Energy levels and wave functions of Bloch electrons in rational and irrational magnetic fields*

Douglas R. Hofstadter†
Physics Department, University of Oregon, Eugene, Oregon 97403
(Received 9 February 1976)

An effective single-band Hamiltonian representing a crystal electron in a uniform magnetic field is constructed from the tight-binding form of a Bloch band by replacing $\hbar\vec{k}$ by the operator $\vec{p} - e\vec{A}/c$. The resultant Schrödinger equation becomes a finite-difference equation whose eigenvalues can be computed by a matrix method. The magnetic flux which passes through a lattice cell, divided by a flux quantum, yields a dimensionless parameter whose rationality or irrationality highly influences the nature of the computed spectrum. The graph of the spectrum over a wide range of "rational" fields is plotted. A recursive structure is discovered in the graph, which enables a number of theorems to be proven, bearing particularly on the question of continuity. The recursive structure is not unlike that predicted by Azbel', using a continued fraction for the dimensionless parameter. An iterative algorithm for deriving the clustering pattern of the magnetic subbands is given, which follows from the recursive structure. From this algorithm, the nature of the spectrum at an "irrational" field can be deduced; it is seen to be an uncountable but measure-zero set of points (a Cantor set). Despite these features, it is shown that the graph is continuous as the magnetic field varies. It is also shown how a spectrum with simplified properties can be derived from the rigorously derived spectrum, by introducing a spread in the field values. This spectrum satisfies all the intuitively desirable properties of a spectrum. The spectrum here presented is shown to agree with that predicted by A. Rauh in a completely different model for crystal electrons in a magnetic field. A new type of magnetic "superlattice" is introduced, constructed so that its unit cell intercepts precisely one quantum of flux. It is shown that this cell represents the periodicity of solutions of the difference equation. It is also shown how this superlattice allows the determination of the wave function at nonlattice sites. Evidence is offered that the wave functions belonging to irrational fields are everywhere defined and are continuous in this model, whereas those belonging to rational fields are only defined on a discrete set of points. A method for investigating these predictions experimentally is sketched.

I. INTRODUCTION

The problem of Bloch electrons in magnetic fields is a very peculiar problem, because it is one of the very few places in physics where the difference between rational numbers and irrational numbers makes itself felt.[1,2] Common sense tells us that there can be no physical effect stemming from the irrationality of some parameter, because an arbitrarily small change in that parameter would make it rational — and this would create some physical effect with the property of being everywhere discontinuous, which is unreasonable. The only alternative, then, is to show that a theory which apparently distinguishes between rational and irrational values of some parameter does so only in a mathematical sense, and yields physical observables which are nevertheless continuous. It is the purpose of this paper to present a method which effects such a reconciliation of "rational" and "irrational" magnetic fields. The method is illustrated in a maximally simple model of the physical situation, but the ideas which arise are, it is to be hoped, applicable to more realistic models of the physical situation.

II. DERIVATION OF THE DIFFERENCE EQUATION

Briefly, then, the model involves a two-dimensional square lattice of spacing a, immersed in a uniform magnetic field H perpendicular to it. We restrict our considerations to what happens to a single Bloch band when the field is applied. This is one strong simplifying feature of the model; the next is that we postulate the following tight-binding form for the Bloch energy function:

$$W(\vec{k}) = 2E_0(\cos k_x a + \cos k_y a) .$$

Perhaps the most difficult step to justify on physical grounds is the following one, which I shall refer to as the "Peierls substitution"[3]: we replace $\hbar\vec{k}$ in the above function by the operator $\vec{p} - e\vec{A}/c$ ($\vec{A}$ being the vector potential), to create an operator out of $W(\vec{k})$, which we then treat as an effective single-band Hamiltonian. Work to justify this substitution has been done.[4-7]

When this substitution is made, the effective Hamiltonian is seen to contain translation operators $\exp(ap_x/\hbar)$ and $\exp(ap_y/\hbar)$. Depending on the gauge chosen, there are, in addition, certain phase factors dependent on the magnetic field

strength, which multiply the translation operators. If the Landau gauge — $\vec{A} = H(0, x, 0)$ — is chosen, then only the translations along y are multiplied by phases. From now on, we assume this gauge. Now when the effective Hamiltonian is introduced into a time-independent Schrödinger equation with a two-dimensional wave function, the following eigenvalue equation results:

$$E_0[\psi(x+a,y)+\psi(x-a,y)+e^{-ieHax/\hbar c}\psi(x,y+a) + e^{+ieHax/\hbar c}\psi(x,y-a)] = E\psi(x,y).$$

Note how the wave function at (x, y) is linked to its four nearest neighbors in the lattice. It is convenient to make the substitutions

$$x = ma, \quad y = na, \quad E/E_0 = \epsilon.$$

It is furthermore reasonable to assume plane-wave behavior in the y direction, since the coefficients in the above equation only involve x. Therefore, we write

$$\psi(ma, na) = e^{i\nu n} g(m).$$

Finally we introduce the parameter about which all the fuss is made.

$$\alpha = a^2 H/2\pi(\hbar c/e).$$

Notice that α is dimensionless, being the ratio of flux through a lattice cell to one flux quantum. (The author is indebted to Professor F. Bloch for pointing out that this parameter can be interpreted as the ratio of two characteristic periods of this problem: one is the period of the motion of an electron in a state with crystal momentum $2\pi\hbar/a$, which is $a^2 m/2\pi\hbar$; the other is the reciprocal of the cyclotron frequency eH/mc.) A value of $\alpha = 1$ implies an enormous magnetic field (on the order of a billion gauss), if the lattice spacing is typical of real crystals (on the order of 2 Å). Despite this, we are going to be interested in the results for such values of α (for a treatment of smaller values of α in this same equation, see Ref. 8.)

With all these substitutions, our Schrödinger equation turns into a one-dimensional difference equation:

$$g(m+1) + g(m-1) + 2\cos(2\pi m\alpha - \nu)g(m) = \epsilon g(m). \tag{1}$$

This equation is sometimes called "Harper's" equation, and has been studied by a number of authors.[8-11]

III. CALCULATION OF THE SPECTRUM AND THE RATIONALITY CONDITION

Another way of writing Eq. (1) is

$$\begin{pmatrix} g(m+1) \\ g(m) \end{pmatrix} = \begin{pmatrix} \epsilon - 2\cos(2\pi m\alpha - \nu) & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} g(m) \\ g(m-1) \end{pmatrix}.$$

The 2×2 matrix is called "$A(m)$." When a product of m successive A matrices is multiplied with the two vector $\langle g(1), g(0)\rangle$, the result is the two vector $\langle g(m+1), g(m)\rangle$. The physical condition which must be imposed on the wave function (i.e., the function g) is boundedness, for all m. This translates into a condition on the products of successive A matrices. Now if the A matrices are periodic in m (which they may very well be, since m enters only under a cosine), then long products of A matrices consist essentially in repetitions of one block of A matrices, whose length is the period in m. Let us assume that the A matrices are indeed periodic in m, with period q. This is a requirement on α, namely that there should exist an integer p such that

$$2\pi\alpha(m+q) - \nu = 2\pi\alpha m - \nu + 2\pi p.$$

Algebra reveals the fact that this condition on α is precisely that of rationality[1]:

$$\alpha = p/q.$$

We now proceed, making full use of this somewhat bizarre ansatz. (Presently, we will consider the case when α is irrational.) The product of q successive A matrices will be called "Q." The condition of physicality is now transferred from the g's to the matrix Q. It can be shown without trouble that the correct condition on Q is that its two eigenvalues be of unit magntiude. That condition can then be shown to be equivalent to requiring its trace to be less than or equal to 2, in absolute value. Hence, a concise test for the boundedness of the g's is the following:

$$|\mathrm{Tr}Q(\epsilon;\nu)| \leq 2.$$

Trace conditions of this type have been found by other authors.[2,12] This one was discovered by Professor G. Obermair, and extensively used by the author. Now it can be shown that the only way that ν affects the value of TrQ is additively, i.e., that as ν changes, the shape of the graph of TrQ, plotted against ϵ, is unchanged — it merely moves as a whole, up and down. (A proof of essentially this fact can be found in Ref. 2.) Therefore $\mathrm{Tr}Q(\epsilon;\nu) = \mathrm{Tr}Q(\epsilon) + 2f(\nu)$, where $f(\nu)$ is a periodic function of unit amplitude, and $Q(\epsilon)$ is defined as $Q(\epsilon; 1/2q)$. We are interested in all values ϵ which, for some ν, yield bounded g's. (Such values will be called "eigenvalues" of the difference equation.) Therefore, we want to form the union of all eigenvalues ϵ, as ν varies. Since $2f(\nu)$ ranges between $+2$ and -2, the condition on the trace can be rewritten as follows:

$$|\mathrm{Tr}Q(\epsilon)| \leq 4.$$

The trace of Q is always a polynomial of degree

q; hence one might expect the above condition to be satisfied in roughly q distinct regions of the ϵ axis (one region centered on each root). This is indeed the case, and is the basis for a very striking (and at first disturbing) fact about this problem: when $\alpha = p/q$, the Bloch band always breaks up into precisely q distinct energy bands. Since small variations in the magnitude of α can produce enormous fluctuations in the value of the denominator q, one is apparently faced with an unacceptable physical prediction. However, nature is ingenious enough to find a way out of this apparent anomaly. Before we go into the resolution, however, let us mention certain facts about the spectrum belonging to any value of α. Most can be proven trivially: (i) Spectrum(α) and spectrum ($\alpha + N$) are identical. (ii) Spectrum(α) and spectrum($-\alpha$) are identical. (iii) ϵ belongs to spectrum(α) if and only if $-\epsilon$ belongs to spectrum(α). (iv) If ϵ belongs to spectrum (α) for any α, then $-4 \leq \epsilon \leq +4$. The last property is a little subtler than the previous three; it can be proven in different ways. One proof has been published.[13]

From properties (i) and (iv), it follows that a graph of the spectrum need only include values of ϵ between $+4$ and -4, and values of α in any unit interval. We shall look at the interval $[0, 1]$. Furthermore, as a consequence of properties, the graph inside the above-defined rectangular region must have two axes of reflection, namely the horizontal line $\alpha = \frac{1}{2}$, and the vertical line $\epsilon = 0$. A plot of spectrum(α), with α along the vertical axis, appears in Fig. 1. (Only rational values of α with denominator less than 50 are shown.)

IV. RECURSIVE STRUCTURE OF THE GRAPH

This graph has some very unusual properties. The large gaps form a very striking pattern somewhat resembling a butterfly; perhaps equally striking are the delicacy and beauty of the fine-grained structure. These are due to a very intricate scheme, by which bands cluster into groups, which themselves may cluster into larger groups, and so on. The exact rules of formation of these hierarchically organized clustering patterns (Π's) are what we now wish to cover. Our description of Π's will be based on three statements, each of which describes some aspect of the structure of the graph. All of these statements are based on extremely close examination of the numerical data, and are to be taken as "empirically proven" theorems of mathematics. It would be preferable to have a rigorous proof but that has so far eluded capture. Before we present the three statements, let us first adopt some nomenclature. A "unit cell" is any portion of the graph located between successive integers N and $N+1$—in fact we will call that unit cell the Nth unit cell. Every unit cell has a "local variable" β, which runs from 0 to 1; in particular, β is defined to be the fractional part of α, usually denoted as $\{\alpha\}$. At $\beta = 0$ and $\beta = 1$, there is one band which stretches across the full width of the cell, separating it from its upper and lower neighbors; this band is therefore called a "cell wall." It turns out that certain rational values of β play a very important role in the description of the structure of a unit cell; these are the "pure cases"

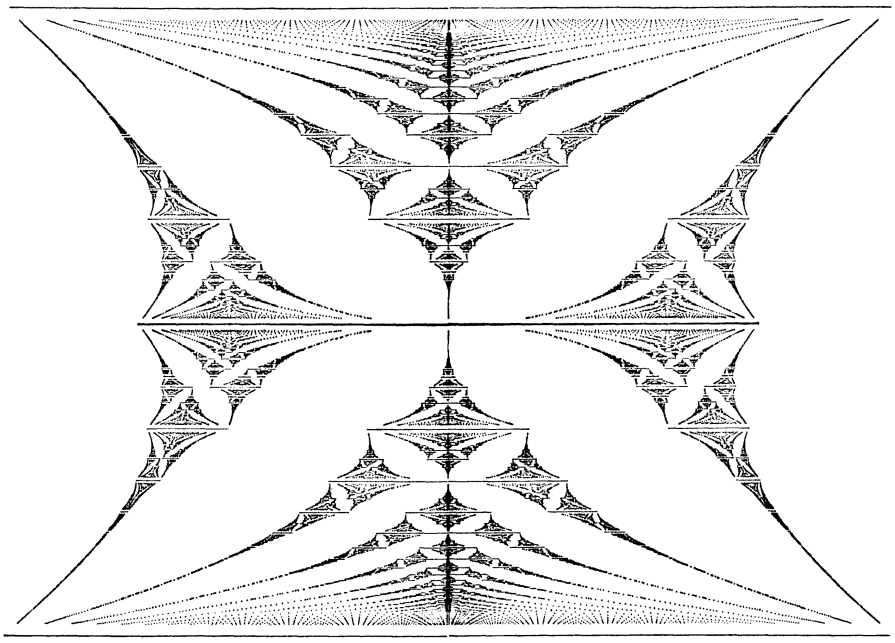


FIG. 1. Spectrum inside a unit cell. ϵ is the horizontal variable, ranging between $+4$ and -4, and $\beta = \{\alpha\}$ is the vertical variable, ranging from 0 to 1.

$1/N$ and $1-1/N$ $(N \geq 2)$;

and the "special cases"

$N/(2N+1)$ and $(N+1)/(2N+1)$ $(N \geq 2)$

(of the special cases, those with numerator N are the "lower" special cases, and those with numerator $N+1$ are the "upper"). The spectra belonging to these rational values form a "skeleton" on which the rest of the graph is hung. Figure 2 shows that skeleton; in it are shown the bands belonging to pure cases (up to $N=37$); in addition, one out of the $2N+1$ bands per special case is included, the centermost (i.e., the $N+1st$, counting from either end). The rest of the graph can be built up from this skeleton by a recursive process. Roughly, that process amounts to compressing the skeleton down to a small fraction of its size, distorting its vertical and horizontal scale in the process, and inserting this shrunken skeleton in between neighboring "ribs" of the large skeleton. When appropriately shrunken skeletons have been inserted between each pair of ribs, then the process is reiterated on the next level down; and this must continue indefinitely. Our goal is to turn this picturesque description into a precise description, and then to extract physical consequences from this weird structure. For this, we need the three statements:

Statement I. At the height inside any cell where its local variable equals a pure case $1/N$ or $1-1/N$, there are N bands between the left and right borders of the cell. (In unit cells, when N is even, there seem to be only $N-1$ bands, because the two centermost bands touch in the middle, where $\epsilon = 0$.) As N goes to infinity, the ratio of band size to gap size goes to zero (in other words, the bands become negligibly thin, compared to the gaps). Furthermore, the pure-case bands are distributed in such a way that the entire length of each cell roof and cell floor is approached, in the limit that N goes to infinity. Moreover, the number of pure-case bands per unit energy interval is a slowly varying and roughly constant function; that is, there is no clustering of the bands belonging to a pure case.

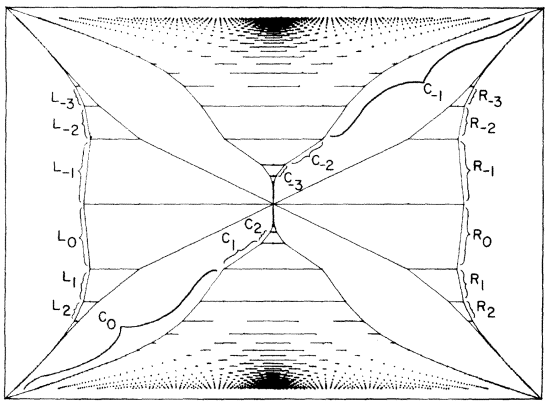


FIG. 2. Unit cell, shown with a "skeleton": the spectra belonging to pure cases $\beta = 1/N$ and $\beta = 1-1/N$, as well as the center band belonging to special cases $\beta = N/(2N+1)$ and $\beta = (N+1)/(2N+1)$. The L chain, C chain, and R chain are shown, all consisting of subcells formed by joining bands in the "skeleton" by straight-line segments. The labeling scheme for subcells in the three chains is indicated.

At heights where the local variable equals a special case, there is a set of bands, of which only the centermost is of interest here. The width of these center bands approaches zero as N goes to infinity. When upper special cases are considered, these bands approach a limit point, which is the inner edge of one of the two bands at $\frac{1}{2}$; when lower special cases are considered, the limit-point is the inner edge of the other band at $\frac{1}{2}$.

The next two statements involve the concept of "subcells," which are at the core of the recursive description of the graph's structure; but the concept of subcells can only be defined after the "skeleton" has been introduced (statement I). This is the reason that the following definition has been sandwiched between statements. It is best understood with the help of Fig. 2.

Rules for Subcell-Construction. The L and R subcells of any cell are formed as follows: Connect the edges of the outermost bands of neighboring pure cases by straight lines. The trapezoidal boxes thus created form the "L chain" and the "R chain" (on the left- and right-hand sides of the cell).

The C subcells of any cell are formed as follows: Connect the outer edges of the next-to-outermost bands of pure cases with $N>2$ by straight lines. This will produce two large boxes whose sides are unions of infinitely many straight-line segments. The remaining C subcells are formed by joining the centermost bands of neighboring special cases by straight lines. All the C subcells taken together form the "C chain." Each subcell has a unique label; the labeling scheme is shown in Fig. 2. We now affirm the existence of large empty swaths crossing the graph.

Statement II. The regions of a cell outside its subcells are gaps (contain no bands or portions of bands).

Finally, statement III contains the essence of the recursive nature of this graph.

Statement III. Each subcell of any cell can be given its own local variable, defined in terms of the local variable of the parent cell. (See below.) Each subcell, when indexed by its own local vari-

able, is a cell in its own right, in that it satisfies statements I, II, and III.

The subcell's local variable is defined as follows: Let β be the "outer" local variable (i.e., that of the parent cell), and β' be the "inner" local variable. Assume first that $\beta \leq \frac{1}{2}$. Then let N be defined by

$$N = [1/\beta] .$$

(Note: The notation $[x]$ stands for the greatest integer less than or equal to x; it follows that N is the denominator of the pure case just above β.)

If the subcell is of L type or R type, then the equation relating β and β' is

$$\beta = (N + \beta')^{-1} .$$

(Note how this forces β' to lie between 0 and 1.) Let us denote the function of β which yields this value of β' by "$\Lambda(\beta)$."

If the subcell is of C type, then the relation between inner and outer local variables is

$$\beta = (2 + 1/\alpha')^{-1} ,$$

$$\beta' = \{\alpha'\} \quad \text{(fractional part of } \alpha') .$$

Let us denote this function of β by "$\Gamma(\beta)$."

Finally, if β is between $\frac{1}{2}$ and 1, then β' is equal to the value of β' which belongs to $1 - \beta$.

The statements are a little startling; they need evidence. In Figs. 3 and 4 are plotted two "rectangularized" subcells of a unit cell, namely L_2 and C_2. A "rectangularized" cell is made from the cell itself by a family of one-dimensional linear transformations. There is a linear stretching at each height, which makes the effective width of the cell be the same at every height (like a unit cell); and the bands as stretched in that way are then plotted using the cell's own local variable, rather than that of its parent cell, as the vertical axis. The characteristic butterfly pattern of the large gaps is very obvious in the rectangularized

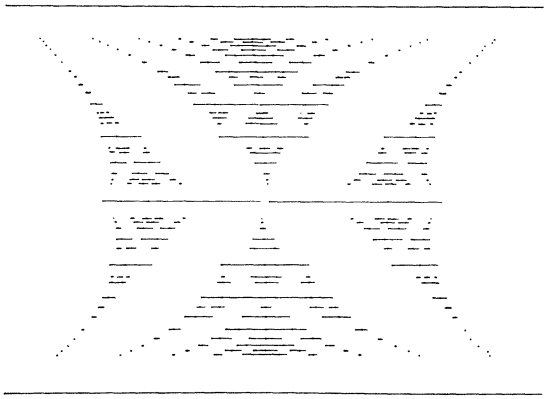


FIG. 3. Rectangularization of L_2.

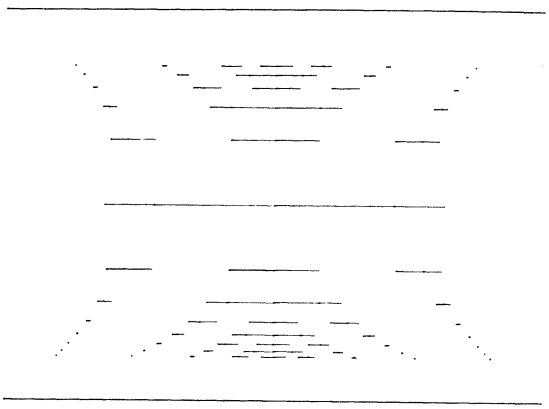


FIG. 4. Rectangularization of C_2. The number of bands calculated was much smaller, which explains why so little detail is visible. All the bands shown belong to the pure-case part of the skeleton of this subcell. (Compare Fig. 2.)

graphs. Note, however, that pure cases with even denominators inside the L cell do not possess the "degeneracy" property (of having two bands which "kiss" at the center).

The recursive structure as here presented confirms in the main (but differs in detail with) the important but extremely difficult article by Azbel',[10] which states that the spectrum is entirely determined by the continued fraction of α. The connection is through the Λ function. If the local variable function Λ is iterated, one obtains the following representation for β:

$$\beta = \cfrac{1}{N_1 + \cfrac{1}{N_2 + \cfrac{1}{N_3 + \ddots}}} ,$$

which is unique, and will terminate for any rational α. Azbel' predicts that spectrum(α) will consist of N_1 bands, each of which breaks up into N_2 subbands, each of which breaks up into N_3 subbands, and so on. This is approximately the same as our result, when all of the N's are large. Our prediction is that the L and R cells will each contain N_1 bands, but the number inside the C cell is not given by this expansion. As the nesting continues, N_2 subbands are indeed found in the L and R subcells of each of the L and R cells, but in the C subcells, once again there is no simple prediction based on the continued-fraction expansion. Qualitatively, though, Azbel''s prediction contains the essence of the structure, and is very intuitively appealing.

From this recursive breakdown of the graph there

follow a number of theorems, most of which involve somewhat tedious topological reasoning (the proofs in complete detail are worked out in the author's thesis[14]). First of all it is important to be able to pinpoint any particular cell, no matter how deeply it is nested inside other cells. A simple notation will do this for us: the outermost cell is written first, followed by successively shrinking cells inside it .... For example, "$U_7L_{-2}C_0R_3L_1$" stands for a cell-in-a-cell-in-a-cell-in-a-cell-in-a-cell. The subscripts are to be interpreted as shown in Fig. 2. ("U_N" stands for the unit cell where $[\alpha]=N$. However, the notation for the unit cell is usually omitted, since all unit cells are identical.)

A result which is quite difficult to establish is the simple fact that all cells are (nearly) homeomorphic to each other. (Homeomorphisms are the topological version of isomorphisms: a homeomorphism is a one-to-one continuous mapping between two manifolds whose inverse is also continuous.) The "nearly" has to be included since there is a feature which could not be preserved under a continuous mapping, and that is the "degeneracy" at rationals with even denominators which exists in unit cells, but not in L or R cells. This means that there is a "branch cut" across which the homeomorphism does not carry. To be precise, each cell can be cut into two pieces — a left and a right half. For unit cells, the dividing line is merely the vertical line at $\epsilon=0$; for other cells, the dividing line can be defined in terms of the center bands of rationals with odd denominators. The left and right halves of any cell, as determined by its dividing line, are homeomorphic to each other and to the halves of every other cell as well. However, the homeomorphism can only be extended over the line in case both cells are of the same type, in the sense that they share the property of degeneracy, or share the property of its absence.

V. HOW THE BANDS ARE CLUSTERED

We now can make a precise definition of the cluster patterns. Suppose we wish to describe the distribution of bands at the value $\alpha=p/q$. Let $\beta=\{\alpha\}$, so that β is the local variable for the unit cell to which α belongs. The recursive decomposition tells us that the spectrum at β consists of three parts, which must be separated by gaps: one inside an L subcell, one inside a C subcell, and one inside an R subcell. Furthermore, the L and R subcells contain bands at that height with a Π belonging to $\Lambda(\beta)$, and the C subcell contains bands at the height with a Π belonging to $\Gamma(\beta)$. In other words, the Π at α consists of three Π's, from right to left, belonging to $\Lambda(\beta)$, $\Gamma(\beta)$, $\Lambda(\beta)$, respectively. Let us take the example of the value $\alpha=\frac{5}{17}$. Its spectrum is shown below:

--•-- -- --- -- --•--

A suggestive symbolic representation for the cluster pattern is

(2-1-2)-(2-3-2)-(2-1-2).

The five bands on either side are located inside the L and R chains; the central seven are located inside the C chain. The reason the breakdown is 5-7-5 is explained recursively as follows:

For the L and R subcells, the local variable is given by

$$\tfrac{5}{17}=(N+\beta')^{-1}=(3+\tfrac{2}{5})^{-1},$$

so that $\beta'=\Lambda(\beta)=\frac{2}{5}$. The demoninator is 5, hence we expect to see 5 bands inside L_1 and R_1.

For the C subcell, the local variable is given by

$$\tfrac{5}{17}=(2+1/\alpha')^{-1}=(2+1/\tfrac{5}{7})^{-1},$$

which yields

$$\beta'=\{\alpha'\}=\{\tfrac{5}{7}\}=\tfrac{5}{7}.$$

The analysis then "predicts" that the spectrum at $\alpha=\frac{5}{17}$ will consist of a set of five bands belonging to the local variable $\frac{2}{5}$; then a gap; then a set of seven bands belonging to the local variable $\frac{5}{7}$; then another gap; then another set of five bands belonging to the local variable $\frac{2}{5}$. But the analysis can be carried further, because the very same operations can be carried out inside the subcells, starting with their local variables and deriving local variables which are even more local. For $\frac{2}{5}$ and $\frac{5}{7}$, this gives

$$\Pi(\tfrac{2}{5})=\Pi(\tfrac{1}{2})\Pi(0)\Pi(\tfrac{1}{2}),$$

$$\Pi(\tfrac{5}{7})=\Pi(\tfrac{1}{2})\Pi(\tfrac{1}{3})\Pi(\tfrac{1}{2}).$$

It is useful to adopt the notation "N" as shorthand for "$\Pi(1/N)$," because, according to statement I, the bands belonging to $1/N$ are smoothly spread out across the cell to which they belong, with no clustering. And $\Pi(0)$ is denoted "1" because at $\beta=0$ there is only one band. With this shorthand, then, we can write

$$\Pi(\tfrac{2}{5})=2\text{-}1\text{-}2,$$

$$\Pi(\tfrac{5}{7})=2\text{-}3\text{-}2.$$

And these Π's can then be stuffed back into the original Π for $\frac{5}{17}$:

$$\Pi(\tfrac{5}{17})=(2\text{-}1\text{-}2)\text{-}(2\text{-}3\text{-}2)\text{-}(2\text{-}1\text{-}2).$$

This coincides with what our eye told us. There is a guarantee that this recursive analysis of Π's will come to an end, because the two operations which

produce new local variables always reduce the numerator or denominator of the input fraction. In the end, one must eventually wind up with pure cases, or zero. The number of levels which one must descend before this process terminates is, however, rather difficult to predict.

VI. SPECTRA BELONGING TO IRRATIONAL FIELDS

The only case in which it is easy to predict what will happen is if you begin with an irrational value of α. In that case, the two operations yield new irrational values, which in turn yield irrational values, etc., ad infinitum. This leads to the very interesting question, "What is left — if anything — in the spectrum of an irrational field, according to this process?" Readers who are familiar with the pathology of point sets may already be anticipating the answer: there is indeed something left, and it is homeomorphic to the Cantor set. (The Cantor set is an uncountable yet measure-zero set of reals in an interval; see Ref. 15 for a detailed exposition of its fundamental properties.)

To demonstrate this starting from the three statements, one looks at the sequences of nested cells which are created by the repeated recursion in statement (iii). That is, given the original irrational α, one knows that its spectrum is confined to some particular unit cell. Statement (iii) says that the confinement can be further specified, as being inside three particular subcells of that cell. Reapplication of statement (iii) creates more deeply nested confining cells; for rational α the process terminates, but for irrational α the end product is uncountably many different infinite sequences of nested cells. It is a well-known theorem of topology that any nested sequence of closed intervals whose lengths tend to zero contains a unique limit point; its two-dimensional generalization to closed sets whose maximum dimension shrinks to zero is immediate, and that theorem is what tells us that the spectrum belonging to any irrational value of α consists of uncountably many points, between any pair of which there is a finite gap. Rigorous topological analysis establishes that the spectrum is indeed homeomorphic to the Cantor set.

It is legitimate to question whether these values of ϵ are actually eigenvalues of the difference equation, i.e., whether, in fact, the wave function $g(m)$ does remain bounded as m goes to infinity. Numerical work suggests that the answer is yes: such values really are the eigenvalues. It would be highly interesting to see a rigorous proof of this fact, or a refutation. Until proven wrong, however, we shall adopt this construction via recursion as the definition of the spectrum belonging to an irrational field value.

VII. MAGNETIC FIELD FLUCTUATIONS CREATE A "BLURRED GRAPH"

When this is done, we have finally achieved an important result: we have found a spectrum for every single value of α. Now the crucial question is, "How physical is this spectrum?" After all, it still remains true that the spectrum at a rational p/q consists of q bands, and q is still a highly fluctuating function of p/q. One can still feel suspicious of the graph. Despite the intellectual misgivings, though, the eye sees something rather continuous. There is something to this visual insight, and it can be stated formally in the following continuity theorem, which has been proven in the author's thesis: For any α, as α' approaches α, then all points of spectrum(α) are approached by points belonging to spectrum(α'); furthermore, only the points of spectrum(α) are so approached.

This theorem confirms the eye's assessment, that vertical motion along the graph is "continuous," in some sense; yet there is something discontinuous about vertical motion as well. Define $M(\alpha)$ to be the Lebesgue measure of spectrum(α). For all rational α, $M(\alpha)$ is positive, since every rational has bands of positive length. But for all irrational α, $M(\alpha)$ is zero. Therefore M has very peculiar behavior: at rational values, M is discontinuous, since there are irrationals arbitrarily near any rational; yet at irrational values, M is continuous. [The proof of this latter statement can be found in the author's thesis; it depends on a careful examination of how spectrum(α') is determined by sequences of nested cells, when α' is taken to be arbitrarily close to irrational values of α.] The function M is continuous at all irrationals, discontinuous at all rationals. This is a direct consequence of our recursive picture, and once again makes one wonder whether the graph is physically meaningful, or not.

Fortunately, there is a very simple resolution to this problem, consisting in the observation that every physical parameter has an experimental uncertainty in it, which smears it over some interval. Thus, the magnetic field, no matter how carefully controlled, has some fluctuations, which may be terribly small. This suggests the following concept: form a union of the spectra of all α within a "window" of height $\Delta\alpha$. This can be thought of as a blurred version of the graph, created by rapid up-down jiggling, where the amplitude of the jiggling is given by $\frac{1}{2}\Delta\alpha$. A blurred graph is shown in Fig. 5, using $\Delta\alpha = \frac{1}{100}$. As you can see, the result of the smearing-process yields a graph with a radically simplified appearance. It can be proven that the number of bands in any smeared graph is bounded by the constant $1/\Delta\alpha + 1$, for all α, and that the band edges change smoothly with

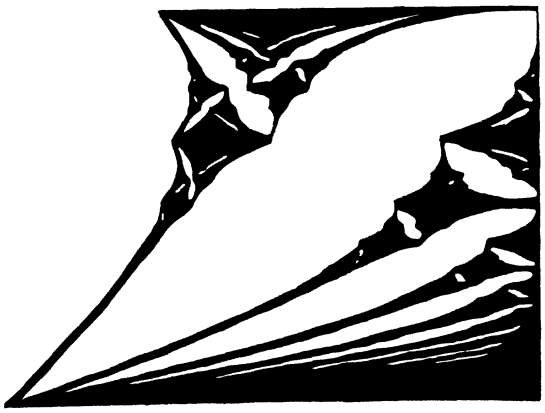


FIG. 5. One quadrant of the smeared graph created by using $\Delta\alpha=\frac{1}{100}$.

α. This establishes a totally continuous behavior for all magnetic field values, as a consequence of the imprecision of the field value. This also gets rid, of course, of the measure anomaly. As $\Delta\alpha$ approaches zero, the fine structure of the graph is bit by bit recovered; the infinitely fine-grained detail never returns (for positive $\Delta\alpha$), but more and more of it is revealed by decreasing the uncertainty $\Delta\alpha$. Of course, at the unphysical value $\Delta\alpha=0$, the entire graph returns.

VIII. CORRESPONDENCE WITH RAUH'S LANDAU-LEVEL APPROACH

One unexpected feature of the recursive nature of the graph is how it corroborates a picture set forth by Rauh concerning the broadening of Landau levels when a periodic potential is "turned on."[11,13] In Rauh's work, the simplest possible two-dimensionally periodic potential, $V(x,y)=2V_0(\cos kx+\cos ky)$, is chosen as a perturbing potential acting on an electron in an initially pure Landau state. The same difference equation arises, with a totally different interpretation: $g(m)$ represents the amplitude of a Landau state of fixed principal quantum number, whose center of localization along the axis of square integrability is ma/α (with α as we have defined it), and ϵ is proportional to the energy splitting. More interesting is the interpretation of α. Instead of measuring the flux in flux quanta, it measures the reciprocal of that quantity. Therefore, a large α in Rauh's equation means a small field. One can use the equations linking inner and outer variables to establish a link between Rauh's conclusions and our graph. This is done as follows. Observe the way the L chain turns into a very thin line as it approaches the Bloch band; it is so thin that it resembles a single level, split by a perturbation. Therefore, we choose to identify the leftmost band with a Landau level, perturbed by the periodic potential of the crystal. The splitting of the band is present in our picture, since in reality the line is composed of very thin L cells. At height β (assuming $\beta<\frac{1}{2}$), the structure inside each of those cells is given by spectrum (x), where

$$\beta=(N+x)^{-1}.$$

Here, N is integral, and x is between 0 and 1. Now by the first symmetry property of the graph,

$$\text{spectrum}\,(x)=\text{spectrum}(N+x)$$
$$=\text{spectrum}(1/\beta).$$

Notice that this says that the split-up of the lowest-lying Landau level is given by the same eigenvalue equation, but with parameter $1/\beta$ instead of β. This is completely consistent with Rauh's work. Moreover, one can identify other chains in the graph with Landau levels, and under this identification, it turns out that each one of them splits up in a pattern given by spectrum $(1/\beta)$. The natural candidate for the 2nd-lowest Landau level is the L chain located inside C_0; the 3rd lowest is the L chain inside C_0C_0, and so on. The number of such levels is essentially $1/\beta$; half of them are L chains inside nested C cells, and the other half are their symmetric counterparts: R chains inside nested C cells. To determine how any one of them is split, we must iterate the formation of local variables. In particular, to derive the splitting of the nth Landau level, we must begin with β, apply the Γ function $n-1$ times to it, and finish by taking Λ of the result. As before, let

$$\beta=(N+x)^{-1},$$

with N integral, and x between 0 and 1. Further, assume N is at least 4. Then by definition, $\Lambda(\beta)=x$. Simple calculation shows also that

$$\Gamma(\beta)=[(N-2)+x]^{-1}.$$

From this expression, we can directly read off $\Lambda(\Gamma(\beta))$: it is also x. Now if $N-2$ is also at least 4, then we can immediately get

$$\Gamma(\Gamma(\beta))=[(N-4)+x]^{-1},$$

and Λ of this is, once again, x. So it will go, with 2 being subtracted from the integer in the denominator over and over again, as long as that integer stays 4 or more. When β is small (i.e., when N is big), then the number of Γ's which can be iterated before the integer ceases to satisfy that condition is roughly $\frac{1}{2}N$. This implies that there are roughly $\frac{1}{2}N$ Landau levels to the left of center, and symmetrically, $\frac{1}{2}N$ to the right of center, all N of which are split according to the pattern of spectrum(x) — but as before, spectrum(x) and spectrum$(1/\beta)$ are identical. Therefore all the Landau levels do split in a similar way. And we have shown that their number is roughly N, which is to

say $1/\beta$. Therefore the separation between Landau levels is roughly $8\beta E_0$. This corresponds to the spacing which one can calculate using an effective-mass approximation at the edges of the band. Altogether, the Landau-level-based theory and the Bloch-band-based theory achieve in this way a satisfying harmony.

IX. WAVE FUNCTIONS AND IRRATIONAL FIELDS

In certain other approaches to this problem, notably those based on the magnetic translation group,[1,2] the rationality of α is forced if one seeks representations of the magnetic translation group by the Frobenius method, which involves finding an invariant subgroup. For some subgroup of magnetic translations to be invariant, all of its members must commute, and this in turn forces certain phase factors, involving the flux through the parallelogram defined by the two translations, to be unity. The end result is that one must choose a rational value p/q for α, and the invariant subgroup consists of "superlattice" translations, where the superlattice consists of lattice points separated from each other by q lattice spacings in both x and y directions. That way, the amount of flux is always an integer, the phase factors are always unity, and the subgroup of magnetic translations is indeed invariant. The problem with this whole approach is that such superlattices can only be defined in the case of rational fields, and there seems to be no obvious way to extend the results to irrational fields.

An alternative type of "superlattice" can be formulated, however, which comes up naturally in the context of our difference equation. One begins with the observation that there are solutions of the Bloch-Floquet type to the difference equation — that is, solutions with the property that

$$g(m+nP)=e^{inkP}g(m),$$

where n is any integer, P is a constant, and k is a wave number. One's first guess might well be that P must be an integer, corresponding to moving through an integral number of lattice spacings. This assumption is erroneous, however; P need not be integral. Indeed, the correct minimal period P is $1/\alpha$, which may be any real whatsoever, rational or irrational. This is proven in exactly the same way as for the Mathieu equation, of which our difference equation is, in some senses, a discrete counterpart.

The crucial fact in the proof is that the coefficients in the difference equation are themselves periodic in the variable m, with period $P=1/\alpha$. Therefore when m is replaced by $m+P$ in the difference equation, $g(m)$ becomes $g(m+P)$ but the coefficients are unaltered, which says that if $g(m)$ is a solution, then so is $g(m+P)$. Now there are two linearly independent solutions to a difference equation of second order (which ours is); let them be $g_1(m)$ and $g_2(m)$. Then $g_1(m+P)$ and $g_2(m+P)$ are also solutions; but since $g_1(m)$ and $g_2(m)$ form a basis, there must be numbers C_{ij} such that

$$\begin{pmatrix} g_1(m+P) \\ g_2(m+P) \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \begin{pmatrix} g_1(m) \\ g_2(m) \end{pmatrix}.$$

Now we can find a linear transformation which will diagonalize the 2×2 matrix; this transformation will mix g_1 and g_2 to produce new functions g_1' and g_2' with the property that

$$g_n'(m+P)=cg_n'(m) \quad (n=1,2),$$

where c is an eigenvalue of the C matrix. This proves the Bloch-Floquet theorem for the difference equation, with period $1/\alpha$. A corollary is that any solution $g(m)$ can be expressed as

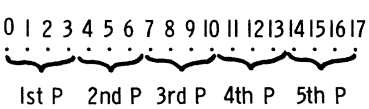
$$g(m)=e^{i\mu m}G(m),$$

where

$$e^{i\mu P}=c,$$

and where $G(m)$ is a periodic function of period P. When c is 1 (which happens, as in the Mathieu equation, at one edge of each band), then $g(m)=G(m)$, so that the difference equation has a purely periodic solution of period P. In any case, the distance $1/\alpha$ plays the role of a fundamental period associated with the difference equation.

The difference equation per se allows us only to determine $G(m)$ when m is an integer. But the periodicity of $G(m)$ allows us to interpolate between integers, and to determine G there also. This comes about because the period $P=1/\alpha$ is (in general) not an integer. Suppose, for instance, that $\alpha=\frac{5}{17}$. Then the period is of length $\frac{17}{5}$. Now $G(0)$, $G(1)$, $G(2)$, and $G(3)$ all fall within one period, but $G(4)$ is beyond $G(\frac{17}{5})$, and hence equals $G(\frac{3}{5})$. Similarly, $G(5)=G(\frac{9}{5})$, $G(6)=G(\frac{13}{5})$, and so on. Finally, $G(17)=G(0)$ and the whole cycle starts over again. Therefore, we can plot the values $G(0)$ through $G(17)$ inside one period of length P; they will appear in some rearranged order. The two orders and their relation are shown below. Integer order:



In the figure below, the five complete periods shown above are superimposed, to give the rearranged order (note that the scale of the two figures is different):

0 7 14 4 11 1 8 15 5 12 2 9 16 6 13 3 10 0

The sequence of integers in the rearranged order is the successive multiples of 7, taken modulo 17. This is because 7 occurs exactly $\frac{1}{17}$ beyond the period boundary in the upper figure, and $\frac{1}{17}$ is the minimum distance possible. The general rule for the rearranged order when $\alpha = p/q$ is to take the multiples of $\tilde{p}$ (modulo q), where $\tilde{p}$ is defined by the congruence

$$p\tilde{p} = 1 \quad (\text{modulo } q).$$

So far we have concentrated on what happens when α is rational; but the same process of folding back all values of $G(m)$ into one period of length P can be carried out. In the irrational case, however, the reordering will create a dense distribution of points inside the whole period. This is one place where irrational fields seem to make more physical sense than rational fields, in that one can determine the values of their wave functions on a dense set, rather than at just a discrete set.

If one takes a sequence of rational values α_n which approach an irrational value (and whose denominators therefore must go to infinity), the various periods $1/\alpha_n$ are all approximately the same, and it is therefore possible to compare the reordered wave functions of these rationals, to see if some trend emerges, pointing the way to the reordered wave function at the irrational field. One must also be sure to choose eigenvalues which are very close to each other; that this can be done is a consequence of the continuity theorem stated above. Such a process of comparison was carried out numerically for the following sequence of fractions (shown with their continued-fraction expansions) and their largest eigenvalues:

$$\alpha = \tfrac{1}{5} \quad (\epsilon = 2.9664),$$

$$\alpha = \tfrac{2}{11} = \cfrac{1}{5+\cfrac{1}{2}} \quad (\epsilon = 3.028\,50),$$

$$\alpha = \tfrac{17}{93} = \cfrac{1}{5+\cfrac{1}{2+\cfrac{1}{8}}} \quad (\epsilon = 3.023\,983\,268).$$

The wave functions in rearranged order are shown in Fig. 6. It appears that an overall shape is established by the fraction with a low denominator (in this case $\frac{1}{5}$), and details of the shape are determined by fractions with higher and higher denominators. Note how these fractions, which are close in value but which have very different denominators, have magnetic periods P of very nearly the same length, which allows the direct comparison of their wave functions. This figure is strong

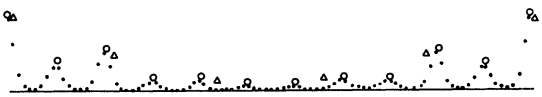


FIG. 6. Values of the wave function inside one magnetic period $P = 1/\alpha$, shown for three values of α (and their largest eigenvalues): triangles: $\alpha = \frac{1}{5}$ ($\epsilon = 2.9664$); circles: $\alpha = \frac{2}{11}$ ($\epsilon = 3.028\,50$); dots: $\alpha = \frac{17}{93}$ ($\epsilon = 3.023\,983\,268$). The x axis represents, in each case, a physical distance of $P (= 1/\alpha)$ lattice spacings. The vertical scale is such that the highest dot represents the value 1.

evidence for the idea that the limiting case — namely, the wave-function for an irrational α — is a continuous function which can be obtained from the discrete points supplied by the difference equation by translating them all into a single magnetic period of length $1/\alpha$.

In this connection, it is also interesting to point out that the one-dimensional "superlattice" of period $1/\alpha$ can be related to the magnetic translation group, in our model. It can be verified easily that all Landau-gauge magnetic translation operators[1]

$$T_{\text{mag}}(\vec{r}) = e^{i\vec{r}\cdot(\vec{p} + e\vec{A}'/c)/\hbar},$$

where

$$\vec{A}' = H(-y, 0, 0)$$

commutes with the effective Hamiltonian defined earlier. (This is not the case with the true Hamiltonian for a crystal electron in a magnetic field.) However, magnetic translation operators do not in general commute with other magnetic translation operators. The condition of commutation is that the parallelogram which they define should intercept an integral number of flux quanta. Now since the Landau gauge leads naturally to a one-dimensional mathematical treatment in which all the interesting phenomena happen along the x axis, it would seem natural to look for a commuting set of magnetic translation operators whose y spacing is "trivial" (i.e., is based on the lattice spacing), and whose x spacing contains information about the field. If we allow any magnetic translation in the y direction as long as it is through an integral number of lattice spacings, then the commutation condition quantizes the allowed magnetic translations along x; and the condition is precisely that they must be through $1/\alpha$ lattice spacings. The reason for this is that a rectangle of dimensions a/α along x, and a along y intercepts precisely one flux quantum.

This observation suggests that the best choice of unit cell for a "magnetic superlattice" may not be a square of q lattice spacings on a side (which only can be done for rational fields), but rather, a rectangle with one side equal to the lattice spacing, and the other side such that exactly one flux

quantum is intercepted. And that is the superlattice defined by the period P for the difference equation.

Somewhat related to these ideas is the article by Chambers,[12] in which orbits, "hyperorbits" (and so on) are discussed. In particular there is a discussion of the correspondence between simple orbits and bands, "hyperorbits" and subbands, and so on.

X. POSSIBLE EXPERIMENTAL TEST

Finally, I would like to comment on the possibility of looking for the features predicted by this model experimentally. At first glance, the idea seems totally out of the range of possibility, since a value of $\alpha = 1$ in a crystal with the rather generous lattice spacing of $a = 2$ Å demands a magnetic field of roughly 10^9 G. It has been suggested, however (by Lowndes among others), that one could manufacture a synthetic two-dimensional lattice of considerably greater spacing than that which characterizes real crystals. The technique involves applying an electric field across a field-effect transistor (without leads). The effect of such a field is to drive electrons (or holes) to one side of the device, where they will crowd together in a thin layer, essentially creating a two-dimensional gas of charged particles. Now if the device is prepared in advance with a dielectric layer which is nonuniform, and which in fact is periodic in each of its two dimensions, then the two-dimensional gas will be moving in a periodic potential that can be manufactured to fit any specifications. In particular, one could make a tight-binding model so that the electronic energy bands are approximately given by our simple sum of two cosines. Moreover — and this is the crux of the idea — one can choose the lattice spacing; thus with a spacing of 200 Å instead of 2, a magnetic field of 100 kG gives a value of α equal to 1. All that remains to be done is to apply a uniform magnetic field perpendicular to the plane of the gas, and to measure the transitions when the sample is irradiated with electromagnetic radiation of various wavelengths. This is not to say that the idea is easy; but such an intriguing spectrum deserves a good experimental test.

ACKNOWLEDGMENTS

Most of the ideas set forth in this article were worked out during the author's stay at the Fachbereich Physik of the Universität Regensburg (W. Germany). In particular, the author enjoyed the hospitality of the Lehrstuhl Obermair, and worked in collaboration with the computer "Rumpelstilzchen," as well as with Professor G. Obermair, Professor A. Rauh, and Professor G. Wannier. The latter was the author's thesis advisor, and contributed many valuable comments and ideas. R. Boeninger and F. Claro provided much help in various ways. Professor R. Donnelly of the University of Oregon generously allowed me to use a small computer without which the work would never have gotten done. I would also like to thank the following individuals for conversations and ideas: Professor F. Bloch, Professor P. Csonka, Professor M. Demianski, Professor R. Feynman, Professor D. Lowndes, Professor J. McClure, Professor M. Moravcsik, Professor A. Nagel, and Professor R. Wallis. Finally, I would like to thank Professor P. Suppes, Director of the Institute for Mathematical Studies in the Social Sciences at Stanford University, for the hospitality afforded me while I was writing this article.

---

*Work supported by the NSF under Grant No. GH 39027.

†Present address: Institute for Mathematical Studies in the Social Sciences (Ventura Hall), Stanford University, Stanford, Calif. 94305.

[1]E. Brown, Solid State Phys. 22, 3313 (1968).
[2]F. A. Butler and E. Brown, Phys. Rev. 166, 630 (1968).
[3]R. E. Peierls, Z. Phys. 80, 763 (1933).
[4]J. M. Luttinger, Phys. Rev. 84, 814 (1951).
[5]W. Kohn, Phys. Rev. 115, 1460 (1959).
[6]G. H. Wannier, Rev. Mod. Phys. 34, 645 (1962).
[7]E. I. Blount, Phys. Rev. 126, 1636 (1962).
[8]W. Y. Hsu and L. M. Falicov, Phys. Rev. B 13, 1595 (1976).
[9]P. G. Harper, Proc. Phys. Soc. Lond. A 68, 874 (1955).
[10]M. Ya. Azbel', Zh. Eksp. Teor. Fiz. 46, 939 (1964) [Sov. Phys.-JETP 19, 634 (1964)].
[11]A. Rauh, Phys. Status Solidi B 65, K131 (1974).
[12]W. G. Chambers, Phys. Rev. 140, A135 (1965).
[13]A. Rauh, Phys. Status Solidi B 69, K9 (1975).
[14]D. R. Hofstadter, Ph.D. thesis (University of Oregon, 1975) (unpublished).
[15]J. G. Hocking and G. S. Young, *Topology* (Addison-Wesley, Reading, Mass., 1961).