaxation at small r is characterized by motion of twin boundaries, when smaller twins are absorbed by larger ones. The slowly changing order parameters for large correlation lengths describe formation of long-range order. This ordering requires more time, which manifests in measurements of physical quantities that are functions of the average carrier concentration in the cuprate plane, such as T_c (Fig. 2), dc conductivity, and the low-frequency ($\sim 5000\,\mathrm{cm}^{-1}$) band in spectra of optical conductivity.[7,19] As for the fast relaxing intensity I_{2ML} of the two-magnon scattering (Fig. 4a), the kinetics observed in experiments correlates with curves in Fig. 4b, which show sums of the order parameters α_1 calculated for small correlation lengths, $\Sigma_{r=1}^{3}\alpha_1^r(t)$, and transformed using function $\hat{f}(\alpha)$ [Eq. (3)]. One can clearly see that there is reasonable agreement between the experimental data and calculations.

The distinguishing feature of our calculated curves is that they cannot be approximated by simple analytical functions like an exponential function whose argument is a power function, as was done in processing experimental data obtained by Veal *et al.*[2] The absence of a simple kinetic function describing relaxation is an additional indication of the complexity of the relaxation process, which generates at the initial stage a structure with a short-range order over lengths of several lattice constants, and at longer times it leads to formation of long chains and twins of large sizes.

Note also the correlations between functions $\alpha_1^r(t)$ with different r. Analyzing the order parameters versus time at different concentrations x and correlation lengths r, we found that the functions $\alpha_1^{r>1}(t)$ can be expressed with good accuracy (within 1%) in terms of $\alpha_1^{r=1}(t)$:

$$\alpha_1^{r>1}(t) \approx A_1^{r=1}(t), \tag{4}$$

where

$$A_1^{r=1}(t) = \Theta(t-\Delta) I_S \alpha_1^{r=1}\left(\frac{t-\Delta}{t_S}\right). \tag{5}$$

Here the step function Θ reflects the presence of a small delay of several steps, during which $\alpha_1^{r>1} \approx 0$ (Δ is a fitting parameter). The further behavior of $\alpha_1^{r>1}(t)$ closely follows the function $\alpha_1^{r=1}(t)$, but with different scales applied to both coordinate axes using multiplicators I_S and t_S. It turned out that the parameter Δ can be approximated for all values of x by the expression

$$\Delta \approx \frac{r-1}{2}. \tag{6}$$

This dependence derives from the fact that only short-range interactions (with nearest neighbors) are taken into account by the model. Thus, correlation between two sites separated by distance r turns up only after $r/2$ steps of the computer simulation. This can be interpreted as propagation of an ''information wave'' with the velocity of one lattice constant per one calculation step for each site. Mutual influence between two sites sets in only when the fronts of two such ''waves'' intersect.

The second fitting parameter in scaling function (5) is the relative intensity of correlations on the initial step I_S. The intensities of such correlations are approximated by the expressions

$$I_S \approx x^{r-1}. \tag{7}$$

The emergence of such expressions indicates that the probabilities of correlations between sites separated by the distance $r>1$ should be calculated by multiplying the probabilities of correlations between neighboring atoms located between the two lattice sites.

The most interesting parameter, which cannot be interpreted in simple terms, is the third fitting parameter t_S. It describes the ratio between the characteristic times of relaxation (ordering) for atoms separated by distance r and the relaxation time for the nearest neighbors. For all oxygen concentrations $x \geq 0.5$, the function $t_S(r)$ is the same to within the uncertainty of the t_S calculation and slightly superlinear: $t_S(r)=1$ at $r=1$, and $t_S \approx 12$ at $r=6$. At $x<0.5$ the relaxation time is considerably longer.

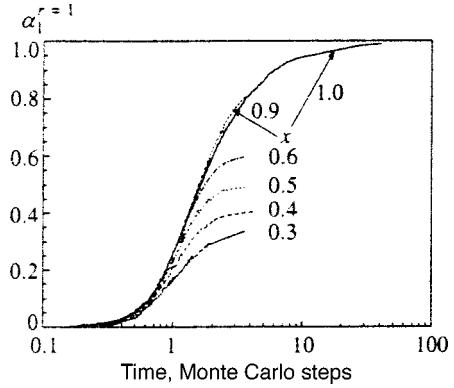

FIG. 10. Calculations of order parameter $\alpha_1^{r=1}$ versus time for different oxygen concentrations x. The number of points was increased by scaling order parameters calculated for different correlation lengths [Eqs. (4) and (5)].

The distance-dependence of t_S is a measure of the strength of connections in the system, which characterizes the velocity of ordering propagation to longer distances. At higher oxygen concentrations $x>0.5$, the contribution of repulsive pair interactions V_1 and V_3 is considerable and increases with x. At $x=0.5$, there is only one configuration of atoms when all interactions in the system are determined by V_2 (the ideal OII phase). This configuration corresponds to the minimal energy per atom. As x becomes smaller, the number of such configurations increases. At $x<0.5$ configurations obtained after 5000 Monte Carlo steps consist of isolated chains (including copper atoms) randomly aligned with respect to the lattice (no large twin clusters have been detected) and separated from one another by distances longer than $2a$, i.e., they do not interact with one another (Fig. 6). In this case, there is ordering only along the chains, which leads to a considerably lower order parameter; hence, it takes much longer time to create long-range order in the system. Thus, we have come to the conclusion that the system kinetics changes radically at $x=0.5$, because strong repulsive interactions are effectively ''switched-off'' at $x<0.5$ owing to the ''low density'' of the system. Another interesting conclusion from our calculations is that the relationship between t_S and correlation length is not affected by x for $x>0.5$; i.e., the system ''stiffness'' does not increase with the oxygen concentration, hence the number of repulsive interactions is constant.

The only component that is required to complete our description of the system kinetics is comparative analysis of the order parameter α_1 evolution at $r=1$ for different oxygen concentrations x. Unfortunately, the order parameter kinetics at $r=1$ is very fast, and changes in the parameter saturate after several initial steps. Therefore, in plotting the kinetic curves for $r=1$ we used the functions $\alpha_1^{r>1}(t)$ for $r>1$ transformed in accordance with Eqs. (4) and (5) using the fitting parameters I_S, Δ, and t_S. This has allowed us to plot more points for the curve of α_1 versus time at $r=1$ and check out the formula fitted to all calculated points (if the points lie on the same curve after the coordinate transformation). Figure 10 shows curves of the order parameter $\alpha_1^{r=1}$ at different x refined using the procedure described above. It is clear that there is a universal complex kinetic function $\alpha_1(t)$, which is plotted over the entire time interval for $x=1.0$. At smaller oxygen concentrations, the curves of the order parameter versus time are described by the universal function until the order parameter approaches x, after which saturation takes place.

Thus, it follows from our calculations that relaxation of order parameters for different correlation lengths can be determined at the initial stage up to 100 steps using the universal functions with the help of a shift in time and variation of scales on both coordinate axes.

4. DISCUSSION OF EXPERIMENTAL RESULTS

In order to understand the difference between the evolution patterns of the two-magnon scattering and electronic continuum, we should clearly identify the origin of these signals and scattering mechanisms in both cases. Every quantity measured in experiments has a characteristic coordinate scale of its changes, which is determined by physical processes involved. When the configuration of oxygen atoms differs from the equilibrium pattern, this brings about changes in various physical parameters. These changes for each specific physical quantity are determined by a corresponding correlation function which relates the instantaneous and equilibrium configurations. This leads, in turn, to the dependence between the relaxation kinetics of a specific physical quantity and the evolution of the correlation function $\alpha^r(t)$, whose space scale r corresponds to the space scale of changes in the physical quantity. Our calculations demonstrate that the relaxation rate of the order parameter α^r depends on its correlation length and decreases as r grows. Thus, comparison between our calculations and kinetic functions measured in experiments allows us to estimate qualitatively characteristic space scales of changes in various physical quantities.

Before proceeding to the experimental data, note that the two-magnon light scattering in quasi-two-dimensional antiferromagnets, as we know, is due to generation of a pair of magnons by an incident photon, and its intensity has a peak at energies corresponding to generation of magnons at the boundary of the magnetic Brillouin zone.[20] The two-magnon Raman scattering of light is therefore most sensitive to short-wave antiferromagnetic correlations in the CuO_2 planes. Despite large quantum fluctuations at $s=1/2$ (Ref. 10), the standard theory of two-magnon scattering can describe quantitatively the line broadening in high-temperature superconducting crystals through phenomenological introduction of phonon damping.[21] This model leads to a conclusion that the loss of the long-range magnetic order has little effect on the two-magnon scattering, because it depends mostly on generation of a pair of two short-wave magnons. Thus, one can assume that the two-magnon Raman scattering of light occurs in a small region with dimensions of about 3×4 lattice constants, which leads to a spin flip of two neighboring atoms and a change in their exchange energy due to interaction with nearest neighbors. From this viewpoint, it is clear why the relaxation time of the two-magnon scattering is so short.

As follows from studies of the two-magnon Raman scattering (Fig. 1) and neutron scattering,[6,22] spin correlations in cuprate planes are also observed in superconducting crystals with index $x \approx 0.4-0.7$. This means that short-range antiferromagnetic correlations ''persist'' in the superconducting state; i.e., antiferromagnetism and superconductivity coexist in $YBa_2Cu_3O_{6+x}$ crystals. It is not clear at this time whether this coexistence is possible in an ideal, homogeneous, two-dimensional electron system or whether it is possible due to decomposition of the CuO_2 plane into microscopic domains with higher and lower conductivities (''metallic'' and ''dielectric'') and dimensions of several lattice constants. This decomposition can be brought about, for example, by either strong electronic correlations in the cuprate plane, which leads to localization of current carriers within a microscopic region,[23] or disorder in $YBa_2Cu_3O_{6+x}$ crystals. If this decomposition really takes place, the two-magnon scattering can occur in the ''dielectric'' microscopic domains, and the coexistence of superconductivity and antiferromagnetism finds a natural explanation.

It follows from our investigation of ordering of mobile oxygen atoms that the local disorder in the chain oxygens has its effect on the CuO_2 planes. In fact, the fast kinetics of the two-magnon scattering, which occurs in cuprate planes, indicates that the change in the short-range order of mobile chain oxygens due to the annealing affects the short-range order in the electron and spin systems of CuO_2 planes without a notable change in the free carrier density. This fast process proceeds concurrently with a slower ordering of oxygen chains, which supplies holes to the CuO_2 plane and manifests in the slow relaxation of T_c. This provides evidence in favor of the microscopically inhomogeneous model of the CuO_2 plane. Moreover, at $0<x<1$ the composition of $YBa_2Cu_3O_{6+x}$ is nonstoichiometric, and there is disorder in the system of chain oxygen both before and after ordering, and this means that the cuprate plane is *a fortiori* locally inhomogeneous.

Before discussing the extended electronic continuum, note that it has been detected in Raman spectra of both superconducting and dielectric HTSC crystals in the form of a flat, almost featureless background extending to energies of about 1 eV. By the present time, no unambiguous interpretation of it nature has been suggested, although a lot of theories have been put forth in attempts to explain its origin. We can stipulate with a degree of certainty that this continuum is due to CuO_2 planes because its basic properties, such as the spectral shape, dependence on the doping level, and temperature dependence are very similar in different HTSC materials. This is also supported by the fact that the electronic continuum has not been detected in *zz* Raman spectra, when the polarizations of both incident and scattered light are perpendicular to the cuprate plane. Moreover, it has been established in numerous studies of features in HTSC Raman spectra related to the presence of the gap in electronic spectra that the low-energy section of Raman spectra is modified below the superconducting transition temperature; namely, the intensity of the electronic continuum reduces and a peak emerges at a frequency approximately corresponding to the superconducting gap width 2Δ (Refs. 24–26), which confirms, undoubtedly, that free carriers contribute to the electronic continuum in the Raman spectra.

However, in superconducting $YBa_2Cu_3O_{6+x}$ crystals the electronic continuum intensity is a rather flat function of the doping level.[26] This fact suggests that, in addition to the Raman light scattering by free carriers,[27,28] light scattering by localized states should also be taken into account.[7,26] The fast kinetics of the electronic continuum intensity observed in $YBa_2Cu_3O_{6.4}$ (Fig. 3) indicates that most of the continuum is due to states of local nature at $x \approx 0.4$. These localized states can be of either magnetic or defect nature.[7] At any rate, the electronic continuum intensity cannot be uniquely related to the average concentration of carriers in the cuprate plane. On the other hand, the increase in the relaxation time with the oxygen content in the $YBa_2Cu_3O_{6+x}$ crystals observed in experiments (Fig. 3) indicates that the characteristic dimension of these states increases with x. Thus, we conclude, therefore, that the electronic continuum in optical Raman spectra has a complex nature, and both localized and extended (free) electronic states contribute to this continuum, moreover, the role of the latter states essentially increases as the doping level approaches the optimal value for $YBa_2Cu_3O_{6+x}$.

5. CONCLUSIONS

In conclusion, we focus the reader's attention on the fact that the kinetic characteristics of the relaxation of the superconducting transition temperature, intensities of the featureless continuum and two-magnon scattering bands in the Raman spectra can be consistently described in terms of the order parameters which were introduced in this paper and which characterize the process of chain oxygen ordering in $YBa_2Cu_3O_{6+x}$ crystals with different doping levels.

The differences between the relaxation times of different physical quantities also find their natural interpretation on the basis of the relatively simple description of the disordered state in terms of the correlation parameters as functions of correlation lengths.

Thus, as a result of experiments and computer simulations reported in this publication, we have established that the characteristic times of the critical temperature relaxation to the equilibrium value of the superconducting transition temperature T_c in crystals with different oxygen contents are relatively long, which is determined by the slow formation of long ordered oxygen chains. In contrast, the intensity of the two-magnon Raman scattering, which is of local nature, has demonstrated a fast relaxation kinetics, because in this case only changes in the close neighborhood of an oxygen atom are essential. As to the extended featureless continuum in Raman spectra of $YBa_2Cu_3O_{6+x}$ crystals, it has a complex nature and is determined by light scattering by localized states and by free carriers. As the content of chain oxygen in $YBa_2Cu_3O_{6+x}$ increases, the contribution of scattering by free carriers to the continuum intensifies.

An important point is that experimental detection of different relaxation times *per se* indicates that breaking of local ordering in oxygen chains in $YBa_2Cu_3O_{6+x}$ leads to local inhomogeneities in the electron and spin systems of CuO_2

planes on the microscopic scale of several lattice constants.

This work was supported by the State R&D Program on HTSC (Grant 96031), NSF DMR No. 91-20000 (STCS), and US DOE, Grant W-31-109-ENG-38.

*)E-mail: maksimov@issp.ac.ru
†)E-mail: tartakov@issp.ac.ru

[1] B. W. Veal, H. You, A. V. Paulikas *et al.*, Phys. Rev. B **42**, 4770 (1990).
[2] B. W. Veal, A. V. Paulikas, Hoydoo You *et al.*, Phys. Rev. B **42**, 6305 (1990).
[3] G. V. Uimin, V. F. Gantmakher, A. M. Neminsky *et al.*, Physica C **192**, 481 (1992).
[4] J. D. Jorgensen, B. W. Veal, A. V. Paulikas *et al.*, Phys. Rev. B **41**, 1863 (1990).
[5] J. Kircher, E. Brücher, E. Schonherr *et al.*, Phys. Rev. B **46**, 588 (1992).
[6] A. A. Maksimov, I. I. Tartakovskii, M. V. Klein, and B. W. Veal, Phys. Rev. B **49**, 15385 (1994).
[7] A. A. Maksimov, D. A. Pronin, S. V. Zaitsev *et al.*, Phys. Rev. B **54**, R6901 (1996).
[8] E. Salomons and D. de Fontaine, Phys. Rev. B **41**, 11159 (1990).
[9] H. F. Poulsen, N. H. Andersen, J. V. Andersen *et al.*, Phys. Rev. Lett. **66**, 465 (1991).
[10] R. R. P. Singh, K. B. Lyons, P. A. Fleury, and P. E. Sulewski, Phys. Rev. Lett. **62**, 2736 (1989).
[11] A. A. Maksimov, I. I. Tartakovskii, and V. B. Timofeev, Physica C **160**, 249 (1989).
[12] D. E. Morris, A. G. Markelz, J. Y. T. Wei *et al.*, Phys. Rev. B **44**, 9556 (1991).
[13] R. McCormack, D. de Fontaine, and G. Ceder, Phys. Rev. B **45**, 12976 (1992).
[14] Ph. Lambin, in *Oxygen Disordered Effects in High-T_c Superconductors*, ed. by J. L. Moran-Lopez and J. K. Schuller, Plenum, New York (1990), p. 101.
[15] B. W. Veal and A. P. Paulikas, Physica C **184**, 321 (1991).
[16] G. Ceder, R. McCormack, and D. de Fontaine, Phys. Rev. B **44**, 2377 (1991).
[17] G. Uimin and J. Rossat-Mignod, Physica C **199**, 251 (1992).
[18] H. F. Poulsen, N. H. Andersen, J. V. Andersen *et al.*, Nature **349**, 594 (1991).
[19] B. W. Veal and Chun Gu, J. of Electron Spectroscopy **66**, 321 (1994).
[20] J. Parkinson, J. Phys. C **2**, 2012 (1969).
[21] W. H. Weber and G. W. Ford, Phys. Rev. B **40**, 6890 (1989).
[22] J. Rossat-Mignod, L. P. Regnault, C. Vettier *et al.*, Physica C **185-189**, 86 (1991).
[23] V. J. Emery and S. A. Kivelson, Physica C **209**, 597 (1993).
[24] A. A. Maksimov, A. V. Puchkov, I. I. Tartakovskiĭ *et al.*, JETP Lett. **56**, 570 (1992).
[25] S. V. Zaĭtsev, A. A. Maksimov, I. I. Tartakovskiĭ *et al.*, JETP Lett. **61**, 865 (1995).
[26] D. Reznik, S. L. Cooper, M. V. Klein *et al.*, Phys. Rev. B **48**, 7624 (1993).
[27] A. Virosztek and J. Ruvalds, Phys. Rev. B **45**, 347 (1992).
[28] C. M. Varma, P. B. Littlewood, S. Schmidt-Rink *et al.*, Phys. Rev. Lett. **63**, 1996 (1989).

Translation was provided by the Russian Editorial office.
Edited by S. J. Amoretty

Theory of coherent generation in resonant-tunneling diodes

V. F. Elesin

Moscow State Engineering-Physics Institute, 115409 Moscow, Russia
(Submitted 31 March 1999)
Zh. Éksp. Teor. Fiz. **116**, 704–716 (August 1999)

A rigorous quantum-mechanical description is given of coherent generation in a resonant-tunneling diode for the simple model of a double-barrier structure. Exact analytical solutions of the Schrödinger equation are found for electrons in the presence of a weak electromagnetic field. The active and reactive polarization currents are calculated over a wide frequency interval and are reduced to a simple and easy-to-visualize form. These expressions satisfy the Kramers–Kronig relations and general criteria of the theory of radiation and other formal requirements. An analysis of the threshold conditions of generation is given. In particular, it is shown that generation is possible at microwave frequencies exceeding the inverse lifetime of an electron in the quantum well. Thus, the ''quasiclassical'' idea of the existence of a limiting frequency, associated with the finite lifetime, turns out to be inapplicable for a resonant-tunneling diode, which occupies an intermediate position between lasers and Gunn diodes.
© *1999 American Institute of Physics.* [S1063-7761(99)02508-1]

1. INTRODUCTION

Resonant tunneling and the negative differential conductivity effect in nanostructures are brought about by purely quantum phenomena of spatial quantization giving rise to resonant energy levels.[1] In what follows we have in mind the simplest nanostructure—a quantum well, which is frequently called a resonant-tunneling diode.

As is well known, negative differential conductivity affords the possibility of generating an electromagnetic field. A number of workers (see Refs. 2 and 3) have demonstrated generation in a resonant-tunneling diode up to 712 GHz.

The question was raised and discussed long ago (see, e.g., Ref. 3) as to whether there exists a fundamental limitation on the generation frequency in a resonant-tunneling diode analogous to the limitations on Gunn diodes and tunneling diodes.[4,5]

At present, there is no consensus in regard to this question. This is due, in particular, to the absence of a generally accepted theoretical model of coherent generation in a resonant-tunneling diode. Most treatments of the theory of generation in resonant-tunneling diodes utilize numerical methods.[3,6–11] They have established a number of important generalizations regarding coherent generation (in the present work we are interested only in the coherent approximation), but contain many contradictory results, the reasons for which are hard to determine (see the discussion in Refs. 3 and 10).

Comparatively little work have been devoted to an analytical approach.[12–15] Almost all of the treatments known to us employ semi-phenomenological methods, in which instead of strictly solving the Schrödinger equation additional approaches are employed (see, e.g., Ref. 12).

According to one widely held point of view (see Ref. 3), the generation frequency is bounded by the inverse of the characteristic lifetime of an electron in the quantum well τ_γ^{-1} ($\tau_\gamma=\hbar/\Gamma$, where Γ is the width of the resonant level). This argument is valid for ''classical'' generators (e.g., tunneling and Gunn diodes,[4,5] in which resonant levels are absent).

Generators based on resonant-tunneling diodes occupy an intermediate place between ''classical'' generators and lasers. They are based on a ''quasiresonant'' interaction between electrons and the electric field. Bear in mind that radiative transitions go between states near one resonant level (and not between two levels, as in a laser). It can be expected that if the frequency of the field ω is small in comparison with the width of the resonant level Γ, then the ''quasiclassical'' description is applicable. In the opposite limit the quantum-mechanical approach is necessary.

The aim of the present work is to give a rigorous quantum-mechanical description of the simplest analytically solvable model of coherent generation in resonant-tunneling diodes. As is well known, this allows a deeper understanding of the physics of the problem. An exact solution of the Schrödinger equation is found for the partial wave functions corresponding to the quasi-energies of the electron in an electromagnetic field. The active (in phase with the field) polarization current J_c and the reactive polarization current J_s are calculated over a wide frequency interval in the small-signal approximation. Using the small parameters ω/ε_R and Γ/ε_R, which are natural for the resonant-tunneling diode problem (ε_R is the energy of the resonant level), expressions for J_c and J_s are reduced without loss of generality to a simple and easy-to-visualize form as a result of cancelation of a number of divergent terms. The resulting expressions for the currents satisfy the Kramers–Kronig relations, general principles of the theory of radiation, and other formal requirements. They differ substantially from previously obtained results, especially for high frequencies. It is shown in particular that generation is possible at microwave frequencies $\omega \gg \Gamma$ if the energy of the electrons (equivalent to the constant voltage on the resonant-tunneling diode) is chosen outside the region of maximum negative conductivity, i.e.,

 © 1999 American Institute of Physics

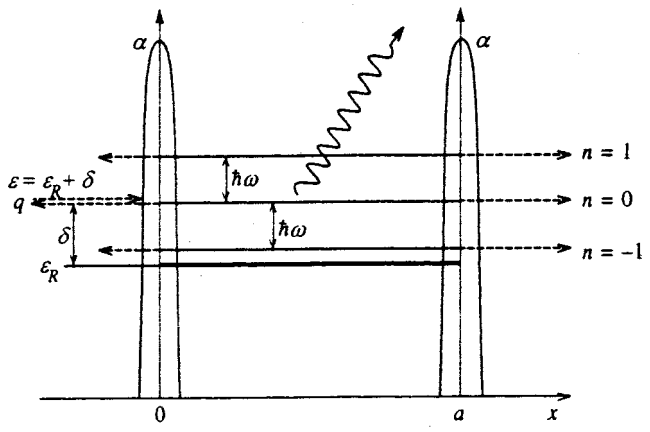

FIG. 1. Double-barrier structure with resonant level ε_R. Other notation explained in text.

the quasiclassical idea of a limiting frequency associated with the width of the level (see Ref. 3) is apparently inapplicable for resonant-tunneling diodes. In addition, the optimal conditions at generation threshold are found as functions of the structure parameters.

2. DESCRIPTION OF THE MODEL AND WAVE FUNCTIONS OF THE RESONANT-TUNNELING DIODE TO FIRST ORDER IN THE FIELD

We examine a model of coherent generation in a resonant-tunneling diode.[12,15,16] For simplicity we consider a one-dimensional quantum well with δ-function barriers at the points $x=0$ and $x=a$ (see Fig. 1). A stationary electron flux is incident on the quantum well from the left ($x=-\infty$), proportional to q^2 with energy ε approximately equal to ε_R. A variable electric field $E(t)$ with potential $V(x,t)$ operates in the region of the quantum well:

$$V(x,t)=\nu(x)\cos(\omega t+\varphi),$$

$$\nu(x)=\begin{cases}\nu_0 x\,\theta(x-a), & x<a\\ \nu_0 a, & x>a\end{cases}, \quad \nu_0=-\frac{eE}{2}. \tag{1}$$

The wave function $\psi(x,t)$ obeys the Schrödinger equation

$$i\frac{\partial\psi}{\partial t}=-\frac{\partial^2\psi}{\partial x^2}+[\alpha\delta(x)+\alpha\delta(x-a)]\psi+V(x,t)\psi, \tag{2}$$

where we have set $\hbar=1$ and $2m=1$. We seek the steady-state solution of Eq. (3) in the form[7,17]

$$\psi(x,t)=e^{-i\varepsilon t}[\psi_0(x)+e^{-i\omega t}\psi_{+1}(x)+e^{i\omega t}\psi_{-1}(x)]. \tag{3}$$

The partial functions $\psi_0(x),\psi_n(x)$ describe electrons with quasi-energies ε and $\varepsilon+n\omega$ ($n=\pm1$), respectively.

The variable field gives rise to polarization currents:

$$J_c(x,t)=J_c(x)\cos(\omega t+\varphi),$$

$$J_s(x,t)=J_s(x)\cos(\omega t+\varphi).$$

The current in phase with the field, J_c, and the reactive current J_s can be expressed in terms of the functions $\psi_n(x)$:

$$J_c(x,t)=J^c_{+1}(x)+J^c_{-1}(x),$$

$$J^c_n(x)=-ie[(\psi_0^*\psi_n'+\psi_0'\psi_n^*)+\text{c.c.}],$$

$$J_s(x,t)=J^s_{+1}(x)-J^s_{-1}(x),$$

$$J^s_n(x)=e[(\psi_0^*\psi_n'-\psi_0'\psi_n^*)+\text{c.c.}]. \tag{4}$$

The zero-approximation function $\psi_0(x)$ in the region $0<x<a$ satisfies the equation

$$\varepsilon\psi_0(x)+\psi_0''(x)=0, \quad \psi_0''\equiv\frac{d^2\psi_0}{dx^2} \tag{5}$$

and boundary conditions (see Ref. 17)

$$\psi_0(0)(1-\beta)+\frac{\psi_0'(0)}{ip}=q,$$

$$\psi_0(a)(1-\beta)-\frac{\psi_0'(a)}{ip}=0, \quad \beta=\frac{\alpha}{ip}, \quad p^2=\varepsilon. \tag{6}$$

The boundary conditions (6) describe the electron flux from the left, proportional to q^2 and their reflection and escape into the region $x>a$.

The corresponding equations for $\psi_n(x)$ have the form

$$p_n^2\psi_n+\psi_n''=\nu(x)\psi_0(x),$$

$$p_n^2=p^2+n\omega. \tag{7}$$

The boundary conditions for ψ_n are essentially analogous to the boundary conditions (6) with $q=0$.

The solutions of Eqs. (5), (7) in the region $0<x<a$ have the form

$$\psi_0(x)=A\exp(ipx)+B\exp(-ipx)$$
$$\equiv\gamma_0\cos(px)+i\delta_0\sin(px),$$

$$\Delta_0 A=q(2-\beta)\exp(-2ipa), \quad \Delta_0\beta=q\beta, \tag{8}$$

$$\Delta_0=\frac{4}{\Gamma}[i(\varepsilon-\varepsilon_R)-\Gamma], \quad \Gamma=\frac{4p^3}{\alpha^2 a}, \tag{9}$$

$$\psi_n(x)=\tilde\psi_n(x)-\frac{\nu_0 x}{\omega_n}\psi_0(x)-\frac{2\nu_0}{\omega_n^2}\psi_0'(x),$$

$$\tilde\psi_n=A_n\exp(ip_n x)+B_n\exp(-ip_n x)$$
$$\equiv\gamma_n\cos(p_n x)+i\delta_n\sin(p_n x), \tag{10}$$

$$\Delta_n A_n=2q_n\exp(-2ip_n a)+\beta_n[\tilde q_n-q_n\exp(-2ip_n a)],$$

$$\Delta_n B_n=2\tilde q_n+\beta_n(\tilde q_n-q_n), \tag{11}$$

$$\Delta_n=\frac{4}{\Gamma}[i(\varepsilon-\omega-\varepsilon_R)-\Gamma],$$

$$q_n=\frac{2\nu_0 ip}{\omega_n^2}\left[(A-B)(1-\beta_n)+(A+B)\right.$$
$$\left.\times\left(1+\frac{\omega_n^2}{4pp_n(p+p_n)^2}\right)\right], \tag{12}$$

$$\tilde{q}_n=\frac{2\nu_0 ip}{\omega_n^2}\Bigg[[A-B\exp(-2ipa)](1-\beta_n)$$

$$-[A+B\exp(-2ipa)]\left(1+\frac{\omega_n^2}{4pp_n(p+p_n)^2}\right)\Bigg]$$

$$\times\exp[2i(p-p_n)a],\quad 3_n=\frac{\alpha}{ip_n},\quad \omega_n=-n\omega,$$

$$\gamma_n=A_n+B_n,\quad \delta_n=A_n-B_n. \tag{13}$$

The quantities δ_n and γ_n are key. The polarization currents (4) are expressed in terms of them. Employing relations (11) and (13), we obtain expressions for δ_n and γ_n

$$\Delta_n\delta_n=\beta_n(\tilde{q}_n-q_n)+(\beta_n-2)[\tilde{q}_n-q_n\exp(-2ip_na)], \tag{14}$$

$$\Delta_n\gamma_n=2[\tilde{q}_n+q_n\exp(-2ip_na)]$$

$$+\beta_nq_n[1-\exp(-2ip_na)]. \tag{15}$$

Formulas (8)–(15) give an exact solution of the problem. After substituting them into Eqs. (4), we obtain exact expressions for the currents $J_c(x)$ and $J_s(x)$. Unfortunately, these expressions are so lengthy and opaque that they are practically useless for analysis.

However, the possibility exists of converting them (without loss of generality) into simple and physically comprehensible expressions by using the small parameter ω/ε_R, which is natural for the resonant-tunneling diode problem. Indeed, the smallness of the frequency ω in comparison with the energy ε_R (and also with the energy difference to the neighboring level) is an intrinsic property of a generator based on resonant-tunneling diode based generators.

In order to carry out this transformation, we represent γ_n and δ_n as sums of two terms

$$\gamma_n=\gamma_n^{(1)}+\gamma_n^{(2)},\quad \delta_n=\delta_n^{(1)}+\delta_n^{(2)}. \tag{16}$$

In the terms $\gamma_n^{(1)}$ and $\delta_n^{(1)}$ we set $p=p_n$ in the exponential factors, and $\gamma_n^{(2)}$ and $\delta_n^{(2)}$ are just the corresponding correction terms. Exact cancellation of a number of terms takes place in the terms $\gamma_n^{(1)}$ and $\delta_n^{(1)}$, so that as a result we obtain

$$\gamma_n^{(1)}=\frac{2\nu_0 ip}{\omega_n^2}\delta_0,\quad \delta_n^{(1)}=\frac{2\nu_0 ip}{\omega_n^2}\gamma_0. \tag{17}$$

Note that the terms $\gamma_n^{(1)}$ and $\delta_n^{(1)}$ diverge as $\omega\to 0$. The residual terms remain finite in the low-frequency limit and are equal to

$$\Delta_n\gamma_n^{(2)}=-\frac{4\nu_0 ip}{\omega_n^2}[Z_n\beta_nA+Z_n^*B\exp(-2ip_na)(2-\beta_n)],$$

$$Z_n=\exp[i(p-p_n)a]-1, \tag{18}$$

$$\delta_n^{(2)}=(\beta_n-1)\gamma_n^{(2)}. \tag{19}$$

In expressions (18) and (19) we have discarded small terms proportional to ω/ε_R and Γ/ε_R. To simplify the expressions, we consider the most interesting case of a quantum well with ''strong barriers,'' where $\Gamma\ll\varepsilon_R$. It is precisely in this limit that important properties of the quantum well are realized.

Taking the smallness of the parameters ω/ε_R and Γ/ε_R into account, expression (18) reduces to

$$\Delta_n\gamma_n^{(2)}=\frac{\nu_0a^2\alpha A}{p^2}, \tag{20}$$

from which it follows that the quantity remains finite in the limit $\omega\to 0$. Finite also is the function $\psi_n(x)$, in particular

$$\psi_n(0)=\psi_n(a)=\gamma_n^{(2)}. \tag{21}$$

Hence it is not hard to show that the implemented weak-field approximation is valid if $eEa\ll\Gamma$.

3. POLARIZATION CURRENTS

First let us find the active component of the current $J_c(x)$. Substituting expression (10) into Eqs. (4), we obtain

$$J_n^c(x)=-ie[(\psi_0^*\tilde{\psi}_n'+\psi_0'\tilde{\psi}_n^*)-\text{c.c.}]-2\pi\nu_0e\,\delta(\omega_n)|\psi_0|^2. \tag{22}$$

In Eq. (22) we have retained terms even in ω_n, since the odd terms cancel in the total current $J_c(x)$. We also drop the last term, nonzero only for $\omega=0$, since it is cancelled out by the corresponding contribution from the first term. Neglecting small corrections of order ω/ε_R, we can reduce expression (22) to the form

$$J_n^c(x)=ep\cos[(p-p_n)x][K_n+\text{c.c.}]$$

$$-iep[F_n-\text{c.c.}]\sin[(p-p_n)x], \tag{23}$$

$$K_n=\delta_0^*\gamma_n+\gamma_0^*\delta_n,\quad F_n=\delta_0^*\delta_n+\gamma_0^*\gamma_n. \tag{24}$$

The terms arising from $\gamma_n^{(1)}$ and $\delta_n^{(1)}$ that diverge as $\omega\to 0$ are purely imaginary. Therefore, they do not contribute to $[K_n+\text{c.c.}]$; in $[F_n-\text{c.c.}]$ they drop out because the parity is even with respect to ω_n. The remaining terms are proportional to the combination

$$\frac{1}{\Delta_n}+\frac{1}{\Delta_n^*}=\frac{-\Gamma^2}{2[(\varepsilon-\varepsilon_R+n\omega)^2+\Gamma^2]}. \tag{25}$$

Thus, the current J_c, which describes amplification (absorption) of the field, is proportional to the decay of the electron, as follows from the general theory of radiation. The final expressions for the current $J_c(x)$ and the reduced current J_c (see Ref. 17) have the form

$$J_c(x)=-\frac{eEaQ\Gamma^2(\varepsilon-\varepsilon_R)}{4[(\varepsilon-\varepsilon_R+\omega)^2+\Gamma^2][(\varepsilon-\varepsilon_R-\omega)^2+\Gamma^2][(\varepsilon-\varepsilon_R)^2+\Gamma^2]}\left\{\cos\left(\frac{\omega}{2p}x\right)[(\varepsilon-\varepsilon_R)^2+\Gamma^2+\omega^2]-\sin\left(\frac{\omega}{2p}x\right)\left(\frac{4\omega p}{a}\right)\right\}, \tag{26}$$

$$J_c=\frac{1}{a}\int_0^a J_c(x)dx=-\frac{e^2EaQ\Gamma^2(\varepsilon-\varepsilon_R)}{4[(\varepsilon-\varepsilon_R+\omega)^2+\Gamma^2][(\varepsilon-\varepsilon_R-\omega)^2+\Gamma^2]},\quad Q=q^2p. \tag{27}$$

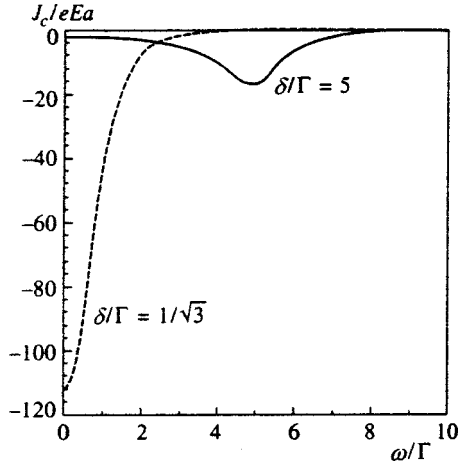

FIG. 2. Dependence of the active polarization current J_c on the frequency ω/Γ for $\delta/\Gamma=1/\sqrt{3}$ and $\delta/\Gamma=5$.

It is clear from Eq. (27) that for $\varepsilon>\varepsilon_R$ the current J_c is negative, which corresponds to amplification over the entire frequency range. In the low-frequency limit $\omega\ll\Gamma$ the current J_c is expressed in terms of the constant-current differential conductivity:

$$J_c=-\frac{e^2Ea\delta\Gamma^2}{4(\delta^2+\Gamma^2)}=\frac{e^2Ea}{4}\frac{\partial}{\partial\varepsilon}J_0(\varepsilon), \tag{28}$$

$$J_0(\varepsilon)=\frac{Q\Gamma^2}{2(\delta^2+\Gamma^2)}, \quad \delta\equiv\varepsilon-\varepsilon_R. \tag{29}$$

In the low-frequency limit expression (27) goes over to the quasiclassical expression and essentially coincides with the results obtained in Refs. 12 and 15; however, for finite frequencies there are important differences from the results of Ref. 12.

Let us analyze the frequency dependence of the current $J_c(\omega,\delta)$, whose character depends critically on the ratio δ/Γ. For $\delta<\Gamma$ the current J_c is maximum at zero frequency (see Fig. 2), and at large frequencies $\omega\gg\Gamma$ it falls off rapidly as $1/\omega^4$:

$$\tilde{J}_c=-\frac{4J_c}{e^2EaQ}\approx-\frac{\delta\Gamma^2}{\omega^4}, \quad \omega\gg\Gamma. \tag{30}$$

If, on the other hand, we have $\delta>\Gamma$, then the current J_c becomes maximum at the frequency (see Fig. 2)

$$\omega_m^2=\delta^2-\Gamma^2 \tag{31}$$

and equal to

$$\tilde{J}_c(\omega_m)=-1/4\delta. \tag{32}$$

The new maximum J_c corresponds to quasiresonant transitions between the states ε and ε_R since for $\omega\gg\Gamma$ equality (31) gives the condition of a ''quasiresonance'' $\omega=\varepsilon-\varepsilon_R$.

Let us now examine the dependence of the current J_c on δ. It is not hard to show that the current J_c reaches its maximum if $\delta=\delta_m$:

$$\delta_m^2=-\frac{1}{3}(\Gamma^2-\omega^2)+\frac{2}{3}(\Gamma^4+\omega^4+\omega^2\Gamma^2)^{1/2}. \tag{33}$$

In the low-frequency limit we obtain the well-known result

$$\delta_m^2\simeq\Gamma^2/3, \tag{34}$$

at which the low-frequency negative differential conductivity reaches its maximum value.

This, in fact, is the bias voltage that is usually chosen in experimental and theoretical work. In this case, $\delta<\Gamma$ and, according to relations (30), the current J_c falls off rapidly with frequency. Hence it follows that amplification is essentially bounded by the ''limiting'' frequency $\omega\approx\Gamma$.

However, if we chose $\delta>\Gamma$ in accordance with the conditions of a quasiresonant transition (31), then, according to Eq. (32), J_c falls off much slower with frequency:

$$\tilde{J}_c=-1/4\omega. \tag{35}$$

Thus, if we choose $\delta^2=\omega^2+\Gamma^2$ and far from the region of maximum negative differential conductivity, then generation at frequencies significantly greater than the ''limiting frequency'' Γ is possible. Physically, this is because of ''quasiresonant'' radiative electron transitions from the quasilevel $\varepsilon=\varepsilon_R+\omega$ to the resonant level ε_R. We indicated the possibility of such behavior of $J_c(\omega)$ in Ref. 11 on the basis of a numerical solution of equations analogous to Eqs. (5) and (7).

Expression (27) for the current J_c can also be obtained in another way. First of all, note that the power delivered by the electrons to the field per period is equal to

$$P_c=a\int_0^{2\pi/\omega}J_c(t)E(t)dt$$

$$=-\frac{e^2E^2a^2Q\Gamma^2\delta}{4[(\delta+\omega)^2+\Gamma^2][(\delta-\omega)^2+\Gamma^2]}. \tag{36}$$

On the other hand, it is possible to find the power from the balance of the number of electrons escaping the quantum well with energies $\varepsilon\pm\omega$:

$$P_T=\omega[(T_{+1}(a)-T_{-1}(a))+(T_{+1}(0)-T_{-1}(0))], \tag{37}$$

where $T_n(x)$ is the electron flux from the quantum well at the points $x=0$ and $x=a$ with energies $\varepsilon\pm\omega$.

Using these solutions for ψ_n, we can show that P_T exactly coincides with P_c over the entire frequency interval (in contrast to the situation in Ref. 12, where it coincides only in the limit $\omega\to0$).

The current J_c can be expressed in another form

$$J_c=\frac{e^2Ea}{4}\frac{J_0(\varepsilon+\omega)-J_0(\varepsilon-\omega)}{2\omega}, \tag{38}$$

where $J_0(\varepsilon)$ is given by formula (29).

It is quite probable that, in view of its simplicity and symmetry, formula (38) is also valid in more complicated models of a resonant-tunneling diode.

It is interesting that expression (38) has the same form as the expression for the high-frequency current in a superlattice, obtained in Ref. 18. The latter describes the current arising due to resonant electron transitions between resonant levels in two neighboring quantum wells in the presence of a constant electric field. The current is caused by processes of momentum scattering on impurities, and also by energy relaxation from an excited level. The calculation employs the method of the quantum kinetic equation without account of boundary effects.

Although in the present work we are considering a boundary-value problem of the Schrödinger equation with open boundary conditions for one quantum well in the coherent approximation, the fact of the coincidence of the formulas is instructive and is an indication of the generality of expression (38).

Let us now calculate the reactive component of the current. Substituting expression (10) in Eqs. (4) and discarding terms even in ω_n, we find

$$J_s^n(x)=ep\{i[K_n-\text{c.c.}]\cos[(p-p_n)x] +[F_n+\text{c.c.}]\sin[(p-p_n)x]\}, \tag{39}$$

where K_n and F_n are given by formulas (24) and (16)–(20). It can be shown that the terms originating from $\gamma_n^{(1)}$ and $\delta_n^{(1)}$ again drop out, and the remaining terms are proportional to the combination

$$\frac{1}{\Delta_n}-\frac{1}{\Delta_n^*}=\frac{-i\Gamma(\varepsilon-\varepsilon_R+n\omega)}{2[(\varepsilon-\varepsilon_R+n\omega)^2+\Gamma^2]}, \tag{40}$$

i.e., the reactive current J_s is determined by the ''detuning'' of $\varepsilon-\varepsilon_R+n\omega$, as it should be according to the theory of radiation.

Gathering results, we obtain for the reactive current

$$J_s=\frac{1}{a}\int_0^a J_s(x)dx = \frac{e^2EaQ\Gamma\delta(\delta^2-\omega^2-3\Gamma^2)\omega}{4(\delta^2+\Gamma^2)[(\delta-\omega)^2+\Gamma^2][(\delta+\omega)^2+\Gamma^2]}. \tag{41}$$

In the low-frequency limit the current J_s exhibits an inductive character in agreement with Refs. 3, 12, and 19 if $\delta^2<3\Gamma^2$:

$$\tilde{J}_s=\frac{4J_s}{e^2EaQ}=\frac{\Gamma\delta\omega(\delta^2-3\Gamma^2)}{(\delta^2+\Gamma^2)^3}. \tag{42}$$

For $\delta^2>3\Gamma^2$ a change of sign occurs (see Ref. 10). If we choose δ in accordance with the condition of a quasiresonant transition, then the current J_s preserves its inductive character over the entire frequency interval.

Expression (41) for the current J_s differs substantially from the corresponding expression in Ref. 12, in particular in the limit $\omega\to 0$.

It is not hard to show that the currents $J_s(\omega)$ and $J_c(\omega)$ satisfy Kramers–Kronig relations:

$$J_s(\omega)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{d\omega'}{\omega'-\omega}J_c(\omega'). \tag{43}$$

This proves the consistency of the expressions for J_s and J_c and argues in favor of their validity over the entire frequency interval.

4. THRESHOLD CONDITIONS OF GENERATION IN A RESONANT-TUNNELING DIODE

For the purpose of illustration, let us consider a simple model where the equations for the field have the form (see Ref. 16)

$$\frac{\partial E}{\partial t}+\frac{E}{2\tau_0}=-\frac{2\pi}{\chi}J_c,$$
$$\frac{\partial\varphi}{\partial t}E+\frac{(\omega^2-\Omega^2)E}{2\omega}=-\frac{2\pi}{\chi}J_s, \tag{44}$$

where τ_0 is the time characterizing losses in the resonator (circuit) with eigenfrequency Ω and χ is the dielectric constant. Knowing $J_c(\omega\delta)$ and $J_s(\omega\delta)$, it is also possible to analyze solutions for other circuits.

In the stationary case, after substituting J_c and J_s from formulas (27) and (41), we obtain equations for Q_{th}, δ, and ω at the threshold of generation:

$$\frac{1}{\tau_0}=\frac{\tilde{Q}_{\text{th}}\delta\Gamma^2}{[(\delta-\omega)^2+\Gamma^2][(\delta+\omega)^2+\Gamma^2]},\quad \tilde{Q}_{\text{th}}=\frac{\pi e^2aQ_{\text{th}}}{\chi}, \tag{45}$$

$$\frac{\omega^2-\Omega^2}{2\omega}=-\frac{\tilde{Q}_{\text{th}}\Gamma\delta\omega[\delta^2-\omega^2-3\Gamma^2]}{2(\delta^2+\Gamma^2)[(\delta+\omega)^2+\Gamma^2][(\delta-\omega)^2+\Gamma^2]}. \tag{46}$$

Using relations (45), we eliminate $\tilde{Q}_{\text{th}}$ from Eq. (46):

$$\omega^2-\Omega^2=-\frac{\omega^2(\delta^2-\omega^2-3\Gamma^2)S}{\delta^2+\Gamma^2},\quad S=\frac{1}{\Gamma\tau_0}, \tag{47}$$

where S is the analog of the stabilization parameter in laser theory.[20]

We have obtained a biquadratic equation for ω^2, which has the following solution:

$$\omega^2=\frac{1}{2}\left[\frac{\delta^2+\Gamma^2}{S}+\delta^2-3\Gamma^2\right] \pm\sqrt{\frac{1}{4}\left[\frac{\delta^2+\Gamma^2}{S}+\delta^2-3\Gamma^2\right]^2-\frac{(\delta^2+\Gamma^2)\Omega^2}{S}}. \tag{48}$$

In the limit $S\ll 1$ the generation frequency differs only slightly from Ω:

$$\omega^2-\Omega^2=-\frac{S\Omega^2(\delta^2-\Omega^2-3\Gamma^2)}{\delta^2+\Gamma^2}. \tag{49}$$

In this case, substituting $\omega\simeq\Omega$ into the threshold condition, we obtain

$$\frac{1}{\tau_0}=\frac{\tilde{Q}_{\text{th}}\delta\Gamma^2}{[(\delta-\Omega)^2+\Gamma^2][(\delta+\Omega)^2+\Gamma^2]}. \tag{50}$$

The optimal conditions of generation (i.e., maximum gain) at the frequency Ω can be found employing the analysis given in Sec. 3. It follows from such an analysis that the value

$\delta=\Gamma/\sqrt{3}$ is optimal at low frequencies $\Omega\ll\Gamma$. At high frequencies, when $\Omega\gg\Gamma$ holds, one should choose the bias $\delta\simeq\Omega$.

In the reverse limiting case $S\gg1$ it is not hard to obtain from formula (48), dropping corrections $\sim1/S$,

$$\omega^2=\delta^2-3\Gamma^2,\quad \delta^2>3\Gamma^2, \tag{51}$$

i.e., the frequency is determined by the energy difference $\varepsilon-\varepsilon_R$ in quasiresonant transitions. In the interval $0<S<\infty$ the frequency is equal to the weighted mean value between Ω and δ, in analogy with laser theory.[20] The difference is that in a resonant-tunneling diode the role of an energy difference is played by $\delta=\varepsilon-\varepsilon_R$.

All this supports our conclusion that a generator based on a resonant-tunneling diode is a quantum generator with radiative transitions near one resonant level.

If $\delta^2<3\Gamma^2$ holds, then the constraint

$$S<S_c=\frac{\delta^2+\Gamma^2}{(\Omega+\sqrt{\Omega^2+3\Gamma^2-\delta^2})^2} \tag{52}$$

is imposed on the parameter S. For the generation frequency at $S=S_c$ we have

$$\omega^2=\Omega(\Omega+\sqrt{\Omega^2+3\Gamma^2-\delta^2}). \tag{53}$$

5. CONCLUSION

In the Introduction we noted that a significant number of theoretical work has already been devoted to the calculation of polarization currents in the coherent approximation. It is convenient to divide them into work based on a numerical solution of the Schrödinger equation[7–11] or equations for the Wigner function[6] (the first group) and analytical work (the second group). The results for the active component of the current J_c are qualitatively similar in the frequency region up to $\omega\approx\Gamma$: a weak dependence on the frequency in the interval $0\ll\omega\ll\Gamma$ followed by a falloff (in absolute value) at $\omega\sim\Gamma$. In the region $\omega>\Gamma$ either a change in sign of the current J_c or a rapid falloff of the negative current is predicted. Note that in these treatments the bias voltage (or energy of the escaping electrons) was chosen to lie in the region of maximum negative differential conductivity.

The situation with the reactive current J_s is much less clear (see, e.g., the detailed discussion in Refs. 3 and 10). Neither the frequency dependence, nor the dependence on δ, nor the sign of the reactive current J_s agrees. It is hard to determine the reason for this disagreement, especially in treatments of the first group. Therefore, let us compare our results with treatments from the second group (analytical studies).

The most detailed and similar in their statement of the problem are Refs. 12 (Liu) and 3 (Liu and Sollner) (see the references in these two papers to previous work). But the approach used by Liu[12] to solve the problem differs strongly from ours. Liu[12] does not find an explicit solution of the Schrödinger equation in the region of the quantum well. He writes the wave function at the collector boundary of the well ($x=a$) in the form

$$\psi_{\rm tr}=\{t_0\exp(ipx)+t_{+1}\exp(ip_{+1}x-i\omega t)+t_{-1}\exp(ip_{-1}x+i\omega t)\}\exp(-i\varepsilon t)U, \tag{54}$$

$$U=\exp\left[\frac{i\nu_0 q}{\omega}\sin(\omega t)\right]. \tag{55}$$

Here $t_0, t_{\pm1}$ are the amplitudes of the transit of an electron across the quantum well for zero and nonzero field (first order in the variable field), respectively. The amplitudes are found by summing the transmitted and reflected waves (the Fabry–Perot model). The potential of the interaction with the field (1) is taken to be constant over the small intervals into which the quantum well is partitioned. It is also assumed that emission occurs in the center of the well. Other approximations are also made, in particular, with respect to the contribution of the matrix elements. Thus, Liu[12] uses a number of approximations although his approach is quite general (e.g., in the form of the barriers, etc.).

As was mentioned in Sec. 3, the expressions for J_c and J_s obtained by Liu[12] differ substantially from formulas (27) and (39). In particular, for $\delta=\omega$ Liu[12] has $J_c\propto-1/\omega^3$ and $J_c\propto-1/\omega$ (see Ref. 35), i.e., according to Liu[12] generation at high frequencies is practically impossible. The polarization currents J_s differ strongly. In particular, expression (41) changes sign with growth of δ or ω in contrast to the analogous current in Ref. 12.

The amplitudes of the transit of the electron across the quantum well $t_{\pm1}$ in Ref. 12 differ dramatically from our $\psi_{\pm1}(a)$:

$$|\psi_{\pm1}(a)|^2=|\gamma_{\pm}|^2=\frac{(\nu_0 a)^2\Gamma^2}{16(\delta^2+\Gamma^2)[(\delta\pm\omega)^2+\Gamma^2]}, \tag{56}$$

$$|t_{\pm1}|^2=\left(\frac{V_{ac}}{2\omega}\right)^2\frac{\Gamma^2[(\delta\pm\omega/2)^2+\Gamma^2]}{(\delta^2+\Gamma^2)[(\delta\pm\omega)^2+\Gamma^2]}. \tag{57}$$

It should be emphasized that $|t_{\pm1}|^2\propto1/\omega^2\to\infty$, $\omega\to0$, i.e., it diverges in the low-frequency limit. Finally, note that the currents J_c and J_s in Ref. 12 do not satisfy requirements (25) and (40) of the theory of radiation.

In Ref. 13 the polarization currents were found in general form for any field. Unfortunately, a derivation and analytical analysis of the expressions for the currents are completely absent. In addition, J_c and J_s do not satisfy requirements (25) and (40). All this hinders a detailed comparison of our results with the results of Ref. 13. Closed expressions for J_c and J_s are not given in the remaining treatments of the second group that are known to us.

Summing up our analysis of the work known to us and our comparison of their results with ours, we may conclude that the expressions for the currents J_c (27) and J_s (39) are well-founded and reliable. Indeed, they are based on an exact analytical solution of the boundary-value problem for the Schrödinger equation and were reduced to simple expressions with a known degree of accuracy. These expressions satisfy the Kramers–Kronig relations (43) and the requirements of the theory of radiation (25), (40), were confirmed by an independent derivation from the conditions of balance (37), are expressed in a universal form (38), and have the correct low-frequency classical limit (28). On the basis of

Eqs. (27) and (35) it may be concluded that the width Γ is not a limiting frequency. Of course, for a detailed description of a particular experiment it would be necessary to take into account the specific details of the structure of the resonant-tunneling diode, charge accumulation effects, and other phenomena.

The author thanks Yu. V. Kopaev, R. A. Suris, and A. V. Krasheninnikov for helpful discussions of this work.

This work was carried out under the program ''Physics of Solid-State Nanostructures'' of the Russian Ministry of Science and Technology (Project No. 96-1030) and also with the support of the Russian Fund for Fundamental Research (Project No. 96-02-17363) and the State Program ''Integration.''

[1] R. Tsu and L. Esaki, Appl. Phys. Lett. **22**, 562 (1973).
[2] E. R. Brown, T. C. L. G. Sollner, C. D. Parker *et al.*, Appl. Phys. Lett. **55**, 1777 (1989); E. R. Brown, J. R. Söderström, C. D. Parker *et al.*, Appl. Phys. Lett. **58**, 2291 (1991).
[3] H. C. Liu and T. C. L. G. Sollner, Semicond. Semimet. **41**, 359 (1994).
[4] W. F. Chow, *Principles of Tunnel Diode Circuits* (John Wiley, New York, 1964), p. 177.
[5] A. M. Georgievskiĭ, D. V. Gromov, K. V. Dudinov *et al.*, Mikroelektronika **25**, 249 (1996).
[6] W. R. Frensley, Appl. Phys. Lett. **51**, 448 (1987); Rev. Mod. Phys. **62**, 745 (1990).
[7] R. K. Mains and G. I. Haddad, J. Appl. Phys. **64**, 3564 (1988).
[8] R. K. Mains and G. I. Haddad, J. Appl. Phys. **64**, 504 (1988).
[9] C. L. Fernando and W. R. Frensley, Phys. Rev. B **52**, 5092 (1995).
[10] F. A. Buot and A. K. Rajagopal, Phys. Rev. B **48**, 17 217 (1993).
[11] V. F. Elesin, D. V. Melnikov, and A. I. Podlivaev, Phys. Low-Dimens. Semicond. Struct. **1**, 23 (1996).
[12] H. C. Liu, Phys. Rev. B **43**, 12 538 (1991); Erratum **48**, 4977 (1993).
[13] V. Kislov and A. Kamenev, Appl. Phys. Lett. **59**, 1500 (1991).
[14] D. Sokolovski, Phys. Rev. B **37**, 4201 (1988).
[15] V. F. Elesin, Phys. Low-Dimens. Semicond. Struct. **5/6**, 79 (1998).
[16] V. F. Elesin, Zh. Éksp. Teor. Fiz. **112**, 483 (1997) [JETP **85**, 264 (1997)].
[17] V. I. Siforov, *Radio Reception Devices* [in Russian] (Izdat. MO SSSR, Moscow, 1954).
[18] R. F. Kazarinov and R. A. Suris, Fiz. Tekh. Poluprovodn. **6**, 148 (1972) [Sov. Phys. Semicond. **6**, 120 (1972)].
[19] E. R. Brown, C. D. Parker, and T. C. L. G. Sollner, Appl. Phys. Lett. **54**, 934 (1989).
[20] W. E. Lamb, Phys. Rev. **134**, 1429 (1964)

Translated by Paul F. Schippnick

Structural stability of disperse systems and finite nature of a coagulation front

P. B. Dubovski*)

Institute of Numerical Mathematics, Russian Academy of Sciences, 117951 Moscow, Russia
(Submitted 23 February 1999)
Zh. Éksp. Teor. Fiz. **116**, 717–730 (August 1999)

A new discrete model of coagulation, which is a discrete analog of the Oort–van de Hulst–Safronov equation, is derived. It is shown that the familiar version, in contrast with Smoluchowski's equation, can be used to calculate the propagation of a coagulation front. The relationship between compliance to the mass conservation law and the finite nature of the coagulation front is established, and then estimates of the time of violation of the mass conservation law are made for several classes of coagulation kernels. One of the conclusions is that the mass conservation law can be violated in cases where particles of roughly equal mass cannot coagulate, as occurs, for example, in gravitational coagulation. Estimates of the time for the appearance of structural instability of the system are made for multiplicative coagulation kernels. © *1999 American Institute of Physics.* [S1063-7761(99)02608-6]

1. PRELIMINARY REMARKS

Let us consider disperse systems containing particles having different masses, which undergo collisions that lead to changes in their masses. It is usually assumed that coagulation can be regarded as the fusion of two colliding particles. On the basis of this assumption we write Smoluchowski's equation:[1,2]

$$\frac{dc_i}{dt}=\frac{1}{2}\sum_{j=1}^{i-1}K_{i-j,j}c_{i-j}c_j-c_i\sum_{j=1}^{\infty}K_{i,j}c_j. \tag{1}$$

In its continuous form Eq. (1) is written as[3]

$$\frac{\partial c(x,t)}{\partial t}=\frac{1}{2}\int_0^x K(x-y,y)c(x-y,t)c(y,t)\,dy - c(x,t)\int_0^\infty K(x,y)c(y,t)\,dy. \tag{2}$$

However, there is another continuous coagulation model, viz., the Oort–van de Hulst model,[4] which was written in a convenient form by Safronov.[5] This model is used in astronomy to analyze cosmological objects (the formation of stars and planets, the evolution of nebulae, galaxies, clouds of cosmic dust, etc.),[4–7] in geophysics,[8–12] and in technical apparatus.[9,13]

In this paper we derive a new discrete coagulation equation and prove that it is a discrete analog of the Oort–van de Hulst–Safronov coagulation equation (Sec. 3). Since some mathematical properties of the two fundamental coagulation models are different, they can be regarded as complementing one another. New results on the kinetics of coagulation can then be obtained.

As an example of such complementation, we calculate the coagulation front velocity, which is the rate of displacement of the boundary of nonzero values of the distribution function. Such calculations are possible only with the use of the Oort–van de Hulst–Safronov equation.

We note further that violation of the mass conservation law occurs during intense coagulation at the time when the coagulation front escapes to infinity for finite initial data. This observation allows us to reveal several new classes of coagulation kernels that lead to violation of the mass conservation law. Estimates of the time for violation of the mass conservation law in a disperse systems are made for multiplicative coagulation kernels of the form $K(x,y)=x^\alpha y^\alpha$ $(\alpha>1)$. It is concluded that the conservation of mass can be violated in cases where particles of roughly equal mass cannot coagulate, as occurs, for example, during gravitational coagulation.

We note that the term ''violation of the mass conservation law,'' which is generally accepted in the mathematical literature on coagulation, does not precisely reflect the physical essence of the phenomenon. In fact, no violation of the mass conservation law occurs (the system is closed everywhere!), and all the mass (or part of it) is assembled in a single, infinitely large cluster and, thereby, ceases to participate in the coagulation kinetics. In percolation theory this is called the percolation transition, in polymerization theory it is called the gel point, in astrophysics it is called collapse, etc.

2. DISCRETE MODEL OF COAGULATION

Let us consider a disperse system which has the following properties:

1) the system is rarefied enough to presume that the interacting particles do not experience the effects of other particles;

2) the mean collision time (a microscopic time) is significantly smaller than the time for variation of the distribution function;

3) there are random forces which stir the disperse system so that the motions of the particles between collision acts (including their approach) are statistically independent;

 © 1999 American Institute of Physics

4) the masses (volumes) of all the particles are proportional to a certain $m_0>0$.

Let the particles grow as a result of collisions between pairs of particles with the masses im_0 and jm_0 (here and below, to fix ideas we assume that $i\geqslant j$). Particles with the masses im_0 are called i-mers, and m_0 is the mass of the smallest particles in the system.

We assume that a collision of an i-mer and a j-mer leads to splitting of the smaller j-mer into j monomers, which instantaneously attach themselves to i-mers. Thus, as a result of one collision event we have $(i+1)$-mers (the number of them equals j), and the j-mer disappears.

From balance considerations we obtain the kinetic model

$$\frac{dc_i(t)}{dt}=c_{i-1}(t)\sum_{j=1}^{i-1}K_{i-1,j}jc_j(t)-c_i(t)\sum_{j=1}^{i}K_{i,j}jc_j(t)-\sum_{j=i}^{\infty}K_{i,j}c_i(t)c_j(t),\quad i\geqslant 1, \tag{3}$$

where $c_i(t)$ is the concentration of i-mers at the time t, $K_{i,j}$ $(i\neq j)$ is, as always, the coagulation kernel, which characterizes the frequency of the collisions between i-mers and j-mers. The value of $K_{i,i}$ is equal to half of the frequency of collisions between particles with the mass i. This is due to the twofold decrease in the number of these particles in the pairs when they interact.

The first term on the right-hand side of (3) gives the influx of i-mers into the disperse system due to the collisions of $(i-1)$-mers and monomers which appear as a result of the splitting of a j-mer. If $i=1$, this term is set equal to zero. The second term describes the decrease in the number of i-mers as a result of the fusion of monomers with them. The multiplier j in the first and second terms shows that j monomers participate in a single collision event. The third term describes the decrease in the number of i-mers as a result of their splitting.

If we supplement Eq. (3) with the nonnegative initial data $c_i(0)$, we see that its solutions will also be nonnegative. In fact, we can write (3) in the integral form

$$c_i(t)=\exp\left\{-\int_0^t\left(\sum_{j=1}^{i}K_{i,j}jc_j(s)+\sum_{j=i}^{\infty}K_{i,j}c_j(s)\right)ds\right\}\times\left(c_i(0)+\int_0^t\exp\left\{\int_0^s\left[\sum_{j=1}^{i}K_{i,j}jc_j(s_1)+\sum_{j=i}^{\infty}K_{i,j}c_j(s_1)\right]ds_1\right\}\times\left[c_{i-1}(s)\sum_{j=1}^{i-1}K_{i-1,j}jc_j(s)\right]ds\right). \tag{4}$$

If the initial data $c_i(0)$ are strictly positive, for a truncated coagulation kernel $K_{m,j}=0$, $(m,j\geqslant N_0)$ we easily obtain positiveness of the $c_i(t)$ for all $i\geqslant 1$, $t>0$ by assuming that we find a time t_0 and a number i_0 such that $c_{i_0}(t_0)=0$, in contradiction to (4). If the initial data are not strictly positive, we approximate them by positive initial data and obtain nonnegativeness of the solution by going to the limit. These arguments are similar to Refs. 14 and 15. Nonnegativeness of the solution for untruncated coagulation kernels can be obtained along with a solution existence theorem by approximating $K_{m,j}$ by a sequence of finite kernels, which generate a sequence of nonnegative solutions (4), and then the transition to the limit, i.e., to the solution (4), is made.

Let us verify whether Eq. (3) obeys the mass conservation law

$$N_1\overset{\text{def}}{=}\sum_{i=1}^{\infty}ic_i(t)=\text{const}. \tag{5}$$

For this purpose, we multiply (3) by i and sum it over the range $1\leqslant i\leqslant\infty$. We then obtain

$$\frac{dN_1}{dt}=\sum_{i=1}^{\infty}\sum_{j=1}^{i-1}ijK_{i-1,j}c_{i-1}c_j-\sum_{i=1}^{\infty}\sum_{j=1}^{i}ijK_{i,j}c_ic_j-\sum_{i=1}^{\infty}\sum_{j=i}^{\infty}iK_{i,j}c_ic_j.$$

In the third term we interchanged the order of summation and summed over $\Sigma_{j=1}^{\infty}\Sigma_{i=1}^{j}$. Then in the second and third terms we separate the components with $j=i$ and make the replacement $i=i'+j'$, $j=j'$. We then obtain

$$\frac{dN_1}{dt}=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(i+j)jK_{i+j-1,j}c_{i+j-1}c_j-\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(i+j)jK_{i+j,j}c_{i+j}c_j-\sum_{i=1}^{\infty}i(i+1)K_{i,i}c_i^2-\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}jK_{i+j,j}c_{i+j}c_j.$$

After several replacements of the type $i\mapsto i+1$, we have zero, and we thus obtain the mass conservation law.

3. TRANSITION TO THE OORT–VAN DE HULST–SAFRONOV EQUATION

An important observation is that the transition to the limit $m_0\to 0$ in (3) gives the familiar continuous coagulation model

$$\frac{\partial c(x,t)}{\partial t}=-\frac{\partial}{\partial x}\left[c(x,t)\int_0^x yK(x,y)c(y,t)\,dy\right]-\int_x^{\infty}K(x,y)c(x,t)c(y,t)\,dy. \tag{6}$$

In fact, to obtain the continuous form of Eq. (3) we introduce the distribution function $c(x,t)$, which describes the distribution of the particles of mass x at the time t, i.e., $c(x,t)\,dx$ is equal to the number of particles with masses in the interval $(x,x+dx)$ at the time t. The mass of the i-mers equals im_0; therefore, $c_i(t)=c(im_0,t)m_0$. Since $K_{i,j}=K(im_0,jm_0)$, we have

$$K_{i,j}c_i(t)c_j(t)=K(im_0,jm_0)c(im_0,t)c(jm_0,t)m_0^2.$$

Therefore, making the replacement $x=im_0$, we obtain

$$\frac{\partial c(x,t)}{\partial t} = -\frac{1}{m_0}\Bigg[c(x,t)\sum_{j=1}^{x/m_0} K(x,jm_0)c(jm_0,t)jm_0 - c(x - m_0,t)\sum_{j=1}^{x/m_0-1} K(x-m_0,jm_0)c(jm_0,t)jm_0\Bigg]m_0 - c(x,t)\sum_{j=i}^{\infty} K(x,jm_0)c(jm_0,t)m_0.$$

Since these sums are simply Darboux integral sums, we go off to the limit $m_0 \to 0$ and obtain (6). Equation (6) was derived by Oort and van de Hulst using totally different methods[4] and was written in the form (6) by Safronov.[5] Thus, it turns out that Eq. (6) is a continuous form of the new discrete equation (3). It is noteworthy that heretofore there were no discrete analogs of Eq. (6). Continuous limiting equations are usually derived from their discrete analogs. As an example we can point to Smoluchowski's equation, which was first obtained in the discrete form (1) (Refs. 1 and 2) and whose continuous form (2) was derived by Müller.[3] Another, more recent example can be found in Ref. 16, where ben-Avraham *et al.* first derived a discrete monomer-monomer model of heterogeneous catalysis and then went over to the limiting equation in a continuous form. We also mention Refs. 17–19 in this connection. In addition, we note that other approaches to the derivation of coagulation models are possible.[20]

Several arguments, which have something in common with our derivation of Eq. (3), can be found in Ref. 13 (p. 131), where Vasenin *et al.* noted the relationship between (6) and the following process: l particles of mass x interact with smaller particles of mass μ/l ($\mu < x$) during the time Δt. This relationship was derived by expanding several functions into series and ''truncating'' these series without proper justification [compare with Refs. 8 (p. 45) and 9 (p. 154)].

The Oort–van de Hulst–Safronov equation (6) can be regarded as a model of continuous growth.[5,9] In fact, if we assume that all particles grow as a result of the attachment of smaller particles, the first integral on the right-hand side of (6) equals dx/dt and the entire first term is simply the variation of $c(x,t)$ as a result of the addition of particles with masses y ($y<x$). Thus, Eq. (6) without its last term is a one-dimensional continuity equation with a ''density'' $c(x,t)$ and a ''velocity'' dx/dt. The second term in (6) corresponds to the escape of particles of mass x as a result of their ''sedimentation'' on larger particles. Thus, a particle retains its ''individuality'' in collisions with smaller particles, but loses it in collisions with larger particles. In other words, collisions of particles of mass x with smaller particles alter the mass x of the particles, and collisions with larger particles alter the number of particles with the mass x. This procedure gives an averaged and smoothed growth rate for all particles of a definite radius.

It is noteworthy that Eq. (6) was applied in Refs. 6 and 7 to an investigation of the evolution of various cosmological objects. A similar approach to coagulational growth involving a small number of identical large droplets descending through a randomly distributed suspension of smaller droplets was used to calculate coagulation processes in atmospheric clouds by Telford,[10] who solved the Oort–van de Hulst–Safronov equation (6) without the second term on the right-hand side. It was shown numerically and analytically in more recent studies reported in Refs. 11 and 12, respectively, that the method in Ref. 10 [and, therefore, Eq. (6)] gives results which are similar to the more familiar Smoluchowski kinetic approach. It was also pointed out in Refs. 9 and 13 that Eq. (6) is also useful for investigating processes in technical apparatus which use two-phase media (in nozzles and engines).

4. PROPAGATION OF DISTURBANCES

One significant difference between the coagulation models under consideration is the fact that, in contrast with the Oort–van de Hulst–Safronov equation (6), Smoluchowski's equation (2) propagates disturbances with infinite velocity. In order to demonstrate this point, we set $K \equiv 1$ and supplement Eq. (2) with the finite initial data $c_0(x) = \theta(1-x)$, where the step function $\theta(x)$ equals unity at $x \geqslant 0$ and zero at $x<0$. Our purpose is to show that at any positive time, no matter how small, a distribution function which obeys Eq. (2) becomes nonzero at all values of the argument x, no matter how large.

Using the Laplace transformation, we obtain the following expression for the Laplace transform $F(p,t)$:

$$F(p,t) = \frac{4(1-e^{-p})}{(2+t)[p(2+t)+t\exp(-p)-t]}.$$

To find the original function we locate the singular points of $F(p,t)$ on the complex plane of the variable $p = a + ib$, $(a,b \in R^1)$. Introducing the notation

$$\frac{2+t}{t} = \alpha > 1,$$

we obtain the system of equations

$$\begin{cases} \alpha a = 1 - e^{-a}\cos b, \\ \alpha b = e^{-a}\sin b, \end{cases}$$

from which we have

$$a = \frac{1}{\alpha} - b\cot b, \quad \alpha e^{1/\alpha} b = e^{b\cot b}\sin b.$$

Therefore, the singular points are at $p_n = a_n + ib_n$:

$$b_n = 2\pi n + \varepsilon_n, \quad b_{-n} = -2\pi n - \varepsilon_n, \quad n \geqslant 1,$$

$$a_n = a_{-n} = \frac{1}{\alpha} - b_n \cot b_n < 0, \quad n \geqslant 1,$$

$$\varepsilon_n > 0, \quad \varepsilon_n \to 0, \quad n \to \infty.$$

We note that

$$\varepsilon_n \sim \frac{2\pi n}{\ln(2\pi n/e) + \ln\alpha}, \quad n \geqslant 1,$$

$$b_n \cot b_n = b_{-n}\cot b_{-n} \sim 1 + \frac{2\pi n}{\varepsilon_n},$$

$$a_n \sim \frac{1}{\alpha} - \ln\alpha - \ln(2\pi n).$$

Let us ascertain the types of singularities at p_n. For this purpose we find the limit

$$\lim_{p\to p_n}(p-p_n)C(p,t) = \frac{4(1-e^{-p_n})}{(2+t)t}\lim_{p\to p_n}\frac{p-p_n}{\alpha p + e^{-p}-1}.$$

We examine this limit in greater detail:

$$\lim_{p\to p_n}\frac{p-p_n}{\alpha p+e^{-p}-1} = \lim_{p\to p_n}\frac{p-p_n}{\alpha(p-p_n)+\alpha p_n + e^{-(p-p_n)}e^{-p_n}-1} = \frac{1}{\alpha - e^{-p_n}} \neq 0. \tag{7}$$

To obtain the second equality we took into account that $\alpha p_n - 1 + e^{-p_n} = 0$ and used the expansion of $e^{-(p-p_n)}$ into a series in $p-p_n$. Therefore, the points p_n are first-order poles of $F(p,t)$. This allows us to easily determine the residues of $e^{px}F(p,t)$ at these points from Eq. (7) and to write the solution of Eq. (2) in the form

$$c(x,t) = \sum_{n=-\infty}^{+\infty} e^{p_n x}\frac{4(1-e^{-p_n})}{(2+t)t(\alpha - e^{-p_n})} = \frac{4}{(2+t)t}\sum_{n=1}^{\infty}\left[\frac{e^{p_n x}(1-e^{-p_n})}{\alpha - e^{-p_n}} + \frac{e^{p_{-n}x}(1-e^{-p_{-n}})}{\alpha - e^{-p_{-n}}}\right].$$

Performing the replacements $p_n = a_n + 2\pi i n + i\varepsilon_n$ and $p_{-n} = a_n - 2\pi i n - i\varepsilon_n$, we ultimately obtain

$$c(x,t) = \frac{4}{(2+t)t}\sum_{n=1}^{\infty} e^{a_n x}\left[\frac{e^{ib_n x}(1-e^{-a_n}e^{-i\varepsilon_n})}{\alpha - e^{-a_n}e^{-i\varepsilon_n}} + \frac{e^{-ib_n x}(1-e^{-a_n}e^{i\varepsilon_n})}{\alpha - e^{-a_n}e^{i\varepsilon_n}}\right] = \frac{8}{(2+t)} \times \sum_{n=1}^{\infty}\left\{e^{a_n(x+1)}\frac{[(2+t)e^{a_n} - t\cos\varepsilon_n]\cos(b_n x) + t\sin\varepsilon_n\sin(b_n x)}{(2+t)^2 e^{2a_n} - 2t(2+t)e^{a_n}\cos\varepsilon_n + t^2} - e^{a_n x}\frac{[(2+t)e^{a_n} - t\cos\varepsilon_n]\cos(b_n(x-1)) + t\sin\varepsilon_n\sin(b_n(x-1))}{(2+t)^2 e^{2a_n} - 2t(2+t)e^{a_n}\cos\varepsilon_n + t^2}\right\}. \tag{8}$$

Thus, even if the initial data are equal to zero at $x \geqslant 1$, the solution instantaneously becomes positive for all values of x, no matter how large. In fact, if a time $t_0 > 0$ and an interval $[x_1, x_2]$, which are such that $c(x,t_0) = 0$ in the interval would be found, the expansion of this zero function in the basis set[1)]

$$e^{a_n x}\cos(b_n x), e^{a_n x}\sin(b_n x), \quad n \geqslant 1, \tag{9}$$

would give zero coefficients, in contradiction to (8).

Therefore, a nonzero initial value propagates with infinite velocity. This deficiency is similar, for example, to the behavior of the thermal-conductivity equation and mathematically means that Eqs. (1) and (2) have parabolic properties.

In contrast with Smoluchowski's equation, Eq. (6) does not have such a deficiency. In fact, we can rewrite (6) in the form

$$\frac{\partial c(x,t)}{\partial t} + v(x,t)\frac{\partial c(x,t)}{\partial x} = -c(x,t)\left(\int_0^x y\partial_1 K(x,y)c(y,t)\,dy\right) - xK(x,x)c^2(x,t) - c(x,t)\int_x^{\infty} K(x,y)c(y,t)\,dy, \tag{10}$$

where

$$v(x,t) = \int_0^x yK(x,y)c(y,t)\,dy$$

and $\partial_1 K(x,y)$ denotes the differentiation of K with respect to the first argument. Let $x(s)$ be a solution of the characteristic equation $dx/dt = v(x,t)$. Then, the substitution of

$$c(x,t) = \exp\left\{-\int_0^t\left[K(x(s),x(s))c(x(s),s)x(s) - \int_0^{x(s)} y\partial_1 K(x(s),y)c(y,s)\,dy - \int_{x(s)}^{\infty} K(x(s),y)c(y,s)dy\right]ds\right\}u(x,t) \tag{11}$$

gives

$$\frac{\partial u(x,t)}{\partial t} + v(x,t)\frac{\partial u(x,t)}{\partial x} = 0. \tag{12}$$

From (11) we see that the functions $c(x,t)$ and $u(x,t)$ are equal or not equal to zero at the same points. The characteristic equation for (10) and (12) has the form

$$\frac{dx}{dt}=v(x,t). \tag{13}$$

We conclude from the simple equation (12) that if $c_0(x_1)=0$, then $c(x_1,t)$ becomes positive not earlier than t_1, when the first characteristic curve $x(s)$ with a nonzero initial value x_0 reaches x_1.

Thus, Eq. (6) imposes a physically reasonable bound on the propagation rate of disturbances and thereby permits calculation of the coagulation front.[2)] Mathematically, this means that Eq. (6) has definite hyperbolic properties.

5. VIOLATION OF THE MASS CONSERVATION LAW FOR $K(x,y)=(xy)^\alpha$

Using the arguments of the previous section, we can estimate the coagulation front in several cases. Let $c_0(x)=0$ if $x\geqslant x_0$. Then, the characteristic curve originating from x_0 divides the plane into two parts so that $c(x,t)=0$ if the point (x,t) is to the right of the curve. We call this characteristic curve the cutoff characteristic or the coagulation front. From Eq. (13), which describes the characteristics, we see that the coagulation front satisfies the equation

$$\frac{dx}{dt}=\int_0^x yK(x,y)c(y,t)\,dy=\int_0^\infty yK(x,y)c(y,t)\,dy \tag{14}$$

with the initial value $x(0)=x_0$.

Therefore, if $K(x,y)=C=\text{const}$, then

$$x(t)=x_0+CN_1t, \tag{15}$$

where N_1 is the constant first moment of the solution.

Similarly, it can be shown for an additive coagulation kernel [i.e., for $K(x,y)=(x+y)$] that the expression for the coagulation front has the form

$$x(t)=\exp(N_1t)\{x_0+N_2(0)t\}, \tag{16}$$

so that the coagulation front moves faster in the case of additive coagulation kernels than in the case of constant kernels, as is perfectly natural.

Let us estimate the coagulation front for the multiplicative kernel $K(x,y)=xy$. From (14) we see that

$$x(t)=x_0\exp\left(\int_0^t N_2(s)\,ds\right).$$

For the Smoluchowski equation (2) we obtain unboundedness of the second moment $N_2(t)$ at the critical time $t_{cr}=[N_2(0)]^{-1}$:

$$N_2(t)=N_2(0)(1-N_2(0)t)^{-1}.$$

We utilize the relations

$$\int_0^\infty\int_0^x x^k y^k c(x)c(y)dy\,dx=\frac{1}{2}\left(\int_0^\infty x^k c(x)\,dx\right)^2,$$

$$\int_x^\infty y^k c(y)dy\leqslant\frac{1}{x}\int_x^\infty y^{k+1}c(y)\,dy.$$

When they are taken into account, we also obtain unboundedness of the second moment for the multiplicative kernel $K(x,y)=xy$ in the Oort–van de Hulst–Safronov model:

$$\frac{N_2(0)}{1-N_2(0)t/2}\leqslant N_2(t)\leqslant\frac{N_2(0)}{1-N_2(0)t}.$$

In this case $[N_2(0)]^{-1}\leqslant t_{cr}\leqslant 2[N_2(0)]^{-1}$.

Thus, it can be seen that the coagulation front escapes to infinity when $t\to t_{cr}$.

Let us turn our attention to another effect of the influence of an increase without bound, viz., violation of the mass conservation law at the same critical time t_{cr}. As has been reported, for the Smoluchowski equation (2) this effect is caused by the escape of the second moment N_2 of the solution to infinity (see, for example, Refs. 14 and 21–26). Thus, it turns out that the escape of the coagulation front to infinity signifies violation of the mass conservation law. This observation permits establishment of violation of the mass conservation law for several other coagulation kernels, which were not previously open to analysis.

Let us consider the kernels $K(x,y)=x^\alpha y^\alpha$ $(\alpha>1)$. We estimate the time for the appearance of structural instability of the system, at which violation of the mass conservation law occurs. Solving Eq. (14) with these kernels, we obtain

$$x^{\alpha-1}=\left\{\frac{1}{x_0^{\alpha-1}}-(\alpha-1)\int_0^t N_{1+\alpha}(s)\,ds\right\}^{-1}. \tag{17}$$

Here, as usual, $N_k(t)$ denotes the kth moment of the distribution function. From (17) we see that the coagulation front escapes to infinity just as the $(\alpha+1)$th moment of the solution becomes infinite. Let us estimate $N_{1+\alpha}(t)$. For this purpose we integrate (2) with the weight x^k:

$$\frac{N_k(t)}{dt}=\frac{1}{2}\int_0^\infty\int_0^\infty K(x,y)[(x+y)^k - x^k-y^k]c(x,t)c(y,t)\,dx\,dy. \tag{18}$$

We use the following inequality, whose proof can be found in the Appendix:

$$(x+y)^\gamma-x^\gamma-y^\gamma\geqslant(2^\gamma-2)x^{\gamma/2}y^{\gamma/2},\quad \gamma\geqslant 2,\quad x,y\geqslant 0. \tag{19}$$

Then, for $k=\gamma=1+\alpha$ we obtain

$$\frac{dN_{1+\alpha}}{dt}\geqslant(2^\alpha-1)N_{(1+3\alpha)/2}^2\geqslant(2^\alpha-1)N_{1+\alpha}^2(t),\quad t\geqslant 0. \tag{20}$$

It follows from (20) that the escape of the coagulation front to infinity together with the violation of the mass conservation law occur in the system no later than the time

$$t_1=[(2^\alpha-1)N_{1+\alpha}(0)]^{-1},$$

when $N_{1+\alpha}$ equals infinity. It also follows from (20) that $N_{1+\alpha}$ and $N_{(1+3\alpha)/2}$ go to infinity simultaneously. Applying the inequality (19) n times and substituting the result into (18) each time, we obtain

$$\frac{d}{dt}N_{\frac{1+(2^n-1)\alpha}{2^{n-1}}}\geqslant\left(2^{\frac{1+(2^n-1)\alpha}{2^{n-1}}-1}-1\right)N_{\frac{1+(2^{n+1}-1)\alpha}{2^n}},$$

$$n\geqslant 1,\quad t\geqslant 0.$$

All the integral moments in this hierarchy become infinitely large simultaneously. Allowing n to tend to infinity,

$$\lim_{n\to\infty}\frac{1+(2^{n+1}-1)\alpha}{2^n}=2\alpha,$$

we find that violation of the mass conservation law occurs just as $N_{2\alpha}$ goes to infinity. Let us, therefore, estimate $N_{2\alpha}$. We substitute $k=2\alpha$ into (18) and use the inequality (19) again:

$$\frac{dN_{2\alpha}(t)}{dt}\geqslant(2^{2\alpha-1}-1)N_{2\alpha}^2(t),\quad t\geqslant 0.$$

Thus,

$$t_{\rm cr}\leqslant[(2^{2\alpha-1}-1)N_{2\alpha}(0)]^{-1}, \tag{21}$$

where $t_{\rm cr}$ is the critical moment for violation of the mass conservation law of the system.

We note that a result which is reasonable from the physical standpoint was obtained in (21): the smaller is the frequency of fusion of large particles (the smaller is α), the later the mass conservation law is violated. When $\alpha=1$, the estimate (21) becomes the familiar expression

$$t_{\rm cr}=[N_2(0)]^{-1}$$

for $K(x,y)=xy$.

6. VIOLATION OF THE MASS CONSERVATION LAW FOR OTHER COAGULATION KERNELS

Here we develop the approach from the previous section to estimate the time the conservation law is violated for the following coagulation kernels:

$$K(x,y)=\begin{cases}a(x)b(y), & x\geqslant y,\\ a(y)b(x), & x\leqslant y.\end{cases} \tag{22}$$

The equation of the characteristic originating from the maximum point x_0, at which the finite initial data are equal to zero, then takes the form

$$\frac{dx}{dt}=a(x)\int_0^\infty yb(y)c(y,t)\,dy,$$

so that

$$\int_{x_0}^{x}\frac{dx}{a(x)}=\int_0^t\int_0^\infty yb(y)c(y,s)\,dy\,ds. \tag{23}$$

Therefore, if

$$\int_0^\infty\frac{dx}{a(x)}<\infty, \tag{24}$$

we can find the finite critical time $t_{\rm cr}<\infty$, at which the coagulation front $x(t)$ escapes to infinity, signifying, as was shown in the preceding section on a phenomenological level of rigorousness, the occurrence of violation of the mass conservation law of the system.

If the function $a(x)$ is fairly small and does not satisfy the inequality (24), we can conclude from (23) that the critical time corresponds to the moment the right-hand side of (23) becomes infinite. For example, for $b(x)=x^\beta$ the critical time corresponds to the time the $(1+\beta)$th moment of the solution $N_{1+\beta}(t)$ goes to infinity.

An important conclusion drawn from these observations is that if the inequality (24) is satisfied, a critical time appears for any $b(x)$. For example, if

$$a(x)=x^{1+\varepsilon},\quad \varepsilon>0,\quad b(x)=\exp(-x),$$

then

$$K(x,x)=x^{1+\varepsilon}\exp(-x)\to 0,\quad x\to\infty.$$

Nevertheless, the mass conservation law is violated. All the previously known kernels, which allow violation of the mass conservation law, took their dominant values at equal arguments, i.e., $K(x,x)\geqslant K(x-y,y)$ $(0\leqslant y\leqslant x)$. In this case $K(x,x)\to\infty$ as $x\to\infty$. Therefore, the nature of the violation of the conservation law is not confined to the intense coagulation of particles of roughly equal mass. This observation is important in the analysis of, for example, gravitational coagulation, where particles of equal mass do not coagulate.

7. CONCLUSIONS

A new discrete model of coagulation, which is a discrete analog of the previously known continuous Oort–van de Hulst–Safronov equation has been obtained.

If the coagulation kernel $K(x,y)$ grows fairly slowly, the mass conservation law is valid for all the types of coagulation described. A particle dissipation law also holds. As usual, variation of the integrand $K(x,y)c(x)c(y)$ by $K(x,y)c(x)c(y)-F(x,y)c(x+y)$ leads to allowance for splitting processes.

It has been shown that Smoluchowski's equation propagates disturbances with infinite velocity. This deficiency is not displayed by the Oort–van de Hulst–Safronov equation, which can, therefore, be used to estimate the velocity of the coagulation front, as we have demonstrated in several cases.

Finally, we have established that the coagulation front escapes to infinity at the same critical time the violation of the mass conservation law occurs. This observation allows us to calculate the appearance of violation of the mass conservation law for a number of important classes of kernels.

We wish to thank V. I. Agoshkov, A. É. Arinshteĭn, and V. P. Shutyaev for interest in this work and for some useful discussions.

This work was carried out with financial support from the Russian Fund for Fundamental Research (Grant No. 99-01-00336).

APPENDIX

Binomial inequality

Let us prove the binomial inequality (19), which was used in Sec. 5:

$$(x+y)^\gamma-x^\gamma-y^\gamma\geqslant(2^\gamma-2)x^{\gamma/2}y^{\gamma/2},\quad \gamma\geqslant 2,\quad x,y\geqslant 0.$$

We note at once that it also holds for $0\leqslant\gamma\leqslant 1$ and that it changes sign for $\gamma\leqslant 0$ and for $1\leqslant\gamma\leqslant 2$.

We introduce the notation $x=ty$ and assume, with no loss of generality, that $x \geqslant y$. We can then write (19) in the form

$$r(\gamma,t) \stackrel{\text{def}}{=} (1+t)^{\gamma}-1-t^{\gamma}-(2^{\gamma}-2)t^{\gamma/2} \geqslant 0,$$
$$\gamma \geqslant 2, \quad t \geqslant 1. \tag{A1}$$

We note that

$$r(\gamma,1) \equiv 0, \quad r'_t(\gamma,1) \equiv 0, \quad \gamma \geqslant 2. \tag{A2}$$

Differentiating (A1) twice, we obtain

$$r''_{tt}(\gamma,t) = \gamma(\gamma-1)[(1+t)^{\gamma-2}-t^{\gamma-2}] - \frac{1}{2}\gamma(\gamma-2)(2^{\gamma-1}-1)t^{(\gamma-4)/2}. \tag{A3}$$

We thus see that $r''_{tt}(\gamma,1)$ is positive for $\gamma \in [2,\infty)$. We shall show that r''_{tt} increases as a function of t if $\gamma \in [2,\infty)$.

We first consider the interval $\gamma \in [3,4]$. Then, as t increases, the expression in the square brackets in (A3) increases, and the subtrahend on the right-hand side of (A3) decreases. Therefore, the second derivative r''_{tt} is positive if $t \in [1,\infty)$ and $\gamma \in [3,4]$, and thus, by virtue of (A2), we establish that $r(\gamma,t) \geqslant 0$ at $t \geqslant 1$ for $\gamma \in [3,4]$.

To prove that r''_{tt} increases with t if $\gamma \geqslant 4$, we consider the derivative

$$r'''_{ttt} = \gamma(\gamma-1)(\gamma-2)[(1+t)^{\gamma-3}-t^{\gamma-3}] - \frac{1}{4}\gamma(\gamma-2)(\gamma-4)(2^{\gamma-1}-1)t^{(\gamma-6)/2}. \tag{A4}$$

The expression in square brackets in (A4) increases with increasing t if $\gamma > 4$, and the subtrahend does not increase if $\gamma \leqslant 6$. Therefore, the statement that r increases with t is also valid for $\gamma \in [4,6]$.

Similarly, differentiating (A4) the required number of times, we find

$$r(\gamma,t) \geqslant 0, \quad \gamma \in [3,\infty), \quad t \in [1,\infty). \tag{A5}$$

Now, let us consider the half-interval $\gamma \in [2,3)$, in which r''_{tt} decreases as a function of t. We can show, however, that it remains positive. It follows from (A3) that a sufficient condition in this case is the inequality

$$v(\gamma,t) \stackrel{\text{def}}{=} (1+t)^{\gamma-2}t^{2-\gamma/2}-t^{\gamma/2} \geqslant \frac{\gamma-2}{2(\gamma-1)}(2^{\gamma-1}-1),$$
$$\gamma \in [2,3), \quad t \in [1,\infty). \tag{A6}$$

It is clear that (A6) is valid at $t=1$. We can see that the derivative $v'_t(\gamma,t)$ at $t \geqslant 1$ is nonnegative by making the replacement $u=t^{-1}$ $(t \geqslant 1)$. Thus,

$$v'_t(\gamma,t) \geqslant 0, \quad t \geqslant 1, \quad \gamma \in [2,3). \tag{A7}$$

The inequality (19) follows from (A5) and (A7).

*)E-mail: dubovski@inm.ras.ru

1)More precisely, we are referring to the basis set obtained from (9) by adding a constant and all the possible products of the elements in (9), which, on the basis of the Stone–Weierstrass theorem, generates an algebra that is dense everywhere in each compact from $[x_0,\infty)$.

2)We recall that we define the coagulation front as the displacement of the boundary of nonzero values of the distribution function.

[1] M. V. Smoluchowski, Phys. Z. **17**, 557 (1916).
[2] M. Smoluchowski, Z. Phys. Chem. **92**, 129 (1917) [Russ. transl., *Coagulation of Colloids*, A. I. Rabinovich and P. S. Vasil'ev (Eds.), ONTI (1936), p. 7].
[3] H. Müller, Kolloidchemische Beihefte **27**, 223 (1928) [Russ. transl., *Coagulation of Colloids*, A. I. Rabinovich and P. S. Vasil'ev (Eds.), ONTI (1936), p. 74].
[4] J. H. Oort and H. C. van de Hulst, Bull. Astron. Inst. Netherlands **10**, 187 (1946).
[5] V. S. Safronov, *Evolution of the Protoplanetary Cloud and Formation of the Earth and the Planets*, Israel Program for Scientific Translations, Jerusalem (1972) [Russ. original, Nauka, Moscow (1969)].
[6] S. Piotrowsky, Acta Astron., Ser. A **5**, 115 (1953).
[7] J. S. Dohnanyi, Report No. TR-67-340-3 (1967).
[8] V. M. Voloshchuk and Yu. S. Sedunov, *Coagulation Processes in Disperse Systems* [in Russian], Gidrometeoizdat, Leningrad (1975).
[9] L. E. Sternin and A. A. Shraĭber, *Multiphase Flow of Gases with Particles* [in Russian], Mashinostroenie, Moscow (1994).
[10] J. W. Telford, J. Meteorol. **12**, 436 (1955).
[11] S. Twomey, J. Atmos. Sci. **21**, 553 (1964).
[12] W. T. Scott, J. Atmos. Sci. **25**, 871 (1968).
[13] I. M. Vasenin, V. A. Arkhipov, V. G. Butov, A. A. Glazunov, and V. F. Trofimov, *Gas Dynamics of Two-Phase Flows in Nozzles* [in Russian], Izd. Tomsk. Univ., Tomsk (1986).
[14] P. B. Dubovski and I. W. Stewart, Math. Methods Appl. Sci. **19**, 571 (1996).
[15] P. B. Dubovski, in *Lecture Notes Series, Vol. 23*, Seoul National University, Research Institute of Mathematics, Global Analysis Research Center, Seoul (1994), p. 169.
[16] D. ben-Avraham, D. Considine, P. Meakin, S. Redner, and H. Takayasu, J. Phys. A: Math. Gen. **23**, 4297 (1990).
[17] A. É. Arinshteĭn and S. M. Mezhikovskiĭ, Khim. Fiz. **16**, 122 (1997).
[18] A. É. Arinshteĭn and V. I. Gol'danskiĭ, Dokl. Ross. Akad. Nauk **352**, 483 (1997).
[19] E. R. Domilovskiĭ, A. A. Lushnikov, and V. N. Piskunov, Prikl. Mat. Mekh. **44**, 697 (1980).
[20] A. A. Lushnikov, Dokl. Akad. Nauk SSSR **237**, 1122 (1977) [Dokl. Phys. Chem. **237** (1977)].
[21] J. B. McLeod, Proc. London Math. Soc. **14**, 445 (1964).
[22] F. Leyvraz and H. R. Tschudi, J. Phys. A: Math. Gen. **14**, 3389 (1981).
[23] M. H. Ernst, R. M. Ziff, and E. M. Hendriks, J. Colloid Interface Sci. **97**, 266 (1984).
[24] V. A. Galkin, Meteorol. Gidrol. **5**, 33 (1984).
[25] R. M. Ziff, M. H. Ernst, and E. M. Hendriks, J. Phys. A: Math. Gen. **16**, 2293 (1983).
[26] R. M. Ziff and G. Stell, J. Chem. Phys. **73**, 3492 (1980).

Translated by P. Shelnitz
Edited by S. J. Amoretty