Integral quantum Hall effect for nonspecialists

D. R. Yennie

Laboratory of Nuclear Studies, Cornell University, Ithaca, New York 14853,
Institute for Theoretical Physics, University of California, Santa Barbara, California 93106
and Stanford Linear Accelerator Center, Stanford, California 94305

An attempt is made to develop a description of the multielectron quantum state responsible for the integral quantum Hall effect. One goal is to provide intuitive support for the very powerful and general argument of Laughlin that the theoretical relationship is insensitive to complicating details in the interior of the sample. The model the author uses is somewhat more realistic than heretofore in that it is three dimensional, does not ignore the atomic structure of the bulk matter, and does not use an effective-mass approximation. In order to treat the problem quantum mechanically, the complete system, including circuitry external to the system of interest, is replaced by a model closed system consisting of a finite number of electrons. In this model, states with a finite Hall current and voltage are *metastable* against decay caused by interactions outside the model, such as those with bulk matter excitations. Such states describe the true situation well only in the conductivity plateaus; between plateaus, there would be current flow between the Hall voltage probes corresponding to decaying states. Experimental constraints replace this transverse current by a voltage drop along the direction of current flow. The interactions between the electrons are expressed in terms of a self-consistent potential which gives an independent-particle description as a starting point, and residual interactions which are treated by perturbation theory. The self-consistent potential is found to be important in understanding the properties of the quantum state of the system, such as the existence of the plateaus in conductivity and how the electrons in the (effective) two-dimensional region come to equilibrium with the different Fermi levels in the voltage probes. To all finite orders of perturbation theory, the residual interactions are found not to alter the quantized Hall conductivity.

CONTENTS

I. INTRODUCTION

The quantum Hall effect is a remarkable phenomenon discovered experimentally in recent years (von Klitzing *et al.*, 1980) in which the Hall conductivity of a two-dimensional system of electrons is found to have plateaus as a function of variables which determine the number of electrons participating in the effect. Simple theory suggests that the Hall conductivity at these plateaus should be an integral multiple of e^2/h and the experiments agree with that prediction to within an accuracy of nearly 0.1 ppm; intercomparisons between different samples show even better agreement (Delahaye *et al.*, 1986). An application of great importance is to metrology; the quantum Hall effect promises a method of providing very precise resistance standards that are insensitive to the particular sample and the details of its fabrication. Belief in this insensitivity to the details of the sample requires either great confidence in the underlying theory or careful experimental verification. A recent review of the experimental situation in which further references may be found is given by Cage (1986). The discovery of the phenomenon also raised great interest among those working in the areas of fundamental constants and precision quantum electrodynamics (QED). The reason is that, if the theoretical formula can be justified to a precision of better than 0.1 ppm, the quantum Hall effect becomes a very powerful method of providing an accurate value of the fine-structure constant α ($=e^2/\hbar c \approx 1/137$) (since the speed of light is now an exactly defined quantity). A very elegant and general physical argument for the validity of the quantum Hall relationship has been given by Laughlin (1981); it relies only upon some general assumptions about the nature of the state of the system inside the sample. Thus the quantum Hall effect becomes an important ingredient in tests of quantum electrodynamics. The present status of such tests is reviewed briefly in Sec. VI. More recently, plateaus that are fractional multiples of the basic unit have been discovered (Tsui *et al.*, 1982b) and an explanation of them has been given (Laughlin, 1983).

The present paper is concerned only with the integral quantum Hall effect, not with the explanation of the fractional quantum Hall effect. We are interested particularly in the role of the self-consistent potential, which changes as external conditions change, and in the question of the exactness of the quantum Hall relation. It is intended to complement Laughlin's powerful argument by extending and generalizing earlier discussions about the properties of the eigenfunctions of the electrons that participate in the effect (Chalker, 1983; Trugman, 1983; Joynt and Prange, 1984). While we do not claim to improve upon the rigor of Laughlin's argument, we believe it is useful to explicate some of the underlying issues. For example, while Laughlin (1981) does not explicitly mention electron-electron interactions, it should become clear that his proof really does account for such complications, and more. Also, it is hoped that a more thorough understanding of the nature of the quantum state of the system may make it possible to study some of its more detailed properties such as the current distribution in the sample and the plateau width as a function of Hall current. It is even conceivable that this could lead to an understanding of small deviations from equilibrium which give rise to the ordinary resistance of a sample.

The present author has considerable experience in precision calculations in quantum electrodynamics but is an outsider to the field of condensed matter physics. He started out with skepticism about whether the physics of the quantum Hall effect were well enough understood that the results should be put on the same footing as other precision determinations of α, and the present work is an outgrowth of that skepticism. It attempts to develop a more complete microscopic picture of the quantum states responsible for the quantum Hall effect, the hope being that this will enable us to identify the physics that does ultimately limit the precision of the determination of α. It may also be regarded as an elaboration of the microscopic properties which are sufficient for the Laughlin argument to be valid. For example, it is demonstrated to all finite orders of perturbation theory that the electron-electron interactions modify the current and the electrochemical potential difference in the same proportion, so that the exactness of the relationship is not disturbed. This is an important conclusion, since that perturbation theory is not characterized by a particularly small parameter. On the other hand, the use of perturbation theory limits the discussion to the integral quantum Hall effect; and it is of course possible that even in the integer case there are nonperturbative complications. As a result of this effort, the author has developed high respect for the argument of Laughlin and a belief that the quantum Hall effect may indeed provide a competitive value for the fine-structure constant.

A. Semiclassical description

In Sec. II we start with an intuitive discussion of the integral quantum Hall effect. This is based on an understanding of those properties of the quantum state which can be inferred from simple semiclassical considerations. While this method cannot justify the exactness of the relation, it does help us understand why it is insensitive to the existence of impurities, geometrical effects, etc. It also helps in understanding the mechanisms involved in producing the large plateaus seen under certain conditions. Here we describe the role of the self-consistent potential in explaining such features of the quantum Hall effect. The need to understand the self-consistent potential arises because the external circuitry can act as an electron reservoir; and when the filling factor is changed, electrons will tend to move in or out of the "two-dimensional"[1] region to keep the available states filled up to the Fermi energy of the higher potential probe [(−) terminal]. The usual argument that a charge cannot build up in the layer because a strong electrostatic force develops to prevent it is not true in this circumstance. When the situation is analyzed self-consistently, a small charge imbalance can actually occur. The redistribution of charges produces changes of the electrostatic potential that are comparable in size to the Hall voltage. The resulting potential distribution in the sample should vary dramatically as a function of the filling factor. If the residual electron-electron interactions did not produce the fractional quantum Hall effect, the self-consistent potential could produce a plateau by itself without other mechanisms. More realistically, the change in the self-consistent potential probably enhances the conventional mechanism due to localized states. It is likely that the self-consistent potential causes a dramatic change in the current distribution across the sample as the filling factor is varied across the plateau.

An approximate semiclassical treatment of the two-dimensional wave functions is given in Appendix A.

[1] The quotation marks are used to indicate that the region carrying the current is essentially two dimensional because the electrons are typically in the lowest sub-band of excitations perpendicular to the layer. However, in the full quantum-mechanical treatment, the dynamics of the motion normal to the layer is not neglected.

B. Quantum-mechanical models

Having developed some intuitive understanding of the microscopic physics of the phenomenon, we turn in Sec. III to the development of an appropriate quantum-mechanical model. Any discussion of the quantum Hall effect must begin with an effective Hamiltonian which contains more or less of the physics. At this point, we discuss some of the problems and aims in formulating an appropriate model for describing the quantum Hall effect. We believe that it is desirable to incorporate as much of the basic physics as is feasible into this effective Hamiltonian. We hope to avoid the introduction of various unnecessary approximations which may make it impossible to uncover the effects that give the true limitations of the quantum Hall effect relationship. Some of the approximations that are customarily made cannot easily be used as a starting point for a more complete treatment in which corrections could (in principle) be systematically worked out. The starting point of many discussions of the quantum Hall effect is a simplified Hamiltonian in which the dynamics of one dimension has been frozen out so that the electrons move in a smoothed two-dimensional potential in which the effects of individual atoms in the device have disappeared. At the same time the mass of the electron has been replaced by an effective mass. While this can give a qualitatively correct picture of the single-particle eigenfunctions, it is hard to see how it could be made the basis for a rigorous discussion to the required accuracy. Indeed, it is the nature of the general discussions that they do not rely on such a description; and, as will be seen, there is no difficulty in eliminating these particular simplifications.

Even though the results are highly suggestive, the skeptical outsider finds the use of the effective-mass approximation to be a particularly dubious start to a high-precision analysis. It is presumably arrived at by first determining the energy of an electron as a function of wave number and then constructing an effective Hamiltonian using the standard gauge-invariant substitution. There are several obvious qualitative objections to this procedure. First of all, the dynamics of an electron in the presence of a strong magnetic field gives entirely different eigenfunctions from those without a magnetic field, so the extrapolation from one condition to the other may be nontrivial. Second, the underlying bulk matter has imperfections of various kinds, so that the effective parameters might be position dependent. Third, terms in the energy that are not quadratic in the wave number could conceivably modify the quantum Hall effect relationship at an observable level. Fourth, there seems to be no way to confirm that, at the level of precision to which we aspire, the existence of individual atoms in the device does not modify the result. Finally, in principle the use of the gauge-invariant substitution could omit some important physics of the interaction of an electron with the bulk matter in the presence of an external field. For example, a more complete effective Hamiltonian might contain terms that are describable only in terms of field strengths. (In a somewhat analogous situation in quantum electrodynamics—the treatment of the electron propagator by first calculating in free space and then using the gauge-invariant substitution—one would lose some important physical effects such as the Lamb shift and the anomalous moment of the electron.) If physics were lost at this stage, it is hard to see how it could be recovered later by some correction procedure, since the formalism no longer includes a complete description of the system.

It is our objective to treat the quantum mechanics of the system as completely as seems feasible. At the present time, we are primarily concerned with the problems associated with condensed matter physics. Ultimately, one should check that relativity and quantum electrodynamics do not introduce significant corrections; but for the moment, these exotic effects seem less serious than the ones at hand. (In fact, relativistic effects can be incorporated into the discussion without serious change.) More to the point, it does seem impractical and unnecessary to include everything in the system we study. For example, the physics of the external circuitry providing the current and measuring the voltages seems irrelevant to an understanding of the physics of the quantum Hall effect. More problematic is the question of the quantum dynamics of the bulk matter in which the electrons move.[2] Ideally, we should not ignore those dynamics. We believe that in principle it should be possible to carry through a more complete analysis; one is described briefly in Sec. V. It is probably desirable to do so in order to give an absolutely convincing demonstration that, for example, the electron charge appearing in the quantum Hall relationship is the actual physical charge of the electron as measured in free space, unmodified by dielectric effects. Here we take the point of view that if our purpose is to obtain a good understanding of the properties of the state of the electrons involved in the effect, it is not really necessary to study the dynamics of the bulk matter in detail. Furthermore, Laughlin's proof can be used to close the logical gap to produce a rigorous result. Accordingly, it is assumed that the main effects of the bulk matter are to provide an external potential in which the electrons move and to modify the effective interactions between electrons. The potentials are three dimensional, not two dimensional, and the external potential includes contributions from the individual atoms in the device. In our model Hamiltonian, coupling between the electrons and bulk matter excitations is ignored aside from effects which can be absorbed into the effective interactions between electrons. However, this coupling is often invoked as a mechanism that permits the system to relax to an equilibrium state.

[2] We use the term "bulk matter" for want of a better one to describe everything except the electrons involved directly in the process. This includes all nuclei and bound electrons in the sample and voltage probes.

C. Closed-system model

Although the actual experimental system is open, we wish to deal only with the finite number of electrons that are inside the sample at any one time. In place of the external circuitry associated with the current flowing through the sample, we take a sample of finite length and introduce appropriate boundary conditions which permit us to treat the current. The system is also open because electrons are free to flow between the sample and the voltage probes. We assume that the electrons at the edges of the sample actually penetrate into the probes and spend a portion of their time there. We believe that the probes play an important role as reservoirs of electrons, permitting the number of electrons in the current-carrying portion of the sample to change as conditions require. The electrons in donor states can also be regarded as being in reservoirs. In principle, all these electrons are included as part of the physical system; but ones that are permanently attached to the atoms are not.

In order to deal with the system quantum mechanically, we should fix the number of electrons (our assumed Hamiltonian conserves that number). However, it is certainly true that as external conditions (a gate voltage, for example) are modified, the actual number of electrons participating in the process will change. Therefore we imagine putting in the "right" number of electrons for each situation and then holding that number fixed. Our discussion seems to have no special sensitivity to the total number of electrons actually participating. Once the quantum states are well understood, it may be possible to develop a more appropriate statistical treatment, which is certainly essential if nonzero temperature effects are to be treated properly.

We intend to treat the realistic situation of finite Hall current and voltage (rather than the limit as these quantities go to zero). For the true open system, this situation is maintained by the external constraints; but in terms of the model closed system, it cannot correspond to the stable ground state of the system because the energy could be lowered by transporting an electron from the higher-potential probe [$(-)$ terminal] to the lower-potential probe [$(+)$ terminal]. In the absence of residual interactions between the electrons and interactions with the bulk matter excitations, the model has energy eigenstates that have finite current and voltage. When those interactions are taken into account, the original eigenstates may turn into decaying states. However, under plateau conditions, the decay of these states to states of lower current involves the rearrangement of a macroscopic number of electrons and/or a tunneling through a large distance. Thus it is plausible, and seems to be experimentally true, that they must be states with a very large lifetime. We refer to them as *metastable* states. If the picture to be described is basically correct, it would be interesting to find a way to estimate their lifetime, since it should be related to the longitudinal conductivity in the sample. Since we consider only the situation with finite Hall current, we have no obvious way of making contact with certain other approaches, for example, ones based on the Kubo formula. For reviews of such approaches, see the articles by Thouless (1986) for a discussion of topological methods and Pruisken (1986) for field theoretical methods.

The model is described in detail in Sec. III. The one-particle Hamiltonian, including a self-consistent potential contribution, is used to derive the quantum Hall relationship from the resulting one-particle eigenfunctions. Laughlin's argument in the language of this model is also presented there.

D. Electron-electron interactions

One of the important aspects of the problem is the role of the interactions between the electrons. Apparently, there is no exact theory of this interaction. It presumably starts out as the basic Coulomb interaction which is modified by dielectric effects and perhaps screening at long distances. It may also include contributions or modifications due to phonon interactions; however, residual dynamical effects due to phonons are not included in our basic theory. Here we do not find it necessary to treat the precise form of the electron-electron interaction. It need not even be translationally invariant. We simply assume that there is some effective interaction that is basically two body without any particular symmetry, but it will be obvious that our arguments permit also the inclusion of effective many-body interactions which may be present.

Our method is based on a self-consistent formalism in which we try first to include the gross effects of the interaction in an effective single-particle Hamiltonian and then to treat the residual effects by perturbation theory. Fortunately, it is possible to do this in a sufficiently general way that it is not necessary to do any real calculations—provided that one accepts the general behavior of the eigenfunctions which is expected on intuitive grounds. This procedure is based on the belief that the quantum states involved in the integral quantum Hall effect correspond to filling certain of the unperturbed one-particle states in the "two-dimensional" region and that such multiparticle states remain isolated in energy from states that are easily reached by the perturbation. With this approach, it turns out to be possible to relate the Hall current to the differences of the energies to remove electrons at either side of the sample where they are in equilibrium with the reservoirs. This difference of energy is given precisely by the electrochemical potential across the sample. The first-order definition of the self-consistent potential is given in Sec. IV, and the second-order contributions to the current and electrochemical potential are studied and shown not to disturb the quantum Hall relation. This is also extended to a general class of contributions. Appendix B contains a more complete discussion of perturbation theory, and it is shown that to all finite orders the integral quantum Hall relation is not affected. Our treatment gives no indication that the eigenstates are electrodynamically unstable, so we assume that decay takes place only by interactions with bulk matter excitations.

E. Miscellaneous

Section V contains a discussion of some refinements of the main treatment which seem helpful for a more detailed understanding of the physics of quantum Hall devices. The first of these has to do with a speculation on how the equilibrium is established between the electrons in the "two-dimensional" layer and the voltage probes which are serving as electron reservoirs. The main point is that, at the edge of the sample, the "two-dimensional" layer has available states with energies both above and below the Fermi level of the adjacent probe, so that the layer fills precisely to that level. The second point discussed is how two or more levels inside the layer can merge in energy to the Fermi level inside the probe. The third point discusses some possible physics of the breakdown of metastability. The fourth critiques briefly the arguments about an open versus closed system. The fifth returns to Laughlin's argument to show how it can account for some possible subtleties such as the screening of the electron's charge by the dielectric properties of the medium and even energy stored in mechanical stresses in the material. Finally, it is shown that nonminimal terms in the effective Hamiltonian, such as that due to alignment of electron spin, do not affect the general conclusions.

Section VI contains a critique and general appraisal. First, it gives some indication of how the ideas presented in this paper may be elaborated. Next, it discusses some of the physics that might possibly limit the accuracy of the interpretation of the quantum Hall effect. Unfortunately, I am at present unable to specify the theoretical limits of this interpretation. Finally, it incorporates a brief review of how the quantum Hall effect fits into the precision analysis of quantum electrodynamics.

The present article makes no pretense of being a general review. Rather, it attempts to present a coherent microscopic description of the integer quantum Hall effect. In arriving at this description, the author has drawn freely on the published literature, but he has not attempted a balanced presentation of all points of view. Some of the material is new. For example, it is an easy extension of some of the earlier work to free it from the restriction of a literal two-dimensional description and the use of an effective mass. Undoubtedly these points are well understood by the experts in the field, but to my knowledge they have not been explicitly presented. The role of the self-consistent potential is emphasized here. Another example is the analysis of the equilibrium between the electrons in the two-dimensional layer and those in the reservoir, including the merging of different levels at the reservoir. Generally, this topic seems to be ignored in the literature and is perhaps not too well understood. A new formal development is the perturbation-theory proof that the residual electron-electron interactions do not modify the quantum Hall relation.

Finally, let me mention some general reviews of the quantum Hall effect which the reader may find helpful. Halperin (1986) has written a *Scientific American* article that should be a useful introduction to the subject. In particular, it contains a description of the devices and a nicely illustrated discussion of how the localized state mechanism can account for the plateaus. A paper by von Klitzing (1986), based on his lecture on the occasion of the presentation of the 1985 Nobel Prize in Physics, includes an overview of the field. Some articles discuss various aspects at a technical level (Cage and Girvin, 1983a, 1983b; Halperin, 1983). The most comprehensive review presently available is a volume based on a seminar series at the University of Maryland (Prange and Girvin, 1986).

II. INTUITIVE DISCUSSION

The general experimental arrangement and nature of the experimental results are illustrated in Fig. 1. The electrons participating in the Hall effect are constrained to move in a thin layer about 50—100 Å thick which is nearly planar, and a strong magnetic field of order 10 T is applied perpendicular to the plane of the sample. The Hall current of order 10 μA flows in and out of the layer through the current source and current drain. The sample has voltage probes which can measure the Hall voltage of order 100 mV across the sample and the resistive voltage drop along the sample. The system is constrained so that there is no current flow across the sample between Hall voltage probes. The physical dimensions of the Hall sample are in mm, with the length (distance along current) normally several times the width. The energy to excite a higher Landau level (which corresponds to a larger cyclotron orbit) is typically of order 10 meV; the energy to excite a higher electron sub-band (excited electron state associated with motion parallel with the magnetic field) is somewhat larger. Of course there are great variations in these characteristics. For example, in some experiments the Hall voltage is much smaller, so that the energy required to move an electron across the sample is much smaller than the energies to excite higher levels.

A striking feature of the data is the occurrence of plateaus in the Hall resistance at values of h/e^2i as a function of the filling factor, which is defined below. Here i is an integer, but in the fractional quantum Hall effect it is replaced by a fraction with odd denominator. These plateaus occur at temperatures under about 4 K; accordingly, a typical value of kT is of order 0.25 meV. Within the plateaus, the longitudinal resistance drops to exceedingly small values. The plateaus get broader and more precisely flat as the temperature is reduced; at the same time, the longitudinal resistance becomes even smaller. It is a remarkable fact that these plateaus occur under such widely differing experimental conditions. It is one of our aims to develop an understanding of the underlying physics that produces this behavior. Where relevant, other features of the experimental phenomenology are described below as we proceed.

The quasi-two-dimensional layer in which the Hall effect takes place is an inversion layer at a semiconductor interface. There are two principal types of devices. One is known as a silicon MOSFET (metal-oxide-semiconductor field-effect transistor). In this device, the inversion layer is produced by applying a gate voltage (10—50 V) across a thin SiO_2 layer which is attached to a Si substrate. The system acts somewhat like a capacitor with the inversion layer being one of the plates. The carrier density in the layer is varied by changing the gate voltage. The other type is called a GaAs-$Al_xGa_{1-x}As$ heterostructure device. This type has a "fixed" carrier density, so it is the magnetic field that is varied to observe the quantum Hall effect. Actually, there are nearby donors permitting the carrier density to vary somewhat, so the carrier density is not absolutely fixed. We need not say much more about the detailed nature of these devices (if that were necessary, it would be hard to conceive how they could achieve such accuracy) except to remark that they need not be very pure. They can have impurities, disorder, geometrical irregularities, spatially varying layer thickness, etc., without disturbing the effect. In fact, it is generally accepted that impurities which produce localized states play an important role in making possible the plateau that is observed. Later on, I shall argue that the situation is actually a little more subtle in that there could probably be a plateau without impurities were it not for the fractional quantum Hall effect. The impurities probably inhibit that effect, making possible the observation of the integral effect.

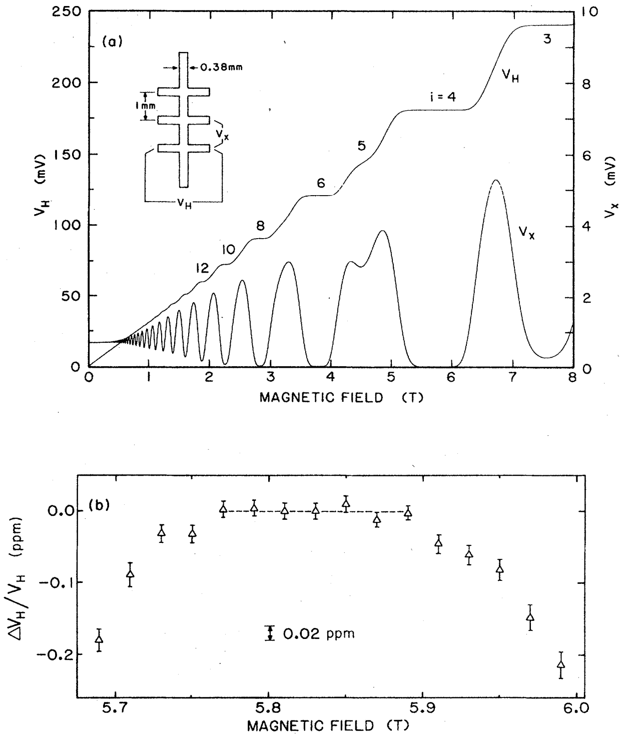

FIG. 1. Schematic of the physical arrangement and examples of results of quantum Hall measurements. (a) Example from Cage *et al.* (1985) showing results for a GaAs-AlGaAs heterostructure device cooled to 1.2 K. The inset shows the geometrical arrangement of the sample in the plane of the layer. The nominal number of electrons per cm^2 is 5.6×10^{11}. A strong magnetic field is applied perpendicular to the plane shown. The source-drain current I of 25.5 μA flows longitudinally (vertically in the figure) through the sample, and the Hall voltage V_H is measured across it. A voltage drop V_x along the direction of current flow is also measured. The plots show V_H and V_x vs B. The Hall resistance R_H is given by V_H/I. This type of plot is remarkable in that it extends from the classical Hall region, where R_H is linear in B, into the quantum Hall region where the plateaus are seen. The numbers on the plateaus correspond to the filling factors at their centers. (b) Example from Cage *et al.* (1985) showing the experimental precision presently attainable; data is for the 6453.20 Ω ($i=4$) plateau for the sample under the same conditions as in (a). At the present time, experimental determination of e^2/h is limited by the problems of comparing with a resistance standard rather than with the accuracy of measurement on the sample.

Actual samples have great variations in composition and geometry. One of the desirable properties of a sample to be used for precision measurements is that it have a small effective electron mass. This makes it possible to achieve the condition for quantization with a smaller magnet. Another is that it have a large zero-field mobility, for example 100 000 cm^2/V sec. The mobility is a measure of how easy it is for the electrons to move through the sample without suffering collisions. To see the fractional quantum Hall effect, still higher mobilities are required. The best precision is produced when the minimum value of the longitudinal resistance at the center of a plateau is as small as possible.

A. Semiclassical discussion of the integral quantum Hall effect

The present aim is to explore qualitatively some of the microscopic features of the integral quantum Hall effect. Therefore, for purposes of exposition, we start with a review of the semiclassical treatment, which is more intuitive and more compact than the quantum-mechanical treatment to be given later. A particularly nice discussion of these ideas, which we generalize slightly, is given by Trugman (1983). We assume that the temperature is low enough so that thermal excitations can be ignored except where explicitly mentioned. The discussion of this section centers on the behavior of the single-particle energy eigenfunctions in the self-consistent potential and ignores the effects of the residual interactions. As usual, we imagine that the electron dynamics normal to the thin layer is frozen into the lowest quantum state and we study the resulting two-dimensional motion.

What must be accepted from quantum mechanics is that the number of available states per unit area in the lowest Landau level is $1/2\pi l^2=eB/h$;[3] in typical situations, the order of magnitude of l is 100 Å. In the full quantum-mechanical treatment, each individual eigenfunction spreads over a much larger area, but the eigenfunctions overlap sufficiently to produce this average density of states. Now we try to understand the qualitative behavior of these one-particle states using semiclassical ideas. At this point, the reader may wish to refer to Appendix A for a brief treatment of the eigenfunctions of electrons in a two-dimensional potential; still more details can be found in the papers by Trugman (1983) and by Joynt and Prange (1984). However, these details are not essential for the immediate discussion.

The electrons move in a combined magnetic and electric field. The magnetic field is perpendicular to the plane of the electrons' motion. Our discussion concentrates on the lowest Landau level, and we ignore the existence of spin and of valley degeneracy (valley degeneracy is a feature of the crystalline structure of the bulk matter which produces a degeneracy of the lowest Landau level). When there is a current flowing through the sample, the magnetic force acting on the electrons tends to push them to one side until an electric field is created and a steady state is established. Semiclassically, the combined electric and magnetic fields then produce a cycloidal motion of the electrons. At different times in the cycle, the magnetic and electric forces do not precisely cancel. However, provided dissipative effects may be ignored, the electric field produced by this rearrangment adjusts itself until the total Lorentz force acting on an electron averages to zero.[4] That is,

$$\mathbf{E}+\langle\mathbf{v}\rangle\times\mathbf{B}=0\ , \tag{2.1}$$

where $\langle\mathbf{v}\rangle$ is the drift velocity of an electron. Since $\langle\mathbf{v}\rangle$ is constrained not to have a component parallel to $\mathbf{B}$, it is uniquely determined to be

$$\langle\mathbf{v}\rangle=\frac{\mathbf{E}\times\mathbf{B}}{\mathbf{B}^2}\ . \tag{2.2}$$

This is of course just the velocity of a Lorentz transformation which would reduce the electric field to zero at one position. However, since in general the electric field is not uniform, there does not exist any Lorentz transformation that can eliminate it everywhere.

Semiclassically, the electrons drift along the equipotential lines and have a cycloidal motion about them of radius l, also known as the magnetic length. Translated into quantum-mechanical terms, this means that the energy eigenfunctions are extended along the equipotentials, but have a distance across them of order l. We may speak of the equipotentials as "guiding" the eigenfunctions through the layer. In a region where the concept of equipotentials is meaningful—recall that we are projecting out one dimension of the physics and there may be complications from impurities, etc.—there are two types of eigenfunctions: those whose guiding potential extends from one end of the sample to the other in the direction of current flow and those which are trapped on contours that close upon themselves. The former are known as extended or current-carrying eigenfunctions and the latter as localized eigenfunctions. The central guiding equipotentials of localized eigenfunctions enclose (approximately) an integer number of flux units.

We separate the complete two-dimensional region into an extended-state region consisting of equipotentials that extend from one end of the sample to the other and a collection of localized-state regions consisting of equipotentials that close upon themselves. Wave functions associated with one of these regions can of course extend spatially into the other. In a specific model, Trugman (1983) discusses the fraction of the area occupied by the extended eigenfunctions. This vanishes for zero Hall voltage and increases with voltage; here we study only the situation with a finite Hall voltage and Hall current.

Next we wish to argue that there is considerable overlap of the extended eigenfunctions, which leads to the conclusion that fluctuations due to the individual properties of these eigenfunctions should be unimportant. The length of the guiding potential of an extended eigenfunction is at least L, the length of the sample in the direction of the current, so that the area covering the eigenfunction is of order lL. Since the area available to each one is $2\pi l^2$, the average transverse separation between them is of order $2\pi l^2/L$, which is clearly much smaller than the transverse extent of the individual eigenfunctions ($l/L\approx 10^{-5}$). Accordingly, the reasons that we are able to understand the quantum Hall effect using semiclassical ideas are that a very large number of eigenfunctions contribute to the current density at most points in the sample, and that the properties of each eigenfunction can be understood with semiclassical ideas (most of the time they are not in regions of such strong macroscopic electric

[3] This equality is strictly valid only for a uniform electric field. Even though it is not exact in general, that has no effect on the precision of the quantum Hall relation. Although this result should be derived by a quantum-mechanical treatment, it can be understood qualitatively with a simple uncertainty principle argument. A classical orbit of radius R corresponds to a particle whose momentum is eBR, while quantum mechanically its momentum must be of order $\hbar/R$. Balancing these to give the orbit of minimum radius/energy gives $R^2\approx\hbar/eB$, which corresponds to the above area per state. It is interesting to note that this area contains just one flux unit. (A flux unit is h/e; thus the number of flux units within an area A is eBA/h.)

[4] Dissipative effects might be represented by a viscous drag term proportional to $\langle\mathbf{v}\rangle$ on the right side of the following equation. In the absence of $\mathbf{B}$, this would produce the usual longitudinal resistivity. With $\mathbf{B}$, there is a combination of Hall and longitudinal resistivity. A significant longitudinal resistance is observed between plateaus, but within the plateaus, it drops to an exceedingly small value [see Fig. 1(b)]. The generally accepted explanation for this behavior is that within the plateaus all available one-particle states are filled and there is no possibility for electrons to scatter inelastically. There is further discussion of this point at the end of this subsection and in footnote 8.

fields that such approximations break down).

A portion of the sample showing some localized-state regions and an extended state is illustrated in Fig. 2. Writing the charge on the electron as $-e$, we see that the Hall current is

$$I=-e\int n(x,y)\langle\mathbf{v}\rangle\cdot(\hat{\mathbf{B}}\times d\mathbf{s})$$
$$=\frac{e}{B}\int n(x,y)\mathbf{E}\cdot d\mathbf{s}\ , \qquad (2.3)$$

where $n(x,y)$ is the density of carriers in the two-dimensional layer; it is limited by

$$n(x,y)\le n_0\equiv\frac{eB}{h}\ . \qquad (2.4)$$

The integral is carried out from one voltage probe to the other. Two examples of integration paths are shown in the figure. The potential energy of an electron along each path is shown in Fig. 3. The $\times$'s in Fig. 3(b) represent the potential at the position of localized states which lie on the path.

While any path is permissible, it is convenient to choose a path that does not pass through any regions of localized states or impurities where the semiclassical approximation is inadequate. If there is a situation in which all the extended states are precisely filled along this path, then

$$I=\frac{e^2}{h}\Delta V\ \text{(filled)}\ , \qquad (2.5)$$

where ΔV is the electric potential across the sample. (In the quantum-mechanical treatment to be given later, ΔV becomes the electrochemical potential difference between the two voltage probes, which is just the quantity measured by a voltmeter.) This is the quantum Hall relation.

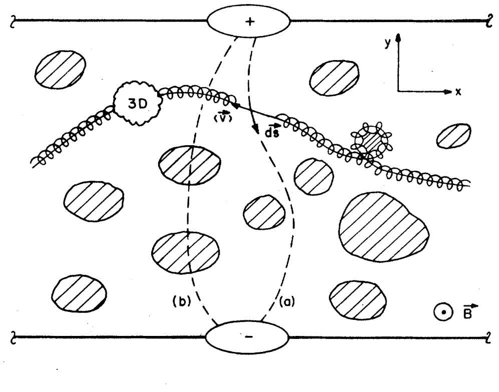

FIG. 2. A section of the sample showing the voltage probes (+ and −), regions of localized states (shaded), and an example of an extended eigenfunction following an equipotential. In this example, the extended eigenfunction encounters some obstacles such as an impurity (3D), and it splits at a saddle point in the equipotential to pass around both sides of a localized region. After encountering the obstacles, it returns to the same equipotential. Two integration paths between the probes are shown. Path (a) always lies in the extended-state region between obstacles.

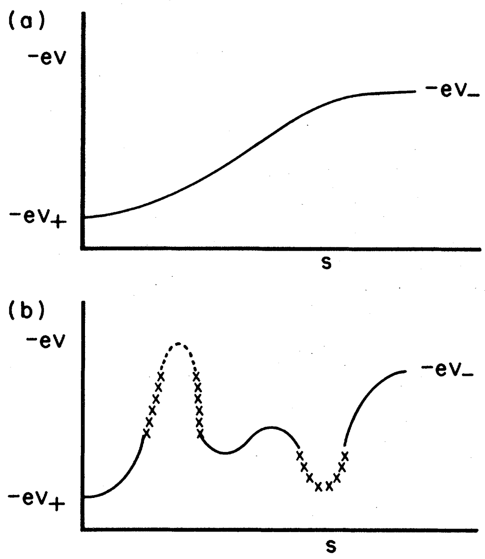

FIG. 3. The potential along the two paths shown in Fig. 1. In both cases, the solid lines lie in the extended-state region. For path (b), the $\times$'s represent a localized region where the states are filled, and the dotted line represents a region of vacant localized states.

No one would be convinced by this argument about the exactness of this relation; such conviction requires the powerful argument of Laughlin (1981), in which microscopic details are shown to be irrelevant provided they have certain general features. However, it does provide intuitive understanding. Notice that no assumption has been made about how the electric field varies across the sample; for example, it need not be uniform. Thus the question of whether the current is distributed over the whole surface or along the edges depends on the self-consistent potential which evolves in a given situation. Clearly the localized eigenfunctions do not give any contribution to the total current.[5] The fact that the existence of localized states does not disturb the quantum Hall relation was first pointed out by Prange (1981), using a quantum-mechanical treatment. The present discussion also indicates why geometrical effects, such as a hole through the sample, do not alter the result.

These considerations are also helpful in understanding why the longitudinal resistance becomes very small within the quantum Hall plateaus (footnote 4). To have longitudinal resistance, it is necessary that the conducting elec-

[5] As far as the discussion has gone, there is of course no difficulty if the integration path does pass through a localized region, since the limiting potentials on the two sides of the region are the same (Kazarinov and Luryi, 1982). However, we shall argue in the following subsection that it is desirable to avoid possible complications that may occur in these regions.

trons lose energy by inelastic scattering. But since all the one-particle states of lower energy are filled under those conditions, it is impossible to have such inelastic scattering and the longitudinal resistance should vanish. However, Rendell and Girvin (1984) argue that there must be places in the two-dimensional layer where the simple description without dissipation must be invalid. These occur near the current source and drain. Because the density of current-carrying electrons in these circuit elements is much larger than in the layer, the Hall voltage across them is much smaller. As far as the layer is concerned, they effectively short out the Hall voltage. Figure 4, taken from their paper, shows an example of equipotential lines calculated under the simplifying assumption that the longitudinal and transverse resistivities are independent of position. Note the resulting crowding of the equipotential lines in diagonally opposite corners of the sample. If the current were to follow the equipotentials strictly, the crowding of the current in the corners would probably lead to a condition that would violate the semiclassical assumptions we have been using so far (for example, in a high electric field the drift velocity might become comparable to the velocity of the cyclotron motion). One might expect that in the regions near the source and drain there would be more dissipation than in other regions and the current would cross equipotential lines.

The present author suggests a possible refinement of the discussion of Rendell and Girvin (1984), based partly on the following material in this section. In the vicinity of the Hall probes, the extended states may be completely filled, giving the ideal quantum Hall relation (2.5). But near the source and drain, where electrons are being removed and injected, there may exist vacancies and/or excitation to higher levels which give a much more complicated situation, causing the resistivity tensor to vary with position within the sample. How to go about analyzing this situation is not presently obvious to the author. It is probably important to do so because it may tell us something about the practical limitations of the accuracy of the quantum Hall device as a method of determining the fine-structure constant.

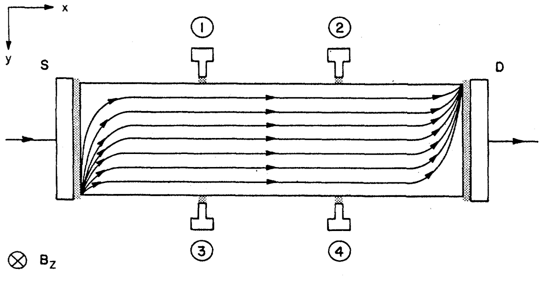

FIG. 4. Qualitative picture of the equipotential lines in the two-dimensional layer for a uniform sample in which σ_{xx} and σ_{xy} are constant. The current source and drain effectively short out the Hall voltage at the ends of the sample, causing a concentration of current in the corners. The numbers on the sides indicate the approximate potentials (in mV) when the source-drain current is 5 μA and the filling factor is 4 (from Rendell and Girvin, 1984).

B. Some properties of the eigenfunctions

The quantum-mechanical treatment is elaborated in more detail in subsequent sections. While there is no obvious direct connection between it and the formulation of Eq. (2.3), there are some general correspondences that are described here briefly. It turns out that the current carried by a one-particle state is, in a loose sense, given by the derivative of its energy with respect to an appropriately defined transverse coordinate. On the other hand, the discussion of the two-dimensional eigenfunction given in Appendix A shows that this energy is approximately $\frac{1}{2}\hbar\omega_c - eV(x,y)$, where ω_c is the cyclotron frequency and $V(x,y)$ is the value of the guiding electrostatic equipotential. Thus for a given eigenfunction, the current carried by an electron is proportional to a weighted average of the transverse electric field that the electron sees along its path. Notice that even though the electric field is not constant along the path, the normalization of the eigenfunction changes in an appropriate way to ensure that the current does not vary along the path. Provided all these one-particle states are occupied, the total current then turns out to be given in terms of the difference in the energies of electrons at opposite sides of the sample, giving back the result (2.5) with the interpretation of ΔV as the electrochemical potential difference.

One may now visualize situations in which the behavior of the eigenfunctions seems too complicated to support the previous conclusions. I want to argue that such apparent difficulties need not alter the results. For example, impurities seem to be outside the realm of a two-dimensional description. If we assume that the concept of a guiding potential is meaningful for some spatial regions, what happens when the actual eigenfunction gets into the region of a complicated three-dimensional impurity? Such a situation is illustrated in Fig. 2. There our semiclassical considerations, as well as the discussion of Appendix A, break down. One should in principle calculate the actual three-dimensional eigenfunction in such a region and match it somehow to the appropriate eigenfunctions in the two-dimensional formalism. However, the connection between the more correct three-dimensional formalism and its two-dimensional model appears to be quite obscure, to say the least. This particular situation presents no special difficulty in the three-dimensional quantum-mechanical discussion to be given later, and I mention it here partly to show the inadequacies of a two-dimensional treatment and partly to argue why it does not matter. If we accept that between impurities the eigenfunctions are primarily two dimensional in character (the dynamics perpendicular to the layer is frozen in its lowest quantum level), then the impurities become "black boxes" which somehow connect pieces of the two-dimensional eigenfunctions. Were it not for the magnetic field, we would expect an impurity to produce scattering, leading to undirected outgoing waves. However, because of the energy constraint, the eigenfunction actually continues along the same guiding potential after it leaves the region of the impurity (Prange, 1986). Furthermore, current conservation must be valid through an impurity. It is also clear that a huge number of extended eigenfunctions (of order 10^5, perhaps) are affected in a very similar way by any given impurity. Since we examine the current in the region between impurities, we expect that quantum fluctuations tend to average out and that therefore the semiclassical description remains valid there under these conditions. Of course, if the impurities become so dense that it is not possible to find an integration path which is never too close to an impurity (say within a magnetic length), then one might expect the description to become questionable. There ought to be some similar criterion in the quantum-mechanical treatment. Perhaps the condition for insignificance of the quantum fluctuations here is analogous to the validity criteria for the approximation of replacing a sum by an integral in the quantum-mechanical treatment.

Another possible situation that can occur without a breakdown of the two-dimensional description is one in which an extended equipotential crosses itself to produce a localized-state region, as illustrated in Fig. 2. The place where the equipotential crosses itself is a saddle point of the potential, and our semiclassical discussion must clearly break down there. Any eigenfunction spanning this particular equipotential obviously divides at the saddle point and part of the current goes on each side of the localized region. The important thing to observe is that, because of energy conservation, the current must continue along the same equipotential after the saddle point, and the discussion is very similar to that for an impurity. It would require a full quantum-mechanical treatment to study the eigenfunction within such complicated regions, but it is unnecessary to do so because the current carried by the eigenfunction is conserved and we can get the required information about the total current by considering only regions between impurities and saddle points. A similar discussion could be given for a hole through the sample (Tsui and Allen, 1981). In that case, a large number of eigenfunctions have the property of splitting and passing partially on each side of the hole. Plateaus have been produced in a geometry in which the current flows around both sides of a hole (Woltjer *et al.*, 1986). The purpose of that experiment was to examine certain qualitative aspects of the quantum Hall effect rather than to test the precision of Eq. (2.5) to great accuracy. In general, as long as we can analyze the current in a region not too close to the various types of obstacles, our semiclassical approximations can remain reasonable.

As just described, each of the eigenfunctions passing through the sample may actually be quite complicated because of various imperfections it encounters. As emphasized above, it seems important that it is possible to arrange the path of integration in Eq. (2.3) so as to avoid localized regions and impurities and pass entirely within regions of extended eigenfunctions, as illustrated by path (a) in Fig. 2. Trugman (1986) has considered various model potentials with a potential difference across the sample and found that such a path exists for them. If this were not possible, it would mean that there was a barrier between the two electrodes consisting of a region occupied only by localized states. In general, there is no reason to expect a quantized Hall relation in such a situation (unless there is a reason that the potential jump across such a region should vanish). So long as the extended-state region along which we choose the path of integration has all states filled, Eq. (2.5) results. Since they carry no current. it is immaterial for the present purpose whether the localized states are filled or not.

C. Role of the self-consistent potential

The filling factor ν is defined as the *average* carrier density divided by n_0:

$$\nu=\frac{\langle n \rangle}{n_0} . \tag{2.6}$$

For simplicity, we wish to deal only with one filled level. In particular, we ignore the complications of electron spin, which doubles the number of states available to the electrons,[6] or of possible valley degeneracy, which can double it again. Our following discussion therefore refers to the situation near $\nu=1$, but it can obviously be extended to other plateaus. It is given in terms of the GaAs heterostructure devices (which typically do not have valley degeneracy), but similar features are valid for the MOSFET.

From Eq. (2.3), we note that there is a way in which the extended states need not be completely filled, yet the quantum Hall relation may remain valid. If there is a region where the electric field is zero, it is not necessary for the states in that region to be filled. That is, if $n(x,y)=n_0$ wherever $\mathbf{E}\neq 0$, the integral is unaffected.

The remark of the preceding paragraph provides a clue as to the remarkable stability of the quantum Hall relation, even when conditions are changed that vary the number of electrons in the layer away from perfect filling of all the states. Starting with the situation for $\nu=1$, in which the layer is electrically neutral, the potential may have a fairly uniform rise from one probe to the other. Now suppose we increase the magnetic field so that the number of available states increases and ν decreases from one. If the spatial distribution of electrons could not change, the current given by Eq. (2.3) would decrease and Eq. (2.5) would turn into an inequality. To maintain the quantum Hall relation, it is necessary that electrons move

[6] Typically, the energy associated with electron spin is much smaller than that due to the cyclotron motion because of the suppression of the g_s factor in the solid together with the enhancement of the cyclotron frequency because of the small effective mass inside the solid.

around so as to keep the extended states filled in the regions where $\mathbf{E}\neq 0$. The aim of the following discussion is to persuade the reader that the most stable state of the system actually has this property. The argument may appear to be somewhat involved, but its essence is actually quite simple: There cannot be an empty extended one-particle state where $\mathbf{E}\neq 0$ because if there were it would be possible for an electron from a state of slightly higher energy to decay quickly into it by exciting the bulk matter.[7] The complication arises because when the occupation of the electron states is changed, the self-consistent potential also changes. In discussing the following ideas, the author has found that there exists a natural prejudice that the number of electrons in the layer cannot change from that giving a neutral charge distribution because of the strong effects of the Coulomb interaction.[8] It is argued below that this expectation is not correct under steady-state conditions when there is a potential across the sample.

A few possible mechanisms will be described. In all likelihood, they have a complicated interrelation; but they will at first be described as though each one acted individually. In the first, which is possible only for heterostructure devices, the needed electrons come from nearby donors outside the two-dimensional layer. An estimate of this effect has been made by Baraff and Tsui (1981), who studied the problem self-consistently in the z direction. This mechanism presumably has very little x,y dependence, but it establishes a new equilibrium between the donor states and the conducting electrons. It seems to be able to account for a few-percent change in B from the central value of the plateau; in any case, it does not apply to MOSFET's.

The generally accepted mechanism is that the plateaus are due to the localized states produced by impurities (Laughlin, 1981; Halperin, 1982). Rather than describe it here in the usual language, which seems to ignore the spatial distribution of the states and may not even apply to the situation with a finite Hall voltage, it is presented in a way that makes better contact with our present point of view. Such a description is also given by Halperin (1986) in a *Scientific American* article, where it is nicely illustrated. In terms of the two-dimensional description, the localized states are associated with hills and valleys of the potential. These states can serve as reservoirs for the electrons that participate in the Hall current. Presumably, the heights and depths of these regions can be very large on the scale of the smoothly varying potential associated with the Hall current. As illustrated in Fig. 3, all the valley and extended states are filled, and the hill states are filled up to some level (just what determines that level is not obvious, since there may not be a definite Fermi level throughout the layer).

Now if the magnetic field is increased, the number of states per unit area increases in all these regions. If no electrons are permitted to move into the layer from other sources, this situation is unstable, and electrons can decay from higher-lying states into the newly created lower ones. It is possible that this may produce a new stable state in which the number of electrons trapped on hills is reduced and the extended states are kept filled. This may permit the quantum Hall effect to be maintained for a significant change in B. In this process, it is also possible that electrons can be supplied from potential valleys which were originally filled by electrons in higher Landau levels. (Recall that increasing B raises these energies relative to those of extended states carrying the Hall current because the excitation energy of a Landau level is proportional to B.) Decreasing B leads to changes in the opposite direction; electrons are forced to move further up on the hills and also into higher Landau states in the valleys. In any case, the result (2.5) is unchanged if all the extended states can be kept filled. Without more specific information about the nature of the localized states, it is not clear how much of the plateau can be attributed to this mechanism. Presumably, it becomes less important as the Hall voltage increases, since the fraction of the area containing localized states decreases (Trugman, 1983).

Continuing to study the situation that might evolve as the magnetic field is increased, we next consider the role of additional electrons which can enter the two-dimensional layer from the external circuitry. The effect of these additional electrons on the self-consistent potential should be determined by a quantum-mechanical analysis. However, we can get a rough idea of the size of the effects by considering the change of the electrostatic energy produced by moving the electrons into the sample to produce a small excess negative charge. Recall that if a charge distribution is changed from an original configuration given by $\rho_0(\mathbf{x})$, which produces the electrostatic potential V_0, the change of electrostatic energy is given by

$$\Delta U_{\text{es}}=\int\int\frac{\delta\rho(\mathbf{x})\delta\rho(\mathbf{y})}{2\pi\varepsilon\,|\,\mathbf{x}-\mathbf{y}\,|}d^3x\,d^3y+\int\delta\rho(\mathbf{x})V_0(\mathbf{x})d^3x\ .$$

If this is applied to the situation where δN electrons are

[7] A more precise statement is that the one-particle states in a region where $\mathbf{E}\neq 0$ cannot be partially filled; they must be either empty or completely filled. However, the electric field need not vanish at the boundary between a region where the states are empty and a region where they are filled, for example, on a hill of the potential.

[8] One way of presenting this argument is to recall the phenomenological equation connecting the current density and electric field in two dimensions: $\mathbf{j}=\sigma_{xx}\mathbf{E}-\sigma_{xy}\mathbf{E}\times\hat{\mathbf{B}}$. In the steady state, the divergence of this must be zero. If the coefficients σ were independent of position, this would lead to $\nabla\cdot\mathbf{E}=0$, which corresponds to zero net charge distribution inside the layer. However, in the ideal quantum Hall limit, the longitudinal conductivity σ_{xx} vanishes, and the argument fails. To understand the physics of σ_{xx} when the state deviates slightly from the ideal quantum Hall limit, we note that it arises from the viscous drag term mentioned in footnote 4. Therefore it seems likely that this term is proportional to the number of local vacancies of the level, and it will remain small and *position dependent* so that $\nabla\cdot\mathbf{j}=0$ does not imply $\nabla\cdot\mathbf{E}=0$.

moved from the high-potential reservoir into the two-dimensional layer, the change of electrostatic energy is estimated to be

$$\Delta U_{\rm es}\approx\frac{1}{2}\left\langle\frac{e^2}{\kappa\varepsilon_0 r}\right\rangle(\delta N)^2-\lambda e\Delta V\delta N\ , \tag{2.7}$$

where κ is the dielectric constant and r is the separation between two of the additional electrons. The last term arises from the rearrangement of the electrons in the original electrostatic potential; λ is some number less than 1. The average $\langle 1/r\rangle$ is $2G/L$, where G is a geometric factor of order 1, which depends on how the electrons are distributed in the layer. Minimizing this expression with respect to δN, one finds with reasonable parameters that of order 10^5—10^6 electrons can move into the layer if enough states are available below the Fermi energy of the high-potential probe [(—) terminal].[9] Since the total number of electrons in the layer is of order 10^{10}, the relative change in number is minute. Clearly, however, the change of electrostatic potential produced under these conditions is of order ΔV. If fewer states are available, the minimum is not reached, and the change in electrostatic potential is smaller.[10]

The purpose of the preceding paragraph was to convince the reader that when one-particle states are available, electrons flow into the two-dimensional layer and the additional charge will cause a change of electrostatic potential which is appreciable on the scale of other important potentials. These changes should be quite dramatic experimentally under appropriate conditions. It might be thought that in the absence of the previous two mechanisms the plateau length would correspond to a change of B of order one part in 10^4 or 10^5, since the additional electrons would fill all the available eigenfunctions and change the electrostatic potential enough to prevent more electrons from entering the layer. We now argue that the situation is more subtle in that the eigenfunctions are filled in such a manner that the self-consistent potential that evolves is able to maintain a plateau for an observable change of B, even in the absence of localized states. It turns out that the self-consistent potential is so arranged that the current is crowded near one edge of the sample where all the eigenfunctions are filled and that the rest of the sample is in equilibrium with the higher-potential terminal. Only the current-carrying states need be filled. When localized eigenfunctions are taken into account, the self-consistent potential experiences a smaller change as a function of B, but it acts to enhance the role of those eigenfunctions. The possible behavior of the potential as a function of ν is illustrated in Fig. 5(a).

We shall now switch from the semiclassical language and continue the discussion in terms of the energies of the one-particle states. It is clear that the energies of all states are pushed up by the change of potential. It might seem that this would give a much larger energy change than Eq. (2.7), but one should remember that the additional potential also acts on the positively charged lattice, giving a negative contribution to the energy.

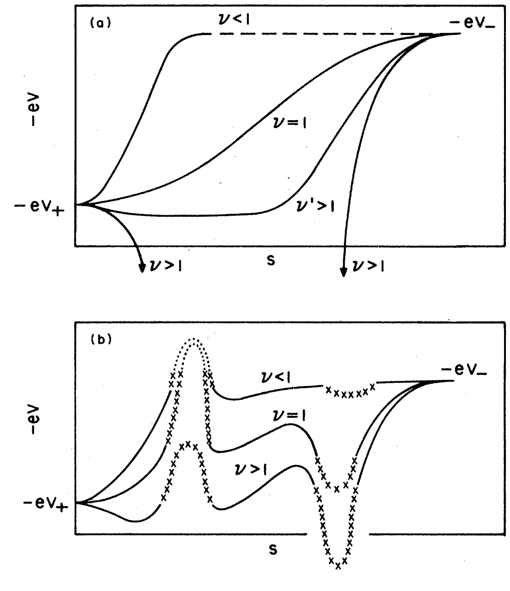

FIG. 5. Electrostatic potential for different values of B. Curves labeled $\nu=1$ refer to B in the center of the plateau, and those labeled $\nu<1$ ($\nu>1$) refer to a B that is larger (smaller) than the central value. (a) The situation when there are no localized states available to act as reservoirs. The dashed line occurring for $\nu<1$ represents a region where the states are partially filled in just such a way that the potential there is independent of position. Without localized reservoirs, the $\nu>1$ curve drops to very negative values. Even a change as small as 1% might produce variations as significant as those indicated here. If a higher level can absorb electrons (not shown), the curve $\nu'>1$ may replace the $\nu>1$ curve. This curve could also describe the potential along path (a) of Fig. 2 when localized states are available. (b) The situation when there are localized states available to act as reservoirs, as illustrated by the potential along path (b) of Fig. 2. Now fewer electrons are required to move in or out from the external circuitry. At the same time, the change in effective potential enhances the ability of the localized states to serve as reservoirs. The potential along path (a) would be similar to (a) except that the changes would be more gradual as a function of B.

[9] The present discussion is for intuitive purposes only. In a more correct analysis, one should take into account that G and λ depend weakly on δN as well. G is of the order of $\ln(L/W')$, where W' is the width into which the extra electrons are squeezed, and λ depends on the region into which they are placed. Nevertheless, the order-of-magnitude estimate should be reasonable.

[10] In the case of a MOSFET, the argument may require technical modification, since there it is a gate voltage that reduces the number of electrons relative to the number of states. Moreover, this is a situation in which the total number of active electrons changes. However, I claim that the practical result is the same. Once the gross equilibrium has been established by the gate voltage, the internal distribution of electrons will fine-tune in a manner similar to that described in the following paragraphs.

Again starting with $\nu=1$, we suppose that all the extended one-particle states are filled and that their energies rise more or less uniformly from one probe to the other as a function of distance across the sample. Now we try to visualize the situation that evolves as B is gradually increased, creating additional states uniformly across the sample. When energetically possible, electrons will flow into the layer from the higher-potential probe to fill these states. As they do so, the repulsive electrostatic potential will build up until a steady state is established. How the states fill is given by obvious physical considerations. In a region where the single-particle energies are varying with position, there can be no partially occupied states because the system can relax to lower the energy, possibly by emitting a phonon. This relaxation time should be very rapid because of the large overlap of two neighboring eigenfunctions. Thus any holes corresponding to states below the Fermi energy of the higher-potential probe will move across the sample until they are filled by an electron from that probe. As the electrostatic potential changes, states near the higher-potential probe will reach the Fermi energy of that probe and finally not accept additional electrons (if they did, their energy would be pushed higher than that of the probe's Fermi energy, which is not permitted).

Thus, as B increases, there develops a region in contact with the lower-potential probe in which all states are filled up to the energy of the higher-potential probe and a region of partially filled states that is in equilibrium with the higher-potential probe. The first region will shrink in size as B increases, and the latter will expand. Most of the excess electrons go into the first region, which carries all the Hall current. Since we expect the electrostatic potential estimate to give a reasonable guide to the order of magnitude of the number of additional electrons that can be accepted in the layer, we see that the filled-state region must shrink rather quickly as B increases. In our example, a 1% increase of B might require the current-carrying region to shrink to 10^{-3} of the original width. Presumably, so long as the current-carrying width remains large compared to a magnetic radius, our general picture remains valid and the quantization condition (2.5) is correct. The situation is expected to be as in Fig. 5(a) ($\nu<1$ curve). In effect, the experimental $\Delta\nu$ is permitted to be large because the current-carrying region can shrink to a small width where all the states are filled. The conclusion is that Eq. (2.3) continues to produce the quantum Hall relation.

In the more realistic situation with localized states, one would expect the two mechanisms to work together. Since the additional electrostatic potential raises the energies of all the states, it makes it easier for electrons in localized states on a hill to drop down to fill unoccupied extended or valley states. The existence of this mechanism reduces the number of electrons that must enter the layer from the probes, which in turn reduces the crowding of the current suggested in the preceding paragraph. The role of the localized states is suggested by the behavior of the potential along path (b) of Fig. 2, as illustrated in Fig. 5(b). A larger change of B is necessary to reach the point where partially filled states are in equilibrium with the higher-potential reservoir. Detailed model calculations should be done to study these mechanisms in more detail.

The localized states may also play an important role in suppressing the fractional quantum Hall effect. It is an experimental fact that as the "two-dimensional" layer is made very pure, the integer plateaus shrink in width and the fractional ones occur between them (Störmer, Tsui, and Gossard, 1982). Clearly, the fractional quantum Hall effect is associated with the residual electron-electron interactions that produce quasiparticles having lower energy than the independent-particle system (Laughlin, 1983). This leads to new plateaus and thus reduces the length of the integer plateaus. We would expect a new plateau to start whenever the partially filled extended states reach the correct filling to lower the energy. Since in fact there are a large number of such fractional-filling plateaus, it is possible that the integer plateaus would become insignificant if there were not some mechanism to inhibit the fractional ones. This mechanism may be that the impurities somehow dynamically prevent the residual interactions from forming the quasiparticle states.

In addition to the possibility of impurities' dynamically suppressing the interactions that produce the fractional plateaus, there is the likelihood that the role of localized states as reservoirs of electrons is important. They provide electrons that permit the extended states to remain filled for a larger change in B than would otherwise be possible. Thus, as B increases, the crowding of the current takes place more slowly than in the case without impurities, and the partially filled region in equilibrium with the higher-potential reservoir requires a larger change in B before its fractional filling reaches the condition for a prominent fractional plateau. We expect a special stability for filled states because of the relatively large gap (compared to kT) necessary to involve higher Landau levels or higher sub-band states. The situation with localized states may then be somewhat as illustrated in Fig. 5(b). Here it is assumed that there are enough electrons available on hill states so that the extended states may be kept filled. However, the current is still concentrated largely on the low-potential side of the sample [(+) terminal]. Perhaps the important lesson here is the way in which the total quantum state in the layer is able to adapt to various changes so as to maintain the plateau.

Without localized states, the situation in which ν is increased seems very different. In this case, there are not enough states available to hold all the original electrons. Some electrons may be returned to their donors and others may move into the potential probes. Referring to Eq. (2.7) with δN negative now (possibly reversing the sign of the second term since the electrons are pushed into the low-potential reservoir), we see that a very large energy is required to force a substantial fraction of the electrons out of the layer—perhaps hundreds of electron volts per electron. The situation might be as suggested by the curve labeled $\nu>1$ in Fig. 5(a). Since this is totally out of scale with other potentials in the problem, something else must happen. The resulting self-consistent potential in the layer will be lower than the original one. To prevent the large distortion in the potential, it is necessary that new states be available for some of the electrons. One obvious way this can happen is that localized states which previously had too high an energy to be occupied now become available, somewhat along the same lines as described earlier, but now enhanced by the overall negative change in potential. This is suggested by the $\nu > 1$ curve of Fig. 5(b).

It also seems conceivable that extended states from the next-higher Landau level could absorb some of the electrons without disturbing the quantum Hall relationship. In some cases several higher levels lie in the energy range between the Fermi energies of the two probes. Usually they remain unoccupied because of the exceedingly small spatial overlap between the occupied states and states of higher levels of nearly the same energy (energy separation $\lesssim kT$); in the central region of the plateau, the leakage rate to states of a higher level is then so small that they remain negligibly occupied. If the potential acting on the electrons goes low enough, a higher level may accept electrons that can be in equilibrium with the lower-potential probe without filling sufficiently to attain equilibrium with the higher-potential probe. In the energetically most favorable situation, the occupied states will have zero voltage across them and will not yield a net contribution to the Hall current. The possible result is indicated in Fig. 5(a) ($\nu' > 1$ curve), where it is assumed that the totality of these mechanisms leads to a modest net decrease of the potential. With further decrease of the magnetic field, the higher Landau level may go down in energy sufficiently that it can be in equilibrium (locally) with both voltage probes. In that case, a higher conductivity plateau is produced.

D. Some possible experimental consequences of the effective potential

Now we may discuss possible experimental consequences of this picture. Conceivably, a very wide plateau might not be accounted for by the donor and localized-state mechanisms alone, so that a more dramatic redistribution of the potential distribution may be required, as indicated by the $\nu \neq 1$ curves in Fig. 5. It should be possible to observe the effects suggested here by experimentally studying the potential or current distribution across the sample. Actually such experiments have already been carried out at 4.2 K by Ebert *et al.* (1985), Sichel *et al.* (1985), and Zheng *et al.* (1985). These experiments see a dramatic shift of the potential distribution as the magnetic field is varied through the region of filling factor 2. An example of this variation is shown in Fig. 6. For a magnetic field of 7.8 T, the current is somewhat concentrated near the $(-)$ terminal, while for 8.2 T, most of the current flows near the $(+)$ terminal. This is just the behavior anticipated in Fig. 5. However, the present explanation is not the correct one since, when the current is reversed, the largest voltage drop remains on the same physical side of the sample. Apparently, the correct explanation is that the shift is due to a small (of order a few percent) variation of the carrier density across the sample. This variation results from the manner in which the sample is prepared using molecular-beam epitaxy. In effect, the complete sample consists of several quantum Hall devices in parallel, and the current chooses to flow in the one where the carrier density most nearly matches the number of available states.

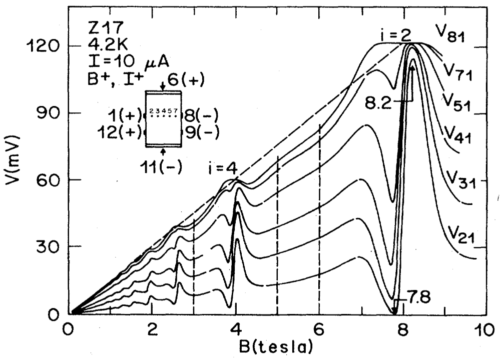

FIG. 6. Potential distribution across the sample as a function of B (from Zheng *et al.*, 1985). The sketch of the sample indicates several small voltage probes placed between the outer ones. The potential difference between each of the probes and probe 1 is plotted vs B. Between plateaus (e.g., $B = 6$ T), the voltage drop across the sample is rather uniformly distributed, but inside the plateaus the distribution changes quite rapidly as a function of B.

Under the conditions of these experiments, the plateau is rather narrow and the new mechanism proposed here may not be required. It would be interesting to carry out a similar experiment at lower temperature such that a wider conductivity plateau would be produced. If the ideas presented here are correct, the system should become very unstable against having the current on the wrong side of the sample when the magnetic field is far from the center of the plateau. Of course, the observation of the expected effect might be difficult, since the relaxation time becomes very long when the source impedance of the interior contacts becomes very large (Chang, 1985). The necessary dynamical analysis of how several parallel devices establish a steady state has not yet been carried out.

The behavior of the ordinary resistance and its linear relationship to the deviation of the quantum Hall resistance from its ideal value may be regarded as possible indirect evidence for the change of the potential distribution as a function of B. The main point is that at the center of the plateau the measured longitudinal resistance R_x has a very strong minimum as a function of B (Cage *et al.*, 1984), with a value that depends strongly on the temperature. One possible explanation of R_x is known as variable-range hopping between localized levels, but it does not lead to a linear relationship (Wysokinski and

Brenig, 1983). Instead, I hypothesize that this resistance comes about because the occupied Landau levels have a small leakage to higher levels wherever the higher eigenfunctions are sufficiently close spatially to a lower-level one whose energy is within kT. The small occupancy of the higher levels leads to a small modification of the quantum Hall resistance and introduces a small longitudinal resistance. A steady state is maintained because of the constraint that there be no net flow of current across the sample. The deviation of the quantum Hall resistance from its ideal value is found to be linear in R_x (as predicted by such a model). The present point is that as the B field is varied, the E field rapidly becomes large in certain regions, as we have seen. Thus the leakage to higher levels increases very rapidly as B is varied away from the minimum. Therefore it is expected that R_x will increase strongly with that variation in B. The weakness of this argument is that it is not possible to distinguish experimentally on which side the $\mathbf{E}$ field is large.

E. Overview of the quantum-mechanical discussion

In the following sections most of our attention will be spent in trying to understand the nature of the multielectron eigenfunction in the "two-dimensional" region. We conclude this section by summarizing the picture that has been developed so far. The most important thing is that we are dealing with an open system to which quantum mechanics is not readily applied. However, the behavior of the system in the region where the Hall voltage is measured is most simply understandable in terms of a multielectron state dominated by single-particle behavior. Therefore we make the assumption that we may isolate a finite piece of the sample and replace it by a closed system with a finite number of electrons. The voltage probes as reservoirs also appear to be an essential part of the system, but we disconnect them from the voltmeter in order to have a closed system. When conditions (such as current, magnetic field, or gate voltage) are changed, the number of electrons in the system actually changes. Here we imagine inserting the correct number of electrons, which are then held fixed, to discuss the eigenfunctions. In a state of finite Hall current, the electrons are distributed asymmetrically between the reservoirs to produce the correct Hall voltage (the distribution of electrons is given by this complete quantum state, of course).

We are imagining that the current is imposed from the outside—corresponding to the experimental arrangement—and the Hall voltage is observed. However, this imposition is not part of the Hamiltonian that provides the information about the possible states; we must select the state that has the desired properties. In our description, such states are assumed to be electrodynamically stable. The true ground state of the system would have zero Hall voltage and current. The current-carrying states can decay only by mechanisms outside the model, such as couplings which permit excitations of the bulk matter. If the resulting decay rate were appreciable, it would correspond to a current flow between reservoirs. In an actual sample, this transverse current is instead constrained to vanish; the physics of the decay in the model then manifests itself as a longitudinal voltage drop.

So far, we have imagined that for each value of the current through the sample there is necessarily a unique quantum state for the system, which we take to be the state of lowest energy under the constraint that the current have a fixed value. States of higher energy would decay to this state quickly by the mechanisms outside the model. We think we gave a convincing case that the extended states must be occupied in the manner claimed. However, it is not necessary that localized eigenfunctions be filled to the higher Fermi level in order to produce the quantum Hall relation. It is entirely possible that some such eigenfunctions on a potential hill might be so isolated spatially that the time scale for them to equilibrate may be made very long because of a small tunneling probability. Thus there may be many system quantum states that have the same physical content, differing primarily in how the localized eigenfunctions are occupied. The different states may also have slightly different self-consistent potentials. Away from Hall plateaus, it is not possible to keep all the single-particle states in the "two-dimensional" region filled, and electrons can easily flow across the sample. Under those conditions, it is necessary to maintain the Hall voltage externally. Further qualitative discussion of the conditions for metastability appears in Sec. V.

III. FORMULATION OF THE QUANTUM-MECHANICAL DESCRIPTION

We turn now to the details of our model. Since we have eliminated the outside world by taking a finite sample length and by opening the circuit involving the voltage probes, we must introduce some other properties of the model that permit it to have behavior similar to that of the actual sample. Because our system is confined to a finite length ($0 \leq x \leq L$) in the direction of current flow, it is appropriate to introduce periodic boundary conditions in that direction. This emphasizes the role of the long-range phase correlations in the eigenfunctions, as was originally made clear by the argument of Laughlin (1981). Practically, it makes the single-particle eigenfunctions discrete and countable. No similar boundary conditions are necessary in the other two directions, since the sample itself confines the electrons, and there is no current flow out of the voltage probes.

As already emphasized, the probes are reservoirs of electrons, and only the tiny fraction of the electrons that are in the "two-dimensional" layer actually participate in the Hall current. All the electrons that are free to move, including those in the reservoirs, are part of the quantum-mechanical system to be studied. This model is schematized in Fig. 7. We refer to the thin layer that carries the Hall current as the "two-dimensional" region to

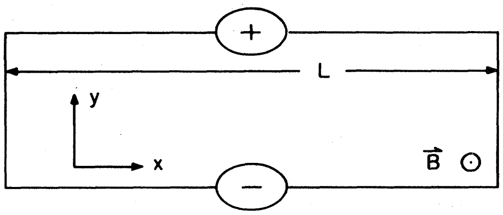

FIG. 7. Idealized schematic of the model. In place of the external circuitry which specifies the current and measures the voltage, a fixed current is assumed to flow through the sample in the x direction, and the voltage probes [the regions labeled (+) and (−)] have become electron reservoirs.

emphasize that the dynamics there is nearly two dimensional, but that we have not ignored three-dimensional physics (footnote 1). We believe that the electrons in the reservoirs come to equilibrium with separate Fermi energies, but emphasize that electrons are free to pass between regions. In fact, electrons near the edge of the sample will have eigenfunctions that span the "two-dimensional" region and one of the probes. It does not seem to be necessary to discuss the physics of the reservoirs in great detail. Basically, each reservoir provides an electrochemical potential with which nearby current-carrying electrons must be in equilibrium. A description of how this equilibrium might be established is given in Sec. V.

In the "two-dimensional" region, the three-dimensional potential confines the electrons most of the time to the lowest quantum state in the direction perpendicular to the two-dimensional layer (about 50—100 Å thick). Such a thickness would lead to an energy scale for excitations in this degree of freedom of order 10—20 meV (taking the effective electron mass for this consideration to be about $\frac{1}{10}$ the electron's mass); typical Landau excitation energies are somewhat smaller. In some cases, Hall voltages are of order 100 mV, so that there are several levels of Landau and sub-band excitation available within the energy range between the Fermi energies of the two probes. In some measurements, more than one of these levels is occupied; for example, the first two Landau levels may be occupied (in MOSFET's, there can also be "valley degeneracy" in which the lowest sub-band is degenerate). Often, the spin degeneracy is not resolved, leading to another factor of 2 in the occupancy. Clearly, the existence of the quantum Hall effect depends on having certain levels fully involved while higher levels remain unoccupied. Since $kT \lesssim 0.4$ meV, there is a significant gap in the energy necessary for thermal excitation of an electron to a higher sub-band or to a higher Landau level (than the ones participating at a given plateau). As discussed in the previous section, it is also important that there not be significant "sideways" leakage to higher levels. This is suppressed by the very small overlap between states of the same energy in different levels. For such decays to be strong, the electric field would have to produce an energy difference of order 10 meV in a distance of about l (i.e., in 100 Å), corresponding to an electric field of order 10^4 V/cm. The average electric field is of course much smaller than this, but under some circumstances we may have to worry about this possibility.

We assume that the magnetic field is large enough and the layer thickness small enough so that the occupation of higher levels can largely be ignored. After understanding more about the nature of the quantum Hall effect, it should be possible to estimate the necessary limits on these parameters. However, in trying to obtain an accuracy of 0.1 ppm, it is not advisable to ignore completely small effects that could arise from interactions which excite (virtually) higher levels. In our discussion, such virtual excitations are treated perturbatively.

A. The Hamiltonian and boundary conditions

Figure 7 illustrates the geometry of the system just described. The total number of electrons is fixed, and the space available to them is divided into three regions. Two of these represent the voltage probes and are treated here as three-dimensional electron reservoirs for the "two-dimensional" region. Since we aim to study individual states, our considerations are restricted to zero temperature. No assumptions are made about the symmetries of the sample (other than those implied by the periodic boundary condition) either at a microscopic or a macroscopic level. Once a sufficiently good understanding of the physics of the states is obtained, it should be possible to move toward a still more realistic model.

To study this system, we introduce a second-quantized formalism for the electrons in the active region. The bulk matter (footnote 2) provides a given external potential. For simplicity, we treat the interaction between the electrons as a two-body interaction but may easily generalize it to many-body interactions. This interaction can be influenced by the bulk matter, for example, by a dielectric constant; and it may even vary with position within the sample (i.e., it need not be translationally invariant). However, the dynamics of the bulk matter is suppressed other than for these effects.

The electron operators are taken to be periodic in a distance L along the direction of the current according to

$$\psi(x+L,y,z)=\psi(x,y,z) . \tag{3.1}$$

The main role of this condition is to make the eigenfunctions discrete and countable so that we do not have to deal with continuum normalization and with the properties of the (unquantized) eigenfunctions outside the region of interest; we do not believe that it distorts the physics. It would be desirable to free the discussion from such artificialities, or at least to prove that they do not influence the conclusions.[11] Here we assume any gauge for the vector potential which permits this condition.

We take the Hamiltonian to be

$$H=\int \psi^\dagger(\mathbf{x})\left[\frac{\left[\mathbf{p}+e\mathbf{A}+\frac{\hbar\beta}{L}\hat{\mathbf{x}}\right]^2}{2m}+V_e\right]\psi(\mathbf{x})d^3x$$
$$+\tfrac{1}{2}\int \psi^\dagger(\mathbf{x})\psi^\dagger(\mathbf{x}')U_2(\mathbf{x},\mathbf{x}')\psi(\mathbf{x}')\psi(\mathbf{x})d^3x\,d^3x' . \tag{3.2}$$

Here β is an analytic tool introduced in order to define a convenient expression for the current. It might appear that it could be transformed away with a gauge transformation, but that is not so. Only gauge transformations that maintain Eq. (3.1) are admissible. However, two values of β which differ by integer multiples of 2π are related by gauge invariance and hence are physically equivalent. Evidently, this model is equivalent to those of Laughlin (1981) and Halperin (1982) which have the sample closing on itself and an adjustable flux threading the hole (here β plays the same role as the threading flux). Note that, in such models, it is not necessary to have any particular symmetry—the boundaries need not be circles and the voltage probes can be localized on the circumference. As emphasized in the Introduction, m is the physical mass of the electron, not an effective mass. However, it turns out that the energy scales are more simply expressed using an effective mass, e.g., $\hbar\omega_c=eB/m^*$. Of course, nothing depends critically on the precise magnitude of this energy scale. But, as described above, the observed scale is important for the existence of the quantum Hall effect. If it were based on the electron mass (i.e., a factor of 10 smaller), the onset of leakage between levels would be greatly enhanced and it would undoubtedly have been more difficult to observe the effect. In fact, one of the considerations in the choice of an experimental sample is to make the effective mass as small as possible. The one-particle potential V_e is the external potential supplied by the bulk matter. In principle, it is nonlocal (to ensure orthogonality to the eigenfunctions of electrons bound in the bulk matter), but we ignore this for the moment. Interactions between electrons are represented by U_2. They may easily be generalized to include many-body interactions.

We introduce an effective one-particle potential U_1 which is to be determined in some self-consistent way to represent the average effect of the interaction between the electrons. To avoid complications in the discussion, it is important that U_1 be independent of β. U_1 and the provided external potential are nonlocal, but we suppress this complication temporarily and treat them both as local. The precise definition of U_1 is rather involved; it is discussed in Sec. IV and Appendix B. For the present, we treat U_1 as known and incorporate it into the one-particle part of the Hamiltonian, which is then given by

$$H_1=\int \psi^\dagger(\mathbf{x})h_1\psi(\mathbf{x})d^3x\ , \tag{3.3a}$$

where

$$h_1=\frac{\left[\mathbf{p}+e\mathbf{A}+\frac{\hbar\beta}{L}\hat{\mathbf{x}}\right]^2}{2m}+V_e+U_1\ . \tag{3.3b}$$

This defines the unperturbed problem, while the perturbation is given by

$$H'=\tfrac{1}{2}\int \psi^\dagger(\mathbf{x})\psi^\dagger(\mathbf{x}')U_2(\mathbf{x},\mathbf{x}')\psi(\mathbf{x}')\psi(\mathbf{x})d^3x\,d^3x'$$
$$-\int \psi^\dagger(\mathbf{x})U_1\psi(\mathbf{x})d^3x\ . \tag{3.3c}$$

The unperturbed problem is exactly solvable by expressing the operators ψ in terms of eigenfunctions of h_1 satisfying

$$h_1\psi_{\hat{\alpha}}(\mathbf{x})=\varepsilon_{\hat{\alpha}}\psi_{\hat{\alpha}}(\mathbf{x}) \tag{3.4a}$$

with the usual normalization. Anticipating that we shall perturb about reference states in which certain of these one-particle eigenfunctions are occupied and the rest empty, we introduce the following convention to label the eigenfunctions: $\hat{\alpha}$ represents a general eigenfunction, α represents an eigenfunction occupied in the reference state and to be treated by hole theory, and $\bar{\alpha}$ represents an eigenfunction unoccupied in the reference state. We emphasize that the labels run over all types of one-particle eigenfunctions, including Landau excitations, z-confinement excitations, etc., so that they form a complete set. No distinctions are made (or can be made) between eigenfunctions in the reservoirs and those in the "two-dimensional" region, and some eigenfunctions span two adjacent regions. However, we think of the "coupling" between adjacent regions as being sufficiently weak that in each reservoir a Fermi level is established (to good approximation).

With this convention, we write

$$\psi=\sum_{\alpha} b_\alpha^\dagger\psi_\alpha+\sum_{\bar{\alpha}} a_{\bar{\alpha}}\psi_{\bar{\alpha}}\ , \tag{3.4b}$$

where the operators $a^\dagger$, a, $b^\dagger$, and b satisfy the usual anticommutation rules. The a's are referred to as particle operators, and the b's as hole operators. The metastable reference state with no particles or holes is not the complete ground state of the system; rather it is the ground state for a fixed value of current. We refer to it as the unperturbed state, and assume that its metastability is main-

[11] Discussions with Marvin Weinstein (1985) shed some light on the meaning of the periodic boundary conditions. He pointed out to me that if we restrict our universe to length L in the x direction (∞ in the other directions) and require $p_x=-i\hbar\partial/\partial x$ to be Hermitian, then we must require quasiperiodic boundary conditions with an arbitrary phase function of y and z. Starting from that point, it is possible to change the gauge so as to make the phase factor unity (the original phase is just the difference of the required gauge function at the two boundaries), as we have assumed here.

tained under perturbations. To be precise: the perturbations do modify the value of the current, but we assume that it is possible to find an original unperturbed state which produces a state with the desired current. Using (3.4b) in (3.3a), we find

$$H_1 = \sum_{\alpha} \varepsilon_\alpha - \sum_{\alpha} \varepsilon_\alpha b_\alpha^\dagger b_\alpha + \sum_{\bar{\alpha}} \varepsilon_{\bar{\alpha}} a_{\bar{\alpha}}^\dagger a_{\bar{\alpha}} \ . \tag{3.5}$$

The statement that the unperturbed state has a lower energy than other states with the same current is that $\varepsilon_{\bar{\alpha}} > \varepsilon_\alpha$ for pairs that must be created in order to obtain the other states from the unperturbed one.

The operator giving the Hall current may be written

$$\hat{I} = -\frac{e}{mL} \int \psi^\dagger \left[p_x + eA_x + \frac{\hbar\beta}{L} \right] \psi \, d^3x \ . \tag{3.6}$$

A convenient expression for the current is then given by[12]

$$I = -\frac{e}{\hbar} \left\langle \frac{\partial H(\beta)}{\partial \beta} \right\rangle = -\frac{e}{\hbar} \frac{\partial E(\beta)}{\partial \beta} \ , \tag{3.7}$$

where we have used the Feynman-Hellman theorem in the last step. The expectation value is for any exact eigenstate of the Hamiltonian H of eigenvalue $E(\beta)$. As mentioned previously, the current-carrying states must be metastable. That means we assume that they are eigenstates of the Hamiltonian (3.2), but other physics not yet taken into account, such as bulk matter excitations, could permit them to decay. For present purposes, we assume that the "other physics" is sufficiently weak that it may be ignored in discussing the eigenstates; yet it can be invoked as a mechanism for bringing about decay into these states. If it is true that the current-carrying states are eigenstates of H, then it is at least plausible that they can be obtained by perturbation theory from the unperturbed states defined by H_1, provided the potential U_1 is chosen with sufficient care. Here we understand β as a mathematical parameter, and all functions are assumed to depend continuously upon it. $E(\beta)$ is the total energy of the state under consideration, including that of the electrons in the reservoirs and taking into account interactions between electrons. More will be said about the general behavior of $E(\beta)$ in the discussion of Laughlin's argument below.

B. Qualitative properties of the eigenfunctions and energy spectrum

We expect the general picture of the eigenfunctions as described in Sec. II to be correct, but now we plan to discuss them without any unnecessary approximations. Since V_e includes contributions from individual atoms, the eigenfunctions satisfying Eq. (3.4a) must be very complicated because the potential varies rapidly on a scale of 1 Å. Fortunately, it is not necessary to calculate or analyze the wave functions at that level. We need only make a very straightforward and plausible assumption. This crucial assumption is that each eigenfunction has a phase that can be defined continuously as a function of position. That is, it may be expressed

$$\psi_{\hat{\alpha}}(\mathbf{r}) = e^{i\varphi_{\hat{\alpha}}(\mathbf{r})} R_{\hat{\alpha}}(\mathbf{r}) \ , \tag{3.8a}$$

where both $\varphi_{\hat{\alpha}}$ and $R_{\hat{\alpha}}$ are real. This appears to correspond to the idea of long-range phase rigidity, which is important in Laughlin's argument. Provided an eigenfunction has no two-dimensional zeros (i.e., lines of zeros that go completely through the eigenfunction in the z direction) this phase will be unique within an overall additive constant. (Zeros at the sites of atoms cause no ambiguity.) If an eigenfunction has two-dimensional zeros, the phase will not be unique but will be path dependent; however, single valuedness of the eigenfunction requires that phases determined by different paths differ by integer multiples of 2π.

As in Sec. II, some eigenfunctions in the "two-dimensional" region are localized; i.e., they do not extend continuously from one end of the sample to the other (they may exist artificially at both ends due to our use of periodic boundary conditions). Then single valuedness gives a quantization condition that restricts the eigenfunctions. The total phase change around any closed curve must be an integral multiple of 2π (zero if it does not enclose two-dimensional zeros).[13]

[12] In carrying out these derivatives, note that ψ and $\psi^\dagger$ are independent of β. They are restricted only by the boundary conditions and the anticommutation relations.

[13] The author wishes to call attention here to a mathematical point that may be worth further investigation. As Brenig (1983) and Joynt and Prange (1984) point out, it is not normal for discrete, locally normalizable states to have energies within the continuum. (In our model, of course, there is no true continuum, but we may think of the extended states as playing that role.) Normally, the "would-be" discrete states become decaying states. They argue that discrete states lie outside the continuum, i.e., at either higher or lower energies. With the electrons confined to a finite width, there is actually no upper bound to the continuum, so the discrete states would lie only below the continuum. If this point of view is correct, and the author believes so, each would-be localized state mixes with a set of extended states having nearly the same energy to produce extended states that have a large relative probability of being in the would-be localized-state region. Each of these states would have a very small, but nonvanishing, amplitude to extend between the two original regions. This means that in the following discussion the localized states, but not the relabeling gaps, would disappear. However, the author contends that the physics of the discussion still makes sense. It is likely that most of the would-be localized states have such a long lifetime that in any practical situation we may as well treat them as individual eigenstates. From now on, we ignore this possible complication in the discussion.

Eigenfunctions confined to either of the reservoirs are also regarded as localized eigenfunctions in the present discussion. There is some subtlety here, since there may be (technically) extended eigenfunctions whose probability is concentrated inside a reservoir, but whose amplitude elsewhere is very small and does extend from one end of the sample to the other. If there are such eigenfunctions, a tiny fraction of the Hall current may actually pass through the reservoirs. Thus there is a very gradual changeover from extended eigenfunctions to ones that are actually localized in a reservoir. As we go from one extended eigenfunction to the next, the probability to be outside the reservoir and the current carried through the sample diminish to some point where they are both so small that we may ultimately regard an eigenfunction as being attached to the reservoir and not as an extended eigenfunction. While the place where we define the changeover to localized eigenfunctions is arbitrary, it should be possible to do this in such a way that the total current carried by the states defined to be localized is completely negligible compared to the total Hall current. The nature of the equilibrium near the edge of the sample is discussed further in Sec. V, and the ideas of this paragraph are needed in order to understand the discussion of Laughlin's treatment of the quantum Hall relationship.

For eigenfunctions extending from one end of the sample to the other, the total phase change φ must be an integer multiple of 2π (if there are two-dimensional zeros, the multiple is of course path dependent). For each eigenfunction, we select a particular path for defining the total phase change between $x=0$ and $x=L$ (in case there are two-dimensional zeros). It seems appropriate, and turns out to be convenient, to use this total phase change as one of the parameters to label the eigenfunctions. We reexpress this phase in terms of an average wave number by writing

$$\varphi \equiv kL = \varphi_{\hat{\alpha}}(x+L,y,z) - \varphi_{\hat{\alpha}}(x,y,z) \ , \tag{3.8b}$$

where

$$k = 2\pi n/L \ . \tag{3.8c}$$

For each value of this average wave number, we expect there to be several states which require additional labels. In the simplest case, where only the first Landau level is filled in the unperturbed state, we use k in place of α to label extended states. The spacing in k values is very small $(2\pi/L)$, and the set of eigenfunctions corresponding to two successive values of k are very similar (having only an overall phase difference of 2π over the total length of the sample and a tiny transverse separation of order $2\pi l^2/L$). We assume that for each value of k (associated with a definite path for defining the phase), the energy eigenvalues have low degeneracy and that the energy spectrum does not change character as a function of k. The same set of additional labels may then be used for all k, and a given set defines a "level." To the extent possible, the same labeling scheme may be applied to the localized states, but with k replaced by some other suitable label. Usually, the different levels correspond to differing Landau excitations, with approximate energy separation $\hbar\omega_c$, or different excitations of the z-confinement modes. As discussed in Sec. V, this labeling scheme may break down at the edge of the sample, where two states with the same k can see a very different environment and become degenerate even though they belong to different energy levels in the interior of the "two-dimensional" region.

While it is possible to solve the one-particle eigenvalue equation explicitly for certain simple potentials, and that is the usual procedure, our present aim is to develop some general intuitive understanding of the properties of the eigenfunctions. Some instructive examples of interactions with a scattering potential are described by Joynt and Prange (1984) and Prange (1986), and the time-dependent development of a wave packet in the presence of a strong scatterer is given by Joynt (1982, 1984). Intuitive understanding is also provided by a semiclassical analysis of the eigenfunctions in two dimensions (Trugman, 1983; Joynt and Prange, 1984); this is described briefly in Appendix A. Here we give a description that is three dimensional and does not require any important specialization of the external potential. While the author believes that it provides insight into some properties of the wave function, he does not claim that it would serve as a useful starting point for their actual calculation. Initially, the description ignores the complications associated with two-dimensional zeros of the wave function; in any case, it can be applied to small regions which can later be pieced together.

Substituting Eq. (3.8a) into the eigenvalue equation (3.4a), and separating the resulting eigenvalue equation into real and imaginary parts, we find

$$\left[\frac{\mathbf{p}^2}{2m} + \mathscr{V}_{\hat{\alpha}} + V_e + U_1\right] R_{\hat{\alpha}} = \varepsilon_{\hat{\alpha}} R_{\hat{\alpha}} \tag{3.9a}$$

and

$$\nabla \cdot [(\hbar\nabla\varphi_{\hat{\alpha}} + e\mathbf{A})R^2] = 0 \ , \tag{3.9b}$$

where

$$\mathscr{V}_{\hat{\alpha}} = \frac{(\hbar\nabla\varphi_{\hat{\alpha}} + e\mathbf{A})^2}{2m} \ .$$

This gives a pair of equations for the functions $R_{\hat{\alpha}}$ and $\varphi_{\hat{\alpha}}$. The first has the form of a simple eigenvalue equation with an additional effective potential $\mathscr{V}_{\hat{\alpha}}$ depending on $\varphi_{\hat{\alpha}}$. The reader should note that this additional potential is gauge invariant. The second is an expression of current conservation.[14]

[14] It is interesting to note the structure of the longitudinal current as a function of distance across the eigenfunction. It changes sign near the center of the eigenfunction. This corresponds classically to the cyclotron motion of the electrons. If there is a potential gradient across the eigenfunction, the current does not quite cancel out; this corresponds to the drift velocity of the electrons.

In other situations, these equations could serve as the starting point for the development of the eikonal approximation. Here that is not possible because of the rapid variation contained in the external potential V_e and in $\mathscr{V}_{\hat{\alpha}}$, which do not permit us to treat the first term of (3.9a) as small. Here we use them qualitatively to obtain an intuitive understanding of the properties of the eigenfunctions, following somewhat the reasoning of Appendix A. It should be emphasized that, in general, $\varphi_{\hat{\alpha}}$ depends on the label designating the level; i.e., states of the same k need not follow exactly the same path through the sample. In fact, this observation turns out to be crucial in discussing how different levels can merge in energy at the edge of the sample.

Writing out the additional effective potential in detail in the Landau gauge, we find

$$\mathscr{V}_{\hat{\alpha}}=\frac{\left[\hbar\frac{\partial\varphi_{\hat{\alpha}}}{\partial x}-eyB\right]^2+\left[\hbar\frac{\partial\varphi_{\hat{\alpha}}}{\partial y}\right]^2+\left[\hbar\frac{\partial\varphi_{\hat{\alpha}}}{\partial z}\right]^2}{2m} . \tag{3.9c}$$

All of this looks quite intractable analytically, but it does provide some tools for intuitive understanding. The phase varies systematically along the length of the sample to accumulate its total change, but that variation need not be uniform. It may also fluctuate on a microscopic level to produce the results of an effective mass. In simple potential models, $\mathscr{V}_{\hat{\alpha}}$ provides a strong potential "trough" through the sample; it is this trough which confines the wave function to a transverse distance of order l. Assuming that the systematic phase change along the length of the sample dominates the behavior, one sees that the potential minimum occurs at

$$y_0(x,z)\equiv l^2\frac{\partial\varphi_{\hat{\alpha}}}{\partial x}(x,y_0(x,z),z) , \tag{3.9d}$$

which describes a curve through the sample. In the vicinity of this minimum, the potential has the form $(y-y_0)^2\hbar^2/(2ml^2)$. For an arbitrarily specified $\varphi_{\hat{\alpha}}$, the trough will not occur in the right place throughout the sample to produce an appropriate eigenfunction. In the simplest example of a smooth two-dimensional potential, the trough should follow an equipotential. In general, the phase function must adjust itself to put the trough in the right place to produce an eigenfunction and to have the correct total change along the sample. For intuitive purposes, it is helpful to define an average value of y_0 using Eq. (3.9d) by neglecting the y,z dependence on the right-hand side. This gives

$$y_k=kl^2 . \tag{3.9e}$$

This average lateral position is *not* weighted by R^2; aside from that, it does correspond to the approximate density of states $n_0=1/2\pi l^2$. For eigenfunctions in the interior of the "two-dimensional" region, the center tends to move along two-dimensional equipotentials, which are obtained by averaging out the ordinary three-dimensional potentials (in this averaging, it may also be necessary to take into account changes in the sub-band energy if the layer is not quite uniform). As y_k changes, the associated eigenfunctions displace across the sample in a corresponding way.

If we were to plot the distributions of energies of a level, they would smear out into a band. The appearance of a band would depend on the size of the fluctuating component of V_e relative to the systematic effect of the Hall potential. Although it is not explicitly stated, I believe that the plots of energy distributions that one frequently sees in discussions of the quantum Hall effect correspond to the situation in which the systematic Hall voltage is small relative to the fluctuations or is somehow subtracted out. In other situations, the Hall voltage may correspond to an energy difference several times the local energy differences between levels. Bands from different levels (including contributions from the localized states) would then have considerable overlap. Even in those plots, it would be better if one were to plot the distributions for extended, valley, and hill eigenfunctions separately, since they might actually overlap if the variation of the "smooth" component of the potential were large enough. These remarks are inserted here to point out that such plots lack an important piece of information, namely the spatial location of the eigenfunctions. If one plots the energies of eigenfunctions as a function of k, which is a measure of their transverse location, it is seen that eigenfunctions of the same energy in different levels usually have very different values of k and hence very different spatial locations. In order to obtain the quantum Hall effect, it is necessary that the spatial overlap of such eigenfunctions be exceedingly small so that a lower level cannot easily decay into a higher one.

As just remarked, if an extended eigenfunction has two-dimensional zeros, the labeling by k is ambiguous. This happens when an eigenfunction divides to avoid a region containing localized states. A two-dimensional example with a saddle point is shown in Fig. 8(a). One type of eigenfunction labeled γ lies entirely on one side of the localized-state region, another type σ lies entirely on the other side, and a type δ passes partly on each side. For eigenfunctions of the γ or δ types, we may define the total phase change by taking a path along the eigenfunction on the γ side of the region. Similarly, for eigenfunctions of the σ or δ types, we may use the σ side for the definition. These two possible choices produce the two curves in the plot of ε_k vs k in Fig. 8(b); in the region of the δ-type eigenfunctions, the two curves are simply displaced by a horizontal distance $2\pi N/L$, where N is approximately the number of flux units contained within the localized state area. According to a semiclassical argument by Prange (1986), it is also the number of localized states associated with that area. [Perhaps this is connected to an argument by Brenig (1983) which relates the total number of localized states of a level to properties of a level shift function, in a sort of generalized Levinson (1949) theorem. The present author has not been able to understand Brenig's argument well enough to relate it to the present work.] It is clear that as k steps through successive values on one

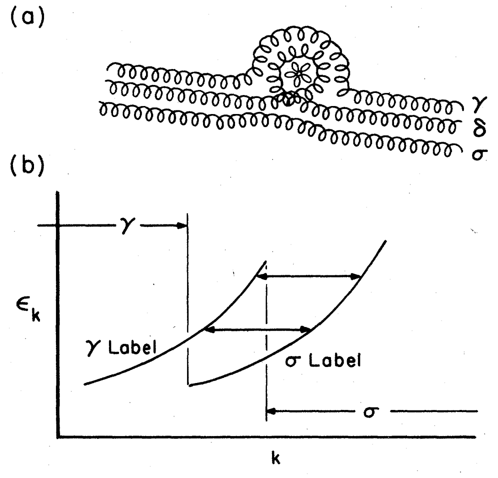

FIG. 8. Illustration of the origin of "relabeling gaps." (a) Examples of extended eigenfunctions in the vicinity of a localized region. Those of type γ pass entirely on one side of the localized region (ignoring exponentially small tails of the wave functions), while those of type σ pass entirely on the other side. Those of type δ split and pass partially on each side; they may be labeled continuously with either of the other two types. A localized state is shown inside the loop of a δ-type eigenfunction. (b) The single-particle energies labeled with the wave number k of the extended eigenfunction. The range in k for each type of state is illustrated. There is a unique assignment of k for eigenfunctions of type γ or σ. With those of type δ, we may elect to choose either method of labeling, as shown by the two curves. The double-arrowed lines indicate examples of places where we can shift from one label to another while introducing a relabeling gap. The eigenfunctions corresponding to opposite ends of these lines are the same, but the label has changed.

of these curves, the single-particle eigenfunctions change quite gradually. At some point, we may choose to relabel the eigenfunctions by switching from one curve to the other. Since the situation changes gradually with k, there is no unique way to specify where the relabeling should take place; there is no physical discontinuity, even though certain values of k do not appear in the labeling of the states. Two possibilities for the location of the relabeling gap are illustrated in Fig. 8(b). The actual situation is undoubtedly very complicated, with many such relabeling gaps (even several for individual eigenfunctions). However, it seems essential (in order to obtain the quantized current) that there be a continuous progression through eigenfunctions which can be labeled by k. This is analogous to the point made in Sec. II that it should be possible to find an integration path for the evaluation of the current which lies entirely within the extended region. Wherever gaps occur in the values of k used for the labels, there is no physical gap in the succession of eigenfunctions. We could, in fact, label the eigenfunctions without a gap in k, at the expense of having k lose its original meaning as giving the total change in phase whenever that is unique. We do this below in the discussion of Eq. (3.13), where we refer to the variable that labels the states successively as k'. (The mathematically inclined reader may wish to refer back to footnote 13 at this point.)

Since some (perhaps most) of the eigenfunctions do not have a unique value of k, the average transverse displacement refers to the particular path chosen to define the overall phase. Obviously, one should not take y_k too literally as giving the position of the orbit. Nevertheless, it is a useful parameter for visualization, particularly when the magnetic field is changed. Then the density of states in y_k changes as a function of B, but the general nature of an eigenfunction as a function of y_k does not change if the effective potential is held fixed.

Now we must describe briefly the properties that the external potential V_e must have in order that the eigenfunctions and energy spectrum have the general characteristics described above. What sort of structure is actually contained in V_e? We know that on a very small scale (of order 1 Å) it has a very rapid and large spatial dependence associated with the atomic structure of the bulk matter. However, the scale of the wave function is much larger (of order 100 Å), so the effect of these small-scale fluctuations should average out while giving rise to an effective mass dependence of the energy spectrum. Prange (1986) discusses in detail three other components of V_e. One of these is a "smooth" component, which varies slowly with position ($l\,|\nabla V_e| \ll \hbar\omega_c$). However, one should note that the total variation of this component over the sample can be several times larger than $\hbar\omega_c$. It is for this part of the potential that we may think of the eigenfunctions as being "guided" by equipotentials as in Sec. II. The condition ensures that eigenfunctions of different Landau levels are never close enough in space so that a filled lower level can decay into a higher one. In semiclassical terms, this condition is equivalent to the requirement that the drift velocity given by Eq. (2.2) be smaller than the cyclotron velocity. Another part of the potential varies rapidly with position on the scale of l but it always remains very small in magnitude ($\ll \hbar\omega_c$). Prange gives arguments that this potential does not introduce extended eigenfunctions into the gap between Landau levels. We take this to mean that extended eigenfunctions in a given region of the sample outside impurities have a well-defined hierarchy of Landau and sub-band levels, as discussed above. The other type of potential is associated with impurities and is referred to as a scattering potential. Prange discusses how this gives only forward scattering because of energy conservation. We refer the reader to the excellent review by Prange (1986) for more details.

For our treatment of perturbation theory, it is useful to introduce new eigenfunctions defined by

$$\psi'_{\hat{\alpha}} = e^{i\beta x/L}\psi_{\hat{\alpha}} \,, \tag{3.10a}$$

which are eigenfunctions of

$$h_1^0 \equiv h_1\,|_{\beta=0} \,. \tag{3.10b}$$

Localized eigenfunctions are independent of β, while for extended eigenfunctions, a change in β corresponds to a change in label k according to

$$\delta\beta \leftrightarrow L\,\delta k\ . \qquad (3.10c)$$

Later on, when it becomes necessary to differentiate the β dependence contained in the eigenfunctions, it is more convenient to use these new functions even though they are not periodic. The reason is that for localized eigenfunctions the derivatives vanish, and for extended eigenfunctions we may use the relation (3.10c) to replace derivatives with respect to β by derivatives with respect to k.

C. How the states are filled

In Sec. II we described the general features of localized and extended single-particle eigenfunctions in the "two-dimensional" region and how these states should be occupied in a self-consistent way. Here we refine that discussion in the light of a somewhat better idea of the nature of the eigenfunctions. Each reservoir can be considered approximately as an isolated system with all states filled up to some Fermi level. Because the reservoirs are not thin in the z direction, these states have an almost continuous range of energies associated with kinetic energy in that direction and also have many levels of excitation of the magnetic states. On the other hand, the electrons in the "two-dimensional" region exist in the lowest state in the z direction and have relatively high energy because of confinement to the thin layer. For the present discussion, higher states associated with this confinement play a role only in perturbation theory; they may have a more significant role for the higher plateaus. Under plateau conditions, an equilibrium is established between the reservoirs and the adjacent "two-dimensional" region. We accept this here and defer a more complete discussion to Sec. V.

In preparation for discussing how the quantization of the conductivity comes about when we may ignore interactions between the electrons, we first describe various possibilities for the dependence of the one-particle energy levels on y_k. Suppose that conditions are such that we need consider only the lowest Landau level and lowest z-confinement state. In addition, for the moment, ignore the self-consistency question and imagine that we have a good starting assumption for U_1, which leads to a set of extended and localized eigenfunctions and their eigenvalues. Between relabeling gaps, the extended eigenfunction energies are represented by points lying on a smooth curve as a function of y_k. If we delete the gaps, the separate curves join into one smooth one. The density of points is proportional to the magnetic field, and the ends of the curve will tie into the Fermi energies at the two reservoirs. From our previous discussion in Sec. II, we expect that all eigenfunctions with energies below the higher Fermi energy are occupied, but that there may be partially filled eigenfunctions in equilibrium with the higher Fermi energy.

Figure 9 illustrates the possible behavior of ε_k as a function of y_k for various magnetic fields. These curves are not based on actual self-consistent analyses; rather, they are impressionistic representations of the plausibility arguments given here and in Sec. II. With one exception, it is assumed that the next-higher Landau level or electron sub-band has an energy gap large enough so that such excited eigenfunctions play no direct role and they are not shown. For each case, one plot shows the eigenfunction energies with representative relabeling gaps and the other shows them with the gaps removed. Solid horizontal lines in the gap regions indicate filled localized eigenfunctions

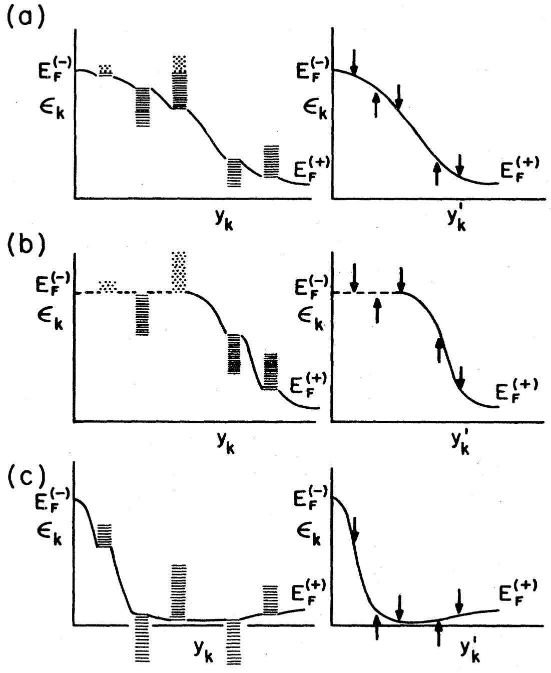

FIG. 9. Energies of one-particle states. The left side of each figure shows extended-state energies plotted as a smooth function of wave number and examples of localized-state energies introduced at the positions of the sample relabeling gaps. Filled localized states are represented by solid horizontal lines and empty ones by dotted horizontal lines, under the assumption that these states are filled to the higher Fermi energy. Partially filled extended states are represented by dashed lines. The right side of each figure shows the extended states with the relabeling gaps deleted at positions indicated by vertical arrows. (a) might represent the situation in the middle of a plateau. (b) might represent the situation where the magnetic field has been increased from that of (a), causing the self-consistent potential acting on the electrons to increase. The resulting curve for the extended states must not go higher than the Fermi energy of the negative voltage probe. In the example shown, this results in a region in equilibrium with that probe where the extended states are only partially filled. (c) might represent the situation when the magnetic field is reduced from that of (a), causing the self-consistent potential to decrease. In all cases, the occupation of the localized states changes to accommodate the overall change of the filling factor, but some electrons flow in or out from the probes as well.

on hills or valleys, while dotted horizontal lines represent empty localized eigenfunctions. Figure 9(a) might represent the situation in which the magnetic field is in the middle of a plateau and the energy drop of the extended eigenfunctions is fairly uniform across the sample as illustrated in the plot at the right.

Figure 9(b) shows the possible effect of increasing the magnetic field, following the discussion of Sec. II. Recall that the number of eigenfunctions per unit area is proportional to B. We assume that the extended and localized eigenfunctions remain filled to the energy $E_F^{(-)}$. This means that electrons must flow into the layer and change the self-consistent potential until the proper equilibrium is established. As described in Sec. II, the repulsive Coulomb energy increases the potential acting on the electrons and raises the energies of all the eigenfunctions. Hill eigenfunctions that are pushed above $E_F^{(-)}$ need not remain filled, so that the total number of electrons flowing into the layer from the reservoirs is less than it would be without localized eigenfunctions. This may make it possible for all the extended eigenfunctions to remain filled for a modest change in B. In any case, since each ε_k for a filled eigenfunction should not exceed $E_F^{(-)}$, the potential increase is limited and the drop in energy is necessarily more precipitous near the $(+)$ terminal. As B increases further, we may reach the situation illustrated here in which the extended eigenfunctions at the energy $E_F^{(-)}$ are no longer completely filled and there is a very rapid drop in energy near the $(+)$ terminal. The argument for this configuration is exactly the same as in Sec. II.

Figure 9(c) shows how the system may evolve when the magnetic field is reduced. In this case, with the original self-consistent potential there are not enough eigenfunctions to hold all the original electrons, and some electrons are pushed out of the layer into the reservoirs. As a result, the energies of all the eigenfunctions are lowered. As described in Sec. II, this permits new regions of potential hills to accept electrons, which reduces the number that might otherwise be pushed out of the layer. If the area that can now accept electrons is large enough, it is possible that with a modest change of B the situation would evolve as illustrated here. Now more of the current is near the $(-)$ terminal than the $(+)$ terminal. However, the number of eigenfunctions within the first Landau level is limited, and at some point enough electrons will be pushed out of the layer that the eigenfunction energies may be pushed down significantly on the scale of the potential difference. If a potential minimum in the middle of the sample results, this leads to a current greater than the net Hall current near the $(-)$ terminal together with some reverse current near the $(+)$ terminal. With further decreases of B higher Landau levels would begin to play a role. At first, valley eigenfunctions from higher levels could accept electrons, and this would tend to stabilize the potential. Ultimately higher extended eigenfunctions would drop down, and perhaps one would come to equilibrium with the $(+)$ terminal. If so, it could accept electrons; but not having a net difference of energy across it, it would not contribute to the current. Ultimately it would fill up, leading to a transition to another plateau. We defer a discussion of higher plateaus to Sec. V.

The preceding discussion is undoubtedly simplistic. Providing that the essential ideas are reasonably correct, however, it may lead to a more detailed understanding of the microscopic mechanisms involved in the quantum Hall effect. The condition that all levels be filled to the Fermi level of the $(-)$ terminal, while at first sight plausible, is probably too rigid. Eigenfunctions near the top of potential hills may not be in good communication with nearby filled eigenfunctions if the distance separating them is more than a few magnetic lengths. In that case, they may not come to equilibrium completely as the magnetic field is swept through the plateau. While this could affect the amount by which the self-consistent potential must adapt by the flow of electrons between the reservoirs and the "two-dimensional" layer, it does not affect the quantum Hall relation to be (re)derived below. It is important only that the extended eigenfunctions be filled in the region where ε_k actually depends on k. The same argument given in Sec. II can be restated here. If there were an empty eigenfunction in such a region, a neighboring eigenfunction with very good spatial overlap could decay into it by exciting the bulk matter. This implies that any quantum state with reasonable occupation of the localized eigenfunctions has the extended eigenfunctions appropriately filled to produce the correct quantum Hall relationship within a plateau. Therefore the number of states that have the correct properties to produce a plateau may actually be quite large. Generally, in the remainder of this paper, we ignore this possible generalization and pretend that there is a unique state, secure in the knowledge that other possible states have the same macroscopic behavior. However, these considerations certainly play an important role in understanding the length of the plateaus and the mechanism of the transition between them.

D. Connection with Laughlin's argument

We may now indicate briefly the argument of Laughlin in the present context using Eq. (3.7). In the case of the integer quantum Hall effect, we vary β continuously from 0 to 2π while the current remains nearly fixed. While the Hamiltonian can be restored to its original form by a gauge transformation, the continuously deformed state does not return to the original one since its energy has decreased by $-h\langle I\rangle/e$ (since I will in principle vary a tiny amount as a function of β, we use its average $\langle I\rangle$ over the 2π range here). How can this happen? The answer is that as the state deforms, an integer number of electrons (in this case one) moves out of the high-potential reservoir [$(-)$ terminal] and the same number moves into the low-potential reservoir [$(+)$ terminal]. At the same time, electrons in the "two-dimensional" region gradually shift across the region, but at the end the state of the electrons there returns to its original form (and energy). Because the number of electrons that move across is a tiny fraction of all the electrons in the three-dimensional regions, this change results in a totally unobservable change in the current and in the potential difference across the "two-dimensional" region. The decrease in the total energy of the state can then be attributed to the difference of the electrochemical potentials across the sample. The desired result is obtained by equating the two expressions for the energy change of the system,

$$-h\langle I\rangle/e=-e\Delta V \Longrightarrow \langle I\rangle=\frac{e^2}{h}\Delta V\ . \tag{3.11}$$

One aim of this paper is to make intuitively plausible that the electron states actually have the properties that justify this conclusion. We do this first for the situation in which all extended eigenfunctions for the lowest Landau level are filled, which we expect to be true near the center of a plateau. We also assume that the current is small enough that leakage to higher Landau and sub-band levels can initially be ignored. If we consider the unperturbed problem, holding U_1 fixed during the variation of β, we see that the single-particle eigenfunctions do have just the property required by Laughlin's argument. As β increases by 2π, each of the extended eigenfunctions transforms gradually into the next according to Eq. (3.10c). The presence of localized eigenfunctions does not affect this argument; they are unaffected by β. A very large number of extended eigenfunctions spans regions of localized eigenfunctions, or even a hole through the sample, and each one individually experiences a very tiny change. Yet the net effect is to move one electron across each relabeling gap. Similarly, near a probe, the probability for an extended eigenfunction to be outside the probe becomes vanishingly small as k goes to its highest or lowest value near a probe. At some point, one electron is simply counted as being in the probe rather than in an extended eigenfunction. This is not an abrupt transition, but the net effect is that the total charge in one probe increases by e and that in the other decreases by e after the 2π change in β. The behavior near the edge is discussed in more detail in Sec. V. (Note: for the fractional quantum Hall effect, β must increase by a higher multiple of 2π before the state in the "two-dimensional" region returns to its original form.)

Now let us examine what this implies for the complete state of the system, including the effects of electron-electron interactions. Recall that we are using β as a label for the set of quantum states associated with a given Hamiltonian rather than as a physical flux parameter as in Laughlin (1981) and Halperin (1982), but the correspondence should be obvious. Values of β that differ by 2π have the same set of states. As β is varied continuously, we assume that each of the associated eigenstates of Eq. (3.2) also transforms continuously provided that the parameters are in the center of the plateau region. This assumption is based on the similar behavior of the unperturbed states of the system: There are no sudden jumps as β is varied; what happens at relabeling gaps or at the edge of the sample is actually very gradual. (We are implicitly assuming that the residual electron-electron interactions do not change the essential features of the spectrum of the system; that is, they do not make the states metastable and we are presently ignoring the fractional quantum Hall effect. Should they cause metastability, the present picture would break down unless the lifetimes were so long that decays would not occur on an experimental time scale.) Now consider a plot of $E(\beta)$ vs β for fixed B starting from the ground state and going to some large value of $|\beta|$ [see Fig. 10(a)]. We assume that this is a smoothly varying function of β (roughly parabolic in shape for small β) since varying β moves electrons across the sample, producing a potential difference between the probes and, according to Eq. (3.7), a current that increases with $|\beta|$. When B has a value within a plateau, the system does not return to the original state whenever β changes by 2π because there is no easy mechanism for it to do so. The electric and magnetic forces are in balance, so individual eigenfunctions do not tend to move across the sample. The only way the energy of the system could return to its original value is for an electron to tunnel a considerable distance (through filled eigenfunctions). The eigenfunctions reached by this continuous variation of β

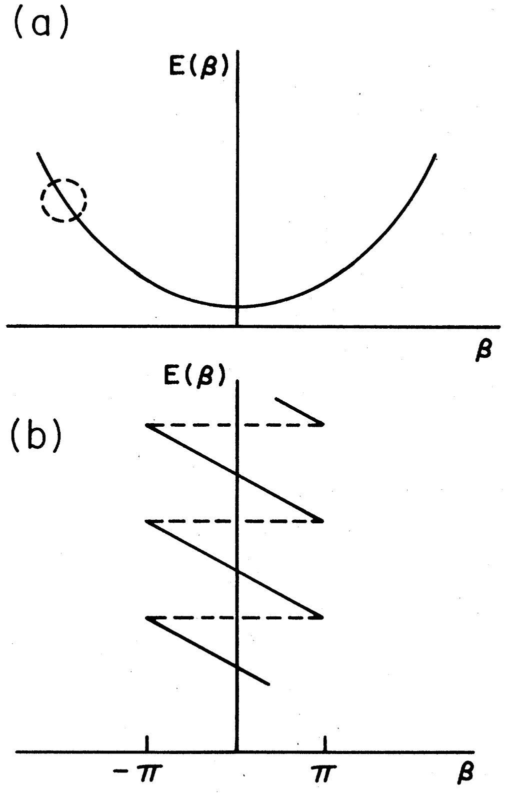

FIG. 10. The total energy of the system as a function of β. At the minimum, the Hall current would be zero by virtue of Eq. (3.7). For finite currents realized in the laboratory, our model requires that the system be very far away from this minimum. In (a), the scale is very large; i.e., the values of β are huge. The small region indicated in (a) is greatly enlarged and folded into a 2π region of β in (b).

are assumed to be stable within the model defined earlier, but are metastable against decay by bulk matter excitation. When B has a value between plateaus, the lifetime of the states becomes very short and the present description breaks down. It also breaks down when the current becomes very large.

Once we have forced the system to be closed, it has many possible states corresponding to different potential differences and their associated currents. They differ primarily in the number of electrons in each of the reservoirs; this asymmetry is responsible for the potential difference between the reservoirs. The true ground state has no potential difference or current. From the point of view of the present paper, it is uninteresting; but in other treatments [see reviews by Pruisken (1986) and Thouless (1986)], it is the object of study. The energy of the states may be displayed in another way, which emphasizes their metastability. A region of the curve of Fig. 10(a) (corresponding to positive current) is magnified in Fig. 10(b), with all states plotted within a 2π range of β. A continuous increase of β by 2π moves the state down one of these line segments where it shifts to the beginning of the next one. The curvature of the lines is negligible as seen on such a plot, and I is practically constant within such a range. This change of β corresponds to the motion of one electron across the sample, as described above.

There are some ways the preceding argument may break down, but the possible resulting effect on the quantum Hall relation is hard to quantify. Away from the center of the plateau, the one-particle states are not necessarily filled. As argued in the preceding subsection, there may be a partially filled region in which these states have the Fermi energy of the high-potential probe. This cannot be an exact statement; there must be small variations in energy away from the mean. Perturbation theory in the residual electron-electron interactions must be suspect in this region. Presumably it produces a superposition of the various possibilities for occupying the one-particle states. In some sense, this region becomes an extension of the probe. Also, away from the center of the plateau, the electric fields in the layer may become large locally, permitting electrons to leak between Landau levels (invoking the bulk matter interactions) or into the low-potential reservoir. This would lead to longitudinal resistivity and a concomitant deviation from the ideal Hall conductivity. In the center of the plateau, this leakage should be exceedingly small, but not necessarily zero. It is suppressed by the very small overlap between wave functions of the same energy in two different levels (typically, their spatial separation is many magnetic lengths).

E. Lowest-order estimate of the Hall current

Now we are prepared to discuss the Hall current in the absence of perturbations. In this approximation, the total energy of the system is given by

$$E^{(0)}(\beta)=\sum_{\alpha}\varepsilon_{\alpha}\ . \tag{3.12}$$

The energy of a localized eigenfunction is independent of β; recall that U_1 is defined for $\beta=0$.

Next we note that for extended eigenfunctions the β and k dependences are linked through Eq. (3.10c). The current then simplifies to

$$I^{(0)}=-\frac{e}{\hbar L}\sum_{k}\frac{\partial\varepsilon_k}{\partial k}\ . \tag{3.13}$$

The sum extends over the filled ground-state Landau levels for any of the situations illustrated in Fig. 9. Since there is no physical discontinuity at relabeling gaps, they present no special difficulty. In fact, we may simply replace k by a new label k' which labels the eigenfunctions without any gaps. Then, as pointed out in Sec. II, the lateral separation between the eigenfunctions in the "two-dimensional" region is much smaller than their internal extensions. Thus the energies vary extremely slowly as a function of k', and it should be legitimate to replace the sum by an integral with a totally negligible error. At this stage, we have

$$I^{(0)}=-\frac{e}{h}\int_{\text{filled}}\frac{\partial\varepsilon_{k'}}{\partial k'}dk'\ . \tag{3.14}$$

Clearly, this integral is just the difference of Fermi energies between the two reservoirs, which is equivalent to the electrochemical potential difference between the two voltage probes. Hence we find

$$I^{(0)}=\frac{e^2}{h}\Delta V\ . \tag{3.15}$$

The replacement of the sum by the integral can be made more rigorous if we note that it corresponds to averaging the expression with respect to β over the range 2π. However, this introduces the new assumption that $U_1(\beta)$ is independent of β over this small range.

It is the aim of the next section and Appendix B to demonstrate that this relationship is not disturbed by taking into account the residual interactions between the electrons. This is done in each order of perturbation theory: a modification of $E(\beta)$ which produces a modification of I is exactly matched by a modification of the difference of electrochemical potentials.

Most of the previous microscopic descriptions of the quantum Hall effect have used basically the sort of argument just given, but with somewhat less generality. Some of these start with the properties of Landau wave functions in a uniform electric field and then study the modifications produced by impurities. As we have seen, it is not necessary to tie the discussion to such an unperturbed starting point. The main requirement is that it be possible to label successive one-particle eigenfunctions by an average wave number k without any physical gap, even if there are relabeling gaps. It is possible that there are *no* regions where the unperturbed wave functions are a good approximation. Others have observed the generality of Eq. (3.13) for a simplified two-dimensional system with an effective electron mass. Chalker (1983) gives a derivation of (3.13) in which the phase shifts produced by impurities are analyzed and related to a change in the current. Joynt and Prange (1984) give a derivation somewhat similar to ours in that it is based on the Feynman-Hellman theorem. The derivation of (3.13) alone can be much simpler than the one given here, but it will be seen in Sec. IV that the definition using β is very powerful for studying the effect of perturbations. The discussion given here is an improvement over earlier ones in that fewer assumptions are made about the physics of the eigenfunctions in the "two-dimensional" region. The concept of effective mass is unnecessary; the presence of atoms is taken into account; features of the potential such as impurities need hardly be mentioned since they are automatically taken into account [however, see the discussion about the potential following Eq. (3.8c)]. The essential features of the single-particle eigenfunctions are that some are localized eigenfunctions independent of β and others are extended eigenfunctions which have a definable phase change from one end of the sample to the other, permitting the definition of an average longitudinal wave number. The wave numbers need not be unique if the eigenfunctions have two-dimensional zeros, but this causes no essential difficulties. Also, the practically continuous behavior of successive eigenfunctions in a relabeling gap region is stressed here.

The replacement of the sum by an integral in passing from Eq. (3.13) to (3.14) of course requires that this description not be distorted by the presence of an excessive number of relabeling gaps. This is a subject for further study. Hard boundaries at the edges of the "two-dimensional" region play a secondary role in this discussion. However, it is possible that they provide a feature of the environment that assists in the establishment of equilibrium with the reservoir as described in Sec. V. It seems closer to the real physical situation to introduce high- and low-potential reservoirs to represent the voltage probes. In the present analysis, there is a realistic potential difference (perhaps of order 100 mV), not an infinitesimal one. Incidentally, it is not possible to produce a known Hall current in a sample simply by applying an external electric field across the sample from the outside. Even with such a field, the extended eigenfunctions could fill up in such a way that they have no net energy difference across the sample and hence no net current. If a Hall current is sent through the sample, the appropriate energy difference will be produced across the sample; but it will have no simple relationship to the applied electric field.

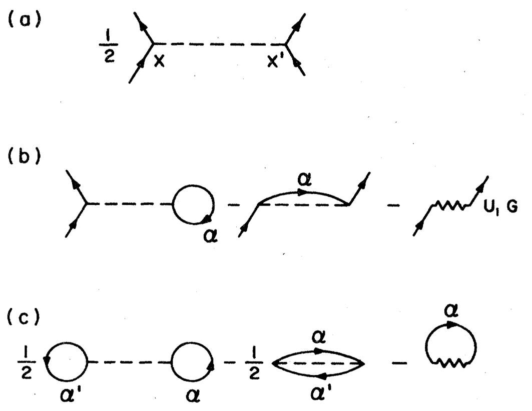

FIG. 11. Graphical elements for the construction of perturbation theory. In (a) is shown the four-operator contribution corresponding to Eq. (4.1). An incoming line represents either a particle state destroyed at the interaction or a hole state created by the interaction; similar remarks hold for the outgoing lines. In (b) is shown the two-operator part of the interaction corresponding to Eq. (4.2). The first term gives a direct contribution and the second an exchange contribution. The third term is the subtracted effective potential. Note that the first two terms have a dependence on β through the eigenfunctions labeled by α, while the U_1 term is calculated for $\beta=0$. (c) shows the simplest contributions to the shift in the reference state energy caused by the perturbation. While these terms do not cancel in the energy, their leading contribution to the current does.

IV. THE SELF-CONSISTENT POTENTIAL AND PERTURBATION THEORY

A. Formulation of the self-consistent potential

When we insert the definition (3.4b) into the perturbation Hamiltonian (3.3c) and rearrange the various terms in normal order, three types of terms are produced. The first has products of combinations of four creation and annihilation operators. It is given by

$$H_2'=\frac{1}{2}\int N[\psi^\dagger(\mathbf{x})\psi^\dagger(\mathbf{x}')U_2(\mathbf{x}',\mathbf{x})\psi(\mathbf{x}')\psi(\mathbf{x})]d^3x\, d^3x' \ , \tag{4.1}$$

where N stands for normal ordering. It is represented graphically by Fig. 11(a).

The second, which is pictured in Fig. 11(b), has two such operators and is given by

$$H_1'=\int N\left\{\sum_\alpha [\psi^\dagger(\mathbf{x})\psi(\mathbf{x})U_2(\mathbf{x},\mathbf{x}')\psi_\alpha^\dagger(\mathbf{x}')\psi_\alpha(\mathbf{x}')-\psi^\dagger(\mathbf{x})\psi_\alpha(\mathbf{x})U_2(\mathbf{x},\mathbf{x}')\psi_\alpha^\dagger(\mathbf{x}')\psi(\mathbf{x}')] \right.$$
$$\left. -\psi^\dagger(\mathbf{x})U_1(\mathbf{x},\mathbf{x}')G(\mathbf{x},\mathbf{x}')\psi(\mathbf{x}')\right\}d^3x\, d^3x' \ . \tag{4.2a}$$

Here we have made U_1 explicitly nonlocal, as is true in the general case. The factor G, which is introduced to assure gauge invariance, is given by

$$G(\mathbf{x},\mathbf{x}')=\exp\left[-\frac{ie}{\hbar}\int_{\mathbf{x}'}^{\mathbf{x}}\mathbf{A}(\mathbf{x}'')\cdot d\mathbf{x}''-i\beta(x-x')/L\right]. \tag{4.2b}$$

It would appear desirable to choose U_1 so that Eq. (4.2a) cancels out. However, to avoid cumbersome complications, it turns out to be advisable to choose U_1 to be independent of β. Therefore we define U_1 in terms of single-particle eigenfunctions defined for $\beta=0$ and for some particular reference gauge $\mathbf{A}_0$. It also proves useful to incorporate some higher-order effects from U_2 into U_1, so we define only the first-order contribution to U_1 at this point:

$$U_1^{(1)}(\mathbf{x},\mathbf{x}')=G_0(\mathbf{x},\mathbf{x}')^{-1}\sum_{\alpha_0}\left[\delta(\mathbf{x}-\mathbf{x}')\int U_2(\mathbf{x},\mathbf{x}'')\psi_{\alpha_0}^{\dagger}(\mathbf{x}'')\psi_{\alpha_0}(\mathbf{x}'')d^3x''-\psi_{\alpha_0}^{\dagger}(\mathbf{x}')\psi_{\alpha_0}(\mathbf{x})U_2(\mathbf{x},\mathbf{x}')\right], \tag{4.2c}$$

where the subscript 0 on G and α refers to the choice of reference gauge and to $\beta=0$. Now H_1' has been constructed so that to first order in U_1 it vanishes when $\beta=0$. This means that when it is used in perturbation theory to work out a contribution to Eq. (3.7), it provides the factor which must be differentiated. Consequently, it can occur only once, since we set $\beta=0$ after differentiating. It might appear that the factor G could as well have been omitted, since the U_1 term must cancel between the two pieces of the Hamiltonian in Eq. (3.3). However, it would then have turned out that the current density defined for the unperturbed part of the Hamiltonian alone would not have been conserved. With G, one can easily identify another piece of the current-density operator which yields a conserved current and where the total current is also generated properly from Eq. (3.7).

Finally, the third is a nonoperator contribution to the energy which must be taken into account in calculating the current. It is pictured in Fig. 11(c) and is given by

$$\begin{aligned}E'=\int\Bigg[&\tfrac{1}{2}\sum_{\alpha,\alpha'}[\psi_{\alpha}^{\dagger}(\mathbf{x})\psi_{\alpha}(\mathbf{x})U_2(\mathbf{x},\mathbf{x}')\psi_{\alpha'}^{\dagger}(\mathbf{x}')\psi_{\alpha'}(\mathbf{x}')-\psi_{\alpha}^{\dagger}(\mathbf{x})\psi_{\alpha'}(\mathbf{x})U_2(\mathbf{x},\mathbf{x}')\psi_{\alpha'}^{\dagger}(\mathbf{x}')\psi_{\alpha}(\mathbf{x}')]\\&-\sum_{\alpha}\psi_{\alpha}^{\dagger}(\mathbf{x})U_1(\mathbf{x},\mathbf{x}')G(\mathbf{x},\mathbf{x}')\psi_{\alpha}(\mathbf{x}')\Bigg]d^3x\,d^3x'.\end{aligned} \tag{4.3}$$

Note that the β dependence still occurs in the functions appearing explicitly here even though it does not occur in U_1. However, when this expression is differentiated with respect to β, which is then set equal to 0, all the terms cancel to first order by virtue of Eq. (4.2c). [The easiest way to see that the β dependence of G causes no difficulty is first to rewrite Eq. (4.3) in terms of the functions ψ' defined in (3.10a).] To first order, this choice of $U_1^{(1)}$ seems to be optimal. Any other choice would be "self-correcting" in the sense that changes induced in the basic single-particle eigenfunctions would be offset by compensating changes in the perturbations due to Eqs. (4.2a) and (4.3). However, that would make the analysis unnecessarily cumbersome. When we turn to perturbation theory (without incorporating higher-order effects into U_1), it will be necessary to take into account at most one order in (4.2a) together with all orders of (4.1).

B. Perturbation theory for a class of contributions

Here we consider only those contributions to the energy $E(\beta)$ which are one-particle—one-hole irreducible. This means that the diagrams representing those contributions cannot be separated into two pieces by cutting two lines. The reason for making this restriction will appear later. Appendix B removes this restriction from the discussion.

Before proceeding to a general discussion of perturbation theory, let us consider the second-order contribution to the energy of the ground state. This will have all the ingredients we need to understand the general case. The two contributions are illustrated by the graphs in Fig. 12. We write out that of Fig. 12(a) explicitly:

$$E^{(2a)}(\beta)=\tfrac{1}{2}\sum_{\alpha,\alpha',\bar{\alpha},\bar{\alpha}'}\int d^3x_1d^3x_2d^3x_3d^3x_4\frac{\psi_{\alpha}'^{\dagger}(\mathbf{x}_1)\psi_{\bar{\alpha}}'(\mathbf{x}_1)U_2(\mathbf{x}_1,\mathbf{x}_2)\psi_{\alpha'}'^{\dagger}(\mathbf{x}_2)\psi_{\bar{\alpha}'}'(\mathbf{x}_2)\psi_{\bar{\alpha}}'^{\dagger}(\mathbf{x}_3)\psi_{\alpha}'(\mathbf{x}_3)U_2(\mathbf{x}_3,\mathbf{x}_4)\psi_{\bar{\alpha}'}'^{\dagger}(\mathbf{x}_4)\psi_{\alpha'}'(\mathbf{x}_4)}{\varepsilon_{\alpha}+\varepsilon_{\alpha'}-\varepsilon_{\bar{\alpha}}-\varepsilon_{\bar{\alpha}'}}. \tag{4.4}$$

We have to evaluate the derivative of this with respect to β, taking into account that we get pairs of equal contributions. For convenience, we have expressed this in terms of the functions ψ' introduced in Eq. (3.10), since these are independent of β for localized states and have simpler derivatives for extended states.

In differentiating the α or α' dependence, we may ignore localized states and use Eq. (3.10c) to change derivatives with respect to β into derivatives with respect to k for the extended states. Thus we find the term

$$\sum_{k}\frac{\partial}{L\partial k}\sum_{\alpha',\bar{\alpha},\bar{\alpha}'}\int d^3x_1d^3x_2d^3x_3d^3x_4\frac{\psi_{k}'^{\dagger}(\mathbf{x}_1)\psi_{\bar{\alpha}}'(\mathbf{x}_1)U_2(\mathbf{x}_1,\mathbf{x}_2)\psi_{\alpha'}'^{\dagger}(\mathbf{x}_2)\psi_{\bar{\alpha}'}'(\mathbf{x}_2)\psi_{\bar{\alpha}}'^{\dagger}(\mathbf{x}_3)\psi_{k}'(\mathbf{x}_3)U_2(\mathbf{x}_3,\mathbf{x}_4)\psi_{\bar{\alpha}'}'^{\dagger}(\mathbf{x}_4)\psi_{\alpha'}'(\mathbf{x}_4)}{\varepsilon_{k}+\varepsilon_{\alpha'}-\varepsilon_{\bar{\alpha}}-\varepsilon_{\bar{\alpha}'}}. \tag{4.5}$$

The inner sum has an obvious interpretation. It is a contribution to the self-energy of a hole, with sign reversed because of the anticommutation relations, as represented by Fig. 13(a). (Note that the internal lines of the self-energy are independent of k when the differentiation is carried out.) A similar contribution arising from differentiating the contribution of Fig. 12(b) is shown in Fig. 13(c).

The terms arising from differentiating the $\bar{\alpha}$ or $\bar{\alpha}'$ dependence do not have such a direct interpretation and require a special trick to rearrange them. Consider the relevant factors containing the $\bar{\alpha}$ dependence in Eq. (4.4). It takes the form

$$\sum_{\bar{\alpha}} \frac{\psi'_{\bar{\alpha}}(\mathbf{x}_1)\psi'^{\dagger}_{\bar{\alpha}}(\mathbf{x}_3)}{K-\varepsilon_{\bar{\alpha}}} F(\mathbf{x}_1,\mathbf{x}_3) \ , \tag{4.6a}$$

where we have suppressed some of the dependence that is irrelevant for the present discussion. Now we take the following steps: Rewrite the sum in Eq. (4.6a) as a sum over all states minus a sum over filled states. In the sum over all states, we replace the energy in the denominator by the single-particle Hamiltonian h^0_1 acting on the eigenfunction and then use closure to find

$$\frac{1}{K-h^0_1}\delta(\mathbf{x}_1-\mathbf{x}_3)F(\mathbf{x}_1,\mathbf{x}_3) \ . \tag{4.6b}$$

With certain cautions to be described in a moment, the resulting operator is independent of β, and the term vanishes when differentiated (recall that the β dependence in F is treated elsewhere). Again dropping localized-state contributions and using Eq. (3.10c) for the extended states, we find that the subtracted term becomes

$$-\sum_k \frac{\partial}{L\partial k} \sum_{\alpha',\alpha,\bar{\alpha}'} \int d^3x_1 d^3x_2 d^3x_3 d^3x_4 \frac{\psi'^{\dagger}_{\alpha}(\mathbf{x}_1)\psi'_k(\mathbf{x}_1)U_2(\mathbf{x}_1,\mathbf{x}_2)\psi'^{\dagger}_{\alpha'}(\mathbf{x}_2)\psi'_{\bar{\alpha}'}(\mathbf{x}_2)\psi'^{\dagger}_k(\mathbf{x}_3)\psi'_{\alpha}(\mathbf{x}_3)U_2(\mathbf{x}_3,\mathbf{x}_4)\psi'^{\dagger}_{\bar{\alpha}'}(\mathbf{x}_4)\psi'_{\alpha'}(\mathbf{x}_4)}{\varepsilon_\alpha+\varepsilon_{\alpha'}-\varepsilon_k-\varepsilon_{\bar{\alpha}'}} \ . \tag{4.7}$$

This term is illustrated in Fig. 13(b). It is the derivative of the negative of another second-order contribution to the self-energy of a hole. The diagrams associated with Fig. 12(b) are shown in Figs. 13(c) and 13(d); they complete the set of second-order contributions to the self-energy of a hole.

It is easy to see that the same procedure may be applied to any one-particle—one-hole irreducible contribution. The only change from Eq. (4.6) is that the energy $\varepsilon_{\bar{\alpha}}$ may occur in several denominators. Diagrams involving the interaction (4.2a) also make their proper contributions. Extension of the argument to incorporate multiparticle interactions is also obvious.

At this point, we state the general result that this result exemplifies. Suppose we calculate the nth-order contribution $E^{(n)}$ to the energy of the ground state and do the same thing for the energy of a hole, which we call $-E^{(n)}_k$. Then their derivatives are related by

$$\frac{\partial E^{(n)}(\beta)}{\partial\beta} = \frac{1}{L}\sum_k \frac{\partial E^{(n)}_k}{\partial k} \ . \tag{4.8}$$

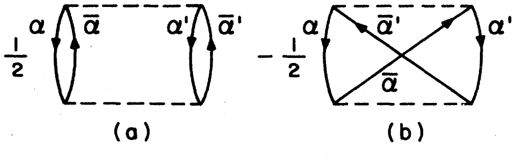

FIG. 12. The two second-order contributions to the energy of the reference state. (a) is the direct term and (b) the exchange term.

Replacing the sum by an integral, we find that the right-hand side can be replaced by $(E^{(n)}_{k_{\max}} - E^{(n)}_{k_{\min}})/2\pi$. This is our objective. To any order of perturbation, the current and the difference of chemical potentials are modified in the same way.

The form of the result (4.8) is important; it gives a useful relationship which is valid whatever the size of the effect, so that it is not necessary actually to calculate corrections. In an early stage of this work, the author attempted to estimate the perturbative corrections to the

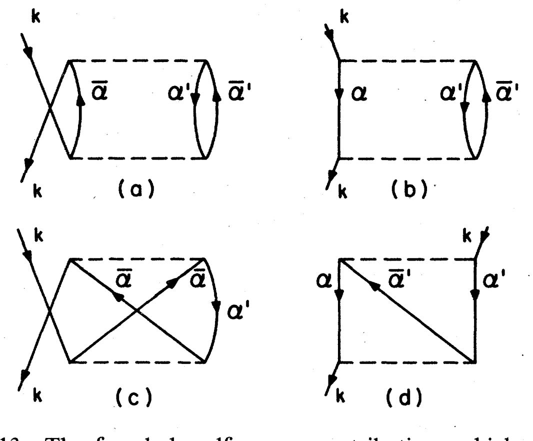

FIG. 13. The four-hole self-energy contributions which arise from differentiating the contributions from Fig. 12. (a) and (c) arise from differentiating the dependence of the α or α' lines. (b) and (d) arise from differentiating the dependence of the $\bar{\alpha}$ or $\bar{\alpha}'$ lines, after a mathematical transformation.

current and found that they are not characterized by a particularly small expansion parameter. This will now be described very briefly. If two pairs are excited by the perturbation (the careful definition of the self-consistent potential avoids single-pair excitation), each electron in an excited state gives a current contribution $\partial\varepsilon_{\bar{k}}/\partial\bar{k}$, while the absence of an electron from the original state gives a contribution $-\partial\varepsilon_k/\partial k$. These are to be multiplied by the probability of the double-pair excitation, of course. These contributions can easily be identified by differentiating the energy denominator of (4.4). The original attempt did not incorporate the other terms from differentiating the wave function or recognize that the whole thing was a perfect derivative. Using a very simplified model for the interaction and the wave functions, it is not hard to estimate the order of magnitude of the correction. It turns out that the current correction for a given k has cancellations, and the estimate gives

$$\delta I_{k_0} \approx -\frac{1}{256}\left[\frac{m^* l}{m\kappa a}\right]^2 l^2 \frac{\partial^2 I_k}{\partial y_k^2} \ ,$$

where I_k is the current of an eigenfunction labeled k and δI_{k_0} is the modification of the current for pairs whose mean wave number is k_0 $(=k)$. The Bohr radius a is introduced as a convenient length scale, κ (≈ 10) is the dielectric constant, and m^*/m (≈ 0.1) is the ratio of the electron's effective to its actual mass. The combination $m^* l/m\kappa a$ is fortuitously approximately 1. This estimate does not include the effects of exchange and finite layer thickness, which reduce the result somewhat. The fact that this is also a perfect derivative should not be taken seriously because of the crudity of the model; presumably, the fact that it is larger when ε_k varies erratically could have significance. It is helpful that the numerical coefficient is somewhat small. The net effect, though, is that one might expect corrections on the order of 0.1%, which are much too large for one to be hopeful about making a systematic perturbative calculation valid to better than 1 part in 10^7. On the other hand, they may be small enough to make us optimistic about the possible convergence of the expansion so that the validity of Eq. (4.8) order by order encourages us to believe that the complete set of corrections does not change the ideal quantum Hall relation.

Now we have to return to a subtlety that was glossed over when we dropped the derivative with respect to β of a sum over a complete set of states, which led to an expression apparently independent of β. The difficulty with this is that the operator still acts in a Hilbert space which incorporates the β-dependent boundary conditions. Thus the legitimacy of dropping this term depends on the context in which it is to be used. If the remaining factors to be summed and integrated over evaluate the expression in a local way—i.e., one that is not sensitive to the boundary conditions—then the neglect of the derivative is justified. However, if the remaining factors can be sensitive to conditions at opposite ends of the sample, then the dropping of this term could be unjustified. For the contributions discussed in this section, there seems to be no problem. However, if we start with a one-particle—one-hole reducible contribution, as illustrated in Fig. 14(a), and blindly follow the procedure given here, we obtain one-hole reducible contributions to the hole energy as illustrated in Figs. 14(b) and 14(c). One of these [Fig. 14(c)] can have a small denominator corresponding to the hole propagating a long distance through the medium between self-interactions. The fact that these are sensitive to the boundary conditions is clear, and our present analysis is not justified for them. To take them into account, we need a more sophisticated procedure in which U_1 is modified to make the hole self-energies vanish on the energy shell. Such a procedure is described in Appendix B.

V. SOME REFINEMENTS

Let us recall quickly the main features of our model. We assume that the "active" electrons move in an external potential provided by the bulk matter and that their interactions with each other may be influenced by the bulk matter. We also assume that the complete eigenstates of the energy, like those of the unperturbed problem with a self-consistent potential, can have a finite Hall current. We have invoked the existence of interactions with bulk matter excitations to permit each voltage probe, which serves an an electron reservoir, to come to an equilibrium internally with some Fermi energy. These interactions also permit the complete quantum state to relax to the most stable one for a given value of the Hall current. Other than that, interactions with bulk matter excitations are outside of the model Hamiltonian. In the "two-dimensional" region, we have (almost) a pure quantum state, provided the temperature is sufficiently small.

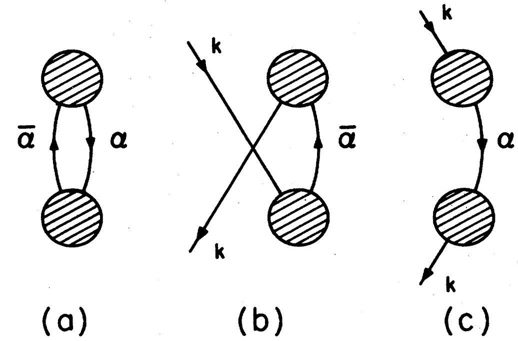

FIG. 14. (a) shows a contribution to the self-energy of the reference state, which is one-particle—one-hole reducible. When differentiated, it produces, among other terms, the contributions of (b) and (c). Contribution (b) arises from differentiating the α dependence of (a). Contribution (c) arises from the $\bar{\alpha}$ dependence by differentiating and making a mathematical transformation. In this case, the mathematical transformation is unjustified.

This state is the lowest energy state that carries the specified current.[15]

This section deals with a number of topics that seem helpful for a detailed understanding of the physics of the quantum Hall devices. Because of the great variety of devices and situations, we try to emphasize a few features that might be more general. Except for the last topic, which has to do with nonminimal electromagnetic coupling, these topics are concerned mainly with the nature of the metastable quantum state in the plateau region, and how that metastability might break down between plateaus. Since we are more interested in enlarging our intuitive understanding than in giving a rigorous development, this discussion is in terms of single-particle eigenfunctions. The residual electron-electron interactions make the actual quantum state very complicated, but we assume that the concept of an electrochemical potential within each reservoir is still meaningful. In any case, the work of Appendix B shows that the effective potential in principle takes into account those residual interactions in the definition of the single-particle eigenfunctions. We assume that the relaxation to states of lower energy can be understood in terms of single-particle transitions in which the initial and final single-particle states have some overlap in space.

A. Speculations on the equilibrium at the edge

In our discussion up to this point, we think of the two probes and the "two-dimensional" layer as three nearly independent dynamical systems aside from the assumption that they come to an equilibrium in which the single-particle energies at the edge of the layer must match the Fermi energy within the nearby probe. A probe is a three-dimensional conducting region in which the electrons establish an equilibrium with all one-particle states filled up to the Fermi energy.

Our present aim is to give a discussion of the equilibrium between extended eigenfunctions that pass near a reservoir and those confined to that reservoir. Strictly speaking, there is no sharp break between the two types of eigenfunctions. Certain eigenfunctions extend from one end of the sample to the other, yet have a significant probability of being inside a reservoir. We may think of the eigenfunctions as tunneling from one region to the other, even though there may not be a potential barrier in the ordinary sense. We refer to these as *transitional eigenfunctions.* In the following, we examine their general properties without attempting an explicit calculation of individual eigenfunctions. The potential that acts on these eigenfunctions must be very complicated, since it involves both the probe and the edges of the sample. Since we are concerned primarily with understanding the qualitative features of the eigenfunctions, it seems reasonable to ignore some of the complications that were discussed previously, such as the possible presence of relabeling gaps and nonlocality of the effective potential. However, our treatment must be three dimensional and it need not ignore the microscopic structure such as the presence of individual atoms. Where necessary, we invoke the exclusion principle, which requires that an extended eigenfunction not penetrate significantly into a reservoir unless its energy be at or above the Fermi energy of the reservoir; technically, this might be incorporated in the nonlocal structure of the effective potential.

In Sec. III arguments are given that an approximate measure of the average transverse displacement of an extended eigenfunction is given by Eq. (3.9e), which we repeat here for convenience,

$$y_k = kl^2 . \tag{5.1}$$

In case there are two-dimensional zeros, this expression is associated with a particular path through the sample. It gives no indication whatever of where the eigenfunction is located at a particular value of x. That location depends on the phase dependence and is given approximately by Eq. (3.9d). Moreover, eigenfunctions with the same k for different levels (defined relative to the same overall path) may pass through somewhat different parts of the sample.

In contrast to the extended eigenfunctions, the eigenfunctions confined to the reservoirs are in localized states. Because this region is three dimensional, there are many states of motion in the z direction, and states of all types are filled up to the Fermi energy of the probe. These eigenfunctions may extend into the semiconductor region where they would generally be exponentially suppressed with distance except possibly where a peak (as a function of z) in the eigenfunction happens to coincide in position with the potential well which provides the channel for the Hall current. There, if the energy of an eigenfunction has the right value, it may be possible for it to extend significantly into the "two-dimensional" region. The fact that it has more nodes in the z direction than normal "two-dimensional" eigenfunctions should be inconsequential, since the eigenfunction is already strongly damped wherever those nodes occur.

As y_k is increased, successive eigenfunctions move transversely across the sample until they come into contact with the probe or the edge of the sample, or both. Various situations may occur, depending on the relationship between ε_k and the Fermi energy of the probe. If $\varepsilon_k > E_F$ for a filled state close to the probe, that state can decay into the probe by exciting the bulk matter. This

[15] Here we ignore the complications of the fractional quantum Hall effect, but make a few remarks about the situation in this footnote. Because of the residual interactions between the electrons, it is possible that several different states of very different nature might carry the same current but have different electrochemical potential differences. Presumably, the interaction with bulk matter excitations would result in the experimental realization only of the one having the lowest energy, which under some circumstances would correspond to the fractional effect.

would correspond to a current flow across the sample, which is not possible for a metastable state. Hence, in the plateau region, there cannot be filled states in the nearby "two-dimensional" region whose energy is more than E_F. "Nearby" means that they have sufficient spatial overlap with the probe that they would be able to decay quickly into it. In practice, this means that filled eigenfunctions that pass within a few magnetic lengths of the probe must have $\varepsilon_k \leq E_F$. We have argued in Secs. II and III that eigenfunctions near the high-potential probe need not be filled provided that their energy is precisely the Fermi energy of that probe. If their energy is higher, they will decay into the probe; if it is lower, they will be filled from the probe.[16] In the metastable states, this partial filling is not expected near the low-potential probe because there are always filled states of higher energy in the "two-dimensional" region which are available to fill them.

Next, let us consider qualitatively the behavior of the transitional eigenfunctions. The main thing that we must assume is that the quantity y_k retains some qualitative validity in this region; that is, two eigenfunctions that have different values of y_k along the same general path will have some overall transverse displacement in the direction indicated by y_k. Suppose then that as y_k increases the first eigenfunctions that overlap the probe slightly have $\varepsilon_k < E_F$. Since they must remain orthogonal to the localized eigenfunctions in the probe, their probability to be inside the probe should be small. Unless ε_k can approach E_F, these eigenfunctions cannot move very far inside of the probe as y_k increases. Therefore the eigenfunctions must move toward the edge (in a region away from the probe) as y_k increases. As they come into contact with the edge, their energy can begin to increase with y_k. As they are in contact with the probe also, all states satisfying $\varepsilon_k \leq E_F$ must then be filled. (In detail, the self-consistent potential may be important in this process.) But since the edge region can support states of arbitrarily large energy on the scale of interest (by pushing y_0 beyond the actual boundary if necessary), this means that eigenfunctions in contact with the probe will fill up precisely to E_F.

This establishes the main assumption we had to make in Sec. III. Figure 15(a) illustrates this for some representative eigenfunctions. The one labeled α is in contact with the probe, but not the edge. After a sufficient increase of y_k, the one labeled β has moved over to the edge but not advanced appreciably into the probe. Finally, with a further increase of y_k, the one labeled γ has pushed sufficiently into the edge region that its energy rises close to the Fermi energy of the probe. We regard this as a possible, but not necessarily the only, scenario for the extended states to be in equilibrium with the probe. The essential point is that the extended region has available single-particle states whose energy spans the Fermi energy of the probe. Provided a mechanism exists (which I have tried to illustrate), the extended states must fill up to that Fermi energy.

It is also of interest to know whether ε_k approaches E_F smoothly as a function of y_k; this could be related to the question of accuracy through the step of replacing the

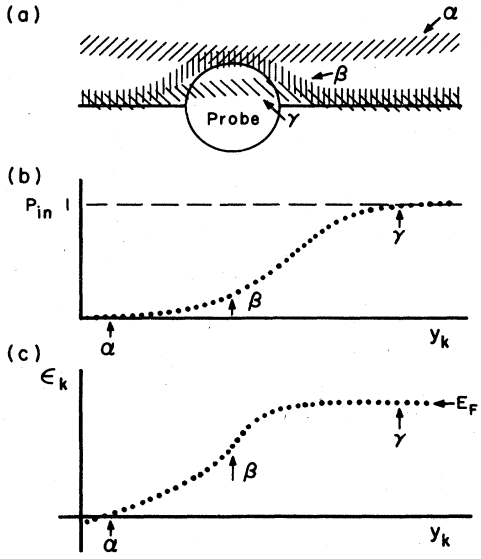

FIG. 15. (a) The possible behavior of eigenfunctions near the probe and sample edge as the average wave number k (or equivalently $y_k = kl^2$) is varied. In the situation shown, the single-particle energies ε_k approach the Fermi energy of the probe from below. The eigenfunction labeled α is for a y_k such that the eigenfunction overlaps the probe very slightly. With an increase of y_k, an eigenfunction labeled β lies close to the edge, but still penetrates the probe only slightly. With further increase of y_k, the eigenfunction labeled γ passes through the probe and has pushed further into the edge of the sample. (b) The probability for an eigenfunction to be *inside* the probe plotted as a function of y_k; the region for calculating the probability extends a short distance outside the probe to include tails of the eigenfunctions. Each dot represents one eigenfunction; a realistic figure might include of the order of 10^5 dots. (c) A plot of the eigenvalues ε_k as a function of y_k.

[16] The preceding two sentences cannot be exactly true, since they do not take into account the self-consistent potential. It is highly unlikely that the single-particle energies are exactly degenerate in this region. Because of the near degeneracy, the residual interactions and thermal fluctuations are likely to be relatively important. This should result in some fluctuations about the ideal quantum Hall relation, but not enough is known about the nature of the states to estimate the size of such fluctuations. If the partial filling were to approach certain fractions, the residual interactions would lead to states of lower energy, and a transition to the fractional quantum Hall effect could occur (Laughlin, 1983). Barring such behavior, Laughlin's argument would appear to suggest that these fluctuations tend to average out over a range of β of 2π. To the extent that localized states play a role as reservoirs, they would tend to prevent a partially filled region and hence help avoid this question. If the general picture being presented here is basically correct, this point bears further investigation.

sum by an integral in Eq. (3.14). This is an interesting question which deserves study. We describe here how it is possible that this approach might be smooth, even though ε_k would be a very steep function of $y_0(x)$ near the edge. We need only consider the case where $\varepsilon_k < E_F$ for an eigenfunction that first makes significant contact with the probe. As y_k increases further, the energy will not change much until the eigenfunction approaches the edge; i.e., at this stage, y_0 varies more rapidly with y_k in the edge region than in the probe. At that point, ε_k will start to increase and the eigenfunction will be able to enter the probe and increase its probability there. This will decrease the probability for the wave function to be in the "two-dimensional" region [i.e., R^2 in Eq. (3.9) gets smaller there]. Since y_0 is increasing both in the probe and elsewhere, the change of ε_k with y_k is smaller than if it increased only at the edge. As the probability to be inside the probe becomes more and more significant, the value of R^2 along the edge gets smaller and the current carried by an eigenfunction decreases also (in spite of the fact that it is in an electric field which is effectively larger as the eigenfunction is pushed toward the edge). From Eq. (3.13), this implies $\partial\varepsilon_k/\partial k \rightarrow 0$ as y_k increases. The argument just given is intended to be suggestive; the actual dependence has not yet been determined.

If this picture of the transitional eigenfunctions is correct, we would expect that the probability that an extended eigenfunction is inside the reservoir gradually increases as y_k moves toward its limiting value, finally approaching unity, corresponding to a state that is completely localized there. In defining this probability we should use a volume slightly larger than that of the actual conductor, so as to include the exponentially decreasing probability distribution outside the conductor. The probability that an extended eigenfunction is inside this volume is indicated schematically in Fig. 15(b), where it is assumed that $\varepsilon_k < E_F$ as the probe is approached, as in Fig. 15(a). Each dot represents one such eigenfunction. As the probability that an eigenfunction is inside the conductor increases, the current carried by that state also decreases unless the local velocity in the "two-dimensional" region increases sharply in a compensating way; as described above, we assume this does not happen. The energy should then behave as a function of y_k as illustrated in Fig. 15(c).

B. Mechanism for higher plateaus

In regions away from the edges, higher Landau levels for a given k are separated in energy from the ground level by approximate multiples of $\hbar\omega_c$, where ω_c is the cyclotron frequency ($\equiv eB/m^*$, where m^* is the effective mass). Although the cyclotron frequency depends on the effective mass, we need not be concerned about its precise description. At the larger magnetic field strengths, this energy separation is of the order of 10 meV. Except for the case of valley degeneracy, the excitation energy of higher sub-bands is of the same order of magnitude but probably a little larger. The spin interaction with the magnetic field is typically much less, both because the Landau excitation energy is enhanced because of the small effective mass and because the g_s factor of the electrons is reduced in the medium of the sample. It should be clear from our prior discussion that the quantum Hall effect does not require that the level separations be independent of k. In fact, the existence of the higher plateaus indicates that several different types of states can be in equilibrium with the reservoir at the same time. How does this happen? First it should be clear that all these types of states exist inside the reservoir and have the same Fermi energy. Of course, inside the reservoir, there is no distinction such as sub-band excitation, but, nevertheless, it should be possible for "two-dimensional" states with that type of excitation to extend into the reservoir.

The highest level that is in equilibrium with the probe must have its ε_k approach E_F from below or be very flat as a function of y_k. Far from the edges and probes, we expect all levels to follow a similar path through the sample for a given value of y_k. The energy difference between different levels should be approximately independent of y_k within the interior of the "two-dimensional" region, but it is not necessary for our considerations that this be exactly constant across the sample. For example, a not perfectly homogeneous magnetic field or sample properties which cause ω_c to depend on position do not disturb our conclusions. However, transitional eigenfunctions may have very different paths through the sample for a given value of y_k, and their energy separations may change very dramatically.

In order to have lower states in equilibrium with the reservoir, it is then necessary that they rise very rapidly in energy near the reservoir before turning over to match the Fermi energy. We have seen how this is possible in the previous subsection. As changing y_k causes the states to reach the reservoir and enter it, the highest state will approach the Fermi energy first in the manner described in the previous subsection. Lower-lying states will be unable to enter the reservoir for the same value of y_k. As y_k increases, these eigenfunctions bend around to avoid the reservoir and are pushed toward the edge of the sample elsewhere. This causes their energy to rise until they can come into equilibrium with the Fermi level inside the reservoir. Thus two energy levels (or more) can merge at a reservoir. This can happen because the paths of the eigenfunctions through the sample can be different for different excitations. An example is shown for two levels in Fig. 16(a).

If this merging of energy levels happens only at the low-potential reservoir, the second Landau level may be partially filled as illustrated in Fig. 16(b). This corresponds to the situation in which the filling factor is increased from 1 slightly and the change of effective potential is negative. The availability of the second Landau level to accept electrons prevents a large negative change of the effective potential. It is possible for the second Landau level to fill in such a way that there is no energy difference between its two limiting values of y_k. (Of

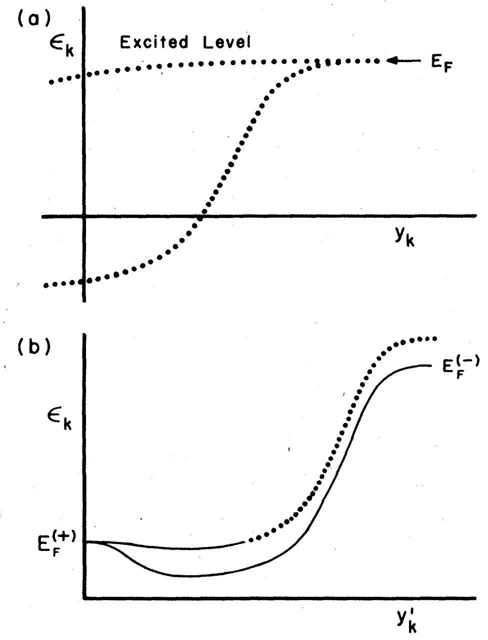

FIG. 16. (a) The energies of two different Landau levels near a voltage probe. When both states with the same y_k are far from the probe, their energy difference is approximately $\hbar\omega_c$. Choosing a condition such that the higher level goes rather smoothly to the probe's Fermi energy as y_k is varied, we find that the lower level would tend to come in below that level. Since it must remain orthogonal to filled eigenfunctions within the three-dimensional region, the associated states take a different path through the sample and ultimately come into equilibrium with the same Fermi level. (b) Illustration of the possible role of a higher Landau level in maintaining a plateau as the magnetic field is decreased; vertical and horizontal scales are reduced as compared to (a). The higher level can accept some electrons near the Fermi energy of the (+) terminal without carrying a net current. The dotted portion of the higher level shows the energies of unoccupied states.

course, this possibility may be inhibited because the relaxation time required to establish equilibrium might be very large.) If this condition is established, the second Landau level should give no contribution to the current from its term in (3.14) (of course, the replacement of the sum by the integral may introduce a more significant error in this case). This is how the second Landau level could conceivably contribute to the ground-state plateau behavior, as mentioned in Sec. III. As the gate voltage or magnetic field is changed further, the second Landau level begins to carry a net current, causing a transition to the next plateau in which it finally comes to equilibrium with the high-potential probe.

C. Criteria for the breakdown of metastability

What is the physical distinction between the plateaus and the transition region between them? Here we must distinguish between our model, which we hope gives a satisfactory description of the physics of the plateaus at zero temperature, and the real physics of the sample, which would be necessary for a description of the dissipative processes that take place away from the plateaus. Our overview is that, if the extended states of a given level in the current-carrying region can be kept filled, there will be no dissipative collisions and no current flow across the sample from one voltage probe to the other. Under these conditions, we believe it should be possible to treat the physics with the idealized quantum states that have been described in this paper. It is a feature of these states that they are metastable, as has been emphasized repeatedly. In the actual sample, the system is constrained not to have a current flow between the voltage probes. If the extended states in a region where ε_k varies with k cannot be kept filled for some reason, there could be current flow across the sample by phonon emission. Because of the constraint, the resulting transverse current flow must be balanced by a Hall current produced by a potential drop along the sample. Thus the breakdown of metastability is directly related to the existence of longitudinal resistance.

In the model Hamiltonian, the system can lower its energy (by exciting the bulk matter) whenever electrons can work their way across the sample from the high-potential reservoir to the low one. If only one level is involved, and if all the (one-particle) states of that level in the "two-dimensional" region are kept filled, it is very difficult for an electron to jump all the way across the sample. However, what is to prevent an electron near the low-potential reservoir from moving into the reservoir to create a hole, which can then work its way over to the high-potential reservoir to produce the same effect? If the energy level is rising as the (+) probe is approached, this could happen only if the electron could absorb energy from the bulk matter. Ultimately, as the electrons cascade down to fill the hole created, more energy would be returned to the bulk matter than was originally required to start the process. This could not happen at zero temperature as a succession of real processes, but it might be a significant effect at finite temperatures, or it might occur as some high-order virtual process. The small size of the longitudinal resistance might be regarded as indirect evidence that, in order to maintain the metastability, the energy levels are either very flat in y_k or rise slightly as the low-potential reservoir is approached.

As the magnetic field is increased or the gate voltage is changed to decrease the filling factor, we have argued that the energy-level dependence on y_k should become very asymmetric across the sample, with a very precipitous drop near the low-potential probe. This means that as y_k is varied to go away from that probe, the energy level must ultimately rise by the Hall voltage within a small fraction of the total width; in some circumstances, this is of order 100 meV. If a potential rise exceeding the Landau and/or sub-band excitation energy gets close enough spatially to the probe, there should be a breakdown of our picture resulting in current flow across the sample. We may refer to this as a *leakage* into higher levels. While this seems qualitatively plausible, no detailed calculations have been made to find out at what point this happens, and it would be expected to be very sensitive to the geometric details of the sample. It seems likely that there would be a considerable difference in the onset of the breakdown, depending on whether the probe is imbedded directly in the "two-dimensional" region or is off to the side on a tab. In the latter case, the three-dimensional probe would be more isolated from the rapid change of energy with y_k and the plateau might persist to larger changes in the filling factor. This mechanism for the breakdown, like other ones, would also make the plateau width smaller for larger Hall currents.

There are other possible mechanisms for a breakdown. For example, there can be variable-range hopping of electrons from one localized state to another (Tsui *et al.*, 1982a; Ebert, von Klitzing, Probst *et al.*, 1983; Wysokinski and Brenig, 1983). We have tentatively guessed that this and exchange of electrons between extended and localized states might tend to fill the localized states up to the Fermi energy of the highest-potential reservoir. But, obviously, localized states that are too far from that probe to have any significant spatial overlap with it could be filled to some other level which depends more on the local environment. Any filled extended (one-particle) states that overlap a localized region will come to equilibrium with it ultimately and fill it to some level that may vary across the sample, corresponding to a Fermi level that varies with position across the sample. Locally this level should not fall below the energy of nearby extended states. The electrons in localized states might be able to move across the sample by hopping from one localized region to another or by jumping to unfilled higher Landau or sub-band states. Once an electron is in a higher level, it could move easily across the sample and ultimately drop into the lower-potential probe. Meanwhile, electrons from the original filled extended states would be dropping into the localized states to try to maintain them at the preferred level. The resulting holes would also contribute to the current flow across the sample and the breakdown. This mechanism would clearly be enhanced as the filling factor is changed in such a way that a precipitous voltage drop occurs in a small transverse region.

We have noted previously that possibly many Landau levels can lie in the energy interval between the Fermi energies of the two probes. Then another mechanism not requiring localized states is that, when the effective potential changes rapidly with y, different Landau or sub-band levels of the same energy may begin to have a good spatial overlap with each other. For example, if $l\partial\varepsilon_k/\partial y_k \approx \hbar\omega_c$, states in the second Landau level will have a large overlap with states of the same energy in the ground level. This corresponds roughly to the situation in the semiclassical picture in which the drift velocity is comparable to the velocity of the cyclotron motion. This condition would require a very large potential gradient, of order 10^4 V/cm. This is much larger than typical applied voltage gradients. Even where a breakdown is caused by imposing large currents, the average gradients are only of order 10^2 V/cm (Cage *et al.*, 1983; Ebert, von Klitzing, Ploog, and Weimann, 1983; Kuchar *et al.*, 1984).

Inside the sample, the voltage gradients may be much larger than the average across the sample. In our intuitive discussion of Sec. II, it was remarked that an increase of 1% of B from its value at the "center" of the plateau might cause the current-carrying width to shrink to 10^{-3} of its original size. The electric field in that region could then easily be 10^3 V/cm. A similar large increase of the electric field would take place with a decrease of B. This local gradient would be a very sensitive function of the filling factor and the geometrical arrangement. Impurities might have a similar effect. Near them, it should be possible for states from two different levels of nearly the same energy to approach each other in space. Even with a much weaker overlap than is suggested by the above criterion, the breakdown might be significant because an electron in the lowest Landau level can decay into any of several states in a higher level, and there may be a numerical enhancement even though the overlap to any individual level may be small. Also, once an electron has decayed to a higher level, it is very quickly swept away by the field, and that may enhance the effect.

Another mechanism that has been mentioned in the literature is that there may be a kind of Cherenkov effect when the drift velocity of the electrons exceeds the velocity of sound in the medium (Heinonen *et al.*, 1984; Středa and von Klitzing, 1984).

D. Open versus closed

The previous discussion sets forth my belief that, for purposes of discussing the plateaus, our model closed system should be adequate. It may also indicate some causes of the breakdown that occurs between the plateaus. It has some awkward features which have been swept under the rug in the previous discussion. The periodic boundary condition has been employed so that we may isolate the sample from a full treatment including the external world. This seems very plausible, and it appears that the main role of this condition is that it permits us to treat the single-particle states as discrete and countable rather than with a continuum normalization. It is also true that when the electrons are not inside the sample they are subject to incoherent inelastic collisions, so that our attempt to treat the problem as describable by a pure quantum state would be totally wrong. Perhaps it is possible to give a treatment that is better in principle, but it is hard to see how such a treatment would improve our actual understanding of the physical mechanisms involved. Provided that the dynamical effects of the bulk matter are suppressed within a sufficiently large region of the sample because of the exclusion principle for the active electrons, we expect that the behavior in that region will be similar to that of a pure state.

There are probably some technical mathematical problems that have been ignored in the use of the periodic boundary condition. We have assumed that since the localized wave functions decay exponentially with distance with a characteristic length l, the periodic boundary condition has negligible impact on them. Localized states that lie near the ends of the sample necessarily occur at both ends, due to the periodic boundary condition. I have not investigated whether this upsets the logic of the perturbation discussion; but if it does, I would regard it as an unphysical artifact which should be ignored. Nevertheless, it may be necessary to understand some of these questions if the question of accuracy (dependence on sample size) is to be properly understood.

E. Laughlin's argument revisited

The connection of the discussion of the equilibrium in the edge region with Laughlin's (1981) argument is straightforward. As β changes by 2π, each of the dots in Figs. 15(b) and 15(c) moves over to the next position. Let us say that the signs are such that this is to the right in these figures. Then the system of dots is restored to its original configuration, but one dot has moved off the end and the associated eigenfunction has been redefined to be a reservoir eigenfunction. At no point is there a real physical discontinuity, since the probability for the electron to be inside the reservoir is already unity well before a dot reaches the point where this redefinition occurs. Another way to look at this is to observe that a large number of eigenfunctions individually change their probabilities to be in the reservoir by a tiny amount, but the net result is that there is one additional electron charge in the reservoir. The energy change can be viewed in a similar way. The total energy of the particles in extended eigenfunctions has not changed, but the system has changed its energy exactly by the amount required to remove an electron at the Fermi level in one probe and add an electron at the Fermi level in the other probe.

While the ideas of this paper were being formulated, a number of individuals raised some interesting questions, which are dealt with here. They illustrate the power of the Laughlin argument.[17] Originally, I assumed an effective Hamiltonian that separated the physics into two parts, one consisting of the "active electrons" that reside in the reservoirs and in the "two-dimensional" region and that are not bound to atomic sites, and the other consisting of electrons included as part of the bulk matter, whose dynamics are suppressed except for the external potential that the bulk matter provides for the active electrons. The bulk matter could also influence the interaction between the active electrons, e.g., through a dielectric constant, but none of its dynamical variables occurred in the Hamiltonian describing the active electrons. The same type of criticism can be leveled at this starting point as we leveled earlier against the use of effective mass and literal reduction of the physics to two dimensions. While our model treats the physics more realistically, there is still the question of whether it is adequate to deal with an accuracy of 0.1 ppm. The fact that we have not found any deviation from the precise quantum Hall relation within our model does not logically mean that our conclusions remain valid for a more correct Hamiltonian.

Naively, one can see why the charges bound in the bulk matter should not affect the current. It is true that locally around a moving electron the total charge, including bound charges, is less than the physical electron charge. Therefore the expression for the current density, which is proportional to the physical charge, is incorrect. However, we calculated the total current as the integral of the x component of the current density over the whole material divided by its total length. The total current density is given by the expression we used plus the contribution from bound electrons. Since the volume integral of the latter expression must vanish, our result should remain correct. The electric charge also enters in the electrochemical potential difference. Again, it is correct to use the physical electron charge because that is the meaning of the energy required to move an electron from one conductor to another through an electrochemical potential difference. Nevertheless, this discussion shows that the bulk matter is involved dynamically in the problem, and therefore it requires an extension. For example, it is conceivable that, in suppressing the dynamics of the bulk matter, the resulting effective Hamiltonian for the active electrons contains types of terms that are not adequately treated by our discussion. Perhaps the kinetic energy term in the Hamiltonian is more complicated or perhaps there occur nonminimal couplings. I have a suspicion that, for any such changes, the same results could be attained with a necessary elaboration of the arguments that have been given (as, for example, in the following subsection), but it is probably better to try to deal with the complete system.

Consider the complete system of positively charged nuclei and electrons and its quantum states. The Hamiltonian (3.2) is generalized to include the kinetic energy of the nuclei and Coulomb interactions between nuclei. Now V_e is the Coulomb interaction between an electron and all the nuclei, and U_2 is the Coulomb interaction between electrons without dielectric constant. The new Hamiltonian is more fundamental. This time we want to freeze out the dynamics of the nuclei by asserting that the electrons provide some sort of Born-Oppenheimer potential and that

[17] Both Robert Schrieffer and Walter Kohn suggested that I include all the electrons in the analysis, not just those directly involved in the Hall current. At the time, the problem seemed complicated enough without extending it in that way. To include all the electrons would seem to mean that one would have to deal with the problem of what holds the solid together. However, the question came back in a somewhat more pointed form during a seminar given at SLAC. Leonard Susskind asked why one should use the physical electron charge in defining the current, since the material has dielectric properties and that charge should be partly shielded. Michael Peskin again suggested that one should simply include all the electrons in the treatment; and Marvin Weinstein helped me develop that point of view, which is described here.

the nuclei sit at minima of the potential. (In principle, we imagine that bulk matter excitations permit the quantum states of the electrons to reach a local minimum just as before.) The expression (3.7) is now valid for the current, with the physical electron charge occurring in the coefficient. It is even possible to modify the kinetic energy terms of the nuclei so that Eq. (3.7) incorporates the current density associated with them. At this point, the general properties of the states can be inferred from the discussion in the preceding sections, and Laughlin's argument may then be used to derive the quantum Hall relation. This permits the energy of the bulk matter to play a role in Eq. (3.7); i.e., we have extended the dynamics beyond those of the electrons participating directly in the current. However, we must still assume that when β has increased by 2π, the energy change is accounted for entirely by moving an electron from one electrochemical potential to another. The electrons in the "two-dimensional" region, including those bound to atoms, are assumed to return closely to their original configuration, so that the energy of that part of the system is unchanged. Laughlin's argument is also powerful enough to incorporate changes in the total energy of the system associated with the bulk matter, such as those due to mechanical stresses that result from the electric fields which are present. These would show up as a modification of the electrochemical potential.

Finally, it is useful to give an example of a case in which Laughlin's argument is not exactly valid. Whether it is realized in actual situations is not known; it does shed light on the physics of the argument. Suppose we have the situation of Fig. 16(b) which was described at the end of V.B. If β is increased continuously by 2π, the usual argument applies to the ground-state level. However, as each eigenfunction in the excited level moves one step to the right, there is a (very small) net change in energy because an additional state is occupied in the middle of the sample at an energy slightly greater than the Fermi energy of the higher probe. (Looked at another way, this corresponds to a correction of the leading approximation to the Euler-Maclaurin formula for replacing the sum by the integral in the derivation.) This complete state will ultimately decay to the original one by an irreversible excitation of the bulk matter. It is important for Laughlin's argument that such irreversible processes not be important in the "two-dimensional" region. At the same time, it is essential that they be available inside the probes (Laughlin, 1985).

F. Nonminimal coupling terms including spin interaction

Nonminimal coupling terms in the Hamiltonian correspond to contributions to the current density that are not generated by the prescription (3.7). Moreover, the magnetic moment coupling of the electron to the magnetic field contributes to the energy of the one-particle states. If the magnetic field were not perfectly uniform, one could worry that this coupling might affect the precision of the quantum Hall relation. We argue here that these apparent complications actually cause no difficulties.

A nonminimal coupling contribution to the Hamiltonian density, linear in the magnetic field, takes the form

$$\mathscr{H}_Q = \mathbf{B}\cdot\mathbf{Q}(x) , \tag{5.2}$$

where $\mathbf{Q}$ is constructed from the electron operators and may depend on properties of the bulk matter also. Associated with such a term in the Hamiltonian density, there is a new contribution to the current density,

$$\mathbf{j}_Q = \nabla\times\mathbf{Q} . \tag{5.3a}$$

While the current contribution arising from this term cannot be derived from Eq. (3.7), it is easy to see that it vanishes and hence does not destroy our result:

$$\begin{aligned} I_Q &= \frac{1}{L}\int \hat{\mathbf{x}}\cdot\mathbf{j}_Q d^3x \\ &= \frac{1}{L}\int_S (\mathbf{Q}\times\hat{\mathbf{x}})\cdot\hat{\mathbf{n}}\, da . \end{aligned} \tag{5.3b}$$

This surface integral is easily seen to vanish: for the part of the surface outside the system, $\mathbf{Q}=0$; for the parts at $x=0$ or L, $\hat{\mathbf{x}}$ is parallel to $\hat{\mathbf{n}}$ and the integrand vanishes. Thus (3.7) remains true in the presence of nonminimal terms in the Hamiltonian. The effect of such terms is simply to modify the Hamiltonian by contributions that happen to depend on the magnetic field. If they are one-particle terms, they can be incorporated directly into the original unperturbed problem. If they are multiparticle terms, they become part of the interaction between the electrons, which must be treated following the procedures of Sec. IV and Appendix B.

The case in which $\mathbf{Q}$ is proportional to the magnetic moment of the electron in the layer may seem to require special discussion. In this case, the difference in the energies of the states where the electron spin is oriented parallel or antiparallel to the magnetic field is given by $\hbar\omega_s$, where ω_s is the (possibly spatially dependent) spin-flip frequency. Each electron spin orientation can be treated independently, and this difference in energy shows up as a difference in potential in the two orientations. Since this energy difference tends to be much smaller than the Landau energy separation, the main effect of the electron spin is to give plateaus in which there are two electrons per spatial eigenfunction. However, plateaus corresponding to a separation of spin states can be observed.

VI. CRITIQUES AND DISCUSSION

A. General outlook

Let me remind the reader that the author is not a condensed matter physicist. This has been both an advantage and a handicap in preparing this paper. Having no natural prejudices about how to approach such a study, the author has perhaps been less encumbered by standard lore and more able to look at the problem of understanding the quantum Hall effect in a fresh way. As an outsider to the field, he may have been able to appreciate the difficulties other outsiders were having in comprehending the new developments of the past few years. On the other hand, it is also possible that by not knowing enough condensed matter physics, he may have made egregious blunders. Having discussed the issues with many condensed matter physicists, he feels that the latter has not happened. The present subsection deals with the general features of the quantum Hall effect, while the next has remarks about the question of its accuracy.

As originally conceived and as suggested by the title, the objective of the article was to give a presentation directly primarily toward nonspecialists, such as my colleagues in elementary particles physics. It has evolved into more than that, and I hope it may also be of use to workers in the field. While not strictly a review, much of the paper is based on conventional developments. The main claims to originality in a technical sense are the emphasis on the role of the external probes as an electron reservoir and their influence on the plateaus through the self-consistent potential, the generalization of the single-particle treatment to free it from some of the assumptions of earlier treatments, and the argument showing that the integer quantum Hall relation remains valid to all finite orders of perturbation in the residual electron-electron interactions. Not having understood Laughlin's argument so well at the start, I came to realize that much of the present work actually provides a detailed exposition of some of the underlying physics of that argument without necessarily adding real substance.

On a lesser but perhaps more useful level, it is hoped that the picture presented here may be helpful in creating a more detailed understanding of some aspects of present and future experiments. In the preceding sections, the observant reader may have noted some implied criticisms of the "standard" picture. In that picture, the plateaus are produced as the Fermi level moves through a mobility gap associated with localized states which lie (in energy) between the extended states that carry the current. This seems to be tied to a picture of the system in which the current, and hence Hall voltage, is zero. It is based on the energy spectrum alone, without any reference to the spatial location of the various types of states. Perhaps this language is understood by the specialists to have some meaning in terms of the actual spatial distribution of one-particle states and with a realistic voltage across the sample (with a local Fermi energy), but the only place where I have seen this described explicitly is in the *Scientific American* article by Halperin (1986). In my opinion, it is very important to understand the spatial distribution of the various types of states and how it depends on the background potential, which varies with the filling factor.

There has been some controversy in the literature about where the Hall current is distributed, primarily along the edge or primarily over the whole surface. Recent experiments (Ebert *et al.*, 1985; Sichel *et al.*, 1985; Zheng *et al.*, 1985) have shown that it in fact has a complicated dependence on the magnetic field. A detailed understanding of this behavior will likely turn out to be very complicated, as it will depend on both the details of the fabrication of the sample (i.e., the density of carriers that is built in) and how this works in combination with the self-consistent potential effects from the external reservoirs. The experimental devices are actually several quantum Hall devices coupled in parallel, and the dynamics of that is not very well understood. The present experiments also suffer from the difficulty that they are not carried out at sufficiently low temperature that the electron state can in any sense be well approximated as a single quantum state. The experimental problem with going to lower temperatures is that the source impedance of the device becomes so large that the time to equilibrate becomes prohibitively large. Ways must be found to improve the theoretical treatment and perhaps to do experiments under conditions in which more meaningful comparisons can be made.

Many theoretical treatments concentrate on conditions inside the layer and ignore exactly how the connection is made to the Hall voltage. In my opinion, this is sensible because that is apparently where the important physics lies; and one is most interested in dealing with questions such as how complicated the potential can be without distorting the effect. However, in the argument of Laughlin and in the one given here, the voltage probes play a crucial role. Not having found a discussion in the literature about how the states inside the layer come to equilibrium with the electrons in the probes, I try to provide one here (see Sec. V). This discussion is at the level of "how it must be" rather than of a detailed self-consistent treatment, which I feel must be very complicated. It makes no attempt at all to understand this equilibrium at finite temperature, where the electron spectrum inside the probes does not cut off sharply at the Fermi energy. Yet I hope it has elements of the correct physics. Perhaps a better argument would help us understand the deviations from the ideal at finite temperatures. An elaboration of the discussion suggests how several levels inside the sample can come to equilibrium with the same Fermi energy at the edge.

The possibility of relating these ideas to the finite-temperature behavior, in which there are both a longitudinal resistivity and an apparently related deviation from the ideal quantum Hall resistance (Cage *et al.*, 1984), is intriguing. As has been repeatedly emphasized here, my belief is that the states of the model described here are metastable. They are likely to be stable under purely electromagnetic interactions, but become decaying states when the interaction with the bulk matter is taken into account. In the model, this decay corresponds to a current flow between Hall voltage probes; in the experimental configuration, that current must be compensated by a Hall current across the sample, which is produced by a voltage drop along the sample. An initial crude attempt to understand that steady-state behavior under these circumstances leads rather naturally to the linear relation between the quantities mentioned above.

Recently an experimental paper appeared under the title "Quantization of the Hall Effect in an Isotropic Three-Dimensional Electronic System" (Störmer *et al.*, 1986). The experiment used a multilayer Hall sample with thirty periods and found a plateau corresponding to $i=48$ (with spin taken into account, this means that 12 of the available states of neighboring energies were unoccupied). I wish to emphasize that there is no difficulty in discussing such experiments within the general framework presented here, which is already three dimensional. It is also likely that it did not cause great consternation among those who like to give a two-dimensional treatment. The point is that there are now many states, corresponding to different sub-bands, with energy separations that are very small. Provided that the carrier density is large enough, many of these sub-bands can be filled at the same time. The fact that a plateau is clearly seen is interesting in view of the fact that a few of the sub-bands are unoccupied even though their energy is presumably very close to that of occupied sub-bands. The Hall voltage is small enough (≈ 0.5 mV) that it may not be able to cause significant leakage from the occupied to the unoccupied levels. More work should be done, perhaps along the lines of the present paper, to understand this in greater detail.

B. The question of accuracy

In discussing the question of accuracy, we should distinguish the ideal situation in the zero-temperature limit from the breakdown in accuracy that occurs at finite temperature. It is found experimentally that there are small deviations from the ideal quantum Hall resistance at finite temperatures, even under conditions in which the observed plateau is accurately flat (Cage *et al.*, 1984). However, one may be able to calibrate away this deviation by taking note of the fact that it is empirically linear in the measured longitudinal resistance, with the linear coefficient depending on the sample. That is, if we extrapolate to zero longitudinal resistance (presumably at zero temperature), we may assume that the result is the ideal quantum Hall resistance. Nothing we have said so far helps us to come to grips with the important question of the accuracy of the quantum Hall relation in that limit. The best argument in favor of great accuracy for the relation is that of Laughlin (1981). The present discussion cannot improve upon that argument, but it lends support by discussing the nature of the microscopic state and the properties it must have in order to produce his result. If this discussion does indeed contribute to the understanding of the basic mechanisms, then quite possibly it will make it feasible to ask questions that might not be obvious in less complete descriptions.

One feels that somehow the fact that there are only a finite number of electrons involved in the process must lead to a limitation in the relation. If corrections are of order $1/N$, they would not matter at present. However, we have no basis for even suggesting what the functional dependence on N should be. Perhaps such limitations could come about through replacing the sum by an integral. If there is a little "roughness" in the sum, perhaps it introduces an error that is not readily appreciated by the usual methods. Laughlin's argument appears to say that this does not happen. For some particular value of β, there might be fluctuations away from the ideal, but these would be averaged out by integrating β over the range 2π. So long as there is not an obscure mathematical pathology, integrating β *replaces* the sum by the integral. Since any sample experiences small fluctuations in environment, such an averaging seems to be physically plausible.

If we think through Laughlin's argument, it seems to be important only that the total energy of the state depend continuously and reversibly on β. This can even include the energy of the bulk matter, which we treated as inert throughout our discussion (except for its ability to bring the system to metastable equilibrium). Then we have to assume only that we get the same change of energy by varying β by 2π, which moves one electron across the sample for each level participating in the integer plateau, as by removing an electron from one probe and transporting it externally (through a voltmeter or a battery, for example) to the other probe. If moving that one electron itself made a significant change in the electrostatic potentials, then we might have to worry about the finiteness of the number of electrons. But since this is one electron out of the huge number in the probes, such effects should be totally minute.

In the end, (our version of) Laughlin's argument deals with an idealization in which the quantum Hall relation is exact. It has not been possible to give any quantitative estimate of deviations from the ideal and how they depend on the real conditions of the sample. Hopefully, this work provides a starting point for understanding the general properties of the electron state which underly the relation. For the present, the best evidence for the precision is obtained by intercomparisons between different devices (Delahaye *et al.*, 1986).

C. Brief review of relevant QED

The now-standard way to test precision QED in perturbation theory is to use various experiments and their associated theories to predict a value of the fine-structure constant α that may be compared with the values from other types of experiment that are insensitive to the details of QED. The most accurate experiments of the type requiring QED calculations are those concerned with the anomalous moment of the electron (a_e) and the hyperfine structure in muonium (spin splitting in the μ^+-e^- hydrogenic ground state). Although the hyperfine structure in hydrogen is very accurately measured, its interpretation is limited because of uncertainties introduced by the structure of the proton. Other features of atomic structure, such as the Lamb shift, while of intrinsic interest for QED, provide less exacting information about the value of α. The non-QED sources of the fine-structure constant are the ac Josephson effect and the quantum Hall effect.

The most recent review of the status of QED, with emphasis on the anomalous moment of the electron and on the relation of QED to the quantum Hall effect, is given by Kinoshita (1986). Here we give a very brief summary. The most accurate determination of a_e is by an experiment using a Penning trap (Van Dyck, Jr. *et al.*, 1984). Its theoretical interpretation has now reached the level of four-loop QED calculations, which have been in progress for several years (Kinoshita and Lindquist, 1981), with progressively improving accuracy. The most accurate determination of the muonium hyperfine structure is reported in Mariam *et al.* (1982), and a general review of its status is given in Bodwin, Yennie, and Gregorio (1985). The Josephson junction experiments are reported in Williams and Olsen (1979). The experiment actually measures $e/\hbar$. To convert this to a value of $\alpha=e^2/\hbar c$, a chain of other information must be used, including the gyromagnetic ratio of the proton in water. In addition, a very accurate resistance standard must be used. Obviously, the determination of the quantum Hall resistance also requires a very accurate resistance standard. This is not the place to describe these intricacies in detail, but it is worth noting that, in the quantum Hall experiments carried out at the National Bureau of Standards over a period of time, it was discovered that the standard resistors maintained at the National Bureau of Standards actually have a very small drift in value with time (Cage *et al.*, 1985). Thus the quantum Hall effect has already made a very important contribution to metrology. If the quantum Hall effect and Josephson junction experiments can be carried out with the same standard resistance as reference, the dependence on knowledge of that resistance can be calibrated away. Of course, this has not yet been done directly, because the two experiments at the National Bureau of Standards were unfortunately done at different times. However, assuming linearity of the standard resistance as a function of time, a correction can be made.

Assuming that the theoretical understanding of the non-QED experiments is adequate to the present level of precision, we then have the following comparison between all these methods:

$$\begin{aligned}
&\alpha_{a_e}^{-1}=137.035\,994(5)\ ,\\
&\alpha_{\mu\text{-hfs}}^{-1}=137.035\,991(25)\ ,\\
&\alpha_{(\text{ac-Jos})}^{-1}=137.035\,963(15)\ ,\\
&\alpha_{\text{QHE}}^{-1}=137.035\,965(12)\ ,\\
&\alpha_{\text{NBS}}^{-1}=137.035\,981(12)\ .
\end{aligned}\tag{6.1}$$

The numbers in parentheses give the uncertainty in the last digits. For the a_e determination, the uncertainty is primarily theoretical, and it will be improved during the next year or so. The experimental uncertainty is an order of magnitude smaller. The uncertainty of the μ-hfs determination has several components: the most important is that due to the muon's magnetic moment (20); the next most important is the estimate of uncalculated theoretical terms (15); and the uncertainty in the hfs measurement itself is only (5). No allowance for theoretical uncertainties is made in the various condensed matter determinations. We quote only results from the previously mentioned experiments. The NBS determination refers to a composite of National Bureau of Standards results made by Taylor (1985), in which he took into account the drift of the NBS standard of resistance with time and combined the ac-Jos and QHE values in such a way that the standard is eliminated. The agreement between the various types of experiments is remarkably good, and the precision of all the methods is constantly improving.

ACKNOWLEDGMENTS

Encouragement, stimulation, and technical help for this work were given by a large number of friends, some of whom have already been named in the footnotes. Initial interest in the problem was stimulated by an overview talk given by Toichiro Kinoshita, who expressed concern whether the quantum Hall effect was sufficiently well understood to use as an input for precision tests of QED. The first introduction to some of the critical issues was provided by Henry Tye, who continued to provide helpful comment as the work progressed. Over the past two years, Stuart Trugman explained his ideas and asked many probing questions that, in one way or another, have influenced many aspects of this work. During my sabbatical year (1984—1985), I enjoyed the hospitality of Fermilab, the Institute for Theoretical Physics at Santa Barbara, and SLAC. The work at Cornell was supported in part by the National Science Foundation; at Fermilab and SLAC, it was supported in part by the Department of Energy; and at the Institute for Theoretical Physics, it was supported in part by the National Science Foundation under Grant No. PHY82-17853, supplemented by funds from the National Aeronautics and Space Administration. While at Fermilab, I made an abortive attempt to make a perturbative calculation of the current contributions arising from higher states. What I found out was that they are not intrinsically very small. The main positive accomplishment of that period was that conversations with William Bardeen, Geoffrey Bodwin, and Murray Peshkin, contributed to my general understanding in various subtle ways. While in Santa Barbara, I discovered the method based on the use of β which I feel is quite powerful.[18] That idea, as well as the realization of the importance of the dependence of the self-consistent potential on the filling factor, was sharpened by numerous discussions with

[18] Perhaps a more balanced appraisal is that it amounts to a way to translate Laughlin's argument into a form in which you can study the microscopic physics in more detail and carry out the perturbative analysis. But if you accept Laughlin's argument from the beginning as being sufficient (as he does), then it clearly must be possible to carry through such a treatment.

my colleagues Daniel Arovas, Catherine Kallin, Steven Kivelson, Walter Kohn, and Robert Schrieffer, who were all very helpful and supportive of my efforts. The final stage of my sabbatical leave was spent at SLAC. Although the main features of the analysis had already been worked out, many further subtleties were brought to my consciousness (see footnote 17) and the details of Appendix B were worked out. I had many discussions with Richard Blankenbecler, Daniel Boyanofsky, Robert Laughlin, and Marvin Weinstein (at Stanford) which helped sharpen my understanding. At Cornell, several of my condensed matter colleagues have helped educate me on necessary details and given useful critiques of various parts of the work. They are Neil Ashcroft, Michael Fisher, James Krumhansl, James Sethna, and John Wilkins. Marvin Cage, Albert Chang, and Steven Girvin have all given freely of their time to answer various questions about the experiments and theory of the quantum Hall effect, and to discuss my own work.

APPENDIX A: APPROXIMATE TREATMENT OF THE WAVE FUNCTIONS IN TWO DIMENSIONS

To develop better intuition about the eigenfunctions, we study a model two-dimensional problem with an effective mass and smoothed potential. The Hamiltonian for this model takes the form

$$h'_1=\frac{(\mathbf{p}+e\mathbf{A})^2}{2m^*}+\mathscr{V}(x,y)\ , \tag{A1}$$

where m^* is the effective mass. Such eigenfunctions have been described by several authors, in particular by Trugman (1983), and by Joynt and Prange (1984). An important feature is that the magnetic field tends to prevent the spreading out of the eigenfunctions as in a scattering process. This property is of course independent of gauge; however, in order to get an intuitive idea of how these things work, it is convenient to use the Landau gauge given by

$$\mathbf{A}=(-yB,0,0)\ . \tag{A2}$$

We consider a region where there are no zeros of the eigenfunction and where the smoothed potential does not change abruptly in distances of order the magnetic length l (of order 100 Å). To see how this works, we assume the following form for the eigenfunction:

$$\psi=e^{i\varphi(x)}N(x)\mu(x,y)\ . \tag{A3a}$$

Here μ is normalized in the y direction for each fixed value of x, and N allows for a change in the normalization as a function of x.

The most important x dependence is in the phase factor; we shall ignore the other dependence on x at first and deduce it later by simple physical arguments. With these assumptions, we find at fixed x the differential equation in y

$$\left[\frac{-\hbar^2(d/dy)^2+(\hbar\varphi'-eyB)^2}{2m^*}+\mathscr{V}(x,y)\right]\mu(x,y)=\varepsilon\mu(x,y)\ , \tag{A3b}$$

where $\varphi'\equiv d\varphi/dx$.

Now x plays the role of a parameter. We see that the eigenfunction has a strong tendency to peak up at the point

$$y_0(x)=\hbar\varphi'/eB\ , \tag{A3c}$$

and we expand the y dependence of $\mathscr{V}$ about that point and include the linear deviation $[y-y_0(x)]\mathscr{V}'$. The equation then reduces to a harmonic-oscillator problem with length scale l. The ground-state eigenfunction has its center shifted from y_0 by $\delta y_0(x)=-m^*\mathscr{V}'/(e^2B^2)$, and we find the constraint

$$\tfrac{1}{2}\hbar\omega_c+\mathscr{V}[x,y_0(x)+\tfrac{1}{2}\delta y_0(x)]=\varepsilon\ , \tag{A3d}$$

where ω_c is the cyclotron frequency, eB/m^*. Since in practical situations $\delta y_0(x)<<l$, this gives the constraint that $y_0(x)$ must follow an equipotential contour. This is just the "guiding" equipotential referred to in Sec. II. In turn, this determines the function $\varphi(x)$, whose rate of change with x must adapt in this way to the potential. The variable x is also seen to occur in the ground-state eigenfunction through $y_0(x)$ and $\delta y_0(x)$. Locally, the current carried by the state is given by $-N^2\mathscr{V}'/B=-eN^2E_y/B$, in agreement with the discussion of Sec. II. Since the current of a single eigenfunction must be independent of x (an exact result), this tells us how the normalization depends on x. This variation of normalization with distance along the potential maintains the proper total electron density throughout the sample, also in agreement with the requirements of Sec. II.

We may use this result from a local region to understand the properties of eigenfunctions on a larger scale. All the properties expected in Secs. II and III are confirmed. For example, by integrating our expression (A3c) for $y_0(x)$ (note that this is not weighted by N^2), we find that its average value is given by $2\pi nl^2/L$. This confirms the expectation of one state per quantum flux unit and that the average transverse separation between adjacent extended states is much less than their individual widths. This also suggests the parametrization of the average distance of an eigenfunction across the sample in terms of k, namely

$$y_k=kl^2\ . \tag{A4}$$

This parameter should also have meaning for the more realistic case as well, since it is generally the local phase oscillations as a function of x which provides the transverse (i.e., y direction) centering of the eigenfunction, which then has a transverse spread of order l about that center. There is no implication that y_k has a direct significance; however, as k varies from one extreme to the other, the succession of eigenfunctions should move across the sample from one reservoir to the other.

A potential hill or valley together with an overall drop of potential across the sample can produce contour lines like those that guide the eigenfunctions illustrated in Fig. 8(a). Contour lines closing on themselves give rise to localized states. To the extent that $\mathscr{V}$ varies sufficiently slowly with position, these could be studied with the same procedure as in the preceding paragraph. For each small region we could first transform to a local Landau gauge where the x' axis is first taken parallel to a potential contour. If we translate these results back to the original gauge, we find that the total phase change around the path of the eigenfunction (which of course must be an integral multiple of 2π, since the eigenfunction is single valued) is given by the number of flux units contained in a curve that passes along the peak of the eigenfunction. See Prange (1986) for the details of this argument. This is part of the general result that there is one state per flux unit in the "two-dimensional" region.

APPENDIX B: PERTURBATION THEORY INCLUDING PARTICLE-HOLE REDUCIBLE DIAGRAMS

In Sec. IV we showed that the quantum Hall relationship remains valid to every order of perturbation theory for contributions to $E(\beta)$ that do not have intermediate states consisting only of a particle-hole pair. This appendix describes a generalization of that proof to include all contributions; however, it is still limited to perturbation theory. The technique is to define U_1 so that the complications mentioned at the end of Sec. IV do not occur. We recall that the origin of the complications is that one-hole energies can have small energy denominators whenever single-hole intermediate states occur in the perturbation theory. It is essential to our whole discussion that the energies of the extended states are almost a continuous function of k, so that we may replace a sum by an integral. The difficulty is obviously a consequence of the fact that we should be doing degenerate rather than nondegenerate perturbation theory. To avoid this would seem to be an incredible chore, since we are dealing with a huge number of electrons. Our technique is based on the fact that we have the freedom to choose U_1, and hence our basis states, in such a way that the matrix elements to nearly degenerate intermediate states vanish to all orders.

It may be helpful to give first a brief qualitative motivation for the following technical discussion. Refer back to Eq. (4.2), where the leading contribution (i.e., first order in U_2) to U_1 was defined. It has the important property that the unperturbed eigenfunctions are not altered by first-order perturbation corrections since the matrix element giving such superpositions is zero. If there were superpositions involving higher Landau levels, the analysis would become quite complicated because the operator $\hat{I}$ has large matrix elements between different levels (for example, its matrix element between the ground and first levels is much larger than its expectation value in the ground state by the ratio of l to δy_0 of Appendix A). The cancellation of the first-order contribution to the current is made clear by the discussion following Eq. (4.3). The aim of our general discussion is to extend this property to all orders. The self-consistent potential is to be defined in such a way that the lowest-order single-particle eigenfunctions are not modified by the perturbations. The result is that the current calculated using Eq. (3.13) becomes exact to all orders of perturbation theory.

In Eq. (4.2b), we defined the first-order contribution to U_1 to compensate the effects of H'_1 as closely as possible. Now we add a hierarchy of additional contributions designed to compensate higher-order effects from H'_1 and H'_2:

$$U_1(\mathbf{x},\mathbf{x}')=U_1^{(1)}(\mathbf{x},\mathbf{x}')+\delta U_1(\mathbf{x},\mathbf{x}') \;, \tag{B1}$$

where

$$\delta U_1(\mathbf{x},\mathbf{x}')=\delta_1 U_1(\mathbf{x},\mathbf{x}')+\delta_2 U_1(\mathbf{x},\mathbf{x}')+\cdots \;.$$

The successive terms in δU_1 will be designed to compensate increasingly complicated contributions from the perturbation theory. It hardly seems desirable to write out this procedure in complete detail, since we do not wish to carry through a genuine calculation. However, it will be carried far enough so that the approach should be convincing. Although U_1 cancels in the Hamiltonian, it is clearly necessary to require it to yield a Hermitian contribution, so that the eigenfunctions defined by h_1 form a complete orthonormal set. This requirement is simply

$$\bar{U}_1(\mathbf{x},\mathbf{x}')=U_1(\mathbf{x}',\mathbf{x}) \;, \tag{B2}$$

where the overbar denotes complex conjugation. This is satisfied by the first term in (B1). We also require U_1 to be independent of β so that it does not complicate the analysis when we differentiate with respect to β. Using the completeness of the set of eigenfunctions defined by h_1, we may write

$$\delta U_1(\mathbf{x},\mathbf{x}')=\left[\sum_{\hat{\alpha},\hat{\alpha}'}\psi_{\hat{\alpha}}(\mathbf{x})Q_{\hat{\alpha}\hat{\alpha}'}\psi^{\dagger}_{\hat{\alpha}'}(\mathbf{x}')\right]_{\beta=0} \tag{B3}$$

with

$$\bar{Q}_{\hat{\alpha}\hat{\alpha}'}=Q_{\hat{\alpha}'\hat{\alpha}}$$

to assure Hermiticity.

For our present purposes, time-ordered perturbation theory seems to work more efficiently than old-fashioned perturbation theory. We use the interaction picture defined by the Hamiltonian H_1, so that the operators become

$$\psi(\mathbf{x},t)=\sum_{\alpha}b^{\dagger}_{\alpha}\psi_{\alpha}(\mathbf{x})e^{-i\varepsilon_{\alpha}t}+\sum_{\bar{\alpha}}a_{\bar{\alpha}}\psi_{\bar{\alpha}}(\mathbf{x})e^{-i\varepsilon_{\bar{\alpha}}t} \;. \tag{B4}$$

In our analysis of perturbation theory, it will be useful to use the wave functions ψ' defined by Eq. (3.9) and the corresponding operators. A pairing of these operators yields the factor

$$\langle 0\,|\,T[\psi'(\mathbf{x}_a,t_a),\psi'^{\dagger}(\mathbf{x}_b,t_b)]\,|\,0\rangle=\sum_{\bar\alpha}\psi'_{\bar\alpha}(\mathbf{x}_a)\psi'^{\dagger}_{\bar\alpha}(\mathbf{x}_b)e^{-i\varepsilon_{\bar\alpha}(t_a-t_b)}\theta(t_a-t_b)-\sum_{\alpha}\psi'_{\alpha}(\mathbf{x}_a)\psi'^{\dagger}_{\alpha}(\mathbf{x}_b)e^{-i\varepsilon_{\alpha}(t_a-t_b)}\theta(t_b-t_a)$$

$$=\sum_{\hat\alpha}\psi'_{\hat\alpha}(\mathbf{x}_a)\psi'^{\dagger}_{\hat\alpha}(\mathbf{x}_b)e^{-i\varepsilon_{\alpha}(t_a-t_b)}\theta(t_a-t_b)-\sum_{\alpha}\psi'_{\alpha}(\mathbf{x}_a)\psi'^{\dagger}_{\alpha}(\mathbf{x}_b)e^{-i\varepsilon_{\alpha}(t_a-t_b)}\,. \tag{B5}$$

The perturbation Hamiltonian is now time dependent, and we may write the energy of the perturbed state as

$$E(\beta)=\sum_{\alpha}\varepsilon_\alpha(\beta)+\langle 0\,|\,\mathscr{E}\,|\,0\rangle_{\text{conn}}\,, \tag{B6}$$

where

$$\mathscr{E}=\sum_n\frac{1}{(n+1)!}\left[\frac{-i}{\hbar}\right]^n\int_{-\infty}^{\infty}\int\cdots\int_{-\infty}^{\infty}dt_1dt_2\cdots dt_n T(H'(0),H'(t_1),H'(t_2),\ldots,H'(t_n))\,.$$

Only connected diagrams occur in the expectation value in Eq. (B6). Note that one of the times is held fixed in the definition of $\mathscr{E}$. The c-number part of the perturbation given by Eq. (4.3) contributes only for $n=0$.

The β dependence of the $\mathscr{E}$ term in the energy occurs explicitly through the dependence of the perturbation Hamiltonian on the functions ψ'_α and also through the pairings of operators as expressed by (B5). Consider the derivative of one of these pairing factors with respect to β. In differentiating a factor (B5), we note that the first term of the second form is formally independent of β; this may be seen by the same type of rearrangement as described in Sec. IV: replace $\varepsilon_{\hat\alpha}$ by the operator h_1 acting on the unprimed eigenfunction, then use closure on the sum, to produce

$$e^{i\beta x_a/L}e^{-ih_1(t_a-t_b)}\delta(\mathbf{x}_a-\mathbf{x}_b)e^{-i\beta x_b/L}\theta(t_a-t_b)$$
$$=e^{-ih_1^0(t_a-t_b)}\delta(\mathbf{x}_a-\mathbf{x}_b)\theta(t_a-t_b)\,.$$

The justification for neglecting the β derivative of this factor depends on the structure of the rest of the expression, but we shall define δU_1 so that this is always legitimate.

To obtain the modification in the current, we must differentiate the second term of Eq. (B6) with respect to β. The contributions arise from Eq. (4.3), where now only the δU_1 piece contributes, from the explicit β dependence in (4.2a), and from the second term of one of the factor pairings as given in (B5). It turns out that all of these can be combined into the expression

$$\frac{\partial\langle 0\,|\,\mathscr{E}\,|\,0\rangle_{\text{conn}}}{\partial\beta}=-\sum_\alpha\frac{\partial\langle 0\,|\,b_\alpha\mathscr{E}b_\alpha^\dagger\,|\,0\rangle_{\text{conn}}}{[\partial\beta]_\alpha}\,, \tag{B7}$$

where the expansion of $\mathscr{E}$ uses H'_2 and the δU_1 term of H'_1. On the right-hand side it is understood that we differentiate only the β dependence which enters through α. As in our previous work, we may replace β by the label k and introduce a factor $1/L$. We note that the quantity differentiated on the right-hand side (rhs) is the hole propagator with incident and final label k and with the external lines amputated. If this could be identified with the energy to create a hole, our work would be finished. In fact, if we consider only the class of graphs discussed in Sec. IV, the expectation value on the rhs would be simply the contribution of a proper self-energy, and we would have reproduced the results of that section in a more compact way.

Now we must deal with the complications of degenerate perturbation theory. It manifests itself here by the presence of small (even zero) energy denominators in the rhs of Eq. (B7) when we consider contributions with higher numbers of self-energies. (The name self-energy is not quite correct, since the bulk matter is capable of scattering a hole from one state to another, but we continue to use it.) Calling this self-energy $\widetilde{M}$, we find that the contribution to the expectation value illustrated in Fig. 17 takes the form

$$\widetilde{M}_{\alpha\alpha}(\varepsilon_\alpha)+\sum_{\hat\alpha_1}\frac{\widetilde{M}_{\alpha\hat\alpha_1}(\varepsilon_\alpha)\widetilde{M}_{\hat\alpha_1\alpha}(\varepsilon_\alpha)}{\varepsilon_{\hat\alpha_1}-\varepsilon_\alpha}$$
$$+\sum_{\hat\alpha_1,\hat\alpha_2}\frac{\widetilde{M}_{\alpha\hat\alpha_1}(\varepsilon_\alpha)\widetilde{M}_{\hat\alpha_1\hat\alpha_2}(\varepsilon_\alpha)\widetilde{M}_{\hat\alpha_2\alpha}(\varepsilon_\alpha)}{(\varepsilon_{\hat\alpha_1}-\varepsilon_\alpha)(\varepsilon_{\hat\alpha_2}-\varepsilon_\alpha)}+\cdots\,. \tag{B8}$$

$\widetilde{M}$ includes a term from $-\delta U_1$, and it may have $-\delta U_1$ appearing internally as well. If vanishing denominators actually occur in Eq. (B8), the expression becomes mean-

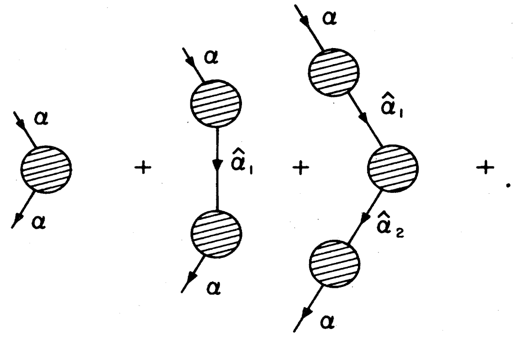

FIG. 17. Contributions to the hole self-energy, with varying numbers of proper self-energy contributions.

ingless. Since the left-hand side of (B7) is well defined in perturbation theory, we conclude that the neglect of the derivative of the first term of the second form of (B5) is at fault in this case. We have introduced terms that give a singular sensitivity to the boundary conditions because a hole can have a self-energy interaction, propagate the length of the "two-dimensional" region (signaled by the small-energy denominator), and have additional self-energy interactions.

It is straightforward, in principle, to define δU_1 so that this difficulty does not occur. We must do this in a way that not only makes Eq. (B8) finite but also makes its derivative with respect to the β dependence contained in α finite. Actually, we must make the series in (B8) vanish term by term. The simplest illustration of this is that if we set $\hat{\alpha}_1=\alpha$ in the second term, the result blows up unless the numerator factor is zero. This requires the first term of (B8) to vanish. We do not require the derivative of the first term to vanish, because the derivative of the second term with $\beta=0$ remains finite. In the same way, the second term must vanish in order to avoid difficulties in the third term, etc. We must also avoid the singularity that occurs when $\varepsilon_{\hat{\alpha}_1}$ approaches ε_α in the second term. To do this, we may require $M_{\hat{\alpha}_1\alpha}$ to vanish at least to first order in the energy difference.

Let $M^{(1)}_{\hat{\alpha}_1\hat{\alpha}_2}(\varepsilon)$ be a proper self-energy part without any overall δU_1 subtractions. This quantity satisfies the Hermiticity property

$$\overline{M}^{(1)}_{\hat{\alpha}_1\hat{\alpha}_2}(\varepsilon)=M^{(1)}_{\hat{\alpha}_2\hat{\alpha}_1}(\varepsilon) \ . \tag{B9}$$

The proof of this relationship follows from considering a T-bracket expression which produces a term in $M^{(1)}$. Taking the complex conjugate reverses the time order. The anti-time-ordered expression can be rearranged algebraically into a sum of terms that are products of time-ordered and anti-time-ordered factors, including one that is the completely time-ordered expression. The latter term gives a contribution to the rhs of Eq. (B9) (including the correct sign). If β is below the threshold to produce real intermediate states, all the other terms vanish.

Now we may define $\delta_1 U_1$ by using Eq. (B3) with

$$Q^{(1)}_{\hat{\alpha}_1\hat{\alpha}_2}=M^{(1)}_{\hat{\alpha}_1\hat{\alpha}_2}[\tfrac{1}{2}(\varepsilon_{\hat{\alpha}_1}+\varepsilon_{\hat{\alpha}_2})] \ , \tag{B10}$$

where the right-hand side is evaluated at $\beta=0$, in accordance with Eq. (B3). The choice here is not unique. We might also have used the average of $M^{(1)}$ evaluated for the two energies. The important point is that $Q^{(1)}$ satisfies the Hermiticity property and gives

$$\widetilde{M}^{(1)}_{\hat{\alpha}_1\hat{\alpha}_2}(\varepsilon_\alpha)=M^{(1)}_{\hat{\alpha}_1\hat{\alpha}_2}(\varepsilon_\alpha)-(\delta_1 U_1)_{\hat{\alpha}_1\hat{\alpha}_2} \tag{B11}$$

with the properties we need. These properties are as follows. (i) $\widetilde{M}^{(1)}_{\alpha\alpha}(\varepsilon)$ vanishes for $\beta=0$. Its derivative does not vanish, however. (ii) $\widetilde{M}^{(1)}_{\alpha_1\alpha}(\varepsilon)$ goes to zero as $\varepsilon_{\hat{\alpha}_1}$ goes to ε_α for $\beta=0$. Again, the derivative does not vanish. Now the second term of (B8) and its derivative with respect to β are well behaved.

The rest of the discussion will be somewhat sketchy. We want to define a term $-\delta_2 U_1$ which will appear in the first term of (B8) and cancel its second term for $\beta=0$, when only the $\widetilde{M}^{(1)}$ parts appear in the sum. As in the discussion of $\delta_1 U_1$, we first generalize the second term to have two different indices and a general energy ε. A Hermiticity requirement like Eq. (B9) is easily seen to be satisfied, and we define a new piece for Q by an expression like (B10), with the sum now appearing on the rhs. This automatically cancels the second term of (B8) when $\beta=0$, to a consistent order. In studying the third term of (B8), we must take $M^{(1)}$ in all places and combine it with pieces of the second term which have $M^{(1)}$ in one factor and $-\delta_2 U_1$ in the other. The result has the property of being well behaved and having a well-behaved derivative. Following the same pattern, we define an additional contribution $\delta_3 U_1$ which makes this set vanish for $\beta=0$. The procedure can obviously be extended systematically to more and more terms of (B8).

Our conclusion is that the self-consistent potential may be chosen in such a way that the second term of Eq. (B6), which represents the perturbative corrections to the energy of the ground state, has vanishing derivative with respect to β at $\beta=0$. At the same time, the perturbations do not shift the energy required to create a hole from the value $-\varepsilon_\alpha$, so the chemical potentials in the reservoirs are unchanged. Of course, the interactions do affect the current and the electrochemical potentials through the choice we have been forced to make for U_1; the point is that, with this choice, the perturbations do not shift these quantities from their effective single-particle values. The consequence is that when perturbation theory is valid, only the integer quantum Hall effect is obtained. The fractional quantum Hall effect occurs when the situation cannot be described perturbatively. Our present discussion sheds no light on the conditions that make perturbation theory valid. Our intuition is that when the filling factor is near unity the quantum state in the "two-dimensional" region has a certain stability against the effect of the perturbations.

REFERENCES

Baraff, G. A., and D. C. Tsui, 1981, Phys. Rev. B **24**, 2274.

Bodwin, G. T., D. R. Yennie, and M. A. Gregorio, 1985, Rev. Mod. Phys. **57**, 723.

Brenig, W., 1983, Z. Phys. B **50**, 305.

Cage, M. E., 1986, in *The Quantum Hall Effect*, edited by R. E. Prange and S. M. Girvin (Springer, New York), Chap. II.

Cage, M. E., R. F. Dziuba, and B. F. Field, 1985, IEEE Trans. Instrum. Meas. **IM-34**, 301.

Cage, M. E., R. F. Dziuba, B. F. Field, E. R. Williams, S. M. Girvin, A. C. Gossard, D. C. Tsui, and R. J. Wagner, 1983, Phys. Rev. Lett. **51**, 1374.

Cage, M. E., R. F. Dziuba, S. M. Girvin, A. C. Gossard, D. C. Tsui, and R. J. Wagner, 1984, Phys. Rev. B **30**, 2286.

Cage, M. E., and S. M. Girvin, 1983a, Comments Solid State Phys. **11**, 1.

Cage, M. E., and S. M. Girvin, 1983b, Comments Solid State

Phys. **11**, 47.

Chalker, J. T., 1983, J. Phys. C **16**, 4297.

Chang, A. M., 1985, private communication.

Delahaye, F., D. Dominguez, F. Alexandre, J. P. Andre, J. P. Hirtz, and M. Razeghi, 1986, Metrologia **22**, 103.

Ebert, G., K. von Klitzing, K. Ploog, and G. Weimann, 1983, J. Phys. C **16**, 5441.

Ebert, G., K. von Klitzing, C. Probst, E. Shuberth, K. Ploog, and G. Weimann, 1983, Solid State Commun. **45**, 625.

Ebert, G., K. von Klitzing, and G. Weimann, 1985, J. Phys. C **18**, L257.

Halperin, B. I., 1982, Phys. Rev. B **25**, 2185.

Halperin, B. I., 1983, Helv. Phys. Acta **56**, 75.

Halperin, B. I., 1986, Sci. Am. (April), 52.

Heinonen, O., P. L. Taylor, and S. M. Girvin, 1984, Phys. Rev. B **30**, 3016.

Joynt, R., 1982, Ph.D. thesis, University of Maryland.

Joynt, R., 1984, J. Phys. C **27**, 4807.

Joynt, R., and R. E. Prange, 1984, Phys. Rev. B **29**, 3303.

Kazarinov, R. F., and S. Luryi, 1982, Phys. Rev. B **25**, 7626.

Kinoshita, T., 1986, presented at the 1986 Conference on Precision Electromagnetic Measurements, National Bureau of Standards, Gaithersburg, Maryland, unpublished.

Kinoshita, T., and W. B. Lindquist, 1981, Phys. Rev. Lett. **47**, 1573.

Kuchar, F., G. Bauer, G. Weimann, and H. Burkhard, 1984, Surf. Sci. **142**, 196.

Laughlin, R. B., 1981, Phys. Rev. B **23**, 5632.

Laughlin, R. B., 1983, Phys. Rev. Lett. **50**, 1395.

Laughlin, R. B., 1985, private communication.

Levinson, N., 1949, K. Dan. Vidensk, Selsk. Mat. Fys. Medd. **25**, 1.

Mariam, F. G., W. Beer, P. R. Bolton, P. O. Egan, C. J. Gardner, V. W. Hughes, D. C. Lu, P. A. Souder, H. Orth, J. Vetter, U. Moser, and G. zu Putlitz, 1982, Phys. Rev. Lett. **49**, 993.

Prange, R. E., 1981, Phys. Rev. B **23**, 4802.

Prange, R. E., 1986, in *The Quantum Hall Effect,* edited by R. E. Prange and S. M. Girvin (Springer, New York), Chaps. I and III.

Prange, R. E., and S. M. Girvin, 1986, Eds., *The Quantum Hall Effect* (Springer, New York).

Pruisken, A. M. M., 1986, in *The Quantum Hall Effect,* edited by R. E. Prange and S. M. Girvin (Springer, New York), Chap. 5.

Rendell, X., and Y. Girvin, 1984, in *Precision Measurements and Fundamental Constants II,* Natl. Bur. Stand. (U.S.) Spec. Publ. No. 617, edited by B. N. Taylor and W. D. Phillips (U.S. GPO, Washington, D.C.), p. 557.

Sichel, E. K., H. H. Sample, and J. P. Salerno, 1985, Phys. Rev. B **33**, 1190.

Störmer, H. L., J. P. Eisenstein, A. C. Gossard, W. Wiegmann, and K. Baldwin, 1986, Phys. Rev. Lett. **56**, 85.

Störmer, H. L., D. C. Tsui, and A. C. Gossard, 1982, Surf. Sci. **113**, 32.

Středa, P., and K. von Klitzing, 1984, J. Phys. C **17**, L483.

Taylor, B. N., 1985, J. Res. Natl. Bur. Stand. **90**, 91.

Thouless, D. J., 1986, in *The Quantum Hall Effect,* edited by R. E. Prange and S. M. Girvin (Springer, New York), Chap. 4.

Trugman, S. A., 1983, Phys. Rev. B **27**, 7539.

Trugman, S. A., 1986, private communication.

Tsui, D. C., and S. J. Allen, 1981, Phys. Rev. B **24**, 4028.

Tsui, D. C., H. L. Störmer, and A. C. Gossard, 1982a, Phys. Rev. B **25**, 1405.

Tsui, D. C., H. L. Störmer, and A. C. Gossard, 1982b, Phys. Rev. Lett. **48**, 1559.

Van Dyck, R. S., Jr., P. B. Schwinberg, and H. G. Dehmelt, 1984, in *Atomic Physics,* edited by R. S. Van Dyck, Jr. and E. N. Fortson (World Scientific, Singapore), Vol. 9.

von Klitzing, K., 1986, Rev. Mod. Phys. **58**, 519.

von Klitzing, K., G. Dorda, and M. Pepper, 1980, Phys. Rev. Lett. **45**, 494.

Weinstein, M., 1985, private communication.

Williams, E. R., and P. T. Olsen, 1979, Phys. Rev. Lett. **42**, 1575.

Woltjer, R., R. Eppenga, J. Mooren, C. E. Timmering, and J. P. André, 1986, Europhys. Lett. **2** (2), 149.

Wysokinski, K. I., and W. Brenig, 1983, Z. Phys. B **54**, 11.

Zheng, H. Z., D. C. Tsui, and A. M. Chang, 1985, Phys. Rev. B **32**, 5506.