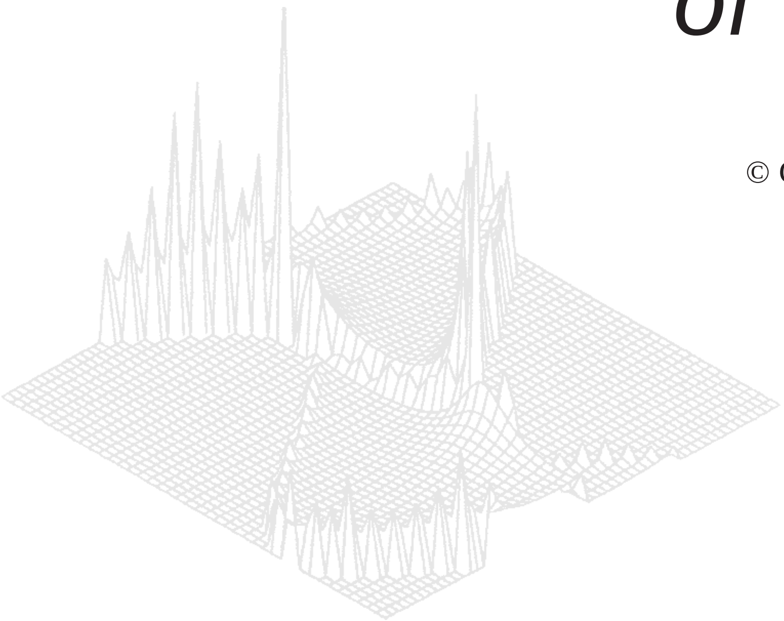
C S I R O   P U B L I S H I N G

Australian Journal of Physics



Volume 50, 1997

A journal for the publication of
original research in all branches of physics

www.publish.csiro.au/journals/ajp

All enquiries and manuscripts should be directed to
Australian Journal of Physics
CSIRO PUBLISHING
PO Box 1139 (150 Oxford St)
Collingwood
Vic. 3066
Australia
Telephone: 61 3 9662 7626
Facsimile: 61 3 9662 7611
Email: peter.robertson@publish.csiro.au

Published by **CSIRO** PUBLISHING
for CSIRO Australia and
the Australian Academy of Science

Aust. J. Phys., 1997, **50**, 1061–81

Gauge Invariant Magnetism*

A. M. Stewart

Department of Applied Mathematics,
Research School of Physical Sciences and Engineering,
Australian National University, Canberra, ACT 0200, Australia.
email: andrew.stewart@anu.edu.au

Abstract

An introduction is given to features of gauge invariance in classical and quantum mechanics that are of importance for magnetism in condensed matter systems. A version of quantum mechanics is described in which full electromagnetic gauge arbitrariness is displayed explicitly at every stage. The division of orbital magnetism into paramagnetism and diamagnetism is examined and it is shown that only by treating both of these on an equal footing can a gauge invariant treatment of magnetism be constructed.

1. Introduction

The intention of this paper is to provide a simple introduction to gauge invariance in quantum mechanics and statistical mechanics in areas that are important to magnetism in condensed matter and to summarise some recent work in this field (Stewart 1996*a*, 1996*b*, 1997). We take a semi-classical point of view usually suitable for condensed matter systems in which charged particles move according to the laws of quantum mechanics in a prescribed external electromagnetic field. We consider here only non-relativistic systems but most of the arguments may be generalised to relativistic ones.

A version of the quantum mechanics of such systems is described in which full electromagnetic gauge arbitrariness is maintained explicitly throughout. The division of orbital magnetism into paramagnetism and diamagnetism is examined and it is shown that only by treating both of these on an equal footing can a gauge invariant treatment of magnetism be constructed. It is shown how static linear response theory may be extended to deal simultaneously with interactions that are both linear and quadratic in the perturbing fields, such as those responsible for paramagnetism and diamagnetism. A discussion of the Aharonov–Bohm effect, which involves gauge effects essentially, is also given. Excellent introductions to the subject of this paper are given by Cohen-Tannoudij *et al.* (1977) and particularly by Sakurai (1985). A comprehensive and accessible mathematical development of many of the issues raised in this paper is given by Felsager

* Refereed paper based on a contribution to the Sixth Gordon Godfrey Workshop on Recent Developments in Magnetism and Magnetic Materials, held at the University of New South Wales, Sydney, in October 1996.

10.1071/PH97024 0004-9506/97/061061$10.00

(1981) who extends the arguments further to review the quantum mechanics of magnetic monopoles in which gauge ideas play an important part.

2. What is Gauge?

The electric $\boldsymbol{E}$ and magnetic $\boldsymbol{B}$ fields described by Maxwell's equations,

$$\nabla\boldsymbol{.}\boldsymbol{B}=\mathbf{0}, \qquad \nabla\boldsymbol{.}\boldsymbol{E}=\rho/\epsilon_0\,,$$

$$\nabla\times\boldsymbol{E}+\partial\boldsymbol{B}/\partial t=\mathbf{0}, \qquad \nabla\times\boldsymbol{B}-c^{-2}\partial\boldsymbol{E}/\partial t=\mu_0\,\boldsymbol{J}\,,$$

which are functions of position $\boldsymbol{r}$ and time t and where ρ and $\boldsymbol{J}$ are the charge and current densities, may be said to have some degree of physical reality. This is because of the observable acceleration that results from the Lorentz force $\boldsymbol{F}$ that they exert on a particle of charge e and mass m moving with velocity $\boldsymbol{v}$:

$$\boldsymbol{F}=e(\boldsymbol{E}+\boldsymbol{v}\times\boldsymbol{B})=m\,\mathrm{d}\boldsymbol{v}/\mathrm{d}t\,. \tag{1}$$

However, Maxwell's theory may be formulated more conveniently in terms of the scalar and vector potentials ϕ and $\boldsymbol{A}$, also functions of $\boldsymbol{r}$ and t, which give rise to the electromagnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$ through their space and time derivatives,

$$\boldsymbol{B}=\nabla\times\boldsymbol{A} \quad \text{and} \quad \boldsymbol{E}=-\nabla\phi-\partial\boldsymbol{A}/\partial t\,, \tag{2}$$

with ∇ being the vector differential operator $\nabla=\hat{\boldsymbol{x}}\partial/\partial x+\hat{\boldsymbol{y}}\partial/\partial y+\hat{\boldsymbol{z}}\partial/\partial z$ and $\boldsymbol{A}$ is a polar vector.

The equations that relate these potentials to their sources ρ and $\boldsymbol{J}$ are obtained by substituting equations (2) into the Maxwell equations (see for example Cohen-Tannoudij *et al.* 1977; Craig and Thirunamachandran 1984; Sakurai 1985; Doughty 1990) and are

$$\partial^2\boldsymbol{A}/\partial(ct)^2-\nabla^2 A+\nabla(\nabla\boldsymbol{.}\boldsymbol{A}+c^{-2}\partial\phi/\partial t)=\mu_0\,\boldsymbol{J}\,, \tag{3a}$$

$$\partial^2\phi/\partial(ct)^2-\nabla^2\phi-\partial/\partial t(\nabla\boldsymbol{.}\boldsymbol{A}+c^{-2}\partial\phi/\partial t)=\rho/\epsilon_0\,. \tag{3b}$$

The potentials cannot be measured directly and have a degree of arbitrariness associated with them. For example, from equation (2), it is clear that if either potential is changed by an amount that is constant in space and time the fields are unchanged. Because of the mathematical identity that curl of the gradient of any scalar function is identically zero, it follows from $\boldsymbol{B}=\nabla\times\boldsymbol{A}$ that if $\boldsymbol{B}$ is zero in some region of space then $\boldsymbol{A}$ may be expressed as the gradient of some scalar function which may be non-zero and spatially varying as may $\boldsymbol{A}$ itself. In other words $\boldsymbol{A}$ may be finite and may vary in space and time even when $\boldsymbol{B}$ is zero. As an example, consider an infinite cylindrical solenoid of radius R concentric with the z axis (see Fig. 2 later). Inside the solenoid the magnetic field has the constant value B along the z axis, outside it is zero. The magnetic flux Φ within the solenoid is $\pi R^2 B$. By applying Stokes' theorem to the first of equations (2) the relation

$$\oint \boldsymbol{A}\,.\,\mathrm{d}\boldsymbol{I} = \int \boldsymbol{B}\,.\,\boldsymbol{s} \tag{4}$$

is obtained, where the line integral of $\boldsymbol{A}$ is around the boundary of an open surface $\boldsymbol{s}$ over which the surface integral is taken. By taking this line integral around a circle of radius r concentric with the z axis it is readily found that $A_\theta = \Phi r/2\pi R^2$ for $r \leq R$ and $A_\theta = \Phi/2\pi r$ for $R \geq R$, where (r, θ, z) are the cylindrical coordinates. The other components of $\boldsymbol{A}$ are zero. It is easy to verify from (2) that this potential does indeed give $\boldsymbol{B} = \hat{\boldsymbol{z}}B$ for $r \leq R$ and $\boldsymbol{B} = 0$ for $r \geq R$. For $r \geq R$, where $\boldsymbol{B}$ is zero, it is apparent that $\boldsymbol{A}$ may be expressed as the gradient of the multi-valued scalar function $g(r, \theta, z, t)$ by means of $\boldsymbol{A} = \nabla g$ with $g = \Phi\theta/2\pi$, which increases by Φ each time the angular coordinate θ winds rounds the origin.

This arbitrariness is quite general in classical electrodynamics. To be precise, if the vector and scalar electromagnetic potentials $\boldsymbol{A}$ and ϕ are transformed to $\boldsymbol{A}_\chi$ and ϕ_χ, where

$$\boldsymbol{A}_\chi = \boldsymbol{A} + \nabla_\chi \quad \text{and} \quad \phi_\chi = \phi - \partial_\chi/\partial t\,, \tag{5}$$

then it is easy to confirm that the electromagnetic fields given by (2) are unchanged by the transformation. The quantity $\chi(\boldsymbol{r},\ t)$ which is an arbitrary scalar function of position and time is known as the gauge function and the transformation as a gauge transformation of the potentials. Because the integral in (4) gives the magnetic flux that threads the loop it is seen to be necessary for the condition

$$\oint \nabla\chi\,.\,\mathrm{d}\boldsymbol{l} = 0$$

to hold for any closed path of integration. This means that χ must be single-valued; it must also be continuously differentiable for the fields to be physically realisable. The values of $\boldsymbol{A}$ and ϕ for the gauge function taking on a particular value is known as the gauge and $\boldsymbol{A}$ and ϕ are known as gauge potentials and $\boldsymbol{B}$ and $\boldsymbol{E}$ as gauge fields. If $\boldsymbol{B} = 0$, as for the case of the region outside the infinite solenoid, $\boldsymbol{A}$ can be expressed as the gradient of a scalar function and in that case is known as a pure gauge potential.

The gauge function is seen to be a generalisation of the constant of integration that occurs in calculus, for example the differential equation $B(x) = \mathrm{d}A(x)/\mathrm{d}x$ has the solution

$$A(x) = \int B(x)\,\mathrm{d}x + C\,,$$

where C is the constant of integration.

3. Gauge Invariance of Quantum Mechanics

The main objective of quantum mechanics is to find the solutions of the Schrödinger equation $\mathcal{S}_0\,\Psi_0(\boldsymbol{r},\ t) = 0$ where the Schrödinger wave equation operator $\mathcal{S}_0 = \mathcal{H}_0 - \mathrm{i}\hbar\partial/\partial t$ and $\mathcal{H}_0$ is the Hamiltonian written in the gauge with $\chi = 0$, which is denoted by the subscript 0,

$$\mathcal{H}_0 = (\boldsymbol{p} - e\boldsymbol{A})^2/2m + e\phi\,, \tag{6}$$

and where $\boldsymbol{p}$ is the canonical momentum operator $\boldsymbol{p} = -\mathrm{i}\hbar\nabla$ and $\Psi_0(\boldsymbol{r}, t)$ is the wavefunction. The reason why the potentials appear in the Hamiltonian in this form, called minimal coupling because the charge is coupled to the potentials and not to their gradients (the fields $\boldsymbol{E}$ and $\boldsymbol{B}$), comes from the Lagrangian and Hamiltonian formulations of particle dynamics which reproduce the equation of motion (1) (see the references given above and Griffith 1961).

The principle of gauge invariance states that all the physical observable predictions of quantum mechanics are independent of the gauge that is used in a calculation. This may be demonstrated in the following way. For general gauge χ the Hamiltonian of equation (6) becomes, using (5),

$$\mathcal{H}_\chi = \{\boldsymbol{p} - e(\boldsymbol{A} + \nabla\chi)\}^2/2m + e(\phi - \partial\chi/\partial t)\,. \tag{7}$$

We show that if the wavefunction of the system that undergoes the gauge transformation is itself transformed by a transformation of the phase to

$$\Psi_\chi(\boldsymbol{r},\ t) = \Psi_0(\boldsymbol{r},\ t)\exp\{\mathrm{i}e\chi(\boldsymbol{r},\ t)/\hbar\}\,, \tag{8}$$

then (a) the transformed Hamiltonian $\mathcal{H}_\chi$ and wavefunction Ψ_χ will obey a time dependent Schrödinger wave equation $\{\mathcal{H}_\chi - \mathrm{i}\hbar\partial/\partial t\}\Psi_\chi(\boldsymbol{r}, t) = 0$ of the same form as the untransformed one involving $\boldsymbol{A}$ and ϕ and (b) the physical predictions of the theory are unchanged. We note that in semi-classical quantum mechanics the gauge function, like the fields, is taken to not have any operator properties (it is sometimes called a c-number, following Dirac).

First consider the operator $\{\boldsymbol{p} - \mathrm{e}(\boldsymbol{A} + \nabla\chi)\}$ acting on Ψ_χ. This gives $\exp(\mathrm{i}e\chi/\hbar)$ $(\boldsymbol{p} - e\boldsymbol{A})\Psi_0$ because, using the chain rule for differentiation, the gradient operator associated with $\boldsymbol{p}$ acting on the phase of the transformed wavefunction produces a term that cancels that coming from the $e\nabla\chi$ term. Applying the operator a second and further times leads to

$$\{\boldsymbol{p} - e(\boldsymbol{A} + \nabla\chi)\}^n\Psi_\chi = \exp(\mathrm{i}e\chi/\hbar)(\boldsymbol{p} - e\boldsymbol{A})^n\Psi_0\,,$$

a relation that is true in particular for the kinetic energy term with $n = 2$. Next let the term $\{e(\phi - \partial\chi/\partial t) - \mathrm{i}\hbar\partial/\partial t\}$ act on Ψ_χ. This gives exp $(\mathrm{i}e\chi/\hbar)\{e\phi - i\hbar\partial/\partial t\}\Psi_0$ because the time derivative operator acting on the phase of the transformed wavefunction produces a term that cancels the $e\partial\chi/\partial\mathrm{t}$ term. We see that $\mathcal{S}_0\,\Psi_0 = 0$ implies $\mathcal{S}_\chi\,\Psi_\chi = 0$ or, in other words, that the Schrödinger equation is invariant in form under a gauge transformation which comprises the transformations of the potentials and of the wavefunction given by equations (5) and (8). The Schrödinger equation is said to be *gauge covariant.*

It is clear from (8) that the charge density $\rho = e\Psi^*\Psi$ is independent of gauge provided that χ is real. Although the gauge function may be complex in classical electrodynamics, quantum mechanics requires it to be real so that the normalisation of the wavefunction is preserved. The expression for the expectation value of the electric current density $\boldsymbol{J}$ which appears in the continuity equation

$$\nabla \cdot \boldsymbol{J} + \partial\rho/\partial t = 0\,, \tag{9}$$

that is obtained from the Schrödinger equation is

$$\langle \boldsymbol{J}_0 \rangle = -\mathrm{i}\{\Psi_0^*(\nabla\Psi_0) - (\nabla\Psi_0^*)\Psi_0\}e\hbar/2m - e^2\boldsymbol{A}\Psi_0^*\,\Psi_0/m \tag{10}$$

(Cohen-Tannoudij *et al.* 1977). If $\langle \boldsymbol{J}_\chi \rangle$ is calculated in gauge χ with the wavefunction of (8) it is found that the two terms arising from the effect of the gradients acting on the phase cancel the extra term in the last term arising from the $\nabla\chi$ of (5). The result is that $\langle \boldsymbol{J}_\chi \rangle = \langle \boldsymbol{J}_0 \rangle$ and the expectation values of both the charge density and current density are independent of gauge. These observable physical properties of the system do not depend on the gauge. A gauge transformation in quantum mechanics consists of the combination of the gauge transformation of the potentials of (5) and the phase transformation of the wavefunction of (8). Of course, if the gauge function is a constant independent of $\boldsymbol{r}$ and t, making (8) a *global* phase transformation, then the above results are trivial. It is the (arbitrary) local space and time dependence of the gauge function that creates the complications.

It is also interesting that the *requirement* that the physical content of the wavefunction be unchanged under the *local* phase transformation of the wavefunction given by (8) can be used to demonstrate the necessity for the existence of the electromagnetic fields described by Maxwell's equations (Kobe 1978). Such notions have been extended to develop the theory of what are known as *gauge fields* (Doughty 1990; Felsager 1981) and all of the fundamental interactions in nature are believed to be of this type. The term 'gauge' used for this collection of ideas stems from the attempt of the mathematician H. Weyl to develop a geometric theory of electromagnetism in which the scale (gauge) of space measurement was varied (Sakurai 1985). 'Phase' would appear to be a more appropriate word to use in a modern context but the use of the obscurely sounding 'gauge' has been hallowed by tradition.

4. Gauge Transformations

Before any calculation can be carried out explicitly it has been believed to be necessary to remove the gauge arbitrariness by setting the gauge to a specific value, or 'fixing' it. Because a calculation will give the same numerical results whatever the gauge that is used, it is clearly best to use the gauge in which the calculation is simplest. This is often the gauge with χ set equal to zero and this is the gauge that is adopted most frequently, invariably without further comment. However, the arbitrary gauge function χ may always be set equal to the sum of a particular function χ^{P} and another arbitrary function χ'. The particular function can often be chosen to have properties that will simplify the calculation even further; the arbitrary function, as before, is eventually set to zero. We discuss several such gauge transformations that are used to simplify the forms of the dynamical equations.

As the simplest example take the potentials $\phi = \text{constant}$, $\boldsymbol{A} = \boldsymbol{B} \times \boldsymbol{r}/2$, where $\boldsymbol{r}$ is the position vector and $\boldsymbol{B}$ a fixed classical field. By equation (2) these give a uniform magnetic field $\boldsymbol{B}$ and zero electric field. Consider the gauge transformation with $\chi = -(\boldsymbol{B} \times \boldsymbol{R}) \cdot \boldsymbol{r}/2 + \chi'$, where $\boldsymbol{R}$ is a time independent vector in real space. In the new gauge this gives rise to the potentials $\boldsymbol{A}_\chi = -\boldsymbol{B} \times (\boldsymbol{r} - \boldsymbol{R})/2 + \nabla\chi'$ and $\phi_\chi = \phi - \partial\chi'/\partial t$. It is seen that this gauge transformation has the effect of shifting the origin of the vector potential. Full freedom to specify an arbitrary gauge remains through the presence of χ'.

In dealing with systems in a uniform magnetic field it is convenient to orient the field along the z axis. With $\boldsymbol{A} = \boldsymbol{B} \times \boldsymbol{r}/2$ the vector potential has components $\boldsymbol{A} = (-y,\ x,\ 0)B/2$ in Cartesian coordinates or $\boldsymbol{A} = (0,\ 1,\ 0)Br/2$ in cylindrical ones. This is called the *symmetric* gauge. Consider now the gauge transformation with $\chi = \eta xyB/2 + \chi'$, where η is an arbitrary number. In Cartesian coordinates the transformed vector potential is $\boldsymbol{A}_\eta = \{-y(1-\eta),\ x(1+\eta),\ 0\}B/2$, dropping the χ'. For $\eta = 0$ we get the symmetric gauge, for $\eta = +1$ the gauge $\boldsymbol{A} = \{0,\ x,\ 0\}B$ and for $\eta = -1$, $\boldsymbol{A} = \{-y,\ 0,\ 0\}B$. The latter two are called *Landau* gauges, their advantage is that the vector potential only varies in one direction. The wavefunctions obtained from a calculation done in any one of them may be related to those done in another by means of (8).

There are three components of the vector potential and one scalar potential. The freedom to choose the gauge means that one of these four components may be made zero. For example, the *temporal* gauge is obtained by taking

$$\chi = \int_{t_0}^{t} \phi(\boldsymbol{r}, t')\ \mathrm{d}t' + \chi' .$$

This gives $\phi_\chi = -\partial\chi'/\partial t$ so the scalar potential vanishes completely if $\partial\chi'/\partial t = 0$. The *axial* gauge is obtained by setting one of the components of the vector potential to zero in a similar manner. Because of the integrations that are involved these two gauges are non-local in time and space respectively. The *dipole* or *Goppert-Mayer* (1931) gauge is useful when the potentials are expanded about the origin at, say, the centre of an atom. This gauge is obtained by using the gauge function $\chi = -\boldsymbol{r} \cdot \boldsymbol{A} + \chi'$ and gives $\phi_\chi = \phi - (\boldsymbol{r} \cdot \nabla)\phi - \boldsymbol{E} \cdot \boldsymbol{r} - \partial\chi'/\partial t$ and $\boldsymbol{A}_\chi = -\boldsymbol{r} \times \boldsymbol{B} - (\boldsymbol{r} \cdot \nabla)\boldsymbol{A} + \nabla\chi'$. The potentials are seen to be expanded in terms of the position vector $\boldsymbol{r}$, the electric dipole term $-\boldsymbol{E} \cdot \boldsymbol{r}$ being dominant.

The differential equations (3) for the potentials are found to be invariant in form under the gauge transformation of (5); the old values of the potentials are replaced by the new ones. They are gauge covariant. Equations (3) will be simplified considerably if it is possible to set $\nabla \cdot \boldsymbol{A} + c^{-2}\partial\phi/\partial t = 0$ because doing this will separate the equations for $\boldsymbol{A}$ and ϕ. In this case equation (3a) becomes a vector inhomogeneous wave equation and (3b) becomes a scalar one. This desirable separation is achieved in the *Lorentz* gauge in which the gauge function is chosen to be $\chi = \chi^{\mathrm{P}} + \chi'$ with χ^{P} satisfying the inhomogeneous wave equation with $\boldsymbol{A}$ and ϕ specified:

$$\nabla^2\chi^{\mathrm{P}} - c^{-2}\partial^2\chi^{\mathrm{P}}/\partial t^2 = -(\nabla \cdot \boldsymbol{A} + c^{-2}\partial\phi/\partial t) . \tag{11}$$

Gauge arbitrariness remains through χ' but only gauge functions that do not satisfy the homogeneous wave equation $\nabla^2\chi' - c^{-2}\partial^2\chi'/\partial t^2 = 0$ maintain it. In the process of applying this gauge transformation the *gauge freedom*, or the amount of arbitrariness inherent in the gauge is reduced. This may be understood by reference to Fig. 1. In this figure the rectangular box represents the set of all acceptable gauge functions, i.e. those that are single valued and continuously differentiable. The small circle represents the set of functions that are everywhere flat with $\nabla\chi' = 0$ and $\partial\chi'/\partial t = 0$, or symbolically $\partial\chi' = 0$. The oval represents the set of all functions that satisfy the homogeneous wave equation $\nabla^2\chi' - c^{-2}\partial^2\chi'/\partial t^2 = 0$. Since all flat functions satisfy the latter equation trivially the flat functions form a subset of these functions. Before the Lorentz gauge transformation was applied the gauge could only be fixed, or set to zero, by choosing a gauge function from among the flat functions. After the transformation it can be fixed by choosing from a larger set of functions. In this way gauge freedom has been reduced by the transformation as gauge arbitrariness remains only for a smaller set of functions, namely those outside the oval.

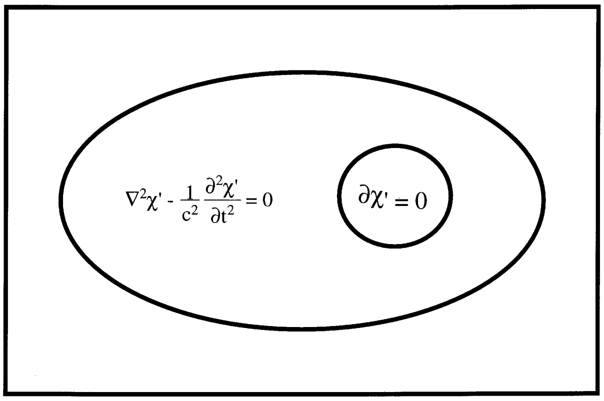

$\nabla^2\chi' - \frac{1}{c^2}\frac{\partial^2\chi'}{\partial t^2} = 0$
$\partial\chi' = 0$

Fig. 1. Illustration of the reduction of gauge freedom resulting from a Lorentz gauge transformation. Within the rectangle is represented the set of all acceptable gauge functions χ', i.e. those that are single valued and continuously differentiable. The oval represents the set of all gauge functions that satisfy the homogeneous wave equation. The small circle represents the set of gauge functions that are everywhere flat, i.e. with $\partial\chi' = 0$ by which is meant $\nabla\chi' = 0$ and $\partial\chi'/\partial t = 0$. All flat functions satisfy the homogeneous wave equation so the circle lies within the oval.

Another gauge that is useful for treating the interaction of radiation with matter (Craig and Thirunamachandran 1984) is the *Coulomb* gauge. In this gauge the condition $\nabla . \boldsymbol{A} = 0$ is imposed by requiring the particular gauge function to satisfy Poisson's equation $\nabla^2\chi^{\mathrm{P}} = -\nabla . \boldsymbol{A}$. Equation (3b) then takes the very simple form $\nabla^2\phi = -\rho/\epsilon_0$; equation (3a) remains more complicated. Gauge freedom remains for functions χ' satisfying Laplace's equation $\nabla^2\chi' = 0$. The *radiation* gauge consists of the combination of the Lorentz gauge and the temporal gauge, the consequence is that $\nabla . \boldsymbol{A} = 0$ also. This gauge is frequently used in treatments of quantum electrodynamics in which the electromagnetic field as well as the particle system is described by quantum mechanics. In Table 1 the gauge transformations discussed above are listed together with the conditions on χ' needed to fix the gauge. Many other gauges are discussed by Leibbrandt (1987).

From Table 1 it can be seen that gauge arbitrariness remains after a gauge transformation. The standard method for proceeding with a calculation is to set χ' to zero and carry out the calculation in that gauge. As shown in Section 3 this will always give the correct value of any physical observable. However, a question of interest that arises is whether the standard operations of quantum mechanics may be carried out in a gauge that remains fully arbitrary.

Table 1. Gauge transformations discussed in Section 4

Each gauge transformation uses a gauge function of the form $\chi = \chi^{\mathrm{P}} + \chi'$, where χ^{P} is a particular function and χ' is an arbitrary one. The gauge fixing condition is the condition on χ' that is sufficient for all gauge arbitrariness to be eliminated from the potentials. Here $\partial\chi' = 0$ indicates both $\nabla\chi' = 0$ and $\partial\chi'/\partial t = 0$, meaning that χ' is flat in space and time

Gauge transformation	Particular gauge function χ^{P}	Gauge fixing condition
Change of origin of $\boldsymbol{A}$ from $\mathbf{0}$ to $\boldsymbol{R}$	$-(\boldsymbol{B} \times \boldsymbol{R}) . \boldsymbol{r}/2$	$\partial\chi' = 0$
Symmetric → Landau	$\eta xyB/2$	$\partial\chi' = 0$
Temporal	$\int_{t_0}^{t} \phi(\boldsymbol{r}, t')\ \mathrm{d}t'$	$\partial\chi'/\partial t = 0$ (for ϕ)
Dipole	$-\boldsymbol{r} . \boldsymbol{A}$	$\partial\chi' = 0$
Lorentz	Equation (11)	$\nabla^2\chi' + c^{-2}\partial^2\chi'/\partial t^2 = 0$
Coulomb	$\nabla^2\chi^{\mathrm{P}} = -\nabla . \boldsymbol{A}$	$\nabla^2\chi' = 0$
Radiation	$\nabla . \boldsymbol{A} = 0$ and $\phi = 0$	Already fixed

A problem closely related to this that has attracted interest, namely how probability amplitudes may be defined in the presence of a time dependent gauge function, was addressed by Yang (1976) following a comment made by Lamb (1952). Contributions to this and similar matters were made subsequently by many authors, see for example Kobe (1978, 1984), Power and Thirunamachandran (1978), Aharonov and Au (1981), Haller (1984), Schlicher *et al.* (1984), Healy (1988) and Power (1989). Clearly such a question can only be answered by carrying out a calculation in which the gauge function is retained explicitly and not set equal to zero.

In the next sections of this paper we describe the formal structure of quantum mechanics when full gauge arbitrariness is preserved throughout. For the purpose of this paper quantum mechanics is taken to be Schrödinger's wave mechanics. Of course, at the end of any calculation all observable quantities must be independent of gauge and so must agree with the results of a conventional calculation which assumes simply that the gauge function is zero. As expected this will always be found to be the case.

5. Basis Functions

We write out explicitly the Schrödinger wave equation $\mathcal{S}_\chi \Psi_\chi = 0$ for a single particle in an arbitrary gauge χ

$$[\{\boldsymbol{p} - e(\boldsymbol{A} + \nabla\chi)\}^2/2m + e(\phi - \partial\chi\partial t) - \mathrm{i}\hbar\partial/\partial t]\Psi_\chi(\boldsymbol{r},\ t) = 0\,. \tag{12}$$

It can be seen that due to the presence of χ the Hamiltonian, which is the sum of the first two operators, is time dependent even if the potentials $\boldsymbol{A}$ and ϕ are static. In this situation the wavefunction cannot be separated into the product of a time dependent part and a space dependent part and so the time-dependent wave equation must be used throughout. However, if $\boldsymbol{E}$ and $\boldsymbol{B}$ have no time dependence it follows from (2) that a set of potentials may be found that has no time dependence either and is a function of $\boldsymbol{r}$ alone. We write these potentials as $\boldsymbol{A}^0(\boldsymbol{r})$ and $\phi^0(\boldsymbol{r})$. One such example, the potentials for a long solenoid, was derived in Section 2. To solve (12) in this case we first set χ to zero. This gives rise to

$$\{(\boldsymbol{p}-e\boldsymbol{A}^0)^2/2m\ +e\phi^0\}\Psi_0(\boldsymbol{r},\ t)=\mathrm{i}\hbar\partial/\partial t\ \ \Psi_0(\boldsymbol{r},\ t)\,. \tag{13}$$

Since the operator on the left-hand side is independent of time the wavefunction may be factored into parts that are respectively space and time dependent. These separate in the usual way to give

$$\Psi_{0,n}(\boldsymbol{r},t)=\psi_n(\boldsymbol{r})\exp(-\mathrm{i}E_nt/\hbar)\,, \tag{14}$$

with E_n and $\psi_n(\boldsymbol{r})$ given by the eigenvalue equation

$$\{(\boldsymbol{p}-e\boldsymbol{A}^0)^2/2m\ +e\phi^0\}\psi_n(\boldsymbol{r})=E_n\,\psi_n(\boldsymbol{r})\,, \tag{15}$$

whose solutions are assumed to be known. They are complete and orthonormal because the Hamiltonian operator is Hermitian. We can now restore gauge dependence by making use of (8) which gives

$$\Psi_{\chi,n}(\boldsymbol{r},\ t)=\psi_n(\boldsymbol{r})\exp\{\mathrm{i}(e\chi-E_n\,t)/\hbar\}\,. \tag{16}$$

We emphasise that we have not fixed the gauge (at a value of zero) in this process. It may be verified by substitution that equation (16) is a solution of (13) with the time independent potentials $\boldsymbol{A}^0$ and ϕ^0. The $\Psi_{\chi,n}$ are time dependent solutions of the wave equation for time independent fields. Their time and space dependences are inseparably linked together by the gauge function. Any linear combination of them $\Psi_\chi(\boldsymbol{r},\ t)=\Sigma_n\,a_n\,\Psi_{\chi,n}(\boldsymbol{r},\ t)$, where the a_n are independent of time, is also a solution of (13) with static fields. In this situation the gauge-explicit probability amplitude for the system to be in state m at time t is defined to be the projection of state m onto the wavefunction $\Psi_\chi(\boldsymbol{r},\ t)$:

$$\text{Probability amplitude}=\int\Psi^*_{\chi,m}(\boldsymbol{r},\ t)\,\Psi_\chi(\boldsymbol{r},\ t)\ \mathrm{d}\boldsymbol{r}\,, \tag{17}$$

which is equal to a_m due to the orthonormality of the $\psi_n(\boldsymbol{r})$ and so is independent of gauge and, in this present case, of time. The probability of the system being in a particular state m is equal to $|a_m|^2$. If the operator on the left side of (15) is the exact Hamiltonian operator the solutions are exact. If it is only part of the total Hamiltonian then the solutions form a basis set with which perturbation theory may be carried out.

6. Gauge Properties of Operators

The operator $\mathcal{O}_\chi$ in gauge χ is defined by the relation $\mathcal{O}_\chi(\boldsymbol{A},\ \phi)=\mathcal{O}_0(\boldsymbol{A}+\nabla\chi,\ \phi-\partial\chi/\partial t)$, where $\mathcal{O}_0$ is the operator with zero gauge function. A gauge *independent* operator $\mathcal{I}$ is defined to be one that is unchanged by a gauge transformation so that $\mathcal{I}_\chi(\boldsymbol{A},\ \phi)=\mathcal{I}_0(\boldsymbol{A},\ \phi)$. An operator such as $\boldsymbol{r}$ or $\boldsymbol{p}$ that does not depend explicitly on the potentials is gauge independent.

As discussed in Section 3 the Schrödinger equation (5) $\mathcal{S}_\chi \mathrm{e}^{\mathrm{i}e\chi/\hbar}\,\Psi_0 = 0$ transforms unitarily to $\mathcal{S}_0\,\Psi_0 = 0$ under a gauge transformation. This requires that

$$\mathcal{S}_\chi = \mathrm{e}^{\mathrm{i}e\chi/\hbar}\,\mathcal{S}_0\,\mathrm{e}^{-\mathrm{i}e\chi/\hbar}\,. \tag{18}$$

The first exponential factor is needed to make the transformation unitary. Any operator that satisfies such a relation is said to be gauge *invariant.* From Section 3 it can be seen that any operator $\mathcal{O}$ that has the functional form $\mathcal{O}_0 = \mathcal{O}_0(\boldsymbol{r}, t, \boldsymbol{p}-e\boldsymbol{A}, \mathrm{i}\hbar\partial/\partial t - e\phi)$ will be gauge invariant if the operator can be expanded in sums of powers of its arguments. An important property of gauge invariant operators is that their matrix elements are independent of gauge. This is shown by using equations (8) and (18) so that

$$\langle\Psi_{\chi,m}|\mathcal{O}_\chi|\Psi_{\chi,n}\rangle = \langle\psi_m(\boldsymbol{r})\exp(\mathrm{i}E_m\,t/\hbar)|\mathcal{O}_0|\psi_n(\boldsymbol{r})\exp(-\mathrm{i}E_n\,t/\hbar)\rangle\,. \tag{19}$$

If $\mathcal{O}$ does not contain a time derivative the matrix element is

$$\langle\Psi_{\chi,m}|\mathcal{O}_\chi|\Psi_{\chi,n}\rangle = \exp\{\mathrm{i}(E_m - E_n)t/\hbar\}\int \psi_m^*(\boldsymbol{r})\,\mathcal{O}_0\,\psi_n(\boldsymbol{r})\;\mathrm{d}\boldsymbol{r}\,. \tag{20}$$

It is apparent that an operator that represents a physical observable is required to be gauge invariant but need not be gauge independent.

The semi-classical Hamiltonian operator that we use in this paper is neither gauge invariant nor gauge independent. From examination of equations (6) and (7) it is seen that

$$\mathcal{H}_\chi = \mathrm{e}^{\mathrm{i}e\chi/\hbar}\,\mathcal{H}_0\,\mathrm{e}^{-\mathrm{i}e\chi/\hbar}\; - e\partial\chi/\partial t\,. \tag{21}$$

Its matrix elements are

$$\langle\Psi_{\chi,m}|\mathcal{H}_\chi|\Psi_{\chi,n}\rangle = (E_n\; - e\partial\chi/\partial t)\delta_{m,n} \tag{22}$$

due to the orthonormality of the $\psi_n(\boldsymbol{r})$, where δ is the Kronecker delta. So although there is a gauge dependent shift of individual energies, energy *differences* remain unchanged (Cohen-Tannoudij *et al.* 1977, Vol. 1, p. 326). Spectroscopy is independent of gauge. It should be noted that a treatment of the interaction between charged particles and radiation in which the radiation field as well as the particle system is treated dynamically gives rise to a Hamiltonian that is gauge invariant (Schiff 1968). It is the semi-classical approximation's assumption of prescribed external fields, with their concomitant inability to exchange energy with the radiation field, that destroys the gauge invariance. However, as is shown in Section 10, the lack of gauge invariance of the semi-classical Hamiltonian does not affect the observable predictions of thermodynamics arising from its use.

Table 2. Gauge properties of operators

Operator	$\boldsymbol{r}$	$\boldsymbol{p}$	$(\boldsymbol{A}, \phi)$	$\boldsymbol{v}$, $(\boldsymbol{p}-\mathrm{e}\boldsymbol{A})$, $(\mathrm{i}\hbar\partial/\partial t - e\phi)$	Unit operator	H
Gauge independent	Yes	Yes	No	No	Yes	No
Gauge invariant	Yes	No	No	Yes	Yes	No

The quantum mechanical operator identified with the particle velocity is $\boldsymbol{v} = \mathrm{d}\boldsymbol{r}/\mathrm{d}t = [\boldsymbol{r},\ \mathcal{H}]/\mathrm{i}\hbar$. This is readily verified to be Hermitian as both $\boldsymbol{r}$ and $\mathcal{H}$ are themselves and from (18) it is also gauge invariant. By commuting $\boldsymbol{r}$ with the Hamiltonian of equation (6) the velocity operator is given by $m\boldsymbol{v} = \boldsymbol{p} - e\boldsymbol{A}$. If other terms involving $\boldsymbol{p}$ are present in the Hamiltonian the commutator of $\boldsymbol{r}$ with them will contribute further to the velocity operator. For example, the spin–orbit interaction $\boldsymbol{s} \times \boldsymbol{E}\,\boldsymbol{.}\,(\boldsymbol{p} - e\boldsymbol{A})e\hbar/4m^2c$ (Frohlich and Studer 1993) will add a term $\boldsymbol{s} \times \boldsymbol{E}e\hbar/4mc$ to $m\boldsymbol{v}$.

It is important to distinguish the notions of gauge invariance, defined by (18), and gauge independence, which means that the quantity concerned does not change when the gauge function is changed. In Table 2 these properties are summarised for various operators.

7. Perturbation Theory

If the potentials are separated into two parts so that $\boldsymbol{A} \rightarrow \boldsymbol{A}^0 + \boldsymbol{A}^1$ and $\phi \rightarrow \phi^0 + \phi^1$ then the quantum wave equation $\{\mathcal{H}_\chi - \mathrm{i}\hbar\partial/\partial t\}\Psi_\chi(\boldsymbol{r},\ t) = 0$ may be written in the following form:

$$\{\mathcal{H}^0_\chi + \mathcal{V}_\chi - \mathrm{i}\hbar\partial/\partial t\}\Psi_\chi(\boldsymbol{r},\ t) = 0\,, \tag{23}$$

where $\mathcal{H}^0_\chi = \{\boldsymbol{p} - e(\boldsymbol{A}^0 + \nabla\chi)\}^2/2m + e(\phi^0 - \partial\chi/\partial t)$, which has the form of an unperturbed Hamiltonian and a perturbation

$$\mathcal{V}_\chi = e\phi^1 - \boldsymbol{A}^1\,\boldsymbol{.}\,\{\boldsymbol{p} - e(\boldsymbol{A}^0 + \nabla\chi)\}e/m + e^2(\boldsymbol{A}^1)^2/2m + \mathrm{i}e\hbar(\nabla\,\boldsymbol{.}\,\boldsymbol{A}^1)/2m\,,$$

where the relation $(\boldsymbol{p}\,\boldsymbol{.}\,\boldsymbol{A} - \boldsymbol{A}\,\boldsymbol{.}\,\boldsymbol{p}) = -\mathrm{i}\hbar(\nabla\,\boldsymbol{.}\,\boldsymbol{A})$ has been used. The perturbation depends on χ but it can be seen to satisfy the relation $\mathcal{V}_\chi = \exp(\mathrm{i}e\chi/\hbar)\ V_0\ \exp(-\mathrm{i}e\chi/\hbar)$ because the operator $\boldsymbol{p}$ generates the $e\nabla\chi$ term from the phase factor on the right that contains the gauge function. Any additional perturbation must be gauge invariant to satisfy the relation. To illustrate the importance of this requirement, the spin–orbit interaction is sometimes incorrectly taken to be of the form $\xi(\boldsymbol{r})\boldsymbol{l}\,\boldsymbol{.}\,\boldsymbol{s}$ where $\boldsymbol{l} = \boldsymbol{r} \times \boldsymbol{p}$. However, the correct form of this interaction is gauge invariant as it involves the gauge invariant quantities $\boldsymbol{E}$, $(\boldsymbol{p}-\mathrm{e}\boldsymbol{A})$ and $\boldsymbol{s}$ (Frohlich and Studer 1993) and it is in this form that it must be used as a perturbation. For a perturbative approach to be useful it is necessary for the matrix elements associated with $\boldsymbol{V}_0$ to be smaller than those associated with H^0_0. If the magnetic field is zero it is permissible to choose the gauge $\boldsymbol{A}^0 = \boldsymbol{0}$, $\boldsymbol{A}^1 = \boldsymbol{0}$. The operators then reduce to $\mathcal{H}^0_0 = \boldsymbol{p}^2/2m + e\phi^0$, $\mathcal{V}_0 = e\phi^1$, the usual expressions of the perturbation theory of the scalar potential.

(7a) Time Independent Perturbation Theory

If $\boldsymbol{A}^1$ and ϕ^1 are independent of time, then $\mathcal{V}_0$ is as well, so instead of the eigensolutions $\psi_n(\boldsymbol{r})$ and E_n being determined by (15): $\mathcal{H}^0_0\,\psi_n = E_n\,\psi_n$, where $\mathcal{H}^0_0$ is the unperturbed Hamiltonian in the $\chi = 0$ gauge, they will be determined by the new eigenvalue equation $(\mathcal{H}^0_0 + \mathcal{V}_0)\psi_n{}' = E_n{}'\,\psi_n{}'$. This may be solved for the $E_n{}'$ and $\psi_n{}'$ in terms of the E_n and ψ_n by the standard methods of time independent perturbation theory. New gauge dependent wavefunctions may then be constructed with the primed quantities by means of (16) and the expectation values of operators that represent physical quantities may be obtained.

Generally, the quantities under the control of an experimenter are the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ which are determined by the placement of electrodes and magnets. Because of their gauge arbitrariness the potentials are not under such control. Therefore, when the fields are considered to be applied externally and their operator nature is ignored, an appropriate form of perturbation to use is the derivative of the Hamiltonian with respect to these fields

$$\mathcal{V} = \delta\boldsymbol{E}\,\boldsymbol{.}\,\partial\mathcal{H}/\partial\boldsymbol{E} + \delta\boldsymbol{B}\,\boldsymbol{.}\,\partial\mathcal{H}/\partial\boldsymbol{B} + \text{higher terms}\,, \tag{24}$$

where the partial derivative with respect to a vector means the gradient with respect to that vector and $\delta\boldsymbol{E}$ and $\delta\boldsymbol{B}$ are small variations of these fields. This form of perturbation fits readily into the structure of thermodynamic perturbation theory (Stewart 1996*a*). If the potentials corresponding to static uniform fields $\boldsymbol{E}$ and $\boldsymbol{B}$ are taken to be $\phi = -\boldsymbol{E}\,\boldsymbol{.}\,(\boldsymbol{r}-\boldsymbol{R})$ and $\boldsymbol{A} = \boldsymbol{B} \times (\boldsymbol{r}-\boldsymbol{R})/2$, the origin of the potentials being at $\boldsymbol{R}$, then from (6), $-\partial\mathcal{H}/\partial\boldsymbol{E} = e(\boldsymbol{r}-\boldsymbol{R})$ the electric dipole moment. The term $\boldsymbol{E}\,\boldsymbol{.}\,\boldsymbol{R}$ in the Hamiltonian gives rise to a shift of all energy levels and is unobservable.

The derivative of $\mathcal{H}$ with respect to the ith component of $\boldsymbol{B}$ is

$$\partial\mathcal{H}/\partial B_i = \Sigma_j(\partial\mathcal{H}/\partial A_j)\ (\partial A_j/\partial B_i)\,. \tag{25}$$

The quantity $\partial\mathcal{H}/\partial\boldsymbol{A}$ is given by $\mathcal{H}(\boldsymbol{A}+\delta\boldsymbol{A})\,\boldsymbol{.}\,(\boldsymbol{p}-\mathrm{e}\boldsymbol{A})e/m$ for a gauge with $\nabla\,\boldsymbol{.}\,\boldsymbol{A} = 0$. For a uniform magnetic field, $\partial A_i/\partial B_j = \Sigma_k\,\epsilon_{ijk}\,r'_k/2$, where the anti-symmetric unit tensor ϵ_{ijk} is zero if any two of the subscripts are the same, is unity if they are in cyclic order and zero otherwise and $\boldsymbol{r}' = \boldsymbol{r}-\boldsymbol{R}$. Therefore, we have

$$\frac{\partial\mathcal{H}}{\partial B_x} = \tfrac{1}{2}\left(y'\,\frac{\partial\mathcal{H}}{\partial A_z} - z'\,\frac{\partial\mathcal{H}}{\partial A_y}\right), \tag{26}$$

etc. and so $-\partial\mathcal{H}/\partial\boldsymbol{B} = \boldsymbol{r}' \times (\boldsymbol{p}-\mathrm{e}\boldsymbol{A})e/2m = \boldsymbol{m}$. The latter is the expression for the orbital magnetic moment operator $\boldsymbol{m}$ in non-relativistic quantum mechanics, the first and second terms representing the paramagnetic and diamagnetic contributions respectively. Since the paramagnetic and diamagnetic terms are not individually gauge invariant they are not observable individually, only the sum of them is (Stewart 1996*b*), a matter that will be discussed further in Section 8. The expression for $\partial\mathcal{H}/\partial\boldsymbol{B}$ involves a cross product with $(\boldsymbol{r}-\boldsymbol{R})$. The term containing $\boldsymbol{R} \times (\boldsymbol{p}-\mathrm{e}\boldsymbol{A})$ is the cross product of the constant vector $\boldsymbol{R}$ with an operator that represents a drift current. By Maxwell's equations a drift current is inconsistent with a uniform magnetic field and consequently in this case the expectation value of $\partial\mathcal{H}/\partial\boldsymbol{B}$ is independent of $\boldsymbol{R}$; this is shown in more detail in Section 8. The only remaining non-zero derivative is $\partial^2\mathcal{H}/\partial B_i\,\partial B_j = -\hat{\chi}^d_{ij}$ where the operator (not to be confused with the gauge function) is $\hat{\chi}^d_{ij} = -({r'}^2\delta_{i,j} - r'_i\,r'_j)e^2/4m$ (Stewart 1996*a*, 1996*b*). The derivatives of the Hamiltonian have been calculated in the gauge with $\chi = 0$. Since $(\boldsymbol{p}-e\boldsymbol{A})$ is a gauge invariant operator the derivatives of the Hamiltonian with respect to the fields are gauge invariant.

(7b) Time Dependent Perturbation Theory

When the perturbing fields and potentials depend on time the wave equation will have the form of (23) with $\mathcal{V}$ being a perturbation that is now time dependent. In this situation the wavefunction may be expressed in the form

$$\Psi_\chi(\boldsymbol{r},\ t) = \Sigma_n\, a_n(t)\, \Psi_{\chi,n}(\boldsymbol{r},\ t)\,, \tag{27}$$

where the $\Psi_{\chi,n}(\boldsymbol{r},\ t)$ are the basis functions of (14) containing the $\psi_n(\boldsymbol{r})$ that are solutions of $\mathcal{H}_0^0\, \psi_n(\boldsymbol{r}) = E_n\, \psi_n(\boldsymbol{r})$ given in (15) for the time independent fields $\boldsymbol{A}^0$ and ϕ^0, but now the a_n depend on time. If the explicit form, given by (14), of the wavefunction of (27) is substituted into (23) then, with the help of the result

$$\{\mathcal{H}_\chi^0 - \mathrm{i}\hbar\partial/\partial t\}\Psi_\chi(\boldsymbol{r},\ t) = -\mathrm{i}\hbar\Sigma_n\, \psi_n(\boldsymbol{r}) \exp\{\mathrm{i}(e\chi - E_n\, t)/\hbar\}\, \mathrm{d}a_n/\mathrm{d}t\,, \tag{28}$$

which is a consequence of (23), and after using the relation $\mathcal{V}_\chi = \exp(\mathrm{i}e\chi/\hbar)$ $\mathcal{V}_0 \exp(-\mathrm{i}e\chi/\hbar)$ and making the appropriate cancellations the gauge function disappears altogether and the result

$$\mathrm{i}\hbar\, \frac{\mathrm{d}a_m}{\mathrm{d}t} = \textstyle\sum_n \mathrm{e}^{\mathrm{i}(E_m - E_n)t/\hbar}\, a_n\, \mathcal{V}_{mn}(t) \tag{29}$$

is obtained, where

$$\mathcal{V}_{mn}(t) = \int \psi_m^*(\boldsymbol{r})\, \mathcal{V}_0(t)\, \psi_n(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r}\,. \tag{30}$$

Equation (29) is exactly the same as is obtained in the conventional treatment of quantum mechanics where gauge is ignored (Cohen-Tannoudij *et al.* 1977) and it does not depend upon $\mathcal{V}$ being a small quantity. The probability amplitude in arbitrary gauge for the system to be in state m at time at t is given by (17). By using equations (16) and (27) and the orthonormality of the $\psi_n(\boldsymbol{r})$ this amplitude is found to be simply $a_m(t)$. This also is identical to that obtained when $\chi = 0$. Probability amplitudes are independent of gauge.

8. Paramagnetism and Diamagnetism

The gauge is of significance for the orbital motion of a particle because, by the property that the canonical momentum $\boldsymbol{p} = -\mathrm{i}\hbar\nabla$ has of being a derivative operator, it operates on the gauge function through (8). On the other hand the spin $\boldsymbol{s}$ may be ignored in connection with the gauge because it appears in the Hamiltonian in the form $-\boldsymbol{s}\,.\,\boldsymbol{B}e/m$ and does not operate explicitly on the gauge function. This characteristic of the orbital motion has consequences for the orbital magnetic moment. The classical expression for the orbital magnetic moment $\boldsymbol{m}$ of a particle of charge e about the point $\boldsymbol{R}'$ is $\boldsymbol{m} = (\boldsymbol{r}-\boldsymbol{R}') \times \boldsymbol{v}e/2$. Because we deal with a physical system of finite extent, such as a molecule, the expectation value of the drift velocity $\boldsymbol{v}$ is zero and consequently the expectation value of $\boldsymbol{m}$ is independent of $\boldsymbol{R}'$.

The operator for the orbital moment is therefore composed of two terms $\boldsymbol{m} = \boldsymbol{m}^{\mathrm{p}}+\boldsymbol{m}^{\mathrm{d}}$, with the velocity $\boldsymbol{v}$ taken to be given by $m\boldsymbol{v} = \boldsymbol{p}-e\boldsymbol{A}$, where the diamagnetic moment is $\boldsymbol{m}^{\mathrm{d}} = -(\boldsymbol{r}-\boldsymbol{R}') \times \boldsymbol{A}e^2/2m$ and the paramagnetic moment is $\boldsymbol{m}^{\mathrm{p}} = (\boldsymbol{r}-\boldsymbol{R}') \times \boldsymbol{p}e/2m$. In calling the latter term paramagnetic no assumption is made that the magnetisation associated with it is proportional to the applied field; it could be a permanent moment. Now let us make the general gauge transformation of (5); $\boldsymbol{m}^{\mathrm{d}}$ becomes $\boldsymbol{m}^{\mathrm{d}}_{\chi} = -(\boldsymbol{r}-\boldsymbol{R}') \times (\boldsymbol{A}+\nabla\chi)e^2/4m$ and $\boldsymbol{m}^{\mathrm{p}}_{\chi} = \boldsymbol{m}^{\mathrm{p}}$ since $\boldsymbol{p}$ is gauge independent. However, the wavefunction changes according to (8). When the effect of the operator $\boldsymbol{p}$ acting on the transformed wavefunction is allowed for, the matrix elements of $\boldsymbol{m}^{\mathrm{p}}$ and $\boldsymbol{m}^{\mathrm{d}}$ in the new gauge are found to be

$$\begin{aligned}\langle\Psi_{\chi}{}'|\boldsymbol{m}^{\mathrm{d}}_{\chi}|\Psi_{\chi}\rangle = &-(e^2/2m)\langle\Psi_0{}'|(\boldsymbol{r}-\boldsymbol{R}') \times \boldsymbol{A}|\Psi_0\rangle \\ &-(e^2/2m)\langle\Psi_0{}'|(\boldsymbol{r}-\boldsymbol{R}')|\Psi_0\rangle \times \nabla\chi\,, \end{aligned} \tag{31a}$$

$$\begin{aligned}\langle\Psi_{\chi}{}'|\boldsymbol{m}^{\mathrm{p}}_{\chi}|\Psi_{\chi}\rangle = &(e/2m)\langle\Psi_0{}'|(\boldsymbol{r}-\boldsymbol{R}') \times \boldsymbol{p}|\Psi_0\rangle \\ &+(e^2/2m)\langle\Psi_0{}'|(\boldsymbol{r}-\boldsymbol{R}')|\Psi_0\rangle \times \nabla\chi\,. \end{aligned} \tag{31b}$$

It is seen that under any gauge transformation the matrix elements of the paramagnetic and diamagnetic moments are changed by equal and opposite amounts. The sum of the two is independent of gauge (Stewart 1996*b*). Since, as shown in Section 4, a change of origin of the vector potential of a uniform magnetic field is equivalent to making a gauge transformation it is proved that the matrix elements of the total orbital moment are independent of the origin of the vector potential. This is a more general and simple derivation than previous ones that applied only to the susceptibility and depended on perturbation methods or a particular basis set of wave functions (Geersten 1989; Griffith 1961; Van Vleck 1932).

We may be more specific. Taking, from Section 4, $\boldsymbol{A} = \boldsymbol{B} \times \boldsymbol{r}/2$ for a uniform magnetic field and $\chi = -(\boldsymbol{B} \times \boldsymbol{R}) \cdot \boldsymbol{r}/2 + \chi'$, so that $\nabla\chi = -(\boldsymbol{B} \times \boldsymbol{R})/2$ (dropping χ' for the moment), equations (31) become

$$\begin{aligned}\langle\Psi_{\chi}{}'|\boldsymbol{m}^{\mathrm{d}}_{\chi}|\Psi_{\chi}\rangle = &-(e^2/4m)\langle\Psi_0{}'|(\boldsymbol{r}-\boldsymbol{R}') \times (\boldsymbol{B} \times \boldsymbol{r})|\Psi_0\rangle \\ &+(e^2/4m)\langle\Psi_0{}'|(\boldsymbol{r}-\boldsymbol{R}')|\Psi_0\rangle \times (\boldsymbol{B} \times \boldsymbol{R})\,, \end{aligned} \tag{32a}$$

$$\begin{aligned}\langle\Psi_{\chi}{}'|\boldsymbol{m}^{\mathrm{p}}_{\chi}|\Psi_{\chi}\rangle = &(e/2m)\langle\Psi_0{}'|(\boldsymbol{r}-\boldsymbol{R}') \times \boldsymbol{p}|\Psi_0\rangle \\ &-(e^2/4m)\langle\Psi_0{}'|(\boldsymbol{r}-\boldsymbol{R}')|\Psi_0\rangle \times (\boldsymbol{B} \times \boldsymbol{R})\,. \end{aligned} \tag{32b}$$

Now take the expectation value in the state Ψ, denoted $\langle\,\rangle$. Further, for convenience choose the origin of coordinates to be at the centre of charge so that $\langle\boldsymbol{r}\rangle = 0$. Because $\boldsymbol{A} = \boldsymbol{B} \times \boldsymbol{r}/2$ it follows that $\langle\boldsymbol{A}\rangle = \boldsymbol{0}$, and since $\boldsymbol{p} = m\mathbf{v} - e\boldsymbol{A}$ and the drift velocity $\langle\boldsymbol{v}\rangle = \boldsymbol{0}$ it follows that $\langle\boldsymbol{p}\rangle = \boldsymbol{0}$ too. Under these conditions the only terms of equations (32) that survive are

$$\langle\boldsymbol{m}^{\mathrm{d}}_{\chi}\rangle = -(e^2/4m)\langle\boldsymbol{r} \times (\boldsymbol{B} \times \boldsymbol{r})\rangle - (e^2/4m)\boldsymbol{R}' \times (\boldsymbol{B} \times \boldsymbol{R})\,, \tag{33a}$$

$$\langle\boldsymbol{m}^{\mathrm{p}}_{\chi}\rangle = (e/2m)\langle\boldsymbol{r} \times \boldsymbol{p}\rangle + (e^2/4m)\boldsymbol{R}' \times (\boldsymbol{B} \times \boldsymbol{R})\,. \tag{33b}$$

Although, as noted above, the expectation value of the total moment is independent of $\boldsymbol{R}'$ this is not true for the individual components. Putting $\boldsymbol{R}' = \boldsymbol{R}$ demonstrates that if $\boldsymbol{m}_R$ denotes the component of the orbital magnetism calculated about the point $\boldsymbol{R}$ with the origin of $\boldsymbol{A}$ at that point also, then

$$\boldsymbol{m}_R^{\mathrm{d}} = \boldsymbol{m}_0^{\mathrm{d}} - (e^2/4m)\boldsymbol{R} \times (\boldsymbol{B} \times \boldsymbol{R}), \quad \boldsymbol{m}_R^{\mathrm{p}} = \boldsymbol{m}_0^{\mathrm{p}} + (e^2/4m)\boldsymbol{R} \times (\boldsymbol{B} \times \boldsymbol{R}) \,. \tag{34}$$

A shift of the origin to $\boldsymbol{R}$ is seen to result in the equal and opposite numerical changes of the paramagnetic and diamagnetic components given above. The changes are quadratic in $\boldsymbol{R}$ and, using the identity $\boldsymbol{R} \times (\boldsymbol{B} \times \boldsymbol{R}) = \boldsymbol{R}^2\boldsymbol{B} - (\boldsymbol{R}\,.\,\boldsymbol{B})\boldsymbol{R}$, maximum when $\boldsymbol{R}$ is perpendicular to $\boldsymbol{B}$. The ith Cartesian component of this latter expression is $\Sigma_j\, B_j(R^2\delta_{i,j} - R_i\, R_j)$ and by differentiating with respect to this component of $\boldsymbol{B}$ the corresponding contribution to the susceptibility may be obtained (Stewart 1996*b*). We remember that we still are able to add and subtract a $\nabla\chi'$ term onto each of (33), so equations (34) are still gauge dependent. They only have meaning individually if χ' is fixed at zero but, of course, their gauge independent sum does not depend on χ'.

The reason why equations (34) may be of interest is that at zero temperature the calculation of the diamagnetic moment requires only a ground state expectation value to be taken, but if the molecule is initially unpolarised the paramagnetic moment is harder to obtain as it must be calculated by second order perturbation theory which requires the excited states of the molecule to be known. Because of this it is sometimes useful to calculate the diamagnetic and paramagnetic moments separately. The question then arises of how to combine the separate moments arising from atoms situated at different origins. The answer to this question is provided by equations (34). Of course, it is only valid to do this if the arbitrary gauge function χ' is the same for the diamagnetic and paramagnetic terms.

9. The Aharonov–Bohm Effect

One of the most fascinating and still controversial issues concerning the vector potential is the Aharonov–Bohm effect (Ehrenberg and Siday 1949; Aharonov and Bohm 1959; see Oliaru and Popescu 1985 and Peshkin and Tonomura 1989 for reviews). A solenoid (Fig. 2) of radius R is concentric with the z axis which is directed out of the paper. As discussed in Section 2 the magnetic field inside the solenoid is B, outside it is zero. A beam of electrons coming from the left in a two slit interference experiment is split into two parts that pass either side of the solenoid without passing through it and are recombined at a detector on the right-hand side. The diffraction pattern at the detector indicates that the electron wavefunction acquires a phase whose magnitude is proportional to the magnetic flux in the solenoid and whose sign depends on which way they travelled around it. This phase difference produces measurable interference effects at the detector, the interference pattern shifting as the flux in the solenoid is varied. This is despite the fact that the electrons have not passed through any region of space in which the magnetic field is non-zero and so have not been subjected to any force of the form of equation (1).

The effect may be understood as follows (Griffiths 1994). Consider the Schrödinger equation for a charged particle moving in a pure gauge field, i.e.

in a region in which the magnetic field is zero, which is the case outside the solenoid. As discussed in Section 2 the vector potential is given by the spatial gradient of a time independent scalar $g(\boldsymbol{r})$ so $\boldsymbol{A} = \nabla g$. For the time being we ignore the arbitrary gauge function. The Schrödinger equation is then

$$\{(\boldsymbol{p} - e\nabla g)^2/2m + e\phi\}\Psi_g(\boldsymbol{r}, t) = \mathrm{i}\hbar\,\partial\Psi_g(\boldsymbol{r}, t)/\partial t\,, \tag{35}$$

the wavefunction depending on g. Now substitute the relation $\Psi_g = \Psi_0 \exp(\mathrm{i}eg/\hbar)$ into (35). Using exactly the same arguments involving the effect of the gradient operator that were used in Section 3 we arrive at

$$(\boldsymbol{p}^2/2m + e\phi)\Psi_0 = \mathrm{i}\hbar\,\partial\Psi_0/\partial t\,. \tag{36}$$

This is the equation of motion for a particle moving in zero magnetic potential. In other words the wavefunction for a particle moving in a pure gauge field is obtained from that in zero gauge field by multiplying it by a phase factor proportional to the potential of the pure gauge field. As shown also in Section 3 the charge and current densities are the same in both cases. However, for the particular gauge field considered here the phase that is acquired depends upon the topology of the path taken by the particle. For the particle travelling through the upper slit in Fig. 2 the phase is given by $e/\hbar$ times $\int \nabla g\,.\,\boldsymbol{dl} = \int A\,.\,\boldsymbol{dl}$, the integral taken over the path ABD. As the integrand is a gradient the integral is independent of path providing that the end points and topology of the path are the same. For the path through the lower slit the phase is given by the integral over ACD. The phase *difference* is therefore given by the closed path integral $(e/\hbar)\oint A\,.\,\boldsymbol{dl}$ around the loop ABDCA. By equation (4) this is equal to $(e/\hbar)$

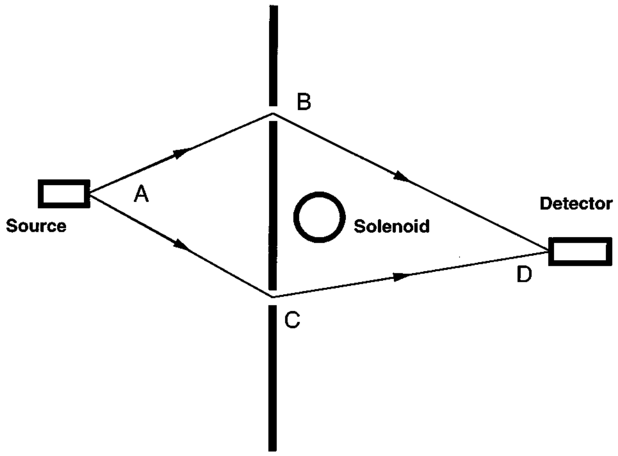


Fig. 2. The Aharonov–Bohm effect. A solenoid of radius R is concentric with the z axis which is directed out of the paper. Inside the solenoid the magnetic field is B, outside it is zero. A beam of electrons coming from the left passes through a two slit interferometer, the two beams pass either side of the solenoid and combine at a detector on the right-hand side of it. By putting suitable electrostatic shields around the solenoid the electron beam is precluded from travelling through its magnetic field. The interference pattern is found to depend on the magnetic flux in the solenoid.

times the magnetic flux enclosed by the loop. This phase difference is responsible for interference observed at the detector, and experimentally the difference is indeed observed to be proportional to the flux. Although only one path is shown in the schematic illustration in Fig. 2, because the phase difference does not depend upon the details of the path the phase difference is the same for all paths of the same topology. It is noted that no interference effect is observable when the phase shift is an integer multiple of 2π, or the flux a multiple of the flux quantum e/h. If the path of integration were taken around a loop that did not enclose the solenoid the phase difference would be zero. The phase factor depends upon the topology of the path that is followed and is said to be *non-integrable*.

What about the gauge function? Since it appears as a gradient it is simply added to g. The integral of its gradient around any closed loop is zero because it is single valued scalar, in contrast to g which is multi-valued. Hence the Aharonov–Bohm phase shift depends only on the flux enclosed by the path and is, as it must be, independent of gauge.

10. Statistical Mechanics

It is assumed that the behaviour of an assembly of particles i, j, is described by the Schrödinger equation $\mathcal{S}\Psi(\boldsymbol{r}_i,\ \boldsymbol{r}_j,\ t) = 0$, where $\mathcal{S} = \mathcal{H} - \mathrm{i}\hbar\partial/\partial t$ and $\mathcal{H}(\boldsymbol{r}_i,\ \boldsymbol{r}_j,\ t)$ is a non-relativistic Hamiltonian consisting of the sum of single-particle terms, for example as in (6), plus an interaction term involving only the coordinates of the particles. The interaction is gauge invariant as it depends only on coordinates so the many particle Schrödinger operator is gauge invariant too. Because the many-body wavefunction is the (in principle infinite) sum of determinants of single particle wavefunctions the gauge function Ξ of the many body wavefunction is the sum of the individual gauge functions of the particles: $\Xi(\boldsymbol{r}_i,\ \boldsymbol{r}_j,\ t) = \Sigma_i\chi(\boldsymbol{r}_i,\ t)$ so that $\Psi_\Xi(\boldsymbol{r}_i,\ \boldsymbol{r}_j,\ t) = \Psi_0(\boldsymbol{r}_i,\ \boldsymbol{r}_j,\ t)\exp(\mathrm{i}e\Xi/\hbar)$. For the time independent fields that are necessary for a thermodynamic treatment the solutions of the wave equations are analogous to those of the single particle situation except that the eigensolutions of (15) are now those for the many particle rather than the single particle Hamiltonian and χ becomes Ξ.

We examine the quantity $\mathrm{Tr}(\mathrm{e}^{-\beta\mathcal{H}_\Xi}\mathcal{O}_\Xi)$ where Tr stands for trace, $\beta = 1/kT$ where T is the temperature and $\mathcal{O}$ is a gauge invariant operator. This quantity is given explicitly by

$$\Sigma_{n,m}\langle n_\Xi|\exp(-\beta\mathcal{H}_\Xi)|m_\Xi\rangle\,\langle m_\Xi|\mathcal{O}_\Xi|n_\Xi\rangle\,. \tag{37}$$

If the states $|n_\Xi(t)\rangle$ over which the trace is taken are the exact solutions of the Hamiltonian (the solutions given in equation 16) then using equations (21) and (22) the trace becomes

$$\mathrm{Tr}\,\mathrm{e}^{-\beta\mathcal{H}_\Xi}\,\mathcal{O}_\Xi = \mathrm{e}^{\beta e\partial\Xi/\partial t}\,\mathrm{Tr}\,\mathrm{e}^{-\beta\mathcal{H}_0}\,\mathcal{O}_0\,, \tag{38}$$

where the trace on the right is taken over states with $\Xi = 0$. If $\mathcal{O}$ is the unit operator we obtain a formally time dependent partition function $Z_\Xi(t) = \mathrm{e}^{\beta e\partial\Xi/\partial t}\,Z_0$, where $Z_0 = \mathrm{Tr}\,\mathrm{e}^{-\beta\mathcal{H}_0}$, the time dependence residing in the gauge function. The free energy is given by the usual expression $F = -kT\ln(Z)$ and is $F_\Xi = -kT\ln(Z_0) - e\partial\Xi/\partial t$. It is gauge dependent, but its derivatives with respect to thermodynamic variables, which give observable thermodynamic quantities, do not contain Ξ and so are gauge independent. The statistical average of any gauge invariant operator $\langle\mathcal{O}_\Xi\rangle = (\mathrm{Tr}\, \mathrm{e}^{-\beta\mathcal{H}_\Xi}\mathcal{O}_\Xi)/Z_\Xi$, for example a spatial correlation function, is gauge independent because the factors involving the gauge function in the numerator and denominator cancel. Also, differences of internal energy, like differences of free energy, are gauge independent. We conclude that the observable quantities calculated with statistical mechanics are independent of gauge.

11. Magnetic Static Linear Response

If a magnetic field is applied to a system in thermal equilibrium its free energy and the statistical average of its magnetic moment will change. For small applied fields the change in magnetic moment is proportional to the field, the constant of proportionality being the susceptibility; the change of free energy is proportional to the square of the field. The description of systems in which the coupling to the external fields in the Hamiltonian is linear, such as spin paramagnetism, is well established. However, from equations (6) or (7) it can be seen that in the case of a static uniform magnetic field $\boldsymbol{B}$ with vector potential $\boldsymbol{A} = \boldsymbol{B} \times \boldsymbol{r}/2$ the coupling has terms that are both linear and quadratic in $\boldsymbol{B}$. In view of the necessity, discussed in Section 8, of maintaining gauge invariance by treating both terms on an equal footing we need to consider the quadratic as well as the linear term. We outline how to do this in this section; more details are given in (Stewart 1996*a*). We work in the gauge with $\chi = 0$ which, as explained before, is permissible so long as we are careful to deal only with gauge invariant operators.

We use the operator identity of Kumar (1965)

$$\frac{\partial}{\partial\mu}\,\mathrm{e}^{-\beta\mathcal{H}} = -\mathrm{e}^{-\beta\mathcal{H}}\int_0^\beta \mathrm{e}^{x\mathcal{H}}\,\frac{\partial\mathcal{H}}{\partial\mu}\,\mathrm{e}^{-x\mathcal{H}}\,\mathrm{d}x\,, \tag{39}$$

where $\mathcal{H}(\mu,\nu)$ is an operator that is a function of the parameters μ and ν which are c-numbers (i.e. not operators themselves, for example the components of the classical magnetic field) but not a function of β. The Hamiltonian $\mathcal{H}$ will not, in general, commute with its derivatives with respect to these parameters. The correctness of (39) is established by noting that if the sides of it are denoted as $Q(\beta)$ then they both satisfy the equation $\partial Q/\partial\beta = -\partial/\partial\mu\{\mathcal{H}\exp(-\beta\mathcal{H})\}$ with $Q(0) = 0$.

We take the trace of (39). The operators on the right-hand side cycle under the trace to eliminate the exponents containing x and the integral is evaluated trivially to give $\partial Z/\partial\mu = -\beta Z\langle\partial\mathcal{H}/\partial\mu\rangle$. The free energy is given by $F = -kT\log Z$ and so

$$\frac{\partial F}{\partial\mu} = \left\langle\frac{\partial\mathcal{H}}{\partial\mu}\right\rangle. \tag{40}$$

The next step is to differentiate (40) with respect to another parameter ν. This second derivative has three terms. The first comes from the derivative of the operator itself and is simply $\langle\partial^2\mathcal{H}/\partial\mu\,\partial\nu\rangle$. The second comes from the derivative of the Z in the denominator of the ensemble average and is $\partial Z/\partial\nu\, Z^{-1}\langle\partial\mathcal{H}/\partial\nu\rangle$.

The third term comes from differentiating the $\exp(-\beta\mathcal{H})$ in the ensemble average with respect to ν using (39); this gives

$$-\int_0^\beta \mathrm{d}y \left\langle \mathrm{e}^{y\mathcal{H}} \frac{\partial\mathcal{H}}{\partial\mu} \mathrm{e}^{-y\mathcal{H}} \frac{\partial\mathcal{H}}{\partial\nu} \right\rangle .$$

Collecting the three terms together we get the second derivative of the free energy

$$\frac{\partial^2 F}{\partial\mu\,\partial\nu} = \left\langle \frac{\partial^2\mathcal{H}}{\partial\mu\,\partial\nu} \right\rangle - \int_0^\beta \mathrm{d}y \left\langle \mathrm{e}^{y\mathcal{H}}\, \delta\frac{\partial\mathcal{H}}{\partial\mu}\, \mathrm{e}^{-y\mathcal{H}}\, \delta\frac{\partial\mathcal{H}}{\partial\nu} \right\rangle , \tag{41}$$

where $\delta\mathcal{O} = \mathcal{O} - \langle\mathcal{O}\rangle$, the fluctuation of the operator from its ensemble average value.

We now express the free energy as a Taylor series in powers of the components of the magnetic field using the derivatives that we calculated in Section 7*a*:

$$F(B_i + \delta B_i,\ B_j + \delta B_j) - F(B_i,\ B_j) = \Sigma_i\, \delta B_i\, \langle \boldsymbol{m}_i \rangle - \tfrac{1}{2}\Sigma_{ij}(\chi^{\mathrm{p}}_{ij} + \chi^{\mathrm{d}}_{ij})\delta B_i\, \delta B_j\ , \tag{42}$$

where we call $\chi^{\mathrm{d}}_{ij} = \langle\hat{\chi}^{\mathrm{d}}_{ij}\rangle$ the diamagnetic susceptibility and χ^{p}_{ij} the paramagnetic susceptibility given by

$$\chi^{\mathrm{p}}_{ij} = \int_0^\beta \mathrm{d}y \langle \mathrm{e}^{y\mathcal{H}}\, \delta\boldsymbol{m}_i\, \mathrm{e}^{-y\mathcal{H}}\, \delta\boldsymbol{m}_j \rangle\, . \tag{43}$$

Although the susceptibilities have been defined here in terms of thermodynamics it can be shown that χ^{d}_{ij} and χ^{p}_{ij} are identical to the diamagnetic and paramagnetic susceptibilities defined with the usual meaning of the linear coefficient of response to an applied field. By using (39) again it is found that for any operator $\mathcal{O}(\lambda)$

$$\frac{\partial\langle\mathcal{O}\rangle}{\partial\lambda} = \left\langle \frac{\partial\mathcal{O}}{\partial\lambda} \right\rangle - \int_0^\beta \mathrm{d}y \left\langle \mathrm{e}^{y\mathcal{H}}\, \delta\frac{\partial\mathcal{H}}{\partial\lambda}\, \mathrm{e}^{-y\mathcal{H}}\, \delta\mathcal{O} \right\rangle . \tag{44}$$

If $\mathcal{O}$ is taken to be the magnetic moment operator $\boldsymbol{m}_i = \boldsymbol{r} \times (\boldsymbol{p} - e\boldsymbol{A})e/2m_i$ and the parameter λ to be B_j then, recalling that with a uniform magnetic field $\langle\partial\boldsymbol{m}_i/\partial B_j\rangle = \chi^{\mathrm{d}}_{ij}$, it follows that $\partial\langle\boldsymbol{m}_i\rangle/\partial B_j = \chi^{\mathrm{p}}_{ij} + \chi^{\mathrm{d}}_{ij}$, the first term of (44) giving the diamagnetic part, the second term the paramagnetic part. In Stewart (1993, 1994, 1996*a*) it is also shown that the susceptibilities satisfy the inequalities $\chi^{\mathrm{p}}_{ii} \geq 0$ and $\chi^{\mathrm{p}}_{ii}\,\chi^{\mathrm{p}}_{jj} > +(\chi^{\mathrm{p}}_{ij})^2$ and $-\chi^{\mathrm{d}}_{ii} \geq 0$ and $\chi^{\mathrm{d}}_{ii}\,\chi^{\mathrm{d}}_{jj} \geq (\chi^{\mathrm{d}}_{ij})^2$. It is interesting that these inequalities hold for quantities that are not individually gauge invariant, and therefore not observable and that no such relations appear to exist for their gauge invariant and observable sum. It is straightforward to show that the inequalities remain valid under a change of the origin of the vector potential of the type discussed in Section 8.

Some other thermodynamic relations may also be obtained from the formalism above. By taking $\mathcal{O}$ in (44) to be the Hamiltonian operator it is found that the derivative of the internal energy $U = \langle\mathcal{H}\rangle$ is given by

$$\partial U/\partial \lambda = \partial F/\partial \lambda - \beta\langle \delta\mathcal{H}\,\delta(\partial\mathcal{H}/\partial\lambda)\rangle\,, \tag{45}$$

and so the derivative of the entropy is $\partial S/\partial\lambda = -\langle\delta\mathcal{H}\,\delta(\partial\mathcal{H}/\partial\lambda)\rangle/kT^2$.

12. Conclusion

It is a general principle of quantum mechanics that its predicted consequences must not depend on the electromagnetic gauge function that is chosen for a calculation. In this paper an examination has been made of those aspects of gauge invariance that impinge on condensed matter magnetism either by exhibiting the gauge function explicitly throughout the calculation or by ensuring that a calculation remains gauge invariant at every stage. One particular issue that has been stressed is the division of orbital magnetism into paramagnetism and diamagnetism. Only by treating both on an equal footing may a gauge invariant treatment of magnetism be constructed.

One question that is inevitably raised by this paper is whether any of the standard results of theoretical magnetism that involve orbital magnetism are invalid because of inadequate treatment of gauge issues. The answer appears to be no. Calculations of atomic paramagnetism and diamagnetism are generally correct because the origin of coordinates, in effect the gauge function, is chosen at the same point, the nucleus, for both. Calculations of orbital magnetism in itinerant systems following Landau (Lifshitz and Pitaevskii 1980) involve differentiating the partition function with respect to magnetic field. This, in turn, requires differentiating the Hamiltonian. As shown in Section 7*a*, its derivatives are gauge invariant so the procedure is justified in this respect.

A further interesting question concerns the arbitrariness implied by the presence of the gauge function. Wu and Yang (1975) have argued as follows. The fields $\boldsymbol{E}$ and $\boldsymbol{B}$ by themselves under-describe the behaviour of a physical system in the sense that they alone are unable to account for the Aharonov–Bohm effect. On the other hand the potentials $\boldsymbol{A}$ and ϕ over-determine the physics because, as discussed in this paper, a degree of gauge arbitrariness always exists in them. Wu and Yang argue that the quantity that describes the physics most succinctly is the phase factor $\exp(\mathrm{i}\varphi)$ where $\varphi = (e/\hbar)\int \boldsymbol{A}\,\boldsymbol{.}\,\mathrm{d}\boldsymbol{l}$ or its relativistic generalisation. Further discussion of this matter may be found in their paper and in Felsager (1981). Discussion of the place of gauge in fundamental physics may be found in the many books on the theory of quantum electrodynamics, quantum fields and gauge fields. An outstanding history of the development of modern fundamental theory and the part in it played by gauge has been written by Pais (1986).

References

Aharonov, Y., and Au, C. K. (1981). *Phys. Lett.* A **86**, 259.

Aharonov, Y., and Bohm, D. (1959). *Phys. Rev.* **115**, 485.

Cohen-Tannoudij, C., Diu, B., and Laloe, F. (1977). 'Quantum Mechanics' (Wiley: New York).

Craig, D. P., and Thirunamachandran, T. (1984). 'Molecular Quantum Mechanics' (Academic: London).

Doughty, N. A. (1990). 'Lagrangian Interaction' (Addison–Wesley: Sydney).

Ehrenberg, W., and Siday, R. E. (1949). *Proc. Phys. Soc. London* B **62**, 8.

Felsager, B. (1981). 'Geometry, Particles and Fields' (Ed. C. Claussen) (Odense University Press).

Frohlich, J., and Studer, U. M. (1993). *Rev. Mod. Phys.* **65**, 733.

Geersten, J. (1989). *J. Chem. Phys.* **90**, 4892.

Goppert-Mayer, M. (1931). *Ann. Phys. (Leipzig)* **9**, 273.

Griffiths, D. J. (1994). 'Introduction to Quantum Mechanics', Ch. 10 (Prentice Hall: Englewood Cliffs).

Griffith, J. S. (1961). 'Theory of Transition Metal Ions', p. 434 (Cambridge Univ. Press).

Haller, K. (1984). *In* 'Quantum Electrodynamics and Quantum Optics' (Ed. A. O. Barut), p. 373 (Plenum: New York).

Healy, W. P. (1988). *Phys. Scripta* T **21**, 90.

Kobe, D. H. (1978). *Am. J. Phys.* **46**, 342.

Kobe, D. H. (1984). *In* 'Quantum Electrodynamics and Quantum Optics' (Ed. A. O. Barut), p. 393 (Plenum: New York).

Kumar, K. (1965). *J. Math. Phys.* **6**, 1928.

Lamb, W. E. (1952). *Phys. Rev.* **85**, 259.

Leibbrandt, G. (1987). *Rev. Mod. Phys.* **59**, 1067.

Lifshitz, E. M., and Pitaevskii, L. P. (1980). 'Statistical Physics', 3rd edn (Pergamon: Oxford).

Oliaru, S., and Popescu, I. I. (1985). *Rev. Mod. Phys.* **57**, 339.

Pais, A. (1986). 'Inward Bound' (Clarendon: Oxford).

Peshkin, T. T., and Tonomura, A. (1989). 'The Aharonov–Bohm Effect', Lecture Notes in Physics, Vol. 340 (Springer: New York).

Power, E. A. (1989). *In* 'New Frontiers in Quantum Electrodynamics and Quantum Optics' (Ed. A. O. Barut), p. 555 (Plenum: New York).

Power, E. A., and Thirunamachandran, T. (1978). *Am. J. Phys.* **46**, 370.

Sakurai, J. J. (1985). 'Modern Quantum Mechanics' (Ed. S. F. Tuan) (Benjamin: Menlo Park).

Schiff, L. I. (1968). 'Quantum Mechanics', Section 57 (Kogakusha: Tokyo).

Schlicher, R. R., Becker, W., Bergou, J., and Scully, M. O. (1984). *In* 'Quantum Electrodynamics and Quantum Optics' (Ed. A. O. Barut), p. 405 (Plenum: New York).

Stewart, A. M. (1993). *Phys. Rev.* B **47**, 11242.

Stewart, A. M. (1994). *Aust. J. Phys.* **47**, 129.

Stewart, A. M. (1996*a*). *J. Phys.* A **29**, 1411.

Stewart, A. M. (1996*b*). *Aust. J. Phys.* **49**, 683.

Stewart, A. M. (1997). *Aust. J. Phys.* **50**, 869.

Wu, T. T., and Yang, C. N. (1975). *Phys. Rev.* D **12**, 3845.

Van Vleck, J. H. (1932). 'The Theory of Electric and Magnetic Susceptibilities' (Oxford Univ. Press).

Yang, K. H. (1976). *Ann. Phys. (New York)* **101**, 62.

Manuscript received 3 April, accepted 4 July 1997