PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Gravity, gauge theories and geometric algebra

A. Lasenby, C. Doran and S. Gull

Phil. Trans. R. Soc. Lond. A 1998 **356**, 487-582
doi: 10.1098/rsta.1998.0178

The Royal Society

Gravity, gauge theories and geometric algebra

By A. Lasenby, C. Doran and S. Gull

MRAO, Cavendish Laboratory, Department of Physics, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK
(c.doran@mrao.cam.ac.uk)

Received 7 April 1996; accepted 18 September 1996

Contents

A new gauge theory of gravity is presented. The theory is constructed in a flat background spacetime and employs gauge fields to ensure that all relations between physical quantities are independent of the position and orientation of the matter fields. In this manner all properties of the background spacetime are removed from physics and what remains are a set of 'intrinsic' relations between physical fields. For a wide range of phenomena, including all present experimental tests, the theory reproduces the predictions of general relativity. Differences do emerge, however, through the first-order nature of the equations and the global properties of the gauge fields and through the relationship with quantum theory. The properties of the gravitational gauge fields are derived from both classical and quantum viewpoints. Field equations are then derived from an action principle and consistency with the minimal coupling procedure selects an action which is unique up to the possible inclusion of a cosmological constant. This in turn singles out a unique form of spin-torsion interaction. A new method for solving the field equations is outlined and applied to the case of a time-dependent, spherically symmetric perfect fluid. A gauge is found which reduces the physics to a set of essentially Newtonian equations. These equations are then applied to the study of cosmology and to the formation and properties of black holes. Insistence on finding global solutions, together with the first-order nature of the equations, leads to a new understanding of the role played by time reversal. This alters the physical picture of the properties of a horizon around a black hole. The existence of global solutions enables one to discuss the properties of field lines inside the horizon due to a point charge held outside it. The Dirac equation is studied in a black hole background and provides a quick (though ultimately unsound) derivation of the Hawking temperature. Some applications to cosmology are also discussed and a study of the Dirac equation in a cosmological background reveals that the only models consistent with homogeneity are spatially flat. It is emphasized throughout that the description of gravity in terms of gauge fields, rather than spacetime geometry, leads to many simple and powerful physical insights. The language of 'geometric algebra' best expresses the physical and mathematical content of the theory and is employed throughout. Methods for translating the equations into other languages (tensor and spinor calculus) are given in appendices.

Keywords: gravity; gauge theory; geometric algebra; Clifford algebra; black holes; cosmology

Part I. Foundations

1. Introduction

In modern theoretical physics particle interactions are described by gauge theories. These theories are constructed by demanding that symmetries in the laws of physics should be local, rather than global, in character. The clearest expositions of this principle are contained in quantum theory, where one initially constructs a Lagrangian containing a global symmetry. In order to promote this to a local symmetry, the derivatives appearing in the Lagrangian are modified so that they are unchanged in form by local transformations. This is achieved by the introduction of fields with certain transformation properties ('gauge fields') and these fields are then responsible for inter-particle forces. The manner in which the gauge fields couple to matter is determined by the 'minimal coupling' procedure, in which partial (or directional) derivatives are replaced by covariant derivatives. This is the general framework that has been applied so successfully in the construction of the 'standard model' of particle physics, which accounts for the strong, weak and electromagnetic forces.

But what of gravity: can general relativity be formulated as a gauge theory? This question has troubled physicists for many years (Utiyama 1956; Kibble 1961; Ivanenko & Sardanashvily 1983). The first work which recovered features of general relativity (GR) from a gauging argument was due to Kibble (1961), who elaborated on an earlier, unsuccessful attempt by Utiyama (1956). Kibble used the ten-component Poincaré group of passive infinitesimal coordinate transformations (consisting of four translations and six rotations) as the global symmetry group. By gauging this group and constructing a suitable Lagrangian density for the gauge fields, Kibble arrived at a set of gravitational field equations—though not the Einstein equations. In fact, Kibble arrived at a slightly more general theory, known as a 'spin-torsion' theory. The necessary modifications to Einstein's theory to include torsion were first suggested by Cartan (1922), who identified torsion as a possible physical field. The connection between quantum spin and torsion was made later (Kibble 1961; Weyl 1950; Sciama 1964), once it had become clear that the stress-energy tensor for a massive fermion field must be asymmetric (Weyssenhoff 1947; Costa de Beauregard 1963). Spin-torsion theories are sometimes referred to as Einstein–Cartan–Kibble–Sciama (ECKS) theories. Kibble's use of passive transformations was criticized by Hehl *et al.* (1976), who reproduced Kibble's derivation from the standpoint of active transformations of the matter fields. Hehl *et al.* also arrived at a spin-torsion theory and it is now generally accepted that torsion is an inevitable feature of a gauge theory based on the Poincaré group.

The work of Hehl *et al.* (1976) raises a further issue. In their gauge theory derivation, Hehl *et al.* are clear that *'coordinates and frames are regarded as fixed once and for all, while the matter fields are replaced by fields that have been rotated or translated'*. It follows that the derivation can only affect the properties of the matter fields and not the properties of spacetime itself. Yet, once the gauge fields have been introduced, the authors identify these fields as determining the curvature and torsion of a Riemann–Cartan spacetime. This is possible only if it is assumed from the outset that one is working in a Riemann–Cartan spacetime and not in flat Minkowski spacetime. But the idea that spacetime is curved is one of the cornerstone principles of GR. That this feature must be introduced *a priori*, and is not derivable from the gauge theory argument, is highly undesirable—it shows that the principle of local gauge invariance must be supplemented with further assumptions before GR is recovered. The conclusions are clear: classical GR must be modified by the introduction

of a spin-torsion interaction if it is to be viewed as a gauge theory and the gauge principle alone fails to provide a conceptual framework for GR as a theory of gravity.

In this paper we propose an alternative theory of gravity which is derived from gauge principles alone. These gauge fields are functions of position in a single Minkowski vector space. However, here we immediately hit a profound difficulty. Parametrizing points with vectors implies a notion of a Newtonian 'absolute space' (or spacetime) and one of the aims of GR was to banish this idea. So can we possibly retain the idea of representing points with vectors without introducing a notion of absolute space? The answer to this is yes—we must construct a theory in which points are parametrized by vectors, but the physical relations between fields are independent of where the fields are placed in this vector space. We must therefore be free to move the fields around the vector space in an arbitrary manner, without in any way affecting the physical predictions. In this way our abstract Minkowski vector space will play an entirely passive role in physics and what will remain are a set of 'intrinsic' relations between spacetime fields at the same point. Yet, once we have chosen a particular parametrization of points with vectors, we will be free to exploit the vector space structure to the full, secure in the knowledge that any physical prediction arrived at is ultimately independent of the parametrization.

The theory we aim to construct is therefore one that is invariant under arbitrary field displacements. It is here that we make contact with gauge theories, because the necessary modification to the directional derivatives requires the introduction of a gauge field. But the field required is not of the type usually obtained when constructing gauge theories based on Lie-group symmetries. The gauge field coupling is of an altogether different, though very natural, character. However, this does not alter the fact that the theory constructed here is a gauge theory in the broader sense of being invariant under a group of transformations. The treatment presented here is very different from that of Kibble (1961) and Hehl *et al.* (1976). These authors only considered infinitesimal translations, whereas we are able to treat arbitrary finite field displacements. This is essential to our aim of constructing a theory that is independent of the means by which the positions of fields are parametrized by vectors.

Once we have introduced the required 'position-gauge' field, a further spacetime symmetry remains. Spacetime fields are not simply scalars, but also consist of vectors and tensors. Suppose that two spacetime vector fields are equated at some position. If both fields are then rotated at a point, the same intrinsic physical relation is obtained. We therefore expect that all physical relations should be invariant under local rotations of the matter fields, as well as displacements. This is necessary if we are to achieve complete freedom from the properties of the underlying vector space—we cannot think of the vectors representing physical quantities as having direction defined relative to some fixed vectors in Minkowski spacetime, but are only permitted to consider relations between matter fields. Achieving invariance under local rotations introduces a further gauge field, though now we are in the familiar territory of Yang–Mills type interactions (albeit employing a non-compact Lie group).

There are many ways in which the gauge theory presented here offers both real and potential advantages over traditional GR. As our theory is a genuine gauge theory, the status of physical predictions is always unambiguous—any physical prediction must be extracted from the theory in a gauge-invariant manner. Furthermore, our approach is much closer to the conventional theories of particle physics, which should ease the path to a quantum theory. A final, seemingly obvious, point is that discarding

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

all notions of a curved spacetime makes the theory conceptually much simpler than GR. For example, there is no need to deal with topics such as differentiable manifolds, tangent spaces or fibre bundles (Eguchi *et al.* 1980).

The theory developed here is presented in the language of *'geometric algebra'* (Hestenes & Sobczyk 1984; Gull *et al.* 1993*a*). Any physical theory can be formulated in a number of different mathematical languages, but physicists usually settle on a language which they feel represents the 'optimal' choice. For quantum field theory this has become the language of abstract operator commutation relations and for GR it is Riemannian geometry. For our gauge theory of gravity there seems little doubt that geometric algebra is the optimal language available in which to formulate the theory. Indeed, it was partly the desire to apply this language to gravitation theory that led to the development of the present theory. (This should not be taken to imply that geometric algebra cannot be applied to standard GR—it certainly can (Hestenes & Sobczyk 1984; Hestenes 1966, 1986*b*; Sobczyk 1981). It has also been used to elaborate on Utiyama's approach (Hestenes 1966).) To us, the use of geometric algebra is as central to the theory of gravity presented here as tensor calculus and Riemannian geometry were to Einstein's development of GR. It is the language that most clearly exposes the structure of the theory. The equations take their simplest form when expressed in geometric algebra and all reference to coordinates and frames is removed, achieving a clean separation between physical effects and coordinate artefacts. Furthermore, the geometric algebra development of the theory is entirely self-contained. All problems can be treated without ever having to introduce concepts from other languages, such as differential forms or the Newman–Penrose formalism.

We realize, however, that the use of an unfamiliar language may deter some readers from exploring the main physical content of our theory—which is, of course, independent of the language chosen to express it. We have therefore endeavoured to keep the mathematical content of the main text to a minimum level and have included appendices describing methods for translating our equations into the more familiar languages of tensor and spinor calculus. In addition, many of the final equations required for applications are simple scalar equations. The role of geometric algebra is simply to provide the most efficient and transparent derivation of these equations. It is our hope that physicists will find geometric algebra a simpler and more natural language than that of differential geometry and tensor calculus.

This paper starts with an introduction to geometric algebra and its spacetime version—the spacetime algebra. We then turn to the gauging arguments outlined above and find mathematical expressions of the underlying principles. This leads to the introduction of two gauge fields. At this point the discussion is made concrete by turning to the Dirac action integral. The Dirac action is formulated in such a way that internal phase rotations and spacetime rotations take equivalent forms. Gauge fields are then minimally coupled to the Dirac field to enforce invariance under local displacements and both spacetime and phase rotations. We then turn to the construction of a Lagrangian density for the gravitational gauge fields. This leads to a surprising conclusion. The demand that the gravitational action be consistent with the derivation of the minimally coupled Dirac equation restricts us to a single action integral. The only freedom that remains is the possible inclusion of a cosmological constant, which cannot be ruled out on theoretical grounds alone. The result of this work is a set of field equations which are completely independent of how we choose to label the positions of fields with a vector x. The resulting theory is conceptually

simple and easier to calculate with than GR, whilst being consistent with quantum mechanics at the first-quantized level. We call this theory *'gauge theory gravity'* (GTG). Having derived the field equations, we turn to a discussion of measurements, the equivalence principle and the Newtonian limit in GTG. We end part I with a discussion of symmetries, invariants and conservation laws.

In part II we turn to applications of gauge theory gravity, concentrating mainly on time-dependent spherically symmetric systems. We start by studying perfect fluids and derive a simple set of first-order equations which describe a wide range of physical phenomena. The method of derivation of these equations is new and offers many advantages over conventional techniques. The equations are then studied in the contexts of black holes, collapsing matter and cosmology. We show how a gauge can be chosen which affords a clear, global picture of the properties of these systems. Indeed, in many cases one can apply simple, almost Newtonian, reasoning to understand the physics. For some of these applications the predictions of GTG and GR are identical and these cases include all present experimental tests of GR. However, on matters such as the role of horizons and topology, the two theories differ. For example, we show that the black hole solutions admitted in GTG fall into two distinct time-asymmetric gauge sectors and that one of these is picked out uniquely by the formation process. This is quite different to GR, which admits eternal time-reverse symmetric solutions. In discussing differences between GTG and GR, it is not always clear what the correct GR viewpoint is. We should therefore be explicit in stating that what we intend when we talk about GR is the full, modern formulation of the subject as expounded by, for example, Hawking & Ellis (1973) and D'Inverno (1992). This includes ideas such as worm-holes, exotic topologies and distinct 'universes' connected by black holes (Kaufmann 1979; Hawking 1993). In short, none of these concepts survive in GTG.

After studying some solutions for the gravitational fields we turn to the properties of electromagnetic and Dirac fields in gravitational backgrounds. For example, we give field configurations for a charge held at rest outside a black hole. We show how these field lines extend smoothly across the horizon and that the origin behaves as a polarization charge. This solution demonstrates how the global properties of the gravitational fields are relevant to physics outside the horizon, a fact which is supported by an analysis of the Dirac equation in a black hole background. This analysis also provides a quick, though ultimately unsound, derivation of a particle production rate described by a Fermi–Dirac distribution with the correct Hawking temperature. We end with a discussion of the implications of GTG for cosmology. A study of the Maxwell and Dirac equations in a cosmological background reveals a number of surprising features. In particular, it is shown that a non-spatially flat universe does not appear homogeneous to Dirac fields—fermionic matter would be able to detect the 'centre' of the universe if $k \neq 0$. Thus the only homogeneous cosmological models consistent with GTG are those which are spatially flat. (This does not rule out spatially flat universes with a non-zero cosmological constant.) A concluding section summarizes the philosophy behind our approach and outlines some future areas of research.

2. An outline of geometric algebra

There are many reasons for preferring geometric algebra to other languages employed in mathematical physics. It is the most powerful and efficient language

for handling rotations and boosts and it generalizes the role of complex numbers in two dimensions, and quaternions in three dimensions, to a scheme that efficiently handles rotations in arbitrary dimensions. It also exploits the advantages of labelling points with vectors more fully than either tensor calculus or differential forms, both of which were designed with a view to applications in the intrinsic geometry of curved spaces. In addition, geometric algebra affords an entirely *real* formulation of the Dirac equation (Hestenes 1975; Doran *et al.* 1993*c*), eliminating the need for complex numbers. The advantage of the real formulation is that internal phase rotations and spacetime rotations are handled in an identical manner in a single unifying framework. A wide class of physical theories have now been successfully formulated in terms of geometric algebra. These include classical mechanics (Hestenes 1974*a*, 1985; Vold 1993*a*), relativistic dynamics (Hestenes 1974*b*), Dirac theory (Hestenes 1975; Doran *et al.* 1993*c*; Gull *et al.* 1993*b*; Doran *et al.* 1996*b*), electromagnetism and electrodynamics (Gull *et al.* 1993*a*, *b*; Vold 1993*b*), as well as a number of other areas of modern mathematical physics (Lasenby *et al.* 1993*a*, *b*, *c*; Doran *et al.* 1993*b*, *d*). In every case, geometric algebra has offered demonstrable advantages over other techniques and has provided novel insights and unifications between disparate branches of physics and mathematics.

This section is intended to give only a brief introduction to the ideas and applications of geometric algebra. A fuller introduction, including a number of results relevant to this paper, is set out in the series of papers: Gull *et al.* (1993*a*, *b*), Doran *et al.* (1993*c*), Lasenby *et al.* (1993*a*), written by the present authors. Elsewhere, the books by Hestenes (1966, 1985) and Hestenes & Sobczyk (1984) cover the subject in detail. The latter, *'Clifford Algebra to geometric calculus'* (Hestenes & Sobczyk 1984), is the most comprehensive exposition of geometric algebra available, though its uncompromising style makes it a difficult introduction to the subject. A number of other helpful introductory articles can be found, including those by Hestenes (1986*a*, 1991) and Vold (1993*a*, *b*). The conference proceedings (Chisholm & Common 1986; Micali *et al.* 1991; Brackx *et al.* 1993) also contain some interesting and useful papers.

Geometric algebra arose from Clifford's attempts to generalize Hamilton's quaternion algebra into a language for vectors in arbitrary dimensions (Clifford 1878). Clifford discovered that both complex numbers and quaternions are special cases of an algebraic framework in which vectors are equipped with a single associative product which is distributive over addition†. With vectors represented by lower-case italic letters (a, b), Clifford's 'geometric product' is written simply as ab. A key feature of the geometric product is that the square of any vector is a scalar. Now, rearranging the expansion

$$(a+b)^2 = (a+b)(a+b) = a^2 + (ab+ba) + b^2 \tag{2.1}$$

to give

$$ab + ba = (a+b)^2 - a^2 - b^2, \tag{2.2}$$

where the right-hand side of (2.2) is a sum of squares and by assumption a scalar, we see that the symmetric part of the geometric product of two vectors is also a scalar.

† The same generalization was also found by Grassmann (1877), independently and somewhat before Clifford's work. This is one of many reasons for preferring Clifford's name (*'geometric algebra'*) over the more usual *'Clifford algebra'*.

We write this 'inner' or 'dot' product between vectors as

$$a\cdot b \equiv \tfrac{1}{2}(ab + ba). \tag{2.3}$$

The remaining antisymmetric part of the the geometric product represents the directed area swept out by displacing a along b. This is the 'outer' or 'exterior' product introduced by Grassmann (1877) and familiar to all who have studied the language of differential forms. The outer product of two vectors is called a *bivector* and is written with a wedge:

$$a\wedge b \equiv \tfrac{1}{2}(ab - ba). \tag{2.4}$$

On combining (2.3) and (2.4), we find that the geometric product has been decomposed into the sum of a scalar and a bivector part,

$$ab = a\cdot b + a\wedge b. \tag{2.5}$$

The innovative feature of Clifford's product (2.5) lies in its mixing of two different types of object: scalars and bivectors. This is not problematic, because the addition implied by (2.5) is precisely that which is used when a real number is added to an imaginary number to form a complex number. But why might we want to add these two geometrically distinct objects? The answer emerges from considering reflections and rotations. Suppose that the vector a is reflected in the (hyper)plane perpendicular to the unit vector n. The result is the new vector

$$a - 2(a\cdot n)n = a - (an + na)n = -nan. \tag{2.6}$$

The utility of the geometric algebra form of the resultant vector, $-nan$, becomes clear when a second reflection is performed. If this second reflection is in the hyperplane perpendicular to the unit vector m, then the combined effect is

$$a \mapsto mnanm. \tag{2.7}$$

But the combined effect of two reflections is a rotation so, defining the geometric product mn as the scalar-plus-bivector quantity R, we see that rotations are represented by

$$a \mapsto Ra\tilde{R}. \tag{2.8}$$

Here the quantity $\tilde{R} = nm$ is called the *'reverse'* of R and is obtained by reversing the order of all geometric products between vectors:

$$\widetilde{(ab\ldots c)} \equiv c\ldots ba. \tag{2.9}$$

The object R is called a *rotor*. Rotors can be written as an even (geometric) product of unit vectors and satisfy the relation $R\tilde{R} = 1$. The representation of rotations in the form (2.8) has many advantages over tensor techniques. By defining $\cos\theta \equiv m\cdot n$ we can write

$$R = mn = \exp\left(\frac{m\wedge n}{|m\wedge n|}\tfrac{1}{2}\theta\right), \tag{2.10}$$

which relates the rotor R directly to the plane in which the rotation takes place. Equation (2.10) generalizes to arbitrary dimensions the representation of planar rotations afforded by complex numbers. This generalization provides a good example of how the full geometric product, and the implied sum of objects of different types, can enter geometry at a very basic level. The fact that equation (2.10) encapsulates a simple geometric relation should also dispel the notion that Clifford algebras are somehow

intrinsically 'quantum' in origin. The derivation of (2.8) has assumed nothing about the signature of the space being employed, so that the formula applies equally to boosts as well as rotations. The two-sided formula for a rotation (2.8) will also turn out to be compatible with the manner in which observables are constructed from Dirac spinors and this is important for the gauge theory of rotations of the Dirac equation which follows.

Forming further geometric products of vectors produces the entire geometric algebra. General elements are called 'multivectors' and these decompose into sums of elements of different grades (scalars are grade zero, vectors grade one, bivectors grade two, etc.). Multivectors in which all elements have the same grade are termed *homogeneous* and are usually written as A_r to show that A contains only grade r components. Multivectors inherit an associative product and the geometric product of a grade r multivector A_r with a grade s multivector B_s decomposes into

$$A_rB_s = \langle AB\rangle_{r+s} + \langle AB\rangle_{r+s-2} + \cdots + \langle AB\rangle_{|r-s|}, \tag{2.11}$$

where the symbol $\langle M\rangle_r$ denotes the projection onto the grade r component of M. The projection onto the grade 0 (scalar) component of M is written $\langle M\rangle$. The '$\cdot$' and '$\wedge$' symbols are retained for the lowest-grade and highest-grade terms of the series (2.11), so that

$$A_r\cdot B_s \equiv \langle AB\rangle_{|r-s|}, \tag{2.12}$$

$$A_r\wedge B_s \equiv \langle AB\rangle_{r+s}, \tag{2.13}$$

which are called the interior and exterior products, respectively. The exterior product is associative and satisfies the symmetry property

$$A_r\wedge B_s = (-1)^{rs}B_s\wedge A_r. \tag{2.14}$$

Two further products can also be defined from the geometric product. These are the scalar product

$$A*B \equiv \langle AB\rangle \tag{2.15}$$

and the commutator product

$$A\times B \equiv \tfrac{1}{2}(AB - BA). \tag{2.16}$$

The scalar product (2.15) is commutative and satisfies the cyclic reordering property

$$\langle A\cdots BC\rangle = \langle CA\cdots B\rangle. \tag{2.17}$$

The scalar product (2.15) and the interior product (2.12) coincide when acting on two homogeneous multivectors of the same grade. The associativity of the geometric product ensures that the commutator product (2.16) satisfies the Jacobi identity

$$A\times(B\times C) + B\times(C\times A) + C\times(A\times B) = 0. \tag{2.18}$$

Finally we introduce some further conventions. Throughout, we employ the operator ordering convention that, *in the absence of brackets, inner, outer, commutator and scalar products take precedence over geometric products*. Thus $a\cdot bc$ means $(a\cdot b)c$, not $a\cdot(bc)$. This convention helps to eliminate unwieldy numbers of brackets. Summation convention is employed throughout except for indices which denote the grade of a multivector, which are not summed over. Natural units ($\hbar = c = 4\pi\epsilon_0 = G = 1$) are used except where explicitly stated. Throughout, we refer to a Lorentz transformation (i.e. a spatial rotation and/or boost) simply as a rotation.

(a) The spacetime algebra

Of central importance to this paper is the geometric algebra of spacetime, the *spacetime algebra* (Hestenes 1966). To describe the spacetime algebra (STA) it is helpful to introduce a set of four orthonormal basis vectors $\{\gamma_\mu\}$, $\mu = 0, \ldots, 3$, satisfying

$$\gamma_\mu \cdot \gamma_\nu = \eta_{\mu\nu} = \mathrm{diag}(+ - - -). \tag{2.19}$$

The vectors $\{\gamma_\mu\}$ satisfy the same algebraic relations as Dirac's γ matrices, but they now form a set of four independent basis vectors for spacetime, not four components of a single vector in an internal 'spin space'. The relation between Dirac's matrix algebra and the STA is described in more detail in Appendix A, which gives a direct translation of the Dirac equation into its STA form.

A frame of timelike bivectors $\{\sigma_k\}$, $k = 1, \ldots, 3$, is defined by

$$\sigma_k \equiv \gamma_k \gamma_0 \tag{2.20}$$

and forms an orthonormal frame of vectors in the space relative to the γ_0 direction. The algebraic properties of the $\{\sigma_k\}$ are the same as those of the Pauli spin matrices, but in the STA they again represent an orthonormal frame of vectors in space and not three components of a vector in spin space. The highest-grade element (or 'pseudoscalar') is denoted by i and is defined as

$$\mathrm{i} \equiv \gamma_0 \gamma_1 \gamma_2 \gamma_3 = \sigma_1 \sigma_2 \sigma_3. \tag{2.21}$$

The symbol i is used because its square is -1, but the pseudoscalar must not be confused with the unit scalar imaginary employed in quantum mechanics. Since we are in a space of even dimension, i *anti*commutes with odd-grade elements and commutes only with even-grade elements. With these definitions, a basis for the 16-dimensional STA is provided by

1	$\{\gamma_\mu\}$	$\{\sigma_k, \mathrm{i}\sigma_k\}$	$\{\mathrm{i}\gamma_\mu\}$	i
1 scalar	4 vectors	6 bivectors	4 trivectors	1 pseudoscalar.

(2.22)

Geometric significance is attached to the above relations as follows. An observer's rest frame is characterized by a future-pointing timelike (unit) vector. If this is chosen to be the γ_0 direction then the γ_0 vector determines a map between spacetime vectors $a = a^\mu \gamma_\mu$ and the even subalgebra of the full STA via

$$a\gamma_0 = a_0 + \boldsymbol{a}, \tag{2.23}$$

where

$$a_0 = a \cdot \gamma_0, \tag{2.24}$$

$$\boldsymbol{a} = a \wedge \gamma_0. \tag{2.25}$$

The 'relative vector' $\boldsymbol{a}$ can be decomposed in the $\{\sigma_k\}$ frame and represents a spatial vector as seen by an observer in the γ_0 frame. Since a vector appears to an observer as a line segment existing for a period of time, it is natural that what an observer perceives as a vector should be represented by a spacetime bivector. Equation (2.23) embodies this idea and shows that the algebraic properties of vectors in relative space are determined entirely by the properties of the fully relativistic STA.

The split of the six spacetime bivectors into relative vectors and relative bivectors is a frame-dependent operation—different observers see different relative spaces. This fact is clearly illustrated with the Faraday bivector F. The 'space-time split' (Hestenes 1974*a*) of F into the γ_0 system is made by separating F into parts which anticommute and commute with γ_0. Thus

$$F = \boldsymbol{E} + \mathrm{i}\boldsymbol{B}, \tag{2.26}$$

where

$$\boldsymbol{E} = \tfrac{1}{2}(F - \gamma_0 F \gamma_0), \tag{2.27}$$

$$\mathrm{i}\boldsymbol{B} = \tfrac{1}{2}(F + \gamma_0 F \gamma_0). \tag{2.28}$$

Both $\boldsymbol{E}$ and $\boldsymbol{B}$ are spatial vectors in the γ_0 frame and $\mathrm{i}\boldsymbol{B}$ is a spatial bivector. Equation (2.26) decomposes F into separate electric and magnetic fields and the explicit appearance of γ_0 in the formulae for $\boldsymbol{E}$ and $\boldsymbol{B}$ shows how this split is observer dependent.

The identification of the algebra of 3-space with the even subalgebra of the STA necessitates a convention which articulates smoothly between the two algebras. Relative (or spatial) vectors in the γ_0 system are written in bold type to record the fact that in the STA they are actually bivectors. This distinguishes them from spacetime vectors, which are left in normal type. No problems arise for the $\{\sigma_k\}$, which are unambiguously spacetime bivectors, so these are also left in normal type. Further conventions are introduced where necessary.

(*b*) *Geometric calculus*

Many of the derivations in this paper employ the vector and multivector derivatives (Hestenes & Sobczyk 1984; Lasenby *et al.* 1993*b*). Before defining these, however, we need some simple results for vector frames. Suppose that the set $\{e_k\}$ form a vector frame. The reciprocal frame is determined by (Hestenes & Sobczyk 1984)

$$e^j = (-1)^{j-1} e_1 \wedge e_2 \wedge \cdots \wedge \check{e}_j \wedge \cdots \wedge e_n e^{-1}, \tag{2.29}$$

where

$$e \equiv e_1 \wedge e_2 \wedge \cdots \wedge e_n \tag{2.30}$$

and the check on $\check{e}_j$ denotes that this term is missing from the expression. The $\{e_k\}$ and $\{e^j\}$ frames are related by

$$e_j \cdot e^k = \delta_j^k. \tag{2.31}$$

An arbitrary multivector B can be decomposed in terms of the $\{e_k\}$ frame into

$$B = \sum_{i<\cdots<j} B_{i\cdots j} e^i \wedge \cdots \wedge e^j, \tag{2.32}$$

where

$$B_{i\cdots j} = B_r \cdot (e_j \wedge \cdots \wedge e_i). \tag{2.33}$$

Suppose now that the multivector F is an arbitrary function of some multivector argument X, $F = F(X)$. The derivative of F with respect to X in the A direction is defined by

$$A * \partial_X F(X) \equiv \lim_{\tau \mapsto 0} \frac{F(X + \tau A) - F(X)}{\tau}. \tag{2.34}$$

From this the multivector derivative ∂_X is defined by

$$\partial_X \equiv \sum_{i<\cdots<j} e^i \wedge \cdots \wedge e^j (e_j \wedge \cdots \wedge e_i) * \partial_X. \tag{2.35}$$

This definition shows how the multivector derivative ∂_X inherits the multivector properties of its argument X, as well as a calculus from equation (2.34).

Most of the properties of the multivector derivative follow from the result that

$$\partial_X \langle XA \rangle = P_X(A), \tag{2.36}$$

where $P_X(A)$ is the projection of A onto the grades contained in X. Leibniz's rule is then used to build up results for more complicated functions (see Appendix B). The multivector derivative acts on the next object to its right unless brackets are present; for example, in the expression $\partial_X AB$, the ∂_X acts only on A, but in the expression $\partial_X(AB)$, the ∂_X acts on both A and B. If the ∂_X is intended to act only on B then this is written as $\dot{\partial}_X A\dot{B}$, the overdot denoting the multivector on which the derivative acts. As an illustration, Leibniz's rule can be written in the form

$$\partial_X(AB) = \dot{\partial}_X \dot{A} B + \dot{\partial}_X A \dot{B}. \tag{2.37}$$

The overdot notation neatly encodes the fact that, since ∂_X is a multivector, it does not necessarily commute with other multivectors and often acts on functions to which it is not adjacent.

The derivative with respect to spacetime position x is called the *vector derivative* and is given the symbol

$$\nabla \equiv \nabla_x \equiv \partial_x. \tag{2.38}$$

In the STA we can therefore write

$$\nabla = \gamma^\mu \frac{\partial}{\partial x^\mu} = \gamma^\mu \partial_\mu, \tag{2.39}$$

so that, just as the γ matrices are replaced by vectors in spacetime, objects such as $x^\mu \gamma_\mu$ and $\nabla = \gamma^\mu \partial_\mu$ become frame-free vectors. The usefulness of the geometric product for the vector derivative is illustrated by electromagnetism. In tensor notation, Maxwell's equations are

$$\partial_\mu F^{\mu\nu} = J^\nu, \quad \partial_{[\alpha} F_{\mu\nu]} = 0, \tag{2.40}$$

which have the STA equivalents (Hestenes 1966)

$$\nabla \cdot F = J, \quad \nabla \wedge F = 0. \tag{2.41}$$

However, we can utilize the geometric product to combine these into the single equation

$$\nabla F = J. \tag{2.42}$$

The great advantage of the ∇ operator is that it possesses an inverse, so a first-order propagator theory can be developed for it (Hestenes & Sobczyk 1984; Gull *et al.* 1993*b*). This is not possible for the separate $\nabla\cdot$ and $\nabla\wedge$ operators.

The derivative with respect to the vector a, ∂_a, is often used to perform linear algebra operations such as contraction. For such operations the following results are

useful:

$$\partial_a a\cdot A_r = rA_r, \tag{2.43}$$

$$\partial_a a\wedge A_r = (n-r)A_r, \tag{2.44}$$

$$\partial_a A_r a = (-1)^r(n-2r)A_r, \tag{2.45}$$

where n is the dimension of the space ($n = 4$ for all the applications considered here).

(*c*) *Linear algebra*

Geometric algebra offers many advantages over tensor calculus in developing the theory of linear functions (Hestenes & Sobczyk 1984; Doran *et al.* 1993*d*; Hestenes 1991). A linear function mapping vectors to vectors is written with an underbar $\underline{f}(a)$. Throughout, the argument of a linear function is assumed to be independent of position, unless stated otherwise.

The adjoint function is written with an overbar, $\overline{f}(a)$, so that

$$a\cdot\underline{f}(b) = \overline{f}(a)\cdot b, \tag{2.46}$$

and hence

$$\overline{f}(a) = \partial_b\langle\underline{f}(b)a\rangle. \tag{2.47}$$

We will frequently employ the derivative with respect to the vectors a, b, etc. to perform algebraic manipulations of linear functions, as in equation (2.47). The advantage is that all manipulations are then frame-free. Of course, the ∂_a and a vectors can be replaced by the sum over a set of constant frame vectors and their reciprocals, if desired.

A symmetric function is one for which $\underline{f}(a) = \overline{f}(a)$. For such functions

$$\partial_a\wedge\underline{f}(a) = \partial_a\wedge\partial_b\langle a\underline{f}(b)\rangle = \underline{f}(b)\wedge\partial_b. \tag{2.48}$$

It follows that for symmetric functions

$$\partial_a\wedge\underline{f}(a) = 0, \tag{2.49}$$

which is equivalent to the statement that $\underline{f}(a) = \overline{f}(a)$.

Linear functions extend to act on multivectors via

$$\underline{f}(a\wedge b\wedge\cdots\wedge c) \equiv \underline{f}(a)\wedge\underline{f}(b)\wedge\cdots\wedge\underline{f}(c), \tag{2.50}$$

so that $\underline{f}$ is now a grade-preserving linear function mapping multivectors to multivectors. In particular, since the pseudoscalar I is unique up to a scale factor, we can define

$$\det(\underline{f}) = \underline{f}(I)I^{-1}. \tag{2.51}$$

Viewed as linear functions over the entire geometric algebra, $\underline{f}$ and $\overline{f}$ are related by the fundamental formulae

$$A_r\cdot\overline{f}(B_s) = \overline{f}[\underline{f}(A_r)\cdot B_s], \quad r\leqslant s, \qquad \underline{f}(A_r)\cdot B_s = \underline{f}[A_r\cdot\overline{f}(B_s)], \quad r\geqslant s, \tag{2.52}$$

which are derived in Hestenes & Sobczyk (1984, ch. 3). The formulae for the inverse functions are found as special cases of (2.52),

$$\underline{f}^{-1}(A) = \det(\underline{f})^{-1}\overline{f}(AI)I^{-1}, \quad \overline{f}^{-1}(A) = \det(\underline{f})^{-1}I^{-1}\underline{f}(IA). \tag{2.53}$$

A number of further results for linear functions are contained in Appendix B. These include a coordinate-free formulation of the derivative with respect to a linear function, which proves to be very useful in deriving stress-energy tensors from action integrals.

3. Gauge principles for gravitation

In this section we identify the dynamical variables which will describe gravitational interactions. We start by reviewing the arguments outlined in the introduction. The basic idea is that all physical relations should have the generic form $a(x) = b(x)$, where a and b are spacetime fields representing physical quantities and x is the STA position vector. An equality such as this can certainly correspond to a clear physical statement. But, considered as a relation between fields, the physical relationship expressed by this statement is completely independent of where we choose to think of x as lying in spacetime. In particular, we can associate each position x with some new position $x' = f(x)$, rewrite the relation as $a(x') = b(x')$ and the equation still has precisely the same content. (A proviso, which will gain significance later, is that the map $f(x)$ should be non-singular and cover all of spacetime.)

A similar argument applies to rotations. The *intrinsic* content of a relation such as $a(x) = b(x)$ at a given point x_0 is unchanged if we rotate each of a and b by the same amount. That is, the equation $Ra(x_0)\tilde{R} = Rb(x_0)\tilde{R}$ has the same physical content as the equation $a(x_0) = b(x_0)$. For example, scalar product relations, from which we can derive angles, are unaffected by this change. These arguments apply to any physical relation between any type of multivector field. The principles underlying gauge theory gravity can therefore be summarized as follows.

(i) The physical content of a field equation in the STA must be invariant under arbitrary local displacements of the fields. (This is called position-gauge invariance.)

(ii) The physical content of a field equation in the STA must be invariant under arbitrary local rotations of the fields. (This is called rotation-gauge invariance.)

In this theory predictions for all measurable quantities, including distances and angles, must be derived from gauge-invariant relations between the field quantities themselves, not from the properties of the STA. On the other hand, quantities which depend on a choice of 'gauge' are not predicted absolutely and cannot be defined operationally.

It is necessary to indicate how this approach differs from the one adopted in gauge theories of the Poincaré group. (This is a point on which we have been confused in the past (Doran *et al.* 1993*a*).) Poincaré transformations for a multivector field $M(x)$ are defined by

$$M(x) \mapsto M' = RM(x')\tilde{R}, \tag{3.1}$$

where

$$x' = \tilde{R}xR + t; \tag{3.2}$$

R is a constant rotor and t is a constant vector. Transformations of this type mix displacements and rotations, and any attempt at a local gauging of this spacetime symmetry fails to decouple the two (Kibble 1961; Hehl *et al.* 1976). Furthermore, the fact that the rotations described by Poincaré transformations include the displacement (3.2) (with $t = 0$) means that the rotations discussed under point (ii) above are not contained in the Poincaré group.

As a final introductory point, whilst the mapping of fields onto spacetime positions is arbitrary, the fields themselves must be well-defined in the STA. The fields cannot be singular except at a few special points. Furthermore, any remapping of the fields in the STA must be one-to-one, else we would cut out some region of physical significance. In later sections we will see that GR allows operations in which regions of spacetime are removed. These are achieved through the use of singular coordinate transformations and are the origin of a number of differences between GTG and GR.

(*a*) *The position-gauge field*

We now examine the consequences of the local symmetries we have just discussed. As in all gauge theories we must study the effects on *derivatives*, since all non-derivative relations already satisfy the correct requirements.

We start by considering a scalar field $\phi(x)$ and form its vector derivative $\nabla\phi(x)$. Suppose now that from $\phi(x)$ we define the new field $\phi'(x)$ by

$$\phi'(x) \equiv \phi(x'), \tag{3.3}$$

where

$$x' = f(x) \tag{3.4}$$

and $f(x)$ is an arbitrary (differentiable) map between spacetime position vectors. The map $f(x)$ should not be thought of as a map between manifolds, or as moving points around; rather, the function $f(x)$ is merely a rule for relating one position vector to another within a single vector space. Note that the new function $\phi'(x)$ is given by the old function ϕ evaluated at x'. We could have defined things the other way round, so that $\phi'(x')$ is given by $\phi(x)$, but the form adopted here turns out to be more useful in practice.

If we now act on the new scalar field ϕ' with ∇ we form the quantity $\nabla\phi[f(x)]$. To evaluate this we return to the definition of the vector derivative and construct

$$\begin{aligned} a\cdot\nabla\phi[f(x)] &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}(\phi f(x+\epsilon a) - \phi f(x)) \\ &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}(\phi[f(x)+\epsilon\underline{f}(a)] - \phi f(x)) = \underline{f}(a)\cdot\nabla_{x'}\phi(x'), \end{aligned} \tag{3.5}$$

where

$$\underline{f}(a) \equiv a\cdot\nabla f(x) \tag{3.6}$$

and the subscript on $\nabla_{x'}$ records that the derivative is now with respect to the new vector position variable x'. The function $\underline{f}(a)$ is a linear function of a and an arbitrary function of x. If we wish to make the position dependence explicit we write this as $\underline{f}(a,x)$ or $\underline{f}_x(a)$. In general, any position-dependent linear function with its position dependence suppressed is to be taken as a function of x. Also—as stated in the introduction—the argument of a linear function should be assumed to be constant unless explicitly stated otherwise.

From (3.5) we see that

$$\nabla_x = \overline{f}(\nabla_{x'}) \tag{3.7}$$

and it follows that

$$\nabla\phi'(x) = \overline{f}[\nabla_{x'}\phi(x')]. \tag{3.8}$$

The bracketed term on the right-hand side, $\nabla_{x'}\phi(x')$, is the old gradient vector $\nabla\phi$ evaluated at x' instead of x. This tells us how to modify the derivative operator

∇: we must introduce a new linear function which assembles with ∇ in such a way that the $\overline{f}$ field is removed when the full object is displaced. The resulting object will then have the desired property of just changing its position dependence under arbitrary local displacements. We therefore introduce the *position-gauge field* $\overline{h}(a,x)$, which is a linear function of a and an arbitrary function of position x. As usual, this is abbreviated to $\overline{h}(a)$ when the position dependence is taken as a function of x. Under the displacement $x \mapsto x' = f(x)$, $\overline{h}(a)$ is defined to transform to the new field $\overline{h}'(a,x)$, where

$$\overline{h}'(a,x) \equiv \overline{h}(\overline{f}^{-1}(a), f(x)) = \overline{h}_{x'}\overline{f}^{-1}(a), \tag{3.9}$$

so that

$$\overline{h}_x(\nabla_x) \mapsto \overline{h}_{x'}\overline{f}^{-1}(\nabla_x) = \overline{h}_{x'}(\nabla_{x'}). \tag{3.10}$$

This transformation law ensures that the vector $A(x)$, say,

$$A(x) \equiv \overline{h}[\nabla\phi(x)], \tag{3.11}$$

transforms simply as $A(x) \mapsto A'(x) = A(x')$ under *arbitrary* displacements. This is the type of behaviour we seek. The vector $A(x)$ can now be equated with other (possibly non-differentiated) fields and the resulting equations are unchanged in form under arbitrary repositioning of the fields in spacetime.

Henceforth, we refer to any quantity that transforms under arbitrary displacements as

$$M(x) \mapsto M'(x) = M(x') \tag{3.12}$$

as behaving *covariantly* under displacements. The $\overline{h}$ field enables us to form derivatives of covariant objects which are also covariant under displacements. When we come to calculate with this theory, we will often fix a gauge by choosing a labelling of spacetime points with vectors. In this way we remain free to exploit all the advantages of representing points with vectors. Of course, all physical predictions of the theory will remain independent of the actual gauge choice.

The $\overline{h}$ field is not a connection in the conventional Yang–Mills sense. The coupling to derivatives is different, as is the transformation law (3.9). This is unsurprising, since the group of arbitrary translations is infinite dimensional (if we were considering maps between manifolds then this would form the group of diffeomorphisms). Nevertheless, the $\overline{h}$ field embodies the idea of replacing directional derivatives with covariant derivatives, so clearly deserves to be called a gauge field.

A remaining question is to find the conditions under which the $\overline{h}$ field can be transformed to the identity. Such a transformation, if it existed, would give

$$\overline{h}_{x'}(a)\overline{f}^{-1}(a) = a \tag{3.13}$$

$$\Rightarrow \quad \overline{h}_{x'}(a) = \overline{f}(a). \tag{3.14}$$

However, from the definition of $\underline{f}(a)$, it follows that

$$\overline{f}(a) = \partial_b\langle ab\cdot\nabla f(x)\rangle = \nabla\langle f(x)a\rangle \tag{3.15}$$

and hence that

$$\nabla\wedge\overline{f}(a) = \nabla\wedge\nabla\langle f(x)a\rangle = 0. \tag{3.16}$$

So, if the $\overline{h}(a)$ field can be transformed to the identity, it must satisfy

$$\nabla_x\wedge\overline{h}_{x'}(a) = 0. \tag{3.17}$$

This condition can be simplified by using equation (3.14) to write ∇_x as $\overline{h}_{x'}(\nabla_{x'})$, giving

$$\overline{h}_{x'}(\nabla_{x'})\wedge\overline{h}_{x'}(a)=0. \tag{3.18}$$

This must hold for all x', so we can equally well replace x' with x and write

$$\overline{h}(\nabla)\wedge\overline{h}(a)=0, \tag{3.19}$$

which implies that

$$\dot{\nabla}\wedge\overline{h}^{-1}\dot{\overline{h}}(a)=-\dot{\nabla}\wedge\dot{\overline{h}}^{-1}\overline{h}(a)=0. \tag{3.20}$$

Thus we finally obtain the 'pure gauge' condition in the simple form

$$\nabla\wedge\overline{h}^{-1}(a)=0. \tag{3.21}$$

An arbitrary $\overline{h}$ field will not satisfy this equation, so in general there is no way to assign position vectors so that the effects of the $\overline{h}$ field vanish. In the light of equations (3.16) and (3.21), it might seem more natural to introduce the gauge field as $\overline{h}^{-1}(\nabla)$, instead of $\overline{h}(\nabla)$. There is little to choose between these conventions, though our choice is partially justified by our later implementation of the variational principle.

(*b*) *The rotation-gauge field*

We now examine how the derivative must be modified to allow rotational freedom from point to point, as described in point (ii) at the start of this section. Here we give an analysis based on the properties of classical fields. An analysis based on spinor fields is given in the following section. We have already seen that the gradient of a scalar field is modified to $\overline{h}(\nabla\phi)$ to achieve covariance under displacements. But objects such as temperature gradients are certainly physical and can be equated with other physical quantities. Consequently, vectors such as $\overline{h}(\nabla\phi)$ must transform under rotations in the same manner as all other physical fields. It follows that, under local spacetime rotations, the $\overline{h}$ field must transform as

$$\overline{h}(a)\mapsto R\overline{h}(a)\tilde{R}. \tag{3.22}$$

Now consider an equation such as Maxwell's equation, which we saw in § 2 *b* takes the simple form $\nabla F=J$ in the STA. Once the position-gauge field is introduced, this equation becomes

$$\overline{h}(\nabla)\mathcal{F}=\mathcal{J}, \tag{3.23}$$

where

$$\mathcal{F}\equiv\overline{h}(F)\quad\text{and}\quad\mathcal{J}\equiv\det(\underline{h})\underline{h}^{-1}(J). \tag{3.24}$$

(The reasons behind these definitions will be explained in § 7. The use of a calligraphic letter for certain covariant fields is a convention we have found very useful.) The definitions of $\mathcal{F}$ and $\mathcal{J}$ ensure that under local rotations they transform as

$$\mathcal{F}\mapsto R\mathcal{F}\tilde{R}\quad\text{and}\quad\mathcal{J}\mapsto R\mathcal{J}\tilde{R}. \tag{3.25}$$

Any (multi)vector that transforms in this manner under rotations and is covariant under displacements is referred to as a *covariant* (multi)vector.

Equation (3.23) is covariant under arbitrary displacements and we now need to make it covariant under local rotations as well. To achieve this we replace $\overline{h}(\nabla)$ by $\overline{h}(\partial_a)a\cdot\nabla$ and focus attention on the term $a\cdot\nabla\mathcal{F}$. Under a position-dependent rotation we find that

$$a\cdot\nabla(R\mathcal{F}\tilde{R})=Ra\cdot\nabla\mathcal{F}\tilde{R}+a\cdot\nabla R\mathcal{F}\tilde{R}+R\mathcal{F}a\cdot\nabla\tilde{R}. \tag{3.26}$$

Since the rotor R satisfies $R\tilde{R} = 1$ we find that

$$a\cdot\nabla R\tilde{R} + Ra\cdot\nabla\tilde{R} = 0 \tag{3.27}$$

$$\Rightarrow \quad a\cdot\nabla R\tilde{R} = -Ra\cdot\nabla\tilde{R} = -\widetilde{(a\cdot\nabla R\tilde{R})}. \tag{3.28}$$

Hence $a\cdot\nabla R\tilde{R}$ is equal to minus its reverse and so must be a bivector in the STA. (In a geometric algebra the bivectors form a representation of the Lie algebra of the rotation group (Doran *et al.* 1993*d*).) We can therefore write

$$a\cdot\nabla(R\mathcal{F}\tilde{R}) = Ra\cdot\nabla\mathcal{F}\tilde{R} + 2(a\cdot\nabla R\tilde{R})\times(R\mathcal{F}\tilde{R}). \tag{3.29}$$

To construct a covariant derivative we must therefore add a 'connection' term to $a\cdot\nabla$ to construct the operator

$$\mathcal{D}_a \equiv a\cdot\nabla + \Omega(a)\times. \tag{3.30}$$

Here $\Omega(a) = \Omega(a, x)$ is a bivector-valued linear function of a with an arbitrary x dependence. The commutator product of a multivector with a bivector is grade preserving so, even though it contains non-scalar terms, $\mathcal{D}_a$ preserves the grade of the multivector on which it acts.

Under local rotations the $a\cdot\nabla$ term in $\mathcal{D}_a$ cannot change and we also expect that the $\mathcal{D}_a$ operator be unchanged in form (this is the essence of 'minimal coupling'). We should therefore have

$$\mathcal{D}'_a = a\cdot\nabla + \Omega'(a)\times. \tag{3.31}$$

However, the property that the covariant derivative must satisfy is

$$\mathcal{D}'_a(R\mathcal{F}\tilde{R}) = R\mathcal{D}_a\mathcal{F}\tilde{R} \tag{3.32}$$

and, substituting (3.31) into this equation, we find that $\Omega(a)$ transforms as

$$\Omega(a) \mapsto \Omega'(a) = R\Omega(a)\tilde{R} - 2a\cdot\nabla R\tilde{R}. \tag{3.33}$$

Of course, since $\Omega(a)$ is an arbitrary function of position, it cannot in general be transformed away by the application of a rotor. We finally reassemble the derivative (3.30) with the $\overline{h}(\partial_a)$ term to form the equation

$$\overline{h}(\partial_a)\mathcal{D}_a\mathcal{F} = \mathcal{J}. \tag{3.34}$$

The transformation properties of the $\overline{h}(a)$, $\mathcal{F}$, $\mathcal{J}$ and $\Omega(a)$ fields ensure that this equation is now covariant under rotations as well as displacements.

To complete the set of transformation laws, we note that under displacements $\Omega(a)$ must transform in the same way as $a\cdot\nabla R\tilde{R}$, so that

$$\Omega_x(a) \mapsto \Omega_{x'}\underline{f}(a) = \Omega(\underline{f}(a), f(x)), \tag{3.35}$$

where the subscript is again used to label position dependence. It follows that

$$\begin{aligned}\overline{h}(\partial_a)\Omega_x(a)\times\mathcal{F}(x) &\mapsto \overline{h}_{x'}\overline{f}^{-1}(\partial_a)\Omega_{x'}\underline{f}(a)\times\mathcal{F}(x')\\ &= \overline{h}_{x'}(\partial_a)\Omega_{x'}(a)\times\mathcal{F}(x'),\end{aligned} \tag{3.36}$$

as required for covariance under local translations.

General considerations have led us to the introduction of two new gauge fields: the $\overline{h}(a, x)$ linear function and the $\Omega(a, x)$ bivector-valued linear function, both of which are arbitrary functions of the position vector x. This gives a total of $4\times4+4\times6 = 40$ scalar degrees of freedom. The $\overline{h}(a)$ and $\Omega(a)$ fields are incorporated into the vector derivative to form the operator $\overline{h}(\partial_a)\mathcal{D}_a$, which acts covariantly on multivector fields.

Thus we can begin to construct equations whose intrinsic content is free of the manner in which we represent spacetime positions with vectors. We next see how these fields arise in the setting of the Dirac theory. This enables us to derive the properties of the $\mathcal{D}_a$ operator from more primitive considerations of the properties of spinors and the means by which observables are constructed from them. First, though, let us compare the fields that we have defined with the fields used conventionally in GR. One might ask, for example, whether the $\overline{h}$ field is a disguised form of *vierbein*. A vierbein in GR relates a coordinate frame to an orthonormal frame. Whilst the $\overline{h}$ function can be used to construct such a vierbein (as discussed in Appendix C), it should be clear that the $\overline{h}$ function serves a totally different purpose in GTG—it ensures covariance under arbitrary displacements. This was the motivation for the introduction of a form of vierbein in Kibble's work (Kibble 1961), although only infinitesimal transformations could be considered there. In addition, the $\overline{h}$ field is essential to enable a clean separation between field rotations and displacements, which again is not achieved in other approaches. Further differences, relating to the existence and global properties of $\overline{h}$, will emerge in later sections.

(*c*) *Gauge fields for the Dirac action*

We now rederive the gravitational gauge fields from symmetries of the Dirac action. The point here is that, once the $\overline{h}$ field is introduced, spacetime rotations and phase rotations couple to the Dirac field in essentially the same way. To see this, we start with the Dirac equation and Dirac action in a slightly unconventional form (Hestenes 1975; Doran *et al.* 1993*c*; Lasenby *et al.* 1993*b*). We saw in § 2 that rotation of a multivector is performed by the double-sided application of a rotor. The elements of a linear space which is closed under single-sided action of a representation of the rotor group are called *spinors*. In conventional developments, a matrix representation for the Clifford algebra of spacetime is introduced and the space of column vectors on which these matrices act defines the spin space. But there is no need to adopt such a construction. For example, the even subalgebra of the STA forms a vector space which is closed under single-sided application of the rotor group. The even subalgebra is also an eight-dimensional vector space, the same number of real dimensions as a Dirac spinor, and so it is not surprising that a one-to-one map between Dirac spinors and the even subalgebra can be constructed. Such a map is given in Appendix A. The essential details are that the result of multiplying the column spinor $|\psi\rangle$ by the Dirac matrix $\hat{\gamma}^\mu$ is represented in the STA as $\psi \mapsto \gamma^\mu \psi \gamma_0$ and that multiplication by the scalar unit imaginary is represented as $\psi \mapsto \psi \mathrm{i}\sigma_3$. It is easily seen that these two operations commute and that they map even multivectors to even multivectors. By replacing Dirac matrices and column spinors by their STA equivalents the Dirac equation can be written in the form

$$\nabla\psi \mathrm{i}\sigma_3 - eA\psi = m\psi\gamma_0, \tag{3.37}$$

which is now representation-free and coordinate-free. Using the same replacements, the free-particle Dirac action becomes

$$S = \int |\mathrm{d}^4x| \langle \nabla\psi \mathrm{i}\gamma_3 \tilde{\psi} - m\psi\tilde{\psi} \rangle \tag{3.38}$$

and, with the techniques of Appendix B, it is simple to verify that variation of this action with respect to ψ yields equation (3.37) with $A = 0$.

It is important to appreciate that the fixed γ_0 and γ_3 vectors in (3.37) and (3.38)

do not pick out preferred directions in space. These vectors can be rotated to new vectors $R_0\gamma_0\tilde{R}_0$ and $R_0\gamma_3\tilde{R}_0$, and replacing the spinor by $\psi\tilde{R}_0$ recovers the same equation (3.37). This point will be returned to when we discuss forming observables from the spinor ψ.

Our aim now is to introduce gauge fields into the action (3.38) to ensure invariance under arbitrary rotations and displacements. The first step is to introduce the $\overline{h}$ field. Under a displacement, ψ transforms covariantly, so

$$\psi(x) \mapsto \psi'(x) = \psi(x'), \tag{3.39}$$

where $x' = f(x)$. We must therefore replace the ∇ operator by $\overline{h}(\nabla)$ so that $\overline{h}(\nabla)\psi$ is also covariant under translations. However, this on its own does not achieve complete invariance of the action integral (3.38) under displacements. The action consists of the integral of a scalar over some region. If the scalar is replaced by a displaced quantity, then we must also transform the measure and the boundary of the region if the resultant integral is to have the same value. Transforming the boundary is easily done, but the measure does require a little work. Suppose that we introduce a set of coordinate functions $\{x^\mu(x)\}$. The measure $|\mathrm{d}^4x|$ is then written

$$|\mathrm{d}^4x| = -\mathrm{i}e_0\wedge e_1\wedge e_2\wedge e_3\,\mathrm{d}x^0\,\mathrm{d}x^1\,\mathrm{d}x^2\,\mathrm{d}x^3, \tag{3.40}$$

where

$$e_\mu \equiv \frac{\partial x}{\partial x^\mu}. \tag{3.41}$$

By definition, $|\mathrm{d}^4x|$ is already independent of the choice of coordinates, but it must be modified to make it position gauge invariant. To see how, we note that under the displacement $x \mapsto f(x)$, the $\{e_\mu\}$ frame transforms to

$$e'_\mu(x) = \frac{\partial f(x)}{\partial x^\mu} = \underline{f}(e_\mu). \tag{3.42}$$

It follows that to ensure invariance of the action integral we must replace each of the e_μ by $\underline{h}^{-1}(e_\mu)$. Thus the invariant scalar measure is

$$-\,\mathrm{i}\,\underline{h}^{-1}(e_0)\wedge\cdots\wedge\underline{h}^{-1}(e_3)\,\mathrm{d}x^0\cdots\mathrm{d}x^3 = \det(\underline{h})^{-1}|\mathrm{d}^4x|. \tag{3.43}$$

These results lead us to the action

$$S = \int |\mathrm{d}^4x|\,\det(\underline{h})^{-1}\langle\overline{h}(\nabla)\psi\mathrm{i}\gamma_3\tilde{\psi} - m\psi\tilde{\psi}\rangle, \tag{3.44}$$

which is unchanged in value if the dynamical fields are replaced by

$$\psi'(x) \equiv \psi(x'), \tag{3.45}$$

$$\overline{h}'_x(a) \equiv \overline{h}_{x'}\overline{f}^{-1}(a) \tag{3.46}$$

and the boundary is also transformed.

(i) *Rotation and phase gauge fields*

Having arrived at the action in the form of (3.44), we can now consider the effect of rotations applied at a point. The representation of spinors by even elements is now particularly powerful because it enables both internal phase rotations and rotations in space to be handled in the same unified framework. Taking the electromagnetic coupling first, we see that the action (3.44) is invariant under the global phase rotation

$$\psi \mapsto \psi' \equiv \psi\mathrm{e}^{\mathrm{i}\sigma_3\phi}. \tag{3.47}$$

Table 1. *The symmetries of the action integral* (3.55)

	transformed fields			
local symmetry	$\psi'(x)$	$\overline{h}'(a,x)$	$\Omega'(a,x)$	$eA'(x)$
displacements	$\psi(x')$	$\overline{h}_{x'}\overline{f}^{-1}(a)$	$\Omega_{x'}\underline{f}(a)$	$e\overline{f}[A(x')]$
spacetime rotations	$R\psi$	$R\overline{h}(a)\tilde{R}$	$R\Omega(a)\tilde{R}-2a\cdot\nabla R\tilde{R}$	eA
phase rotations	$\psi e^{\phi i\sigma_3}$	$\overline{h}(a)$	$\Omega(a)$	$eA-\nabla\phi$

(Recall that multiplication of $|\psi\rangle$ by the unit imaginary is represented by right-sided multiplication of ψ by $i\sigma_3$.) The transformation (3.47) is a special case of the more general transformation

$$\psi \mapsto \psi R, \tag{3.48}$$

where R is a constant rotor. Similarly, invariance of the action (3.44) under spacetime rotations is described by

$$\psi \mapsto R\psi, \tag{3.49}$$

$$\overline{h}(a) \mapsto R\overline{h}(a)\tilde{R}. \tag{3.50}$$

In both cases, ψ just picks up a single rotor. From the previous section we know that, when the rotor R is position dependent, the quantity $a\cdot\nabla R\tilde{R}$ is a bivector-valued linear function of a. Since

$$a\cdot\nabla(R\psi) = Ra\cdot\nabla\psi + (a\cdot\nabla R\tilde{R})R\psi, \tag{3.51}$$

with a similar result holding when the rotor acts from the right, we need the following covariant derivatives for local internal and external rotations:

$$\text{internal:}\quad \mathrm{D}^{\mathrm{I}}_a\psi = a\cdot\nabla\psi + \tfrac{1}{2}\psi\Omega^{\mathrm{I}}(a), \tag{3.52}$$

$$\text{external:}\quad \mathrm{D}_a\psi = a\cdot\nabla\psi + \tfrac{1}{2}\Omega(a)\psi. \tag{3.53}$$

For the case of (internal) phase rotations, the rotations are constrained to take place entirely in the $i\sigma_3$ plane. It follows that the internal connection $\Omega^{\mathrm{I}}(a)$ takes the restricted form $2ea\cdot Ai\sigma_3$, where A is the conventional electromagnetic vector potential and e is the coupling constant (the charge). The full covariant derivative therefore has the form

$$\overline{h}(\partial_a)[a\cdot\nabla\psi + \tfrac{1}{2}\Omega(a)\psi + e\psi i\sigma_3 a\cdot A] \tag{3.54}$$

and the full invariant action integral is now

$$S = \int |\mathrm{d}^4x|\,\det(\underline{h})^{-1}\langle\overline{h}(\partial_a)[a\cdot\nabla + \tfrac{1}{2}\Omega(a)]\psi i\gamma_3\tilde{\psi} - e\overline{h}(A)\psi\gamma_0\tilde{\psi} - m\psi\tilde{\psi}\rangle. \tag{3.55}$$

The action (3.55) is invariant under the symmetry transformations listed in table 1.

(ii) *The coupled Dirac equation*

Having arrived at the action (3.55) we now derive the coupled Dirac equation by extremising with respect to ψ, treating all other fields as external. When applying the Euler–Lagrange equations to the action (3.55), the ψ and $\tilde{\psi}$ fields are not treated

as independent, as they often are in quantum theory. Instead, we just apply the rules for the multivector derivative discussed in § 2 *b* and Appendix B. The Euler–Lagrange equations can be written in the form

$$\partial_\psi \mathcal{L} = \partial_a \cdot \nabla (\partial_{\psi,a} \mathcal{L}), \tag{3.56}$$

as given in Appendix B. Applied to the action (3.55), equation (3.56) yields

$$\begin{aligned}(\overline{h}(\nabla)\psi \mathrm{i}\gamma_3)\tilde{} + \tfrac{1}{2}\mathrm{i}\gamma_3\tilde{\psi}\overline{h}(\partial_a)\Omega(a) + \tfrac{1}{2}(\overline{h}(\partial_a)\Omega(a)\psi\mathrm{i}\gamma_3)\tilde{} - e\gamma_0\tilde{\psi}\overline{h}(A) \\ -(e\overline{h}(A)\psi\gamma_0)\tilde{} - 2m\tilde{\psi} = a\cdot\nabla[\det(\underline{h})^{-1}\mathrm{i}\gamma_3\tilde{\psi}\overline{h}(\partial_a)]\det(\underline{h}).\end{aligned} \tag{3.57}$$

Reversing this equation and simplifying gives

$$\overline{h}(\partial_a)[a\cdot\nabla + \tfrac{1}{2}\Omega(a)]\psi\mathrm{i}\gamma_3 - e\overline{h}(A)\psi\gamma_0 - m\psi = -\tfrac{1}{2}\det(\underline{h})\mathcal{D}_a[\overline{h}(\partial_a)\det(\underline{h})^{-1}]\psi\mathrm{i}\gamma_3, \tag{3.58}$$

where we have employed the $\mathcal{D}_a$ derivative defined in equation (3.30). If we now introduce the notation

$$\mathrm{D}\psi \equiv \overline{h}(\partial_a)[a\cdot\nabla + \tfrac{1}{2}\Omega(a)]\psi, \tag{3.59}$$

$$\mathcal{A} \equiv \overline{h}(A), \tag{3.60}$$

we can write equation (3.58) in the form

$$\mathrm{D}\psi\mathrm{i}\sigma_3 - e\mathcal{A}\psi = m\psi\gamma_0 - \tfrac{1}{2}\det(\underline{h})\mathcal{D}_a[\overline{h}(\partial_a)\det(\underline{h})^{-1}]\psi\mathrm{i}\sigma_3. \tag{3.61}$$

This equation is manifestly covariant under the symmetries listed in table 1—as must be the case since the equation was derived from an invariant action integral. But equation (3.61) is not what we would have expected had we applied the gauging arguments at the level of the Dirac equation, rather than the Dirac action. Instead, we would have been led to the simpler equation

$$\mathrm{D}\psi\mathrm{i}\sigma_3 - e\mathcal{A}\psi = m\psi\gamma_0. \tag{3.62}$$

Clearly, equation (3.61) reduces to equation (3.62) only if the $\overline{h}(a)$ and $\Omega(a)$ fields satisfy the identity

$$\det(\underline{h})\mathcal{D}_a[\overline{h}(\partial_a)\det(\underline{h})^{-1}] = 0. \tag{3.63}$$

(It is not hard to show that the left-hand side of equation (3.63) is a covariant vector; later it will be identified as a contraction of the 'torsion' tensor). There are good reasons for expecting equation (3.63) to hold. Otherwise, the minimally coupled Dirac action would not yield the minimally coupled equation, which would pose problems for our use of action principles to derive the gauged field equations. We will see shortly that the demand that equation (3.63) holds places a strong restriction on the form that the gravitational action can take.

Some further comments about the derivation of (3.61) are now in order. The derivation employed only the rules of vector and multivector calculus applied to a 'flat-space' action integral. The derivation is therefore a rigorous application of the variational principle. This same level of rigour is not always applied when deriving field equations from action integrals involving spinors. Instead, the derivations are often heuristic; $|\psi\rangle$ and $\langle\tilde{\psi}|$ are treated as independent variables and the $\langle\tilde{\psi}|$ is just 'knocked off' the Lagrangian density to leave the desired equation. Furthermore, the action integral given by many authors for the Dirac equation in a gravitational background has an imaginary component (Nakahara 1990; Gockeler & Schucker 1987), in which case the status of the variational principle is unclear. To our knowledge,

only Hehl & Datta (Hehl *et al.* 1976) have produced a derivation that in any way matches the derivation produced here. Hehl & Datta also found an equation similar to (3.63), but they were not working within a gauge theory set-up and so did not comment on the consistency (or otherwise) of the minimal-coupling procedure.

(*d*) *Observables and covariant derivatives*

As well as keeping everything within the real STA, representing Dirac spinors by elements of the even subalgebra offers many advantages when forming observables. As described in Appendix A, observables are formed by the double-sided application of a Dirac spinor ψ to some combination of the fixed $\{\gamma^\mu\}$ frame vectors. So, for example, the charge current is given by $\mathcal{J} \equiv \psi\gamma_0\tilde{\psi}$ and the spin current by $s \equiv \psi\gamma_3\tilde{\psi}$. In general, an observable is of the form

$$M \equiv \psi\Gamma\tilde{\psi}, \tag{3.64}$$

where Γ is a constant multivector formed from the $\{\gamma^\mu\}$. All observables are invariant under phase rotations, so Γ must be invariant under rotations in the $\mathrm{i}\sigma_3$ plane. Hence Γ can consist only of combinations of γ_0, γ_3, $\mathrm{i}\sigma_3$ and their duals (formed by multiplying by i). An important point is that, in forming the observable M, the Γ multivector is completely 'shielded' from rotations. This is why the appearance of the γ_0 and γ_3 vectors on the right-hand side of the spinor ψ in the Dirac action (3.55) does not compromize Lorentz invariance and does not pick out a preferred direction in space (Doran *et al.* 1993*c*). All observables are unchanged by rotating the $\{\gamma^\mu\}$ frame vectors to $R_0\gamma_\mu\tilde{R}_0$ and transforming ψ to $\psi\tilde{R}_0$. (In the matrix theory this corresponds to a change of representation.)

Under translations and rotations the observables formed in the above manner (3.64) inherit the transformation properties of the spinor ψ. Under translations the observable $M = \psi\Gamma\tilde{\psi}$ therefore transforms from $M(x)$ to $M(x')$ and under rotations M transforms to $R\psi\Gamma\tilde{\psi}\tilde{R} = RM\tilde{R}$. The observable M is therefore covariant. These Dirac observables are the first examples of quantities which transform covariantly under rotations, but do not inherit this transformation law from the $\overline{h}$ field. In contrast, all covariant forms of classical fields, such as $\mathcal{F}$ or the covariant velocity along a worldline $\underline{h}^{-1}(\dot{x})$, transform under rotations in a manner that that is dictated by their coupling to the $\overline{h}$ field. Classical GR in fact removes any reference to the rotation gauge from most aspects of the theory. Quantum theory, however, demands that the rotation gauge be kept in explicitly and, as we shall show in § 8, Dirac fields probe the structure of the gravitational fields at a deeper level than classical fields. Furthermore, it is only through consideration of the quantum theory that one really discovers the need for the rotation-gauge field.

One might wonder why the observables are *invariant* under phase rotations, but only *covariant* under spatial rotations. In fact, the $\overline{h}$ field enables us to form quantities like $\underline{h}(M)$, which are invariant under spatial rotations. This gives an alternative insight into the role of the $\overline{h}$ field. We will find that both covariant observables (M) and their rotationally invariant forms ($\underline{h}(M)$ and $\overline{h}^{-1}(M)$) play important roles in the theory constructed here.

If we next consider the directional derivative of M, we find that it can be written as

$$a\cdot\nabla M = (a\cdot\nabla\psi)\Gamma\tilde{\psi} + \psi\Gamma\widetilde{(a\cdot\nabla\psi)}. \tag{3.65}$$

This immediately tells us how to turn the directional derivative $a\cdot\nabla M$ into a covari-

ant derivative: simply replace the spinor directional derivatives by covariant derivatives. Hence we form

$$(\mathrm{D}_a\psi)I\tilde{\psi} + \psi I(\mathrm{D}_a\psi)\tilde{} = (a\cdot\nabla\psi)I\tilde{\psi} + \psi I(a\cdot\nabla\psi)\tilde{} + \tfrac{1}{2}\Omega(a)\psi I\tilde{\psi} - \tfrac{1}{2}\psi I\tilde{\psi}\Omega(a)$$
$$= a\cdot\nabla(\psi I\tilde{\psi}) + \Omega(a)\times(\psi I\tilde{\psi}). \tag{3.66}$$

We therefore recover the covariant derivative for observables

$$\mathcal{D}_a M \equiv a\cdot\nabla M + \Omega(a)\times M. \tag{3.67}$$

This derivation shows that many features of the 'classical' derivation of gravitational gauge fields can be viewed as arising from more basic quantum transformation laws.

Throughout this section we have introduced a number of distinct gravitational covariant derivatives. We finish this section by discussing some of their main features and summarising our conventions. The operator $\mathcal{D}_a$ acts on any covariant multivector and has the important property of being a *derivation*; that is, it acts as a scalar differential operator,

$$\mathcal{D}_a(AB) = (\mathcal{D}_a A)B + A(\mathcal{D}_a B). \tag{3.68}$$

This follows from Leibniz's rule and the identity

$$\Omega(a)\times(AB) = \Omega(a)\times A\,B + A\,\Omega(a)\times B. \tag{3.69}$$

Neither D_a or $\mathcal{D}_a$ are fully covariant, however, since they both contain the $\Omega(a)$ field, which picks up a term in $\underline{f}$ under displacements (3.35). It is important in the applications to follow that we work with objects that are covariant under displacements and to this end we define

$$\omega(a) \equiv \Omega\underline{h}(a). \tag{3.70}$$

We also define the full covariant directional derivatives $a\cdot\mathrm{D}$ and $a\cdot\mathcal{D}$ by

$$a\cdot\mathrm{D}\psi \equiv a\cdot\overline{h}(\nabla)\psi + \tfrac{1}{2}\omega(a)\psi, \tag{3.71}$$
$$a\cdot\mathcal{D}M \equiv a\cdot\overline{h}(\nabla)M + \omega(a)\times M. \tag{3.72}$$

Under these conventions, $\mathcal{D}_a$ and $a\cdot\mathcal{D}$ are *not* the same object—they differ by the inclusion of the $\overline{h}$ field in the latter, so that

$$\underline{h}^{-1}(a)\cdot\mathcal{D} = \mathcal{D}_a. \tag{3.73}$$

The same comments apply to the spinor derivatives $a\cdot\mathrm{D}$ and $\mathrm{D}_a = \underline{h}^{-1}(a)\cdot\mathrm{D}$.

For the $a\cdot\mathcal{D}$ operator we can further define the covariant vector derivative

$$\mathcal{D}M \equiv \partial_a a\cdot\mathcal{D}M = \overline{h}(\partial_a)\mathcal{D}_a M. \tag{3.74}$$

The covariant vector derivative contains a grade-raising and a grade-lowering component, so that

$$\mathcal{D}A = \mathcal{D}\cdot A + \mathcal{D}\wedge A, \tag{3.75}$$

where

$$\mathcal{D}\cdot A \equiv \partial_a\cdot(a\cdot\mathcal{D}A) = \overline{h}(\partial_a)\cdot(\mathcal{D}_a A), \tag{3.76}$$
$$\mathcal{D}\wedge A \equiv \partial_a\wedge(a\cdot\mathcal{D}A) = \overline{h}(\partial_a)\wedge(\mathcal{D}_a A). \tag{3.77}$$

As with the vector derivative, $\mathcal{D}$ inherits the algebraic properties of a vector.

Table 2. *Definitions and conventions*

gauge fields	displacements: $\overline{h}(a)$
	rotations: $\Omega(a), \quad \omega(a) = \Omega\underline{h}(a)$
spinor derivatives	$\mathrm{D}_a\psi = a\cdot\nabla\psi + \frac{1}{2}\Omega(a)\psi$
	$a\cdot\mathrm{D}\psi = a\cdot\overline{h}(\nabla)\psi + \frac{1}{2}\omega(a)\psi$
'observables' derivatives	$\mathcal{D}_a M = a\cdot\nabla M + \Omega(a)\times M$
	$a\cdot\mathcal{D}M = a\cdot\overline{h}(\nabla)M + \omega(a)\times M$
	$\mathcal{D}M = \partial_a a\cdot\mathcal{D}M = \mathcal{D}\cdot M + \mathcal{D}\wedge M$
vector derivative	$\nabla = \gamma^\mu\dfrac{\partial}{\partial x^\mu}$
multivector derivative	$\partial_X = \displaystyle\sum_{i<\cdots<j} e^i\wedge\cdots\wedge e^j(e_j\wedge\cdots\wedge e_i)*\partial_X$

A summary of our definitions and conventions is contained in table 2. We have endeavoured to keep these conventions as simple and natural as possible, but a word is in order on our choices. It will become obvious when we consider the variational principle that it is a good idea to use a separate symbol for the spacetime vector derivative (∇), as opposed to writing it as ∂_x. This maintains a clear distinction between spacetime derivatives and operations on linear functions such as 'contraction' ($\partial_a\cdot$) and 'protraction' ($\partial_a\wedge$). It is also useful to distinguish between spinor and vector covariant derivatives, which is why we have introduced separate D and $\mathcal{D}$ symbols. We have avoided use of the d symbol, which already has a very specific meaning in the language of differential forms. Finally, it is necessary to distinguish between rotation-gauge derivatives ($\mathcal{D}_a$) and the full covariant derivative with the $\overline{h}$ field included ($a\cdot\mathcal{D}$). Using $\mathcal{D}_a$ and $a\cdot\mathcal{D}$ for these achieves this separation in the simplest possible manner.

4. The field equations

Having introduced the $\overline{h}$ and Ω fields, we now look to construct an invariant action integral which will provide a set of gravitational field equations. We start by defining the field strength via

$$\tfrac{1}{2}R(a\wedge b)\psi \equiv [\mathrm{D}_a, \mathrm{D}_b]\psi, \tag{4.1}$$

$$\Rightarrow \quad R(a\wedge b) = a\cdot\nabla\Omega(b) - b\cdot\nabla\Omega(a) + \Omega(a)\times\Omega(b). \tag{4.2}$$

It follows that we also have

$$[\mathcal{D}_a, \mathcal{D}_b]M = R(a\wedge b)\times M. \tag{4.3}$$

The field $R(a\wedge b)$ is a bivector-valued linear function of its bivector argument $a\wedge b$. Its action on bivectors extends by linearity to the function $R(B)$, where B is an arbitrary bivector and therefore, in four dimensions, not necessarily a pure 'blade' $a\wedge b$. Where required, the position dependence is made explicit by writing $R(B, x)$ or $R_x(B)$.

Under an arbitrary rotation, the definition (4.2) ensures that $R(B)$ transforms as

$$R(B) \mapsto R'(B) = R\,R(B)\tilde{R}. \tag{4.4}$$

(This unfortunate double use of the symbol R for the rotor and field strength is the only place where the two are used together.) Under local displacements we find that

$$\begin{aligned} R'(a\wedge b) &= a\cdot\nabla\Omega'(b) - b\cdot\nabla\Omega'(a) + \Omega'(a)\times\Omega'(b) \\ &= a\cdot\overline{f}(\dot{\nabla}_{x'})\dot{\Omega}_{x'}\underline{f}(b) - b\cdot\overline{f}(\dot{\nabla}_{x'})\dot{\Omega}_{x'}\underline{f}(a) + \Omega_{x'}\underline{f}(a)\times\Omega_{x'}\underline{f}(b) \\ &= R_{x'}\underline{f}(a\wedge b). \end{aligned} \tag{4.5}$$

This result rests on the fact that

$$a\cdot\nabla\underline{f}(b) - b\cdot\nabla\underline{f}(a) = [a\cdot\nabla, b\cdot\nabla]f(x) = 0. \tag{4.6}$$

A covariant quantity can therefore be constructed by defining

$$\mathcal{R}(B) \equiv R\underline{h}(B). \tag{4.7}$$

Under arbitrary displacements and local rotations, $\mathcal{R}(B)$ has the following transformation laws:

$$\begin{aligned} \text{translations:}\quad & \mathcal{R}'(B,x) = \mathcal{R}(B,x'), \\ \text{rotations:}\quad & \mathcal{R}'(B,x) = R\mathcal{R}(\tilde{R}BR,x)\tilde{R}. \end{aligned} \tag{4.8}$$

We refer to any linear function with transformation laws of this type as a covariant tensor. $\mathcal{R}(B)$ is our gauge theory analogue of the Riemann tensor. We have started to employ a notation which is very helpful for the theory developed here. Certain covariant quantities, such as $\mathcal{R}(B)$ and $\mathcal{D}$, are written with calligraphic symbols. This helps keep track of the covariant quantities and often enables a simple check that a given equation is gauge covariant. It is not necessary to write all covariant objects with calligraphic symbols, but it is helpful for objects such as $\mathcal{R}(B)$, since both $R(B)$ and $\mathcal{R}(B)$ arise in various calculations.

From $\mathcal{R}(B)$ we define the following contractions:

$$\text{Ricci tensor:}\quad \mathcal{R}(b) = \partial_a\cdot\mathcal{R}(a\wedge b), \tag{4.9}$$

$$\text{Ricci scalar:}\quad \mathcal{R} = \partial_a\cdot\mathcal{R}(a), \tag{4.10}$$

$$\text{Einstein tensor:}\quad \mathcal{G}(a) = \mathcal{R}(a) - \tfrac{1}{2}a\mathcal{R}. \tag{4.11}$$

The argument of $\mathcal{R}$ determines whether it represents the Riemann or Ricci tensors or the Ricci scalar. Both $\mathcal{R}(a)$ and $\mathcal{G}(a)$ are also covariant tensors, since they inherit the transformation properties of $\mathcal{R}(B)$.

The Ricci scalar is invariant under rotations, making it our first candidate for a Lagrangian for the gravitational gauge fields. We therefore suppose that the overall action integral is of the form

$$S = \int |\mathrm{d}^4x|\,\det(\underline{h})^{-1}(\tfrac{1}{2}\mathcal{R} - \kappa\mathcal{L}_m), \tag{4.12}$$

where $\mathcal{L}_m$ describes the matter content and $\kappa = 8\pi G$. The independent dynamical variables are $\overline{h}(a)$ and $\Omega(a)$ and, in terms of these,

$$\mathcal{R} = \langle\overline{h}(\partial_b\wedge\partial_a)[a\cdot\nabla\Omega(b) - b\cdot\nabla\Omega(a) + \Omega(a)\times\Omega(b)]\rangle. \tag{4.13}$$

We also assume that $\mathcal{L}_m$ contains no second-order derivatives, so that $\overline{h}(a)$ and $\Omega(a)$ appear undifferentiated in the matter Lagrangian.

(*a*) *The* $\overline{h}(a)$ *equation*

The $\overline{h}$ field is undifferentiated in the entire action, so its Euler–Lagrange equation is simply

$$\partial_{\overline{h}(a)}[\det(\underline{h})^{-1}(\tfrac{1}{2}\mathcal{R}-\kappa\mathcal{L}_m)]=0. \tag{4.14}$$

Employing some results from Appendix B, we find that

$$\partial_{\overline{h}(a)}\det(\underline{h})^{-1}=-\det(\underline{h})^{-1}\underline{h}^{-1}(a) \tag{4.15}$$

and

$$\partial_{\overline{h}(a)}\mathcal{R}=\partial_{\overline{h}(a)}\langle\overline{h}(\partial_c\wedge\partial_b)R(b\wedge c)\rangle=2\overline{h}(\partial_b)\cdot R(a\wedge b), \tag{4.16}$$

so that

$$\partial_{\overline{h}(a)}(\mathcal{R}\det(\underline{h})^{-1})=2\mathcal{G}\underline{h}^{-1}(a)\det(\underline{h})^{-1}. \tag{4.17}$$

If we now define the covariant matter stress-energy tensor $\mathcal{T}(a)$ by

$$\det(\underline{h})\partial_{\overline{h}(a)}(\mathcal{L}_m\det(\underline{h})^{-1})=\mathcal{T}\underline{h}^{-1}(a), \tag{4.18}$$

we arrive at the equation

$$\mathcal{G}(a)=\kappa\mathcal{T}(a). \tag{4.19}$$

This is the gauge theory statement of Einstein's equation though, as yet, nothing should be assumed about the symmetry of $\mathcal{G}(a)$ or $\mathcal{T}(a)$. In this derivation only the gauge fields have been varied and not the properties of spacetime. Therefore, despite the formal similarity with the Einstein equations of GR, there is no doubt that we are still working in a flat spacetime.

(*b*) *The* $\Omega(a)$ *equation*

The Euler–Lagrange field equation from $\Omega(a)$ is, after multiplying through by $\det(\underline{h})$,

$$\partial_{\Omega(a)}\mathcal{R}-\det(\underline{h})\partial_b\cdot\nabla[\partial_{\Omega(a)_{,b}}\mathcal{R}\det(\underline{h})^{-1}]=2\kappa\partial_{\Omega(a)}\mathcal{L}_m, \tag{4.20}$$

where we have made use of the assumption that $\Omega(a)$ does not contain any coupling to matter through its derivatives. The derivatives $\partial_{\Omega(a)}$ and $\partial_{\Omega(a)_{,b}}$ are defined in Appendix B. The only properties required for this derivation are the following:

$$\partial_{\Omega(a)}\langle\Omega(b)M\rangle=a\cdot b\langle M\rangle_2 \tag{4.21}$$

$$\partial_{\Omega(b)_{,a}}\langle c\cdot\nabla\Omega(d)M\rangle=a\cdot c\,b\cdot d\langle M\rangle_2. \tag{4.22}$$

From these we derive

$$\begin{aligned}\partial_{\Omega(a)}\langle\overline{h}(\partial_d\wedge\partial_c)\Omega(c)\times\Omega(d)\rangle&=\Omega(d)\times\overline{h}(\partial_d\wedge a)+\overline{h}(a\wedge\partial_c)\times\Omega(c)\\&=2\Omega(b)\times\overline{h}(\partial_b\wedge a)\\&=2[\omega(b)\cdot\partial_b)]\wedge\overline{h}(a)+2\partial_b\wedge[\omega(b)\cdot\overline{h}(a)]\end{aligned} \tag{4.23}$$

and

$$\partial_{\Omega(a)_{,b}}\langle\overline{h}(\partial_d\wedge\partial_c)[c\cdot\nabla\Omega(d)-d\cdot\nabla\Omega(c)]\rangle=\overline{h}(a\wedge b)-\overline{h}(b\wedge a)=2\overline{h}(a\wedge b). \tag{4.24}$$

The right-hand side of (4.20) defines the 'spin' of the matter,

$$S(a)\equiv\partial_{\Omega(a)}\mathcal{L}_m, \tag{4.25}$$

where $S(a)$ is a bivector-valued linear function of a. Combining (4.20), (4.23) and (4.24) yields

$$\mathcal{D}\wedge\overline{h}(a)+\det(\underline{h})\mathcal{D}_b[\overline{h}(\partial_b)\det(\underline{h})^{-1}]\wedge\overline{h}(a)=\kappa S(a). \tag{4.26}$$

To make further progress we contract this equation with $\underline{h}^{-1}(\partial_a)$. To achieve this we require the results that

$$\begin{aligned}\underline{h}^{-1}(\partial_a)\cdot[\mathcal{D}\wedge\overline{h}(a)] &= \mathcal{D}_a\overline{h}(\partial_a) - \overline{h}(\partial_b)\underline{h}^{-1}(\partial_a)\cdot[\mathcal{D}_b\overline{h}(a)] \\ &= \mathcal{D}_a\overline{h}(\partial_a) - \overline{h}(\partial_b)\langle\underline{h}^{-1}(\partial_a)b\cdot\nabla\overline{h}(a)\rangle \end{aligned} \tag{4.27}$$

and

$$\langle b\cdot\nabla\overline{h}(a)\underline{h}^{-1}(\partial_a)\rangle = \det(\underline{h})^{-1}\langle b\cdot\nabla\overline{h}(\partial_a)\partial_{\overline{h}(a)}\det(\underline{h})\rangle = \det(\underline{h})^{-1}b\cdot\nabla\det(\underline{h}). \tag{4.28}$$

Hence

$$\underline{h}^{-1}(\partial_a)\cdot[\mathcal{D}\wedge\overline{h}(a)] = \det(\underline{h})\mathcal{D}_b[\overline{h}(\partial_b)\det(\underline{h})^{-1}]. \tag{4.29}$$

It follows that

$$\det(\underline{h})\mathcal{D}_b[\overline{h}(\partial_b)\det(\underline{h})^{-1}] = -\tfrac{1}{2}\kappa\partial_a\cdot\mathcal{S}(a), \tag{4.30}$$

where $\mathcal{S}(a)$ is the covariant spin tensor defined by

$$\mathcal{S}(a) \equiv S\overline{h}^{-1}(a). \tag{4.31}$$

In § 3 *c* we found that the minimally coupled Dirac action gave rise to the minimally coupled Dirac equation only when $\mathcal{D}_b\overline{h}(\partial_b\det(\underline{h})^{-1}) = 0$. We now see that this requirement amounts to the condition that the spin tensor has zero contraction. However, if we assume that the $\Omega(a)$ field only couples to a Dirac fermion field, then the coupled Dirac action (3.55) gives

$$\mathcal{S}(a) = \mathcal{S}\cdot a, \tag{4.32}$$

where $\mathcal{S}$ is the spin trivector

$$\mathcal{S} = \tfrac{1}{2}\psi\mathrm{i}\gamma_3\tilde{\psi}. \tag{4.33}$$

In this case the contraction of the spin tensor does vanish:

$$\partial_a\cdot(\mathcal{S}\cdot a) = (\partial_a\wedge a)\cdot\mathcal{S} = 0. \tag{4.34}$$

There is a remarkable consistency loop at work here. The Dirac action gives rise to a spin tensor of just the right type to ensure that the minimally coupled action produces the minimally coupled equation. *But this is only true if the gravitational action is given by the Ricci scalar*! No higher-order gravitational action is consistent in this way. So, if we demand that the minimally coupled field equations should be derivable from an action principle, we are led to a highly constrained theory. This rules out, for example, the type of '$R+R^2$' Lagrangian often considered in the context of Poincaré gauge theory (Rauch 1982; Hecht *et al.* 1991; Khodunov & Zhytnikov 1992). In addition, the spin sector is constrained to have a vanishing contraction. Some consequences of this are explored in Doran *et al.* (1998*b*). Satisfyingly, these constraints force us to a theory which is first order in the derivatives of the fields, keeping the theory on a similar footing to the Dirac and Maxwell theories.

The only freedom in the action for the gravitational fields is the possible inclusion of a cosmological constant Λ. This just enters the action integral (4.12) as the term $-\Lambda\det(\underline{h})^{-1}$. The presence of such a term does not alter equation (4.26), but changes (4.19) to

$$\mathcal{G}(a) - \Lambda a = \kappa\mathcal{T}(a). \tag{4.35}$$

The presence of a cosmological constant cannot be ruled out on theoretical grounds alone and this constant will be included when we consider applications to cosmology.

Given that the spin is entirely of Dirac type, equation (4.26) now takes the form

$$\mathcal{D}\wedge\overline{h}(a) = \kappa\mathcal{S}\cdot\overline{h}(a). \tag{4.36}$$

This is the second of our gravitational field equations. Equations (4.19) and (4.36) define a set of 40 scalar equations for the 40 unknowns in $\overline{h}(a)$ and $\Omega(a)$. Both equations are manifestly covariant. In the spin-torsion extension of GR (the Einstein–Cartan–Sciama–Kibble theory), $\mathcal{D}\wedge\overline{h}(a)$ would be identified as the gravitational torsion and equation (4.36) would be viewed as identifying the torsion with the matter spin density. Of course, in GTG, torsion is not a property of the underlying spacetime, it simply represents a feature of the gravitational gauge fields. Equation (4.36) generalizes to the case of an arbitrary vector $a = a(x)$ as follows:

$$\mathcal{D}\wedge\overline{h}(a) = \overline{h}(\nabla\wedge a) + \kappa\mathcal{S}\cdot\overline{h}(a). \tag{4.37}$$

(*c*) *Covariant forms of the field equations*

For all the applications considered in this paper the gravitational fields are generated by matter fields with vanishing spin. So, to simplify matters, we henceforth set $\mathcal{S}$ to zero and work with the second of the field equations in the form

$$\mathcal{D}\wedge\overline{h}(a) = 0. \tag{4.38}$$

It is not hard to make the necessary generalizations in the presence of spin. Indeed, even if the spin-torsion sector is significant, one can introduce a new field (Doran *et al.* 1998*b*)

$$\omega'(a) \equiv \omega(a) - \tfrac{1}{2}\kappa a\cdot\mathcal{S} \tag{4.39}$$

and then the modified covariant derivative with $\omega(a)$ replaced by $\omega'(a)$ still satisfies equation (4.38).

The approach we adopt in this paper is to concentrate on the quantities which are covariant under displacements. Since both $\overline{h}(\nabla)$ and $\omega(a)$ satisfy this requirement, these are the quantities with which we would like to express the field equations. To this end we define the operator

$$L_a \equiv a\cdot\overline{h}(\nabla) \tag{4.40}$$

and, for the remainder of this section, the vectors a, b, etc., are assumed to be arbitrary functions of position. From equation (4.38) we write

$$\begin{aligned} &\overline{h}(\dot{\nabla})\wedge\dot{\overline{h}}(c) = -\partial_d\wedge[\omega(d)\cdot\overline{h}(c)] \\ \Rightarrow\quad &\langle b\wedge a\,\overline{h}(\dot{\nabla})\wedge\dot{\overline{h}}(c)\rangle = -\langle b\wedge a\,\partial_d\wedge[\omega(d)\cdot\overline{h}(c)]\rangle \\ \Rightarrow\quad &[\dot{L}_a\underline{\dot{h}}(b) - \dot{L}_b\underline{\dot{h}}(a)]\cdot c = [a\cdot\omega(b) - b\cdot\omega(a)]\cdot\overline{h}(c), \end{aligned} \tag{4.41}$$

where, as usual, the overdots determine the scope of a differential operator. It follows that the commutator of L_a and L_b is

$$\begin{aligned} [L_a, L_b] &= [L_a\underline{h}(b) - L_b\underline{h}(a)]\cdot\nabla \\ &= [\dot{L}_a\underline{\dot{h}}(b) - \dot{L}_b\underline{\dot{h}}(a)]\cdot\nabla + (L_a b - L_b a)\cdot\overline{h}(\nabla) \\ &= [a\cdot\omega(b) - b\cdot\omega(a) + L_a b - L_b a]\cdot\overline{h}(\nabla). \end{aligned} \tag{4.42}$$

We can therefore write

$$[L_a, L_b] = L_c, \tag{4.43}$$

where

$$c = a\cdot\omega(b) - b\cdot\omega(a) + L_a b - L_b a = a\cdot\mathcal{D}b - b\cdot\mathcal{D}a. \tag{4.44}$$

This 'bracket' structure summarizes the intrinsic content of (4.36).

The general technique we use for studying the field equations is to let $\omega(a)$ contain a set of arbitrary functions and then use (4.44) to find relations between them. Fundamental to this approach is the construction of the Riemann tensor $\mathcal{R}(B)$, which contains a great deal of covariant information. From the definition of the Riemann tensor (4.2) we find that

$$\begin{aligned}\mathcal{R}(a\wedge b) &= \dot{L}_a\dot{\Omega}\underline{h}(a) - \dot{L}_b\dot{\Omega}\underline{h}(a) + \omega(a)\times\omega(b)\\ &= L_a\omega(b) - L_b\omega(a) + \omega(a)\times\omega(b) - \Omega(L_a\underline{h}(b) - L_b\underline{h}(a)),\end{aligned} \tag{4.45}$$

hence

$$\mathcal{R}(a\wedge b) = L_a\omega(b) - L_b\omega(a) + \omega(a)\times\omega(b) - \omega(c), \tag{4.46}$$

where c is given by equation (4.44). Equation (4.46) now enables $\mathcal{R}(B)$ to be calculated in terms of position-gauge covariant variables.

(i) *Solution of the 'wedge' equation*

Equation (4.38) can be solved to obtain $\omega(a)$ as a function of $\overline{h}$ and its derivatives. We define

$$H(a) \equiv \overline{h}(\nabla\wedge\overline{h}^{-1}(a)) = -\overline{h}(\dot{\nabla})\wedge\dot{\overline{h}}\,\overline{h}^{-1}(a), \tag{4.47}$$

so that equation (4.36) becomes

$$\partial_b\wedge[\omega(b)\cdot a] = H(a). \tag{4.48}$$

We solve this by first 'protracting' with ∂_a to give

$$\partial_a\wedge\partial_b\wedge[\omega(b)\cdot a] = 2\partial_b\wedge\omega(b) = \partial_b\wedge H(b). \tag{4.49}$$

Now, taking the inner product with a again, we obtain

$$\omega(a) - \partial_b\wedge(a\cdot\omega(b)) = \tfrac{1}{2}a\cdot[\partial_b\wedge H(b)]. \tag{4.50}$$

Hence, using equation (4.48) again, we find that

$$\omega(a) = -H(a) + \tfrac{1}{2}a\cdot(\partial_b\wedge H(b)). \tag{4.51}$$

In the presence of spin generated by matter of spin-$\frac{1}{2}$, the term $\frac{1}{2}\kappa a\cdot\mathcal{S}$ is added to the right-hand side.

(*d*) *Point-particle trajectories*

The dynamics of a fermion in a gravitational background are described by the Dirac equation (3.62) together with the quantum-mechanical rules for constructing observables. For many applications, however, it is useful to work with classical and semi-classical approximations to the full quantum theory. The full derivation of the semi-classical approximation will be given elsewhere, but the essential idea is to specialize to motion along a single streamline defined by the Dirac current $\psi\gamma_0\tilde{\psi}$. Thus the particle is described by a trajectory $x(\lambda)$, together with a spinor $\psi(\lambda)$ which contains information about the velocity and spin of the particle. The covariant velocity is $\underline{h}^{-1}(\dot{x})$ where, for this and the following subsection, overdots are used to denote the derivative with respect to λ. The covariant velocity is identified with $\psi\gamma_0\tilde{\psi}$ and the Lagrange multiplier p is included in the action integral to enforce this identification. Finally, an einbein e is introduced to ensure reparametrization invariance. The resultant action is

$$S = \int \mathrm{d}\lambda\,\langle\dot{\psi}\mathrm{i}\sigma_3\tilde{\psi} + \tfrac{1}{2}\Omega(\dot{x})\psi\mathrm{i}\sigma_3\tilde{\psi} + p(v - me\psi\gamma_0\tilde{\psi}) + m^2e\rangle, \tag{4.52}$$

where

$$v \equiv \underline{h}^{-1}(\dot{x}). \tag{4.53}$$

The equations of motion arising from (4.52) will be presented and analysed elsewhere. (An effect worth noting is that, due to the spin of the particle, the velocity v and momentum p are not collinear.)

We can make a full classical approximation by neglecting the spin (dropping all the terms containing ψ) and replacing $\psi\gamma_0\tilde{\psi}$ by p/m. This process leads to the action

$$S = \int \mathrm{d}\lambda\, [p\cdot\underline{h}^{-1}(\dot{x}) - \tfrac{1}{2}e(p^2 - m^2)]. \tag{4.54}$$

The equations of motion derived from (4.54) are

$$v = ep, \tag{4.55}$$

$$p^2 = m^2, \tag{4.56}$$

$$\partial_\lambda \overline{h}^{-1}(p) = \overset{*}{\nabla}\, p\cdot \overset{*}{\underline{h}}{}^{-1}(\dot{x}), \tag{4.57}$$

where, for this section only, we employ overstars in place of overdots for the scope of a differential operator. The latter equation yields

$$\begin{aligned}\dot{p} &= \overline{h}(\overset{*}{\nabla})p\cdot \overset{*}{\underline{h}}{}^{-1}(\dot{x}) - \dot{x}\cdot \overset{*}{\nabla}\, \overline{h}\, \overset{*}{\overline{h}}{}^{-1}(p)\\ &= \overline{h}[(\overset{*}{\nabla}\wedge \overset{*}{\overline{h}}{}^{-1}(p))\cdot\underline{h}(v)] = H(p)\cdot v/e,\end{aligned} \tag{4.58}$$

where $H(a)$ is defined by equation (4.47) and we have used equation (2.52). From equation (4.51) we see that $a\cdot\omega(a) = -a\cdot H(a)$, hence

$$e\partial_\lambda(v/e) = -\omega(v)\cdot v. \tag{4.59}$$

This is the classical equation for a point-particle trajectory. It takes its simplest form when λ is the proper time τ along the trajectory. In this case $e = 1/m$ and the equation becomes

$$\dot{v} = -\omega(v)\cdot v, \tag{4.60}$$

or, in manifestly covariant form,

$$v\cdot\mathcal{D}v = 0. \tag{4.61}$$

This equation applies equally for massive particles ($v^2 = 1$) and photons ($v^2 = 0$). Since equation (4.61) incorporates only gravitational effects, any deviation of $v\cdot\mathcal{D}v$ from zero can be viewed as the particle's acceleration and must result from additional external forces.

Equation (4.61) is usually derived from the action

$$S = m\int \mathrm{d}\lambda\, \sqrt{\underline{h}^{-1}(\dot{x})^2}, \tag{4.62}$$

which is obtained from (4.54) by eliminating p and e with their respective equations of motion. A Hamiltonian form such as (4.54) is rarely seen in conventional GR, since its analogue would require the introduction of a vierbein. Despite this, the action (4.54) has many useful features, especially when it comes to extracting conservation laws. For example, contracting equation (4.57) with a constant vector a yields

$$\partial_\lambda[a\cdot\overline{h}^{-1}(p)] = a\cdot\overset{*}{\nabla}\, p\cdot \overset{*}{\underline{h}}{}^{-1}(\dot{x}). \tag{4.63}$$

It follows that if the $\overline{h}$ field is invariant under translations in the direction a then the quantity $a\cdot\overline{h}^{-1}(p)$ is conserved. In § 5 *c* we show that this result extends to the case where $\underline{h}^{-1}(a)$ is a Killing vector.

(e) Measurements, the equivalence principle and the Newtonian limit

In the preceeding section we derived the equation $v\cdot\mathcal{D}v = 0$ from the classical limit of the Dirac action. This equation is the GTG analogue of the geodesic equation (see Appendix C). Arriving at such an equation shows that GTG embodies the weak equivalence principle—the motion of a test particle in a gravitational field is independent of its mass. The derivation also shows the limitations of this result, which only applies in the classical, spinless approximation to quantum theory. The strong equivalence principle, that the laws of physics in a freely falling frame are (locally) the same as those of special relativity, is also embodied in GTG through the application of the minimal coupling procedure. Indeed, it is clear that both of these 'principles' are the result of the minimal coupling procedure. Minimal coupling ensures, through the Dirac equation, that point-particle trajectories are independent of particle mass in the absence of other forces. It also tells us how the gravitational fields couple to any matter field. As we have seen, this principle, coupled with the requirement of consistency with an action principle, is sufficient to specify the theory uniquely (up to an unspecified cosmological constant).

The relationship between the minimal coupling procedure (the gauge principle) and the equivalence principle has been noted previously. Feynman (Feynman *et al.* 1995), for example, argues that a more exact version of the equivalence between linear acceleration and a gravitational field requires an equation of the form

$$\text{gravity}' = \text{gravity} + \text{acceleration}, \tag{4.64}$$

which clearly resembles a gauge transformation. What is not often stressed is the viewpoint presented here, which is that if gravity is constructed entirely as a gauge theory, there is no need to invoke the equivalence principle; the physical effects embodied in the principle are simply consequences of the gauge theory approach. This further illustrates the different conceptual foundations of GTG and GR. Similarly, there is no need for the principle of general covariance in GTG, which is replaced by the requirement that all physical predictions be gauge invariant. It is often argued that the principle of general covariance is empty, because any physical theory can be written in a covariant form. This objection cannot be levelled at the statement that all physical predictions must be gauge invariant, which has clear mathematical and physical content.

The simplest, classical measurements in GTG are modelled by assuming that observers can be treated as frames attached to a single worldline. If this worldline is written as $x(\lambda)$, then the covariant velocity is $v = \underline{h}^{-1}(\dot{x})$ and the affine parameter for the trajectory is that which ensures that $v^2 = 1$. The affine parameter models the clock time for an observer on this trajectory. Of course, there are many hidden assumptions in adopting this as a realistic model—quantum effects are ignored, as is the physical extent of the observer—but it is certainly a good model in weak fields. In strong fields a more satisfactory model would involve solving the Dirac equation to find the energy levels of an atom in the gravitational background and use this to model an atomic clock.

Equation (4.61) enables us to make classical predictions for freely falling trajectories in GTG and the photon case ($v^2 = 0$) can be used to model signalling between observers. As an example, consider the formula for the redshift induced by the gravitational fields. Suppose that a source of radiation follows a worldline $x_1(\tau_1)$, with covariant velocity $v_1 = \underline{h}^{-1}(\dot{x}_1)$. The radiation emitted follows a null trajectory with covariant velocity u ($u^2 = 0$). This radiation is received by an observer with a worldline $x_2(\tau_2)$ and covariant velocity $v_2 = \underline{h}^{-1}(\dot{x}_2)$. The spectral shift z is determined

by the ratio of the frequency observed at the source, $u(x_1)\cdot v_1$, and the frequency observed at the receiver, $u(x_2)\cdot v_2$, by

$$1+z \equiv \frac{u(x_1)\cdot v_1}{u(x_2)\cdot v_2}. \tag{4.65}$$

This quantity is physically observable since the right-hand side is a gauge-invariant quantity. This is because each of the four vectors appearing in (4.65) is covariant, which eliminates any dependence on position gauge, and taking the dot product between pairs of vectors eliminates any dependence on the rotation gauge.

A final point to address regarding the foundations of GTG is the recovery of the Newtonian limit. Derivations of this are easily produced by adapting the standard work in GR. Furthermore, in § 6 *c* we show that the description of the gravitational fields outside a static, spherically symmetric star is precisely the same as in GR. The trajectories defined by equation (4.61) are those predicted by GR, so all of the predictions for planetary orbits (including those for binary pulsars) are unchanged. Similarly, the results for the bending of light are the same as in GR. As we show in the applications, the differences between GTG and GR emerge through the relationship with quantum theory and through the global nature of the gauge fields in GTG. These differences have no consequences for classical tests of GR, though they are potentially testable through the interaction with quantum spin and are certainly significant for discussing more fundamental aspects of gravitational physics.

5. Symmetries, invariants and conservation laws

Having determined the gravitational gauge fields and their field equations, we now establish some general results which are applied in the sections that follow. Again, we restrict to the case of vanishing torsion. The approach we adopt in solving the field equations is to let $\omega(a)$ be an arbitrary function and then work with a set of abstract first-order equations for the terms that comprise $\omega(a)$. However, in letting $\omega(a)$ be an arbitrary function, we lose some of the information contained in the 'wedge' equation (4.38). This information is recovered by enforcing various properties that the fields must satisfy, including the symmetry properties of $\mathcal{R}(B)$ and the Bianchi identities. In addition, it is often necessary to enforce some gauge-fixing conditions. For the rotation gauge these conditions are applied by studying $\mathcal{R}(B)$, so it is important to analyse its general structure.

We start with the result that, for an arbitrary multivector $A(x)$,

$$\mathcal{D}\wedge\mathcal{D}\wedge A = \mathcal{D}\wedge\overline{h}[\nabla\wedge\overline{h}^{-1}(A)] = \overline{h}(\nabla\wedge\nabla\wedge\overline{h}^{-1}(A)) = 0, \tag{5.1}$$

where we have made use of equation (4.38). It then follows from the result

$$\begin{aligned}\overline{h}(\partial_a)\wedge\mathcal{D}_a[\overline{h}(\partial_b)\wedge\mathcal{D}_b A] &= \overline{h}(\partial_a)\wedge\overline{h}(\partial_b)\wedge[\mathcal{D}_a\mathcal{D}_b A]\\ &= \tfrac{1}{2}\overline{h}(\partial_a)\wedge\overline{h}(\partial_b)\wedge[R(a\wedge b)\times A],\end{aligned} \tag{5.2}$$

that, for any multivector A,

$$\partial_a\wedge\partial_b\wedge(\mathcal{R}(a\wedge b)\times A) = 0. \tag{5.3}$$

This derivation illustrates a useful point. Many derivations can be performed most efficiently by working with the $\mathcal{D}_a$, since these contain commuting partial derivatives. However, the final expressions take their most transparent form when the $\overline{h}$ field is included so that only fully covariant quantities are employed.

If we now set A in equation (5.3) equal to a vector c, and protract with ∂_c, we find that

$$\partial_c\wedge\partial_a\wedge\partial_b\wedge(\mathcal{R}(a\wedge b)\times c) = -2\partial_a\wedge\partial_b\wedge\mathcal{R}(a\wedge b) = 0. \tag{5.4}$$

Taking the inner product of the term on the right-hand side with c we obtain

$$c\cdot[\partial_a\wedge\partial_b\wedge\mathcal{R}(a\wedge b)] = \partial_b\wedge\mathcal{R}(c\wedge b) - \partial_a\wedge\mathcal{R}(a\wedge c) - \partial_a\wedge\partial_b\wedge[\mathcal{R}(a\wedge b)\times c], \tag{5.5}$$

in which both the left-hand side and the final term on the right-hand side vanish. We are therefore left with the simple expression

$$\partial_a\wedge\mathcal{R}(a\wedge b) = 0, \tag{5.6}$$

which summarizes all the symmetries of $\mathcal{R}(B)$. This equation says that the trivector $\partial_a\wedge\mathcal{R}(a\wedge b)$ vanishes for all values of the vector b, so gives a set of $4\times4 = 16$ equations. These reduce the number of independent degrees of freedom in $\mathcal{R}(B)$ from 36 to the expected 20. It should be clear from the ease with which the degrees of freedom are calculated that the present geometric algebra formulation has many advantages over traditional tensor calculus.

(*a*) *The Weyl tensor*

A good example of the power of the present approach is provided by an analysis of the Riemann and Weyl tensors. To illustrate this point a number of examples of $\mathcal{R}(B)$ for physical systems are included in this section. (These are stated here without derivation.) The first application of geometric algebra to the analysis of the Riemann tensor in classical differential geometry was given by Hestenes & Sobczyk (Hestenes & Sobczyk 1984; Sobczyk 1981). Here the formalism is developed and extended for applications relevant to our gauge theory of gravity.

In GTG, six of the degrees of freedom in $\mathcal{R}(B)$ can be removed by arbitrary rotations. It follows that $\mathcal{R}(B)$ can contain only 14 physical degrees of freedom. To see how these are encoded in $\mathcal{R}(B)$ we decompose it into Weyl and 'matter' terms. Since the contraction of $\mathcal{R}(a\wedge b)$ results in the Ricci tensor $\mathcal{R}(a)$, we expect that $\mathcal{R}(a\wedge b)$ will contain a term in $\mathcal{R}(a)\wedge b$. This must be matched with a term in $a\wedge\mathcal{R}(b)$, since it is only the sum of these that is a function of $a\wedge b$. Contracting this sum we obtain

$$\partial_a\cdot[\mathcal{R}(a)\wedge b + a\wedge\mathcal{R}(b)] = b\mathcal{R} - \mathcal{R}(b) + 4\mathcal{R}(b) - \mathcal{R}(b) = 2\mathcal{R}(b) + b\mathcal{R} \tag{5.7}$$

and it follows that

$$\partial_a\cdot[\tfrac{1}{2}(\mathcal{R}(a)\wedge b + a\wedge\mathcal{R}(b)) - \tfrac{1}{6}a\wedge b\mathcal{R}] = \mathcal{R}(b). \tag{5.8}$$

We can therefore write

$$\mathcal{R}(a\wedge b) = \mathcal{W}(a\wedge b) + \tfrac{1}{2}[\mathcal{R}(a)\wedge b + a\wedge\mathcal{R}(b)] - \tfrac{1}{6}a\wedge b\mathcal{R}, \tag{5.9}$$

where $\mathcal{W}(B)$ is the Weyl tensor and must satisfy

$$\partial_a\cdot\mathcal{W}(a\wedge b) = 0. \tag{5.10}$$

Returning to equation (5.6) and contracting, we obtain

$$\partial_b\cdot(\partial_a\wedge\mathcal{R}(a\wedge b) = \partial_a\wedge\mathcal{R}(a) = 0, \tag{5.11}$$

which shows that the Ricci tensor $\mathcal{R}(a)$ is symmetric. It follows that

$$\partial_a\wedge[\tfrac{1}{2}(\mathcal{R}(a)\wedge b + a\wedge\mathcal{R}(b)) - \tfrac{1}{6}a\wedge b\mathcal{R}] = 0 \tag{5.12}$$

and hence that

$$\partial_a\wedge\mathcal{W}(a\wedge b) = 0. \tag{5.13}$$

Equations (5.10) and (5.13) combine to give the single equation

$$\partial_a \mathcal{W}(a \wedge b) = 0. \tag{5.14}$$

Since the $\partial_a \cdot$ operation is called the 'contraction', and $\partial_a \wedge$ the 'protraction', Hestenes & Sobczyk (1984) have suggested that the sum of these be termed the 'traction'. Equation (5.9) thus decomposes $\mathcal{R}(B)$ into a 'tractionless' term $\mathcal{W}(B)$ and a term specified solely by the matter stress-energy tensor (which determines $\mathcal{R}(a)$ through the Einstein tensor $\mathcal{G}(a)$). There is no generally accepted name for the part of $\mathcal{R}(B)$ that is not given by the Weyl tensor so, as it is entirely determined by the matter stress-energy tensor, we refer to it as the matter or source term.

(i) *Duality*

To study the consequences of equation (5.14) it is useful to employ the fixed $\{\gamma_\mu\}$ frame, so that equation (5.14) produces the four equations

$$\sigma_1 \mathcal{W}(\sigma_1) + \sigma_2 \mathcal{W}(\sigma_2) + \sigma_3 \mathcal{W}(\sigma_3) = 0, \tag{5.15}$$

$$\sigma_1 \mathcal{W}(\sigma_1) - \mathrm{i}\sigma_2 \mathcal{W}(\mathrm{i}\sigma_2) - \mathrm{i}\sigma_3 \mathcal{W}(\mathrm{i}\sigma_3) = 0, \tag{5.16}$$

$$-\mathrm{i}\sigma_1 \mathcal{W}(\mathrm{i}\sigma_1) + \sigma_2 \mathcal{W}(\sigma_2) - \mathrm{i}\sigma_3 \mathcal{W}(\mathrm{i}\sigma_3) = 0, \tag{5.17}$$

$$-\mathrm{i}\sigma_1 \mathcal{W}(\mathrm{i}\sigma_1) - \mathrm{i}\sigma_2 \mathcal{W}(\mathrm{i}\sigma_2) + \sigma_3 \mathcal{W}(\sigma_3) = 0. \tag{5.18}$$

Summing the final three equations, and using the first, produces

$$\mathrm{i}\sigma_k \mathcal{W}(\mathrm{i}\sigma_k) = 0 \tag{5.19}$$

and substituting this into each of the final three equations produces

$$\mathcal{W}(\mathrm{i}\sigma_k) = \mathrm{i}\mathcal{W}(\sigma_k). \tag{5.20}$$

It follows that the Weyl tensor satisfies

$$\mathcal{W}(\mathrm{i}B) = \mathrm{i}\mathcal{W}(B) \tag{5.21}$$

and so is 'self-dual'. This use of the term 'self-dual' differs slightly from its use in the two-spinor formalism of Penrose & Rindler (1984). However, the pseudoscalar i in the STA performs the same role as the Hodge star operation (the duality transformation) in differential form theory, so 'self-duality' is clearly an appropriate name for the relation expressed by equation (5.21).

The fact that tractionless linear functions mapping bivectors to bivectors in spacetime satisfy equation (5.21) was first noted by Hestenes & Sobczyk (1984). Equation (5.21) means that $\mathcal{W}(B)$ can be analysed as a linear function on a three-dimensional complex space rather as a function on a real six-dimensional space. This is why complex formalisms, such as the Newman–Penrose formalism, are so successful for studying vacuum solutions. The unit imaginary employed in the Newman–Penrose formalism is a disguised version of the spacetime pseudoscalar (Lasenby *et al.* 1993*c*). Geometric algebra reveals the geometric origin of this 'imaginary' unit and enables us to employ results from complex analysis without the need for formal complexification. Furthermore, the complex structure only arises in situations where it is geometrically significant, instead of being formally present in all calculations.

Given the self-duality of the Weyl tensor, the remaining content of equations (5.15)–(5.18) is summarized by the relation

$$\sigma_k \mathcal{W}(\sigma_k) = 0. \tag{5.22}$$

This equation says that, viewed as a three-dimensional complex linear function,

$\mathcal{W}(B)$ is symmetric and traceless. This gives $\mathcal{W}(B)$ five complex, or ten real, degrees of freedom. (Since we frequently encounter combinations of the form scalar + pseudoscalar, we refer to these loosely as complex scalars.) The gauge-invariant information in $\mathcal{W}(B)$ is held in its complex eigenvalues and, since the sum of these is zero, only two are independent. This leaves a set of four real intrinsic scalar quantities.

Overall, $\mathcal{R}(B)$ has 20 degrees of freedom, six of which are contained in the freedom to perform arbitrary local rotations. Of the remaining 14 physical degrees of freedom, four are contained in the two complex eigenvalues of $\mathcal{W}(B)$ and a further four in the real eigenvalues of the matter stress-energy tensor. The six remaining physical degrees of freedom determine the rotation between the frame that diagonalizes $\mathcal{G}(a)$ and the frame that diagonalizes $\mathcal{W}(B)$. This identification of the physical degrees of freedom contained in $\mathcal{R}(B)$ appears to be new and is potentially very significant.

(ii) *The Petrov classification*

The algebraic properties of the Weyl tensor are traditionally encoded in its Petrov type. Here we present geometric algebra expressions for the main Petrov types (following the conventions of Kramer *et al.* (1980)). The Petrov classification is based on the solutions of the eigenvalue equation

$$\mathcal{W}(B) = \alpha B, \tag{5.23}$$

in which B is now a bivector (the eigenbivector) and α is a complex scalar. There are five Petrov types: I, II, III, D and N. Type I are the most general, with two independent eigenvalues and three linearly independent orthogonal eigenbivectors. Such tensors have the general form

$$\mathcal{W}(B) = \tfrac{1}{2}\alpha_1(B + 3F_1BF_1) + \tfrac{1}{2}\alpha_2(B + 3F_2BF_2), \tag{5.24}$$

where α_1 and α_2 are complex scalars and F_1 and F_2 are orthogonal unit bivectors ($F_1^2 = F_2^2 = 1$). The eigenbivectors are F_1, F_2 and $F_3 \equiv F_1F_2$, and the corresponding eigenvalues are $(2\alpha_1 - \alpha_2)$, $(2\alpha_2 - \alpha_1)$ and $-(\alpha_1 + \alpha_2)$.

Type D (degenerate) are a special case of type I tensors in which two of the eigenvalues are the same. Physical examples are provided by the Schwarzschild and Kerr solutions. The region outside a spherically symmetric source of mass M has

$$\mathcal{R}(B) = \mathcal{W}(B) = -\frac{M}{2r^3}(B + 3\sigma_r B\sigma_r), \tag{5.25}$$

where σ_r is the unit radial bivector. The eigenbivectors of this function are σ_r, with eigenvalue $-2M/r^3$, and any two bivectors perpendicular to σ_r, with eigenvalue M/r^3. Similarly, $\mathcal{R}(B)$ for a stationary axisymmetric source described by the Kerr solution is (Doran 1994)

$$\mathcal{R}(B) = \mathcal{W}(B) = -\frac{M}{2(r - \mathrm{i}L\cos\theta)^3}(B + 3\sigma_r B\sigma_r). \tag{5.26}$$

This differs from the radially symmetric case (5.25) only in that its eigenvalues contain an imaginary term governed by the angular momentum L. Verifying that (5.25) and (5.26) are tractionless is simple, requiring only the result that

$$\partial_a Ba\wedge b = \partial_a B(ab - a\cdot b) = -bB, \tag{5.27}$$

which employs equation (2.45) from § 2 *b*.

The fact that the Riemann tensor for the Kerr solution is obtained from that for the Schwarzschild solution by replacing r by $r - \mathrm{i}L\cos\theta$ is reminiscent of a 'trick'

used to derive the Kerr solution in the null tetrad formalism (Newman & Janis 1965). This is particularly suggestive given that the unit imaginary employed in the Newman–Penrose formalism is a disguised version of the spacetime pseudoscalar i. The significance of these observations is discussed further in Doran *et al.* (1998*a*).

For tensors of Petrov type other than I, null bivectors play a significant role. Type II tensors have eigenvalues α_1, $-\alpha_1$ and 0 and two independent eigenbivectors, one timelike and one null. Type III and type N have all three eigenvalues zero and satisfy

$$\text{type III:} \quad \mathcal{W}^3(B) = 0, \quad \mathcal{W}^2(B) \neq 0, \tag{5.28}$$

$$\text{type N:} \quad \mathcal{W}^2(B) = 0. \tag{5.29}$$

An example of a type N tensor is provided by gravitational radiation. $\mathcal{R}(B)$ for a plane-polarized gravitational wave travelling in the γ_3 direction is

$$\mathcal{R}^+(B) = \mathcal{W}^+(B) = \tfrac{1}{4} f(t-z)\,\gamma_+(\gamma_1 B\gamma_1 - \gamma_2 B\gamma_2)\gamma_+ \tag{5.30}$$

for waves polarized in the direction of the γ_1 and γ_2 axes, and

$$\mathcal{R}^\times(B) = \mathcal{W}^\times(B) = \tfrac{1}{4} f(t-z)\,\gamma_+(\gamma_1 B\gamma_2 + \gamma_2 B\gamma_1)\gamma_+ \tag{5.31}$$

for waves polarized at 45° to the axes. In both cases, $f(t-z)$ is a scalar function and γ_+ is the null vector

$$\gamma_+ \equiv \gamma_0 + \gamma_3. \tag{5.32}$$

The direct appearance of the null vector γ_+ in $\mathcal{W}(B)$ shows that $\mathcal{W}^2(B) = 0$ and is physically very suggestive. Expressions of the type $\gamma_+ B\gamma_+$ project the bivector B down the null vector γ_+ and such a structure is exhibited in the radiation field generated by an accelerating point charge (Gull *et al.* 1993*a*).

These examples illustrate the uniquely compact forms for the Riemann tensor afforded by geometric algebra. In terms of both clarity and physical insight these expressions are far superior to any afforded by tensor algebra, the Newman–Penrose formalism or differential forms. Only Wahlquist and Estabrook's (3+1) dyadic notation (Estabrook & Wahlquist 1965; Wahlquist 1992) achieves expressions of comparable compactness, although their formalism is of limited applicability.

(*b*) *The Bianchi identities*

Further information from the wedge equation (4.38) is contained in the Bianchi identity. One form of this follows from a simple application of the Jacobi identity:

$$[\mathcal{D}_a, [\mathcal{D}_b, \mathcal{D}_c]]A + \text{cyclic permutations} = 0, \tag{5.33}$$

$$\Rightarrow \quad \mathcal{D}_a R(b\wedge c) + \text{cyclic permutations} = 0. \tag{5.34}$$

Again, use of the $\mathcal{D}_a$ derivatives makes this identity straightforward, but more work is required to achieve a fully covariant relation.

We start by forming the adjoint relation to (5.34), which is

$$\partial_a\wedge\partial_b\wedge\partial_c\langle[a\cdot\nabla R(b\wedge c) + \Omega(a)\times R(b\wedge c)]B\rangle = 0, \tag{5.35}$$

where B is a constant bivector. We next need to establish the result that

$$B_1\cdot\mathcal{R}(B_2) = B_2\cdot\mathcal{R}(B_1). \tag{5.36}$$

This follows by contracting equation (5.6) with an arbitrary vector and an arbitrary bivector to obtain the equations

$$\partial_c\wedge[a\cdot\mathcal{R}(c\wedge b)] = \mathcal{R}(a\wedge b), \tag{5.37}$$

$$(B\cdot\partial_a)\cdot\mathcal{R}(a\wedge b) = -\partial_a B\cdot\mathcal{R}(a\wedge b). \tag{5.38}$$

Protracting the second of these equations with ∂_b and using the first, we obtain

$$\partial_b \wedge [(B \cdot \partial_a) \cdot \mathcal{R}(a \wedge b)] = -2\mathcal{R}(B) = -\partial_b \wedge \partial_a B \cdot \mathcal{R}(a \wedge b). \tag{5.39}$$

Taking the scalar product with a second bivector now gives equation (5.36).

Using this result in equation (5.35), we now obtain

$$\nabla \wedge \overline{h}^{-1}[\mathcal{R}(B)] - \partial_a \wedge \overline{h}^{-1}[\mathcal{R}(\Omega(a) \times B)] = 0. \tag{5.40}$$

Finally, acting on this equation with $\overline{h}$ and using equation (4.38), we establish the covariant result

$$\mathcal{D} \wedge \mathcal{R}(B) - \partial_a \wedge \mathcal{R}(\omega(a) \times B) = 0. \tag{5.41}$$

This result takes a more natural form when B becomes an arbitrary function of position and we write the Bianchi identity as

$$\partial_a \wedge [a \cdot \mathcal{D}\mathcal{R}(B) - \mathcal{R}(a \cdot \mathcal{D}B)] = 0. \tag{5.42}$$

We can extend the overdot notation of § 2 *b* in the obvious manner to write equation (5.42) as

$$\dot{\mathcal{D}} \wedge \dot{\mathcal{R}}(B) = 0, \tag{5.43}$$

which is very compact, but somewhat symbolic and hard to apply without unwrapping into the form of equation (5.42).

The self-duality of the Weyl tensor implies that

$$\dot{\mathcal{D}} \wedge \dot{\mathcal{W}}(\mathrm{i}B) = -\mathrm{i}\dot{\mathcal{D}} \cdot \dot{\mathcal{W}}(B), \tag{5.44}$$

so, in situations where the matter vanishes and $\mathcal{W}(B)$ is the only contribution to $\mathcal{R}(B)$, the Bianchi identities reduce to

$$\dot{\mathcal{D}}\dot{\mathcal{W}}(B) = 0. \tag{5.45}$$

The properties of a first-order equation such as this are discussed in more detail in § 7 *a*.

The contracted Bianchi identities are obtained from

$$(\partial_a \wedge \partial_b) \cdot [\dot{\mathcal{D}} \wedge \dot{\mathcal{R}}(a \wedge b)] = \partial_a \cdot [\dot{\mathcal{R}}(a \wedge \dot{\mathcal{D}}) + \dot{\mathcal{D}}\dot{\mathcal{R}}(a)] = 2\dot{\mathcal{R}}(\dot{\mathcal{D}}) - \mathcal{D}\mathcal{R}, \tag{5.46}$$

from which we can write

$$\dot{\mathcal{G}}(\dot{\mathcal{D}}) = 0. \tag{5.47}$$

An alternative form of this equation is obtained by taking the scalar product with an arbitrary vector and using the symmetry of $\mathcal{G}(a)$ to write

$$\dot{\mathcal{D}} \cdot \dot{\mathcal{G}}(a) = 0. \tag{5.48}$$

Written out in full, this equation takes the form

$$\partial_a \cdot [L_a \mathcal{G}(b) - \mathcal{G}(L_a b) + \omega(a) \times \mathcal{G}(b) - \mathcal{G}(\omega(a) \times b)] = 0. \tag{5.49}$$

(*c*) *Symmetries and conservation laws*

We end this section with some comments on symmetries and conservation laws. These comments are not all directly relevant to the applications discussed in this paper, but concern the general structure of GTG.

The first significant point is that the theory is founded on an action principle in a 'flat' vector space. It follows that all the familiar equations relating symmetries of the action to conserved quantities hold without modification. (A geometric algebra approach to Lagrangian field theory has already been developed in Lasenby

et al. (1993*b*).) Any symmetry transformation of the total action integral which is parametrized by a continuous scalar will result in a vector which is conserved with respect to the vector derivative ∇. To every such vector there corresponds a covariant equivalent, as is seen from the simple rearrangement

$$\mathcal{D}\cdot\mathcal{J} = \mathrm{i}\mathcal{D}\wedge(\mathrm{i}\mathcal{J}) = \det(\underline{h})\mathrm{i}\nabla\wedge[\overline{h}^{-1}(\mathrm{i}\mathcal{J})] = \det(\underline{h})\nabla\cdot[\underline{h}(\mathcal{J})\det(\underline{h})^{-1}]. \tag{5.50}$$

Thus, if J satisfies $\nabla\cdot J = 0$, then the covariant equivalent

$$\mathcal{J} = \underline{h}^{-1}(J)\det(\underline{h}) \tag{5.51}$$

satisfies the covariant equation $\mathcal{D}\cdot\mathcal{J} = 0$. (This explains the definition of $\mathcal{J}$ in § 3 *b*.) Note that if we attempt to form the canonical stress-energy tensor conjugate to translations we obtain the quantity $\mathcal{G}(a) - \kappa\mathcal{T}(a)$, which the field equations set to zero. The overall stress-energy tensor is therefore clearly conserved, but this does not yield any new information.

A second feature of our use of the STA is that all differential equations can be recast in integral form. The integral equation form is not always useful, since it often forces one to deal with non-covariant quantities. But integral equations are particularly well suited to handling singularities in the gravitational fields. Just as Gauss's theorem in electromagnetism can be used to determine the structure of an electric field source, so integral equations can be used to uncover the structure of the matter sources of gravitational fields. An example of this is provided in § 6 *d* where the Schwarzschild solution is shown to arise from a matter stress-energy tensor containing a single δ function source of strength M. A less obvious example is contained in Doran *et al.* (1998*a*), where it is shown that the matter generating the Kerr solution takes the form of a ring rotating at the speed of light, supported by a disk of tension. Such notions are quite different from classical general relativity.

Killing vectors have played a significant role in the analysis of symmetries and conserved quantities in general relativity, and their properties are largely unchanged in GTG. The simplest covariant form of Killing's equation for a Killing vector K is that

$$a\cdot(b\cdot\mathcal{D}K) + b\cdot(a\cdot\mathcal{D}K) = 0, \tag{5.52}$$

for any two vector fields a and b. Contracting with $\partial_a\cdot\partial_b$ immediately yields the result that K is divergenceless:

$$\mathcal{D}\cdot K = 0. \tag{5.53}$$

Killing vectors are frequently obtained when, in some coordinate system, the metric is independent of one of the coordinates. This works as follows. Suppose we introduce a set of four scalar functions $\{x^\mu(x)\}$. These determine a coordinate frame

$$e_\mu \equiv \frac{\partial}{\partial x^\mu}x, \tag{5.54}$$

with dual frame

$$e^\mu \equiv \nabla x^\mu, \quad e_\mu\cdot e^\nu = \delta_\mu^\nu. \tag{5.55}$$

From the coordinate frame $\{e_\mu\}$ one can construct a frame of covariant vectors $\{g_\mu\}$, with reciprocal vectors $\{g^\mu\}$, by

$$g_\mu \equiv \underline{h}^{-1}(e_\mu), \quad g^\mu \equiv \overline{h}(e^\mu). \tag{5.56}$$

From these the metric is defined by (see Appendix C)

$$g_{\mu\nu} \equiv g_\mu\cdot g_\nu. \tag{5.57}$$

Suppose now that the $g_{\mu\nu}$ are all independent of the x^0 coordinate. It follows that

$$\frac{\partial}{\partial x^0} g_{\mu\nu} = g_\mu \cdot (g_0 \cdot \mathcal{D} g_\nu) + g_\nu \cdot (g_0 \cdot \mathcal{D} g_\mu) = 0. \tag{5.58}$$

However, for a coordinate frame,

$$g_\mu \cdot \mathcal{D} g_\nu - g_\nu \cdot \mathcal{D} g_\mu = \underline{h}^{-1}(\partial_\mu e_\nu - \partial_\nu e_\mu) = 0 \tag{5.59}$$

and using this in equation (5.58) we find that

$$g_\mu \cdot (g_\nu \cdot \mathcal{D} K) + g_\nu \cdot (g_\mu \cdot \mathcal{D} K) = 0, \tag{5.60}$$

where $K = g_0$. Equation (5.60) is entirely equivalent to the frame-free equation (5.52).

A further consequence of equation (5.52) is that, for any vector a,

$$a \cdot (a \cdot \mathcal{D} K) = 0. \tag{5.61}$$

So, for a particle satisfying $v \cdot \mathcal{D} v = 0$ (4.61), we see that

$$\partial_\tau (v \cdot K) = v \cdot \mathcal{D}(v \cdot K) = K \cdot (v \cdot \mathcal{D} v) + v \cdot (v \cdot \mathcal{D} K) = 0. \tag{5.62}$$

It follows that the quantity $v \cdot K$ is conserved along the worldline of a freely falling particle.

Part II. Applications

6. Spherically symmetric systems

Our first full application of the formalism we have developed is to time-dependent spherically symmetric fields. For simplicity, we consider only the case where the matter is described by a perfect fluid. The equations derived here are applicable to static and collapsing stars, radially symmetric black holes and many aspects of cosmology, including inflation. Furthermore, in a suitable gauge the relevant equations are essentially Newtonian in form, making their physical interpretation quite transparent. Applications discussed here include an analytic solution to the equations governing collapsing dust and the new understanding of horizons forced by our gauge theory. This section includes an extended version of the work presented in Lasenby *et al.* (1995).

(*a*) *The 'intrinsic' method*

The traditional approach to solving the gravitational field equations in GR is to start with the metric $g_{\mu\nu}$, which is usually encoded as a line element

$$\mathrm{d}s^2 = g_{\mu\nu}\, \mathrm{d}x^\mu\, \mathrm{d}x^\nu. \tag{6.1}$$

The analogous quantity in GTG is derived from the $\overline{h}$ function via

$$g_{\mu\nu} = \underline{h}^{-1}(e_\mu) \cdot \underline{h}^{-1}(e_\nu), \tag{6.2}$$

where the $\{e_\mu\}$ comprise a coordinate frame (see also Appendix C). For a given matter stress-energy tensor, the field equations then yield a set of nonlinear, second-order differential equations for the terms in $g_{\mu\nu}$. These equations are notoriously hard to solve. On the other hand, *any* metric is potentially a solution to the Einstein equations—one where the matter stress-energy tensor is determined by the corresponding Einstein tensor. This approach, in which the tail wags the dog, has recently probably been more popular! Here we present a new approach to solving

the gravitational field equations. The method is closely tied to our gauge-theoretic understanding of gravity, but can always be used to generate a metric which solves the Einstein equations. So, even if one rejects our gauge-theory description of gravity, the techniques developed below can still be viewed as providing a new method for studying the Einstein equations.

Under a local rotation the vector $\underline{h}^{-1}(a)$ transforms as

$$\underline{h}^{-1}(a) \mapsto R\underline{h}^{-1}(a)\tilde{R}. \tag{6.3}$$

It follows that the metric $g_{\mu\nu}$ (6.2) is invariant under rotation-gauge transformations. This is in keeping with our earlier observation that, at the classical level, it is possible to work with a set of equations that are invariant under rotation-gauge transformations, and this is precisely what GR does. This approach has the advantage of removing a number of degrees of freedom from the theory, but one pays a heavy price for this: the equations become second order and one has to deal with complicated nonlinear terms. The approach we develop here is quite different. We keep the rotation-gauge field explicit and work entirely with quantities that transform covariantly under position-gauge transformations. Such quantities include $\overline{h}(\nabla)$, $\omega(a)$ and $\mathcal{R}(B)$. We therefore work with directional derivatives of the form $L_a = a\cdot\overline{h}(\nabla)$ and treat $\omega(a)$ as an arbitrary field. The relationship between $\overline{h}$ and $\omega(a)$ is then encoded in the commutation relations of the L_a. This set-up is achieved by initially making a suitably general ansatz for the $\overline{h}$ function. This trial form is then substituted into equation (4.51) to find the general form of $\omega(a)$. An arbitrary $\omega(a)$ field consistent with this general form is then introduced, resulting in a set of equations relating commutators of the L_a derivatives to the variables in $\omega(a)$.

Next, the Riemann tensor $\mathcal{R}(B)$ is constructed in terms of abstract first-order derivatives of the $\omega(a)$ and additional quadratic terms. The rotation-gauge freedom is then removed by specifying the precise form that $\mathcal{R}(B)$ takes. For example, one can arrange that $\mathcal{W}(B)$ is diagonal in a suitable frame. This gauge fixing is crucial in order to arrive at a set of equations that are not under-constrained. With $\mathcal{R}(B)$ suitably fixed, one arrives at a set of relations between first-order abstract derivatives of the $\omega(a)$, quadratic terms in $\omega(a)$ and matter terms. The final step is to impose the Bianchi identities, which ensure overall consistency of the equations with the bracket structure. Once all this is achieved, one arrives at a fully 'intrinsic' set of equations. Solving these equations usually involves searching for natural 'integrating factors'. These integrating factors provide 'intrinsic' coordinates and many of the fields can be expressed as functions of these coordinates alone. The final step is to 'coordinatize' the solution by making an explicit (gauge) choice of the $\overline{h}$ function. The natural way to do this is to ensure that the coordinates used in parametrizing $\overline{h}(a)$ match the intrinsic coordinates defined by the integrating factors.

The method outlined above is quite general and can be applied to a wide range of problems. Here we employ it in the analysis of time-dependent spherically symmetric systems.

(*b*) *The intrinsic field equations*

We start by introducing a set of spherical polar coordinates. In terms of the fixed $\{\gamma_\mu\}$ frame we define

$$t \equiv x\cdot\gamma_0, \quad \cos\theta \equiv \frac{x\cdot\gamma^3}{r}, \quad r \equiv \sqrt{(x\wedge\gamma_0)^2}, \quad \tan\phi \equiv \frac{x\cdot\gamma^2}{x\cdot\gamma^1}. \tag{6.4}$$

The associated coordinate frame is

$$\left.\begin{aligned} e_t &\equiv \gamma_0, \quad e_r \equiv x\wedge\gamma_0\gamma_0/r, \\ e_\theta &\equiv r\cos\theta(\cos\phi\,\gamma_1 + \sin\phi\gamma_2) - r\sin\theta\gamma_3, \\ e_\phi &\equiv r\sin\theta(-\sin\phi\,\gamma_1 + \cos\phi\,\gamma_2) \end{aligned}\right\} \tag{6.5}$$

and the dual-frame vectors are denoted by $\{e^t, e^r, e^\theta, e^\phi\}$. We will also frequently employ the unit vectors $\hat{\theta}$ and $\hat{\phi}$ defined by

$$\hat{\theta} \equiv \frac{e_\theta}{r}, \quad \hat{\phi} \equiv \frac{e_\phi}{r\sin\theta}. \tag{6.6}$$

Associated with these unit vectors are the unit timelike bivectors

$$\sigma_r \equiv e_r e_t, \quad \sigma_\theta \equiv \hat{\theta} e_t, \quad \sigma_\phi \equiv \hat{\phi} e_t, \tag{6.7}$$

which satisfy

$$\sigma_r\sigma_\theta\sigma_\phi = e_t e_r \hat{\theta}\hat{\phi} = \mathrm{i}. \tag{6.8}$$

The dual spatial bivectors are given by

$$\mathrm{i}\sigma_r = -\hat{\theta}\hat{\phi}, \quad \mathrm{i}\sigma_\theta = e_r\hat{\phi}, \quad \mathrm{i}\sigma_\phi = -e_r\hat{\theta}. \tag{6.9}$$

Throughout we use the abbreviations

$$\partial_r = \partial/\partial r, \quad \partial_t = \partial/\partial t. \tag{6.10}$$

(i) *The $\overline{h}$ function*

Our first step towards a solution is to decide on a general form of the $\overline{h}$ function that is consistent with spherical symmetry. Suppose that B is a constant spatial bivector ($e_t\cdot B = 0$) and define

$$R = \mathrm{e}^{B/2}, \tag{6.11}$$

$$x' = \tilde{R}xR. \tag{6.12}$$

Then, in analogy with electromagnetism, the gravitational fields will be spherically symmetric if rotating $\overline{h}(a)$ to $R\overline{h}(a)\tilde{R}$ and displacing it to the back-rotated position x' leaves $\overline{h}(a)$ unchanged. Hence rotational symmetry is enforced through the requirement that

$$R\overline{h}_{x'}(\tilde{R}aR)\tilde{R} = \overline{h}(a). \tag{6.13}$$

This symmetry immediately implies that the $\{e^r, e^t\}$ and $\{e^\theta, e^\phi\}$ pairs decouple from each other and the action of $\overline{h}(a)$ on the $\hat{\theta}$ and $\hat{\phi}$ vectors is further restricted to the form

$$\overline{h}(\hat{\theta}) = \alpha\hat{\theta} + \beta\hat{\phi}, \quad \overline{h}(\hat{\phi}) = \alpha\hat{\phi} - \beta\hat{\theta}. \tag{6.14}$$

However, the skew-symmetric term parametrized by β can always be removed by a rotation in the $\mathrm{i}\sigma_r$ plane, so we can assume that $\overline{h}(a)$ is diagonal on $\{e^\theta, e^\phi\}$. No such assumption can be made for the $\{e^r, e^t\}$ vectors, so we take $\overline{h}(a)$ as having the general form

$$\overline{h}(e^t) = f_1 e^t + f_2 e^r, \quad \overline{h}(e^r) = g_1 e^r + g_2 e^t, \quad \overline{h}(e^\theta) = \alpha e^\theta, \quad \overline{h}(e^\phi) = \alpha e^\phi, \tag{6.15}$$

where f_1, f_2, g_1, g_2 and α are all functions of t and r only. We retain the gauge freedom to perform a boost in the σ_r direction and this freedom is employed later to simplify the equations. Our remaining position-gauge freedom lies in the freedom to reparametrize t and r, which does not affect the general form of (6.15). A natural parametrization will emerge once the 'intrinsic' variables have been identified.

Table 3. *Covariant derivatives of the polar-frame unit timelike bivectors*

	σ_r	σ_θ	σ_ϕ
$e_t\cdot\mathcal{D}$	0	$G\mathrm{i}\sigma_\phi$	$-G\mathrm{i}\sigma_\theta$
$e_r\cdot\mathcal{D}$	0	$F\mathrm{i}\sigma_\phi$	$-F\mathrm{i}\sigma_\theta$
$\hat{\theta}\cdot\mathcal{D}$	$T\sigma_\theta - S\mathrm{i}\sigma_\phi$	$-T\sigma_r$	$S\,\mathrm{i}\sigma_r$
$\hat{\phi}\cdot\mathcal{D}$	$T\sigma_\phi + S\mathrm{i}\sigma_\theta$	$-S\mathrm{i}\sigma_r$	$-T\sigma_r$

(ii) *The $\omega(a)$ function*

To find a general form $\omega(a)$ consistent with (6.15) we substitute (6.15) into equation (4.51) for $\omega(a)$ as a function of $\overline{h}(a)$. Where the coefficients contain derivatives of terms from $\overline{h}(a)$, new symbols are introduced. Undifferentiated terms from $\overline{h}(a)$ appearing in $\omega(a)$ are left in explicitly. These arise from frame derivatives and the algebra is usually simpler if they are included. This procedure results in the following form for $\omega(a)$:

$$\left.\begin{aligned} \omega(e_t) = Ge_re_t, \quad \omega(e_r) = Fe_re_t, \\ \omega(\hat{\theta}) = S\hat{\theta}e_t + (T-\alpha/r)e_r\hat{\theta}, \quad \omega(\hat{\phi}) = S\hat{\phi}e_t + (T-\alpha/r)e_r\hat{\phi}, \end{aligned}\right\} \tag{6.16}$$

where G, F, S and T are functions of t and r only. The important feature of these functions is that they are position-gauge covariant.

Substituting this definition for $\omega(a)$ into equations (4.43) and (4.44) we find that the bracket relations are as follows:

$$\left.\begin{aligned} &[L_t, L_r] = GL_t - FL_r, && [L_r, L_{\hat{\theta}}] = -TL_{\hat{\theta}}, \\ &[L_t, L_{\hat{\theta}}] = -SL_{\hat{\theta}}, && [L_r, L_{\hat{\phi}}] = -TL_{\hat{\phi}}, \\ &[L_t, L_{\hat{\phi}}] = -SL_{\hat{\phi}}, && [L_\theta, L_\phi] = 0, \end{aligned}\right\} \tag{6.17}$$

where

$$L_t \equiv e_t\cdot\overline{h}(\nabla), \quad L_{\hat{\theta}} \equiv \hat{\theta}\cdot\overline{h}(\nabla), \quad L_r \equiv e_r\cdot\overline{h}(\nabla), \quad L_{\hat{\phi}} \equiv \hat{\phi}\cdot\overline{h}(\nabla). \tag{6.18}$$

The use of unit vectors in these derivatives eliminates the need to calculate irrelevant coordinate derivatives. A set of bracket relations such as (6.17) is precisely what one aims to achieve—all reference to the $\overline{h}$ function is removed and one deals entirely with position-gauge covariant quantities.

(iii) *The Riemann tensor*

Having found a suitable form for $\omega(a)$ we next use equation (4.46) to calculate $\mathcal{R}(B)$. This derivation is simplified by judicious use of the results listed in table 3. The only subtlety in the derivation is the removal of terms involving derivatives of α/r using the bracket relations (6.17). Since $\alpha/r = L_{\hat{\theta}}\theta$, we have

$$L_t(\alpha/r) = L_tL_{\hat{\theta}}\theta = [L_t, L_{\hat{\theta}}]\theta = -S\alpha/r \tag{6.19}$$

and

$$L_r(\alpha/r) = L_rL_{\hat{\theta}}\theta = [L_r, L_{\hat{\theta}}]\theta = -T\alpha/r. \tag{6.20}$$

Application of equation (4.46) is now straightforward and leads to the Riemann tensor

$$\left.\begin{aligned}
\mathcal{R}(\sigma_r) &= (L_rG - L_tF + G^2 - F^2)\sigma_r,\\
\mathcal{R}(\sigma_\theta) &= (-L_tS + GT - S^2)\sigma_\theta + (L_tT + ST - SG)\mathrm{i}\sigma_\phi,\\
\mathcal{R}(\sigma_\phi) &= (-L_tS + GT - S^2)\sigma_\phi - (L_tT + ST - SG)\mathrm{i}\sigma_\theta,\\
\mathcal{R}(\mathrm{i}\sigma_\phi) &= (L_rT + T^2 - FS)\mathrm{i}\sigma_\phi - (L_rS + ST - FT)\sigma_\theta,\\
\mathcal{R}(\mathrm{i}\sigma_\theta) &= (L_rT + T^2 - FS)\mathrm{i}\sigma_\theta + (L_rS + ST - FT)\sigma_\phi,\\
\mathcal{R}(\mathrm{i}\sigma_r) &= (-S^2 + T^2 - (\alpha/r)^2)\mathrm{i}\sigma_r.
\end{aligned}\right\} \tag{6.21}$$

(iv) *The matter field and gauge fixing*

Now that we have found $\mathcal{R}(B)$ in terms of 'intrinsic' functions and their first derivatives, we must next decide on the form of matter stress-energy tensor that the gravitational fields couple to. We assume that the matter is modelled by an ideal fluid so we can write

$$\mathcal{T}(a) = (\rho + p)a\cdot vv - pa, \tag{6.22}$$

where ρ is the energy density, p is the pressure and v is the covariant fluid velocity ($v^2 = 1$). Radial symmetry means that v can only lie in the e_t and e_r directions, so v must take the form

$$v = \cosh u e_t + \sinh u e_r. \tag{6.23}$$

But, in restricting the $\overline{h}$ function to the form of equation (6.15), we retained the gauge freedom to perform arbitrary radial boosts. This freedom can now be employed to set $v = e_t$, so that the matter stress-energy tensor becomes

$$\mathcal{T}(a) = (\rho + p)a\cdot e_te_t - pa. \tag{6.24}$$

There is no physical content in the choice $v = e_t$ as all physical relations must be independent of gauge choices. In particular, setting $v = e_t$ does not mean that the fluid is 'at rest', or that we are 'comoving with the fluid'. An observer comoving with the fluid will have a covariant velocity e_t, but this implies no special relationship with the time coordinate t, since the observer's trajectory would have $\dot{x} = \underline{h}(e_t)$ and nothing has yet been said about the specific form of $\underline{h}(a)$.

In setting $v = e_t$, all rotation-gauge freedom has finally been removed. This is an essential step since all non-physical degrees of freedom must be removed before one can achieve a complete set of physical equations. Note that the rotation gauge has been fixed by imposing a suitable form for $\mathcal{R}(B)$, rather than restricting the form of $\overline{h}(a)$. The reason for working in this manner is obvious—$\mathcal{R}(B)$ deals directly with physically measurable quantities, whereas the algebraic structure of the $\overline{h}$ function is of little direct physical relevance.

From equation (5.9) the source term in $\mathcal{R}(B)$ is given by

$$\mathcal{R}(a\wedge b) - \mathcal{W}(a\wedge b) = 4\pi[a\wedge\mathcal{T}(b) + \mathcal{T}(a)\wedge b - \tfrac{2}{3}\mathcal{T}a\wedge b], \tag{6.25}$$

where $\mathcal{T} = \partial_a\cdot\mathcal{T}(a)$ is the trace of the matter stress-energy tensor. With $\mathcal{T}(a)$ given by equation (6.24), $\mathcal{R}(B)$ is restricted to the form

$$\mathcal{R}(B) = \mathcal{W}(B) + 4\pi[(\rho + p)B\cdot e_te_t - \tfrac{2}{3}\rho B]. \tag{6.26}$$

Comparing this with equation (6.21) we find that $\mathcal{W}(B)$ has the general form

$$\left.\begin{aligned} \mathcal{W}(\sigma_r) &= \alpha_1\sigma_r, & \mathcal{W}(\mathrm{i}\sigma_r) &= \alpha_4\mathrm{i}\sigma_r,\\ \mathcal{W}(\sigma_\theta) &= \alpha_2\sigma_\theta + \beta_1\mathrm{i}\sigma_\phi, & \mathcal{W}(\mathrm{i}\sigma_\theta) &= \alpha_3\mathrm{i}\sigma_\theta + \beta_2\sigma_\phi,\\ \mathcal{W}(\sigma_\phi) &= \alpha_2\sigma_\phi - \beta_1\mathrm{i}\sigma_\theta, & \mathcal{W}(\mathrm{i}\sigma_\phi) &= \alpha_3\mathrm{i}\sigma_\phi - \beta_2\sigma_\theta. \end{aligned}\right\} \tag{6.27}$$

But $\mathcal{W}(B)$ must be self-dual, so $\alpha_1 = \alpha_4$, $\alpha_2 = \alpha_3$ and $\beta_1 = -\beta_2$, and symmetric, which implies that $\beta_1 = \beta_2$. It follows that $\beta_1 = \beta_2 = 0$. Finally, $\mathcal{W}(B)$ must be traceless, which requires that $\alpha_1 + 2\alpha_2 = 0$. Taken together, these conditions reduce $\mathcal{W}(B)$ to the form

$$\mathcal{W}(B) = \tfrac{1}{4}\alpha_1(B + 3\sigma_r B\sigma_r), \tag{6.28}$$

which is of Petrov type D. It follows from the form of $\mathcal{R}(\mathrm{i}\sigma_r)$ that if we set

$$A \equiv \tfrac{1}{4}(-S^2 + T^2 - (\alpha/r)^2), \tag{6.29}$$

then the full Riemann tensor must take the form

$$\mathcal{R}(B) = (A + \tfrac{2}{3}\pi\rho)(B + 3\sigma_r B\sigma_r) + 4\pi[(\rho + p)B\cdot e_t e_t - \tfrac{2}{3}\rho B]. \tag{6.30}$$

Comparing this with equation (6.21) yields the following set of equations:

$$L_t S = 2A + GT - S^2 - 4\pi p, \tag{6.31}$$

$$L_t T = S(G - T), \tag{6.32}$$

$$L_r S = T(F - S), \tag{6.33}$$

$$L_r T = -2A + FS - T^2 - 4\pi\rho, \tag{6.34}$$

$$L_r G - L_t F = F^2 - G^2 + 4A + 4\pi(\rho + p). \tag{6.35}$$

(v) *The Bianchi identity*

We are now close to our goal of a complete set of intrinsic equations. The remaining step is to enforce the Bianchi identities. The contracted Bianchi identity (5.47) for a perfect fluid results in the pair of equations

$$\mathcal{D}\cdot(\rho v) + p\mathcal{D}\cdot v = 0, \tag{6.36}$$

$$(\rho + p)(v\cdot\mathcal{D}v)\wedge v - (\mathcal{D}p)\wedge v = 0. \tag{6.37}$$

Since $(v\cdot\mathcal{D}v)\wedge v$ is the acceleration bivector, the second of these equations relates the acceleration to the pressure gradient. For the case of radially symmetric fields, equations (6.36) and (6.37) reduce to

$$L_t\rho = -(F + 2S)(\rho + p), \tag{6.38}$$

$$L_r p = -G(\rho + p), \tag{6.39}$$

the latter of which identifies G as the radial acceleration. The full Bianchi identities now turn out to be satisfied as a consequence of the contracted identities and the bracket relation

$$[L_t, L_r] = GL_t - FL_r. \tag{6.40}$$

Equations (6.19), (6.20), (6.31)–(6.35), the contracted identities (6.38) and (6.39) and the bracket condition (6.40) now form the complete set of intrinsic equations. The structure is closed, in that it is easily verified that the bracket relation (6.40) is consistent with the known derivatives. The derivation of such a set of equations is the basic aim of our 'intrinsic method'. The equations deal solely with objects that transform covariantly under displacements and many of these quantities have direct physical significance.

(vi) *Integrating factors*

To simplify our equations we start by forming the derivatives of A. From equations (6.19), (6.20) and (6.31)–(6.35) it follows that

$$L_t A + 3SA = 2\pi Sp, \tag{6.41}$$

$$L_r A + 3TA = -2\pi T\rho. \tag{6.42}$$

These results, and equations (6.32) and (6.33), suggest that we should look for an integrating factor for the $L_t + S$ and $L_r + T$ operators. Such a function, X say, should have the properties that

$$L_t X = SX, \quad L_r X = TX. \tag{6.43}$$

A function with these properties can exist only if the derivatives are consistent with the bracket relation (6.40). This is checked by forming

$$[L_t, L_r]X = L_t(TX) - L_r(SX) = X(L_t T - L_r S) = X(SG - FT) = GL_t X - FL_r X, \tag{6.44}$$

which confirms that the properties of X are consistent with (6.40). Establishing the existence of integrating factors in this manner is a key step in our method, because the integrating factors play the role of intrinsically defined coordinates. If the $\overline{h}$ function is parametrized directly in terms of these functions, the physical status of the quantities in it becomes clearer. In the present case, equations (6.19) and (6.20) show that r/α already has the properties required of X, so it is r/α which emerges as the intrinsic distance scale. It is therefore sensible that the position-gauge freedom in the choice of r should be absorbed by setting $\alpha = 1$. This then sets the intrinsic distance scale equal to r, lifting r from the status of an arbitrary coordinate to that of a physically measurable quantity.

Having fixed the radial scale with the position-gauge choice

$$r = X, \quad \alpha = 1, \tag{6.45}$$

we can make some further simplifications. From the form of $\overline{h}(a)$ (6.15) and equations (6.43) and (6.45) we see that

$$g_1 = L_r r = Tr, \tag{6.46}$$

$$g_2 = L_t r = Sr, \tag{6.47}$$

which gives two of the functions in $\overline{h}(a)$. We also define

$$M \equiv -2r^3 A = \tfrac{1}{2} r({g_2}^2 - {g_1}^2 + 1), \tag{6.48}$$

which satisfies

$$L_t M = -4\pi r^2 g_2 p \tag{6.49}$$

and

$$L_r M = 4\pi r^2 g_1 \rho. \tag{6.50}$$

The latter shows that M plays the role of an intrinsic mass.

(vii) *The 'Newtonian' gauge*

So far, a natural distance scale has been identified, but no natural time coordinate has emerged. To complete the solution it is necessary to make a choice for the t coordinate, so we now look for additional criteria to motivate this choice. We are currently free to perform an arbitrary r- and t-dependent displacement along the e_t direction. This gives us complete freedom in the choice of f_2 function. An indication

of how this choice should be made is obtained from equations (6.49) and (6.50) for the derivatives of M (6.49), which invert to yield

$$\frac{\partial M}{\partial t} = \frac{-4\pi g_1 g_2 r^2(\rho + p)}{f_1 g_1 - f_2 g_2}, \tag{6.51}$$

$$\frac{\partial M}{\partial r} = \frac{4\pi r^2(f_1 g_1 \rho + f_2 g_2 p)}{f_1 g_1 - f_2 g_2}. \tag{6.52}$$

The second equation reduces to a simple classical relation if we choose $f_2 = 0$, as we then obtain

$$\partial_r M = 4\pi r^2 \rho, \tag{6.53}$$

which says that $M(r, t)$ is determined by the amount of mass-energy in a sphere of radius r. There are other reasons for choosing the time variable such that $f_2 = 0$. For example, we can then use the bracket structure to solve for f_1. With $f_2 = 0$ we have

$$L_t = f_1 \partial_t + g_2 \partial_r, \tag{6.54}$$

$$L_r = g_1 \partial_r \tag{6.55}$$

and the bracket relation (6.40) implies that

$$L_r f_1 = -G f_1 \quad \Rightarrow \quad \partial_r f_1 = -\frac{G}{g_1} f_1 \quad \Rightarrow \quad f_1 = \epsilon(t) \exp\left(-\int^r \frac{G}{g_1} \, \mathrm{d}r \right). \tag{6.56}$$

The function $\epsilon(t)$ can be absorbed by a further t-dependent rescaling along e_t (which does not change f_2), so with $f_2 = 0$ we can reduce to a system in which

$$f_1 = \exp\left(-\int^r \frac{G}{g_1} \, \mathrm{d}r \right). \tag{6.57}$$

Another reason why $f_2 = 0$ is a natural gauge choice is seen when the pressure is zero. In this case, equation (6.39) forces G to be zero and equation (6.57) then forces $f_1 = 1$. A free-falling particle with $v = e_t$ (i.e. comoving with the fluid) then has

$$\dot{t} e_t + \dot{r} e_r = e_t + g_2 e_r, \tag{6.58}$$

where the dots denote differentiation with respect to the proper time. Since $\dot{t} = 1$, the time coordinate t matches the proper time of all observers comoving with the fluid. So, in the absence of pressure, we are able to recover a global 'Newtonian' time on which all observers can agree (provided all clocks are correlated initially). Furthermore, it is also clear from (6.58) that g_2 represents the velocity of the particle. Hence equation (6.51), which reduces to

$$\partial_t M = -4\pi r^2 g_2 \rho \tag{6.59}$$

in the absence of pressure, has a simple Newtonian interpretation—it equates the work with the rate of flow of energy density. Equation (6.48), written in the form

$$\tfrac{1}{2} {g_2}^2 - M/r = \tfrac{1}{2}({g_1}^2 - 1), \tag{6.60}$$

is also now familiar from Newtonian physics—it is a Bernoulli equation for zero pressure and total (non-relativistic) energy $\frac{1}{2}({g_1}^2 - 1)$.

For these reasons we refer to $f_2 = 0$ as defining the 'Newtonian' gauge. The applications discussed in the following sections vindicate our claim that this is the natural gauge for radially symmetric systems. The full set of equations in the Newtonian

Table 4. *Equations governing a radially symmetric perfect fluid*

the $\overline{h}$ function	$\overline{h}(e^t) = f_1 e^t, \quad \overline{h}(e^r) = g_1 e^r + g_2 e^t$ $\overline{h}(e^\theta) = e^\theta, \quad \overline{h}(e^\phi) = e^\phi$
the ω function	$\omega(e_t) = G e_r e_t, \quad \omega(e_r) = F e_r e_t$ $\omega(\hat{\theta}) = g_2/r \hat{\theta} e_t + (g_1 - 1)/r e_r \hat{\theta}$ $\omega(\hat{\phi}) = g_2/r \hat{\phi} e_t + (g_1 - 1)/r e_r \hat{\phi}$
directional derivatives	$L_t = f_1 \partial_t + g_2 \partial_r, \quad L_r = g_1 \partial_r$
equations relating the $\overline{h}$ and ω functions	$L_t g_1 = G g_2, \quad L_r g_2 = F g_1$ $f_1 = \exp\left\{ \int^r -\frac{G}{g_1} \, \mathrm{d}r \right\}$
definition of M	$M \equiv \frac{1}{2} r ({g_2}^2 - {g_1}^2 + 1)$
remaining derivatives	$L_t g_2 = G g_1 - M/r^2 - 4\pi r p$ $L_r g_1 = F g_2 + M/r^2 - 4\pi r \rho$
matter derivatives	$L_t M = -4\pi r^2 g_2 p, \quad L_t \rho = -(2g_2/r + F)(\rho + p)$ $L_r M = 4\pi r^2 g_1 \rho, \quad L_r p = -G(\rho + p)$
Riemann tensor	$\mathcal{R}(B) = 4\pi[(\rho + p) B \cdot e_t e_t - \frac{2}{3}\rho B]$ $-\frac{1}{2}(M/r^3 - \frac{4}{3}\pi\rho)(B + 3\sigma_r B \sigma_r)$
fluid stress-energy tensor	$\mathcal{T}(a) = (\rho + p) a \cdot e_t e_t - p a$

gauge are summarized in table 4. They underlie a wide range of phenomena in relativistic astrophysics and cosmology. The closest GR analogue of the Newtonian gauge description of a spherically symmetric system is provided by Gautreau's 'curvature coordinates' (Gautreau 1984; see also Gautreau & Cohen 1995). This description employs a set of geodesic clocks in radial freefall, comoving with the fluid. However, such a description can only be applied if the pressure is independent of radius, whereas the Newtonian gauge description is quite general.

One aspect of the equations in table 4 is immediately apparent. Given an equation of state $p = p(\rho)$, and initial data in the form of the density $\rho(r, t_0)$ and the velocity $g_2(r, t_0)$, the future evolution of the system is fully determined. This is because ρ determines p and M on a time slice and the definition of M then determines g_1. The equations for $L_r p$, $L_r g_1$ and $L_r g_2$ then determine the remaining information on the time slice. Finally, the $L_t M$ and $L_t g_2$ equations can be used to update the information to the next time slice and the process can then start again. The equations can thus be implemented numerically as a simple set of first-order update equations. This fact considerably simplifies the study of collapsing matter and should be particularly significant in current studies of the critical phenomena associated with horizon and singularity formation (Choptuik 1993; Abrahams & Evans 1993).

(*c*) *Static matter distributions*

As a simple first application we consider a static, radially symmetric matter distribution. In this case ρ and p are functions of r only. Since $M(r, t)$ is now given

by

$$M(r) = \int_0^r 4\pi r'^2 \rho(r')\,\mathrm{d}r', \tag{6.61}$$

it follows that

$$L_t M = 4\pi r^2 g_2 \rho = -4\pi r^2 g_2 p. \tag{6.62}$$

For any physical matter distribution, ρ and p must both be positive, in which case equation (6.62) can be satisfied only if

$$g_2 = 0 \tag{6.63}$$

$$\Rightarrow \quad F = 0. \tag{6.64}$$

Since $g_2 = 0$, we see that g_1 is given simply in terms of $M(r)$ by

$$g_1{}^2 = 1 - 2M(r)/r, \tag{6.65}$$

which recovers contact with the standard line element for a static, radially symmetric field. It is immediately clear that a solution exists only if $2M(r) < r$ for all r. This is equivalent to the condition that a horizon has not formed.

The remaining equation of use is that for $L_t g_2$, which now gives

$$G g_1 = M(r)/r^2 + 4\pi r p. \tag{6.66}$$

Equations (6.65) and (6.66) combine with that for $L_r p$ to give the famous Oppenheimer–Volkov equation

$$\frac{\partial p}{\partial r} = -\frac{(\rho + p)(M(r) + 4\pi r^3 p)}{r(r - 2M(r))}. \tag{6.67}$$

At this point we have successfully recovered all the usual equations governing a non-rotating star and the description is therefore unchanged from that of GR. The work involved in recovering these equations from the full time-dependent case is minimal and the final form of $\bar{h}(a)$ is very simple (it is a diagonal function). Furthermore, the meaning of the t and r coordinates is clear, since they have been defined operationally.

The solution extends straightforwardly to the region outside the star. We now have M constant and

$$f_1 = 1/g_1 = (1 - 2M/r)^{-1/2}, \tag{6.68}$$

which recovers the Schwarzschild line element. It follows that all predictions for the behaviour of matter in the star's gravitational field, including those for the bending of light and the perihelion precession of Mercury, are unchanged from GR.

(*d*) *Point source solutions: black holes*

The next solution of interest is obtained when the matter is concentrated at a single point ($r = 0$). For such a solution, $\rho = p = 0$ everywhere away from the source and the matter equations reduce to

$$\left.\begin{aligned} L_t M &= 0, \\ L_r M &= 0, \end{aligned}\right\} \Rightarrow M = \text{const.} \tag{6.69}$$

Retaining the symbol M for this constant we find that the equations reduce to

$$L_t g_1 = G g_2, \tag{6.70}$$

$$L_r g_2 = F g_1 \tag{6.71}$$

and

$$g_1{}^2 - g_2{}^2 = 1 - 2M/r. \tag{6.72}$$

No further equations yield new information, so we have an under-determined system of equations and some additional gauge fixing is needed to choose an explicit form of $\overline{h}(a)$. The reason for this is that in the vacuum region the Riemann tensor reduces to

$$\mathcal{R}(B) = -\frac{M}{2r^3}(B + 3\sigma_r B \sigma_r). \tag{6.73}$$

This tensor is now invariant under boosts in the σ_r plane, whereas previously the presence of the fluid velocity in the Riemann tensor vector broke this symmetry. The appearance of this new symmetry in the matter-free case manifests itself as a new freedom in the choice of $\overline{h}$ function.

Given this new freedom, we should look for a choice of g_1 and g_2 which simplifies the equations. If we attempt to reproduce the Schwarzschild solution we have to set $g_2 = 0$, but then we immediately run into difficulties with g_1, which is not defined for $r < 2M$. We must therefore look for an alternative gauge choice. We show in the following section that, when $p = 0$, g_1 controls the energy of infalling matter, with particles starting at rest at $r = \infty$ corresponding to $g_1 = 1$. A sensible gauge choice is therefore to set

$$g_1 = 1, \tag{6.74}$$

so that

$$g_2 = -\sqrt{2M/r}, \tag{6.75}$$

$$G = 0, \tag{6.76}$$

$$F = -M/(g_2 r^2) \tag{6.77}$$

and

$$f_1 = 1. \tag{6.78}$$

In this gauge the $\overline{h}$ function takes the remarkably simple form

$$\overline{h}(a) = a - \sqrt{2M/r}\, a \cdot e_r e_t, \tag{6.79}$$

which only differs from the identity through a single term. From the results of § 2 *c*, the extension to the action of $\overline{h}$ on an arbitrary multivector A is straightforward:

$$\overline{h}(A) = A - \sqrt{2M/r}(A \cdot e_r) \wedge e_t. \tag{6.80}$$

It follows that $\det(\underline{h}) = 1$ and the inverse of the adjoint function, as defined by (2.53), is given by

$$\underline{h}^{-1}(A) = A + \sqrt{2M/r}(A \cdot e_t) \wedge e_r. \tag{6.81}$$

(i) *Point-particle trajectories*

To study the properties of the solution (6.79) we consider the equation of motion for infalling matter. For a particle following the trajectory $x(\tau)$, with τ the proper time, we have

$$v = \dot{t} e_t + (\dot{t}\sqrt{2M/r} + \dot{r}) e_r + \dot{\theta} e_\theta + \dot{\phi} e_\phi. \tag{6.82}$$

Since the $\overline{h}$ function is independent of t we have, from equation (4.63),

$$\underline{h}^{-1}(e_t) \cdot v = (1 - 2M/r)\dot{t} - \dot{r}\sqrt{2M/r} = \text{const}. \tag{6.83}$$

and, for particles moving forwards in time ($\dot{t} > 0$ for $r \to \infty$), we can write

$$(1 - 2M/r)\dot{t} = \alpha + \dot{r}\sqrt{2M/r}, \tag{6.84}$$

where the constant α satisfies $\alpha > 0$. The $\dot{r}$ equation is found from the constraint that $v^2 = 1$, which gives

$$\dot{r}^2 = \alpha^2 - (1 - 2M/r)[1 + r^2(\dot{\theta}^2 + \sin^2\theta\,\dot{\phi}^2)]. \tag{6.85}$$

The horizon lies at $r = 2M$ since, for $r < 2M$, the velocity $\dot{r}$ must be negative. It might appear that an attempt to integrate equation (6.84) will run into difficulties with the pole at horizon, but this not the case. At $r = 2M$ we find that $\dot{r} = -\alpha$ and this cancels the pole. All particles therefore cross the horizon and reach the singularity in a finite coordinate time.

Specialising to the case of radial infall, we see from equation (6.85) that the constant $\alpha^2 - 1$ can be identified with twice the particle's initial energy (for a unit mass particle). Furthermore, equation (6.85) shows immediately that $\ddot{r} = -M/r^2$—a feature of motion in spherically symmetric gravitational field that is ignored in many GR texts. Some possible matter and photon trajectories are illustrated in figure 1. In the case where the particle is dropped from rest at $r = \infty$, equations (6.84) and (6.85) reduce to

$$\dot{r} = -\sqrt{2M/r}, \quad \dot{t} = 1 \tag{6.86}$$

and we recover an entirely Newtonian description of the motion. The properties of a black hole are so simple in the gauge defined by (6.79) that it is astonishing that this gauge is almost never seen in the literature (see Gautreau (1995) for a partial exception). Presumably, this is because the line element associated with (6.79) does not look as natural as the $\overline{h}$ function itself and hides the underlying simplicity of the system. Part of the reason for this is that the line element is not diagonal and relativists usually prefer to find a coordinate system which diagonalizes $g_{\mu\nu}$. Even when the freefall time coordinate t is employed, a different radial coordinate is usually found to keep the metric diagonal (Stephani 1982; Lemaître 1933).

Since the gauge defined by $g_1 = 1$ and $g_2 = -\sqrt{2M/r}$ extends our aim of keeping the equations in a simple Newtonian form, we refer to this solution as defining the 'Newtonian gauge' vacuum solution. We show in § 6 *e* that this gauge arises naturally from the description of collapsing dust. In the Newtonian gauge one hardly needs to modify classical reasoning at all to understand the processes involved—all particles just cross the horizon and fall into the singularity in a finite coordinate time. Furthermore, the horizon is located at $r = 2M$ precisely because we can apply Newtonian arguments! The only departures from Newtonian physics lie in relativistic corrections to the proper time taken for infall and in modifications to the equations for angular motion which lead to the familiar results for orbital precession.

When extracting physical predictions in the Newtonian, or any other, gauge, it is important to ensure that the problem is posed in a gauge-invariant manner. For example, one can envisage a simple experiment with two observers initially at rest outside a black hole at a distance r_0, where this distance is defined in terms of the magnitude of $\mathcal{R}(B)$. One observer can then start free falling and agree to emit photons of a chosen frequency at regular intervals. If one then computes what the remaining, stationary observer sees as a function of their proper time, this is clearly something physically meaningful. It is not hard to show that the predictions for this are gauge invariant, as they must be. Furthermore, if everything takes place outside the horizon, one can work in the 'Schwarzschild' gauge with $g_2 = 0$. However, to see

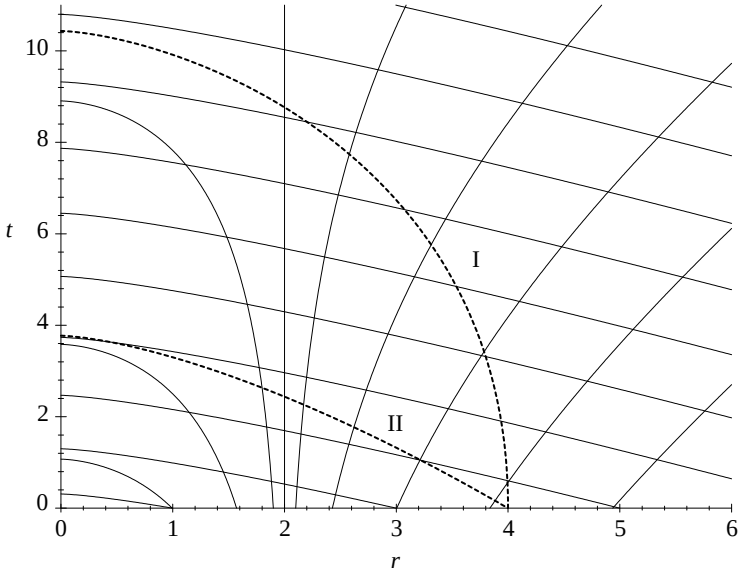

Figure 1. Matter and photon trajectories for radial motion in the in the Newtonian gauge. The solid lines are photon trajectories and the horizon lies at r=2. The broken lines represent possible trajectories for infalling matter. Trajectory I is for a particle released from rest at $r = 4$. Trajectory II is for a particle released from rest at $r = \infty$.

what happens as the free-falling observer crosses the horizon, a global solution such as (6.79) must be used. One still finds that the signal from the free-falling observer becomes successively more red shifted and less intense, as predicted when working with the Schwarzschild metric, but the free-falling observer crosses the horizon in a finite coordinate time.

The fact that all physical predictions should be invariant of the means by which they are computed should, in principle, be as true in standard GR as it is in GTG. However, some authors (see e.g. Logunov & Loskutov 1988) have questioned the uniqueness of predictions made in GR and this remains a controversial issue. While we do not wish to comment on this issue in relation to GR, we would point out that our gauge theory does not permit any such ambiguity.

(ii) *Horizons and time-reversal asymmetry*

Our picture of the gravitational fields due to a radially symmetric point source is rather different from that of GR. These differences are seen most clearly in the effects of time reversal. Time reversal is achieved by combining the displacement

$$f(x) = -e_t x e_t = x' \tag{6.87}$$

with the reflection

$$\overline{h}'(a) = -e_t \overline{h}(a) e_t, \tag{6.88}$$

resulting in the the time-reversed solution

$$\overline{h}^*(a) = e_t \overline{h}_{x'}(e_t a e_t) e_t. \tag{6.89}$$

As an example, the identity function $\overline{h}(a) = a$ is time-reverse symmetric—as it should be. The displacement (6.87) is a gauge transformation and cannot have any physical consequences. The reflection (6.88) is not a gauge transformation, however, and can be used to transform between physically distinct gauge sectors. The reflection (6.88)

is lost when the metric is formed, so GR cannot handle time-reversal in the same manner as GTG.

With the $\overline{h}$ function described by equation (6.15), and with the $\{f_i\}$ and $\{g_i\}$ functions of r only, the effect of (6.89) is simply to change the sign of the off-diagonal elements f_2 and g_2. For example, applied to the solution (6.79), the transformation (6.89) produces the time-reversed solution

$$\overline{h}^*(a) = a + \sqrt{2M/r}a\cdot e_r e_t. \tag{6.90}$$

The result is a solution in which particles inside the horizon are swept out. Once outside, the force on a particle is still attractive but particles cannot re-enter through the horizon.

This lack of time-reversal symmetry is not a feature of the various gauge choices made in arriving at (6.79); it is an inevitable result of the presence of a horizon. To see why, we return to the equations in the form prior to the restriction to the Newtonian gauge. The $\overline{h}$ field is as defined by equation (6.15) with $\alpha = 1$ and the $\{f_i\}$ and $\{g_i\}$ functions of r only. The general set of time-independent vacuum equations still have M constant and

$$g_1{}^2 - g_2{}^2 = 1 - 2M/r, \tag{6.91}$$

with

$$\partial_r g_1 = G, \quad \partial_r g_2 = F. \tag{6.92}$$

The bracket relation (6.40) now gives

$$g_2\partial_r f_2 - g_1\partial_r f_1 = Gf_1 - Ff_2, \tag{6.93}$$

from which it follows that

$$\partial_r(f_1 g_1 - f_2 g_2) = \partial_r \det(\underline{h}) = 0. \tag{6.94}$$

Hence $\det(\underline{h})$ is a constant, with its value dependent on the choice of position gauge. Since $\mathcal{R}(B)$ tends to zero at large r, we can always choose the gauge such that $\overline{h}(a)$ tends to the identity as $r \to \infty$. In this case $\det(\underline{h})$ must be one, so we can write

$$f_1 g_1 - f_2 g_2 = 1. \tag{6.95}$$

If we form the line element derived from our general $\overline{h}$ function we obtain

$$\mathrm{d}s^2 = (1 - 2M/r)\,\mathrm{d}t^2 + 2(f_1 g_2 - f_2 g_1)\,\mathrm{d}t\,\mathrm{d}r - (f_1{}^2 - f_2{}^2)\,\mathrm{d}r^2 - r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2). \tag{6.96}$$

The off-diagonal term here is the one that breaks time-reversal symmetry. However, we must have $g_1 = \pm g_2$ at the horizon and we know that $f_1 g_1 - f_2 g_2 = 1$ globally. It follows that

$$f_1 g_2 - f_2 g_1 = \pm 1 \quad \text{at } r = 2M, \tag{6.97}$$

so the line element (6.96) *must* break time reversal symmetry at the horizon (Doran *et al.* 1993*a*)). In fact, the asymmetry is even more pronounced. Once inside the horizon, equation (6.91) forces a non-zero g_2, so the $\overline{h}$ function cannot be time-reverse symmetric anywhere inside the horizon. This link between the existence of a horizon and the onset of time-reversal asymmetry is one of the most satisfying aspects of GTG. Furthermore, the requirement that a sign be chosen for $f_1 g_2 - f_2 g_1$ at the horizon shows that a black hole has more memory about its formation than simply its mass M—it also remembers that it was formed in a particular time direction. We will see an example of this in the following section.

At the level of the metric the discussion of time reversal is much less clear. For

example, inside the horizon a valid $\overline{h}$ function is obtained by setting f_1 and g_1 to zero. Since f_2 and g_2 are non-zero, this $\overline{h}$ function is manifestly not time-reverse symmetric. However, the line element generated by this $\overline{h}$ function is just the Schwarzschild line element

$$\mathrm{d}s^2 = (1-2M/r)\,\mathrm{d}t^2 - (1-2M/r)^{-1}\,\mathrm{d}r^2 - r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2), \tag{6.98}$$

which is usually thought of as being time-reverse symmetric. Clearly, our gauge theory probes questions related to time-reversal symmetry at a deeper level than GR. The consequences of our new understanding of time reversal will be met again in § 8, where we study the Dirac equation in a black hole background.

In GR, one of the most important results for studying radially symmetric fields is Birkhoff's theorem. This can be stated in various ways, though the most usual statement is that the line element outside a radially symmetric body can always be brought to the form of (6.98). As we have seen, this statement of Birkhoff's theorem is correct in GTG only if no horizon is present. However, the more general statement, that the fields outside a spherically symmetric source can always be made stationary, does remain valid.

(iii) *The Kruskal extension and geodesic completeness*

In GR, the line element (6.98) does not represent the final form of the metric for a radially symmetric black hole. The full 'maximal' solution was obtained by Kruskal (1960), who employed a series of coordinate transformations which mixed advanced and retarded Eddington–Finkelstein coordinates. The Kruskal extension describes a spacetime which contains horizons and is time-reverse symmetric, so can have no counterpart in GTG. Furthermore, the Kruskal solution has two distinct regions for each value of r (Hawking & Ellis 1973) and so is, topologically, quite distinct from the solutions admitted in GTG. This is because any solution of our equations must consist of $\overline{h}(a)$ expressed as a function of the vector position x. The form of this position dependence is arbitrary, but it must be present. So, when the coordinate r is employed in defining the $\overline{h}$ function, this always represents a particular function of the vector x. The point $r = 0$ is, by definition, a single point in space (an unbroken line in spacetime). No fields can alter this fact. The Kruskal solution contains two separate regions with the label $r = 0$, so immediately fails in GTG. Instead of the full Kruskal extension with four sectors (usually denoted I, II, I$'$ and II$'$ (Hawking & Ellis 1973)), GTG admits two distinct solutions, one containing the sectors I and II, and the other containing I$'$ and II$'$. These solutions are related by the discrete operation of time reversal, which is not a gauge transformation. This splitting of a single time-reverse symmetric solution into two asymmetric solutions is typical of the transition from a second-order to a first-order theory. Similar comments apply to the maximal extensions of the Reissner–Nordström and Kerr solutions. The infinite chain of 'universes' GR admits as solutions have no counterpart in our theory. Instead, we use integral equations to determine the nature of the matter singularities, precisely as one would do in electromagnetism (Doran 1998; Doran *et al.* 1998*a*).

In GR, the Kruskal solution is the unique maximal continuation of the Schwarzschild metric. The fact that it has no analogue in GTG means that our allowed solutions are not 'maximal' and forces us to address the issue of geodesic incompleteness. For the solution (6.79), geodesics exist which cannot be extended into the past for all values of their affine parameter. However, we have already seen that the presence of a horizon commits us to a choice of time direction and in the following

section we show how this choice is fixed by the collapse process. So, if we adopt the view that black holes arise solely as the endpoint of a collapse process, then there must have been a time before which the horizon did not exist. All geodesics from the past must therefore have come from a period before the horizon formed, so there is no question of the geodesics being incomplete. We therefore arrive at a consistent picture in which black holes represent the endpoint of a collapse process and the formation of the horizon captures information about the direction of time in which collapse occurred. This picture is in stark contrast with GR, which admits eternal, time-reverse symmetric black hole solutions.

(iv) *Coordinate transformations and displacements*

The coordinate transformations employed in GR have two distinct counterparts in GTG: as passive relabellings of the coordinates employed in a solution, such as changes of variables used for solving differential equations; and as disguised forms of position-gauge transformations. An example of the latter is the transformation between the Schwarzschild and advanced Eddington–Finkelstein forms of the spherically symmetric line element. This is achieved with the coordinate transformations

$$t' - r' = t - (r + 2M\ln(r - 2M)), \tag{6.99}$$

$$r' = r, \tag{6.100}$$

which can be viewed as the result of the displacement defined by

$$f(x) = x' = x - 2M\ln(r - 2M)e_t. \tag{6.101}$$

This displacement is to be applied to the solution

$$\overline{h}(a) = \Delta^{-1/2}a\cdot e_t e^t + \Delta^{1/2}a\cdot e_r e^r + a\wedge\sigma_r\sigma_r, \tag{6.102}$$

where

$$\Delta = 1 - 2M/r. \tag{6.103}$$

Clearly, the gravitational fields are defined only outside the horizon and the aim is to achieve a form of $\overline{h}(a)$ that is globally valid.

Differentiating the definition (6.101) we find that

$$\underline{f}(a) = a + \frac{2M}{r - 2M}a\cdot e_r e_t \tag{6.104}$$

and hence

$$\overline{f}^{-1}(a) = a - \frac{2M}{r - 2M}a\cdot e_t e_r. \tag{6.105}$$

Now, the function (6.102) is independent of t, so $\overline{h}(a, x') = \overline{h}(a, x)$. It follows that the transformed function $\overline{h}'(a)$ is given by

$$\overline{h}'(a) = \overline{h}\,\overline{f}^{-1}(a) = \overline{h}(a) - (2M/r)\Delta^{-1/2}a\cdot e_t e_r. \tag{6.106}$$

This new solution is not yet well defined for all r, but if we now apply the boost defined by the rotor

$$R = \exp(\tfrac{1}{2}\sigma_r\chi), \tag{6.107}$$

where

$$\sinh\chi = \tfrac{1}{2}(\Delta^{-1/2} - \Delta^{1/2}), \tag{6.108}$$

we obtain the solution

$$\overline{h}''(a) = a + (M/r)a\cdot e_- e_-, \tag{6.109}$$

where

$$e_- = e_t - e_r. \tag{6.110}$$

The solution (6.109) is now globally defined. It is the GTG equivalent of the Kerr–Schild form of the Schwarzschild solution and has the property that infalling null geodesics are represented by straight lines on a t–r plot. It is not hard to find a transformation between (6.109) and the Newtonian gauge solution (6.79). This transformation consists of a displacement and a rotation, both of which are globally well defined. On the other hand, if one starts with the solution (6.109) and tries to recover a version of the Schwarzschild solution by working in reverse, it is clear that the process fails. The boost needed is infinite at the horizon and ill defined for $r < 2M$, as is the required displacement. Such transformations fail to meet our requirement that gauge transformations be well defined over the whole region of physical interest.

(v) *Integral equations and the singularity*

The Riemann tensor $\mathcal{R}(B)$ contains derivatives of terms from $\omega(a)$ which fall off as $1/r^2$. When differentiating such terms, one must take account of the fact that

$$\boldsymbol{\nabla}\cdot(\boldsymbol{x}/r^3) = 4\pi\delta(\boldsymbol{x}), \tag{6.111}$$

where $\boldsymbol{x} = x\wedge e_t$. This fact will not affect the fields away from the origin, but will show up in the results of integrals enclosing the origin. To see how, we again return to the set-up before the Newtonian gauge was chosen. From equation (6.34) we see that

$$e_t\cdot\mathcal{G}(e_t) = 8\pi\rho = 2(-L_rT - T^2 + FS) - 4A \tag{6.112}$$

and, using equations (6.46) and (6.92), this gives

$$\begin{aligned} 4\pi\rho &= -(g_1\partial_r g_1 - g_2\partial_r g_2)/r + M/r^3 = \partial_r(M/r)/r + M/r^3 \\ &= (\sigma_r/r)\cdot\boldsymbol{\nabla}(M/r) + (M/r)\boldsymbol{\nabla}\cdot(\sigma_r/r) = \boldsymbol{\nabla}\cdot(M\boldsymbol{x}/r^3). \end{aligned} \tag{6.113}$$

It follows that $\rho = M\delta(\boldsymbol{x})$, so the singularity generating the radially symmetric fields is a simple δ function, of precisely the same kind as the source of the Coulomb field in electrostatics.

The presence of the δ function source at the origin is most easily seen when the solution is analysed in the gauge defined by (6.109). Solutions of this type are analysed in Doran (1998) and we restrict ourselves here to a few basic observations. For the solution (6.109), $\mathcal{R}(B)$ is given by (Doran 1998)

$$\mathcal{R}(\boldsymbol{a} + \mathrm{i}\boldsymbol{b}) = M[\boldsymbol{a}\cdot\boldsymbol{\nabla}(\boldsymbol{x}/r^3) + \mathrm{i}\boldsymbol{\nabla}\cdot(\boldsymbol{b}\wedge\boldsymbol{x}/r^3)] \tag{6.114}$$

and it is simple to see that, away from the origin, (6.114) reduces to (6.73). The significance of (6.114) is that it allows us to compute the integral of the Riemann tensor over a region enclosing the origin simply by converting the volume integral to a surface integral. Taking the region of integration to be a sphere of radius r_0 centred on the origin, we find that

$$\int_{r\leqslant r_0} \mathrm{d}^3x\,\mathcal{R}(\boldsymbol{a}) = M\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\theta\,\sin\theta\,\boldsymbol{a}\cdot\sigma_r\sigma_r = \tfrac{4}{3}\pi M\boldsymbol{a} \tag{6.115}$$

and

$$\int_{r\leqslant r_0} \mathrm{d}^3x\,\mathcal{R}(\mathrm{i}\boldsymbol{b}) = \int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\theta\,\sin\theta\,\mathrm{i}\sigma_r\cdot(\boldsymbol{b}\wedge\sigma_r) = -\tfrac{8}{3}\pi M\mathrm{i}\boldsymbol{b}. \tag{6.116}$$

These results are independent of the radius of the spherical shell, reflecting the spherical symmetry of the solution. The above results combine to give

$$\int_{r\leqslant r_0} \mathrm{d}^3x\,\mathcal{R}(B) = \tfrac{4}{3}\pi MB - 4\pi MB\wedge e_te_t = -\tfrac{2}{3}\pi M[B+3e_tBe_t], \tag{6.117}$$

which makes it clear what has happened. The angular integral of the Weyl component of $\mathcal{R}(B)$ has vanished, because

$$\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\theta\,\sin\theta(B+3\sigma B\sigma_r) = 0, \tag{6.118}$$

and what remains is the contribution from the stress-energy tensor, which is entirely concentrated at the origin. On contracting we find that

$$\int \mathrm{d}^3x\,\mathcal{R}(a) = 4\pi Me_tae_t, \quad \int \mathrm{d}^3x\,\mathcal{R} = -8\pi M, \quad \int \mathrm{d}^3x\,\mathcal{G}(a) = 8\pi Ma\cdot e_te_t \tag{6.119}$$

and, since $\mathcal{R}(a) = 0$ everywhere except for the origin, the integrals in (6.119) can be taken over any region of space enclosing the origin. It is now apparent that the solution represents a point source of matter and we can therefore write

$$\mathcal{T}(a) = M\delta(\boldsymbol{x})a\cdot e_te_t \tag{6.120}$$

for the matter stress-energy tensor. This is consistent with the definition of M as the integral of the density for a static system (6.61).

Analysing singularities in the gravitational fields by means of integral equations turns out to be very powerful in GTG. While the above application does not contain any major surprises, we show in Doran *et al.* (1998*a*) that the same techniques applied to axisymmetric fields reveal that the Kerr solution describes a ring of rotating matter held together by a disk of isotropic tension—a quite different picture to that arrived at in GR. This clearly has implications for the ultimate fate of matter falling onto the singularity and could yield testable differences between GTG and GR.

(*e*) *Collapsing dust*

The equations in table 4 can be used to determine the future evolution of a system given an equation of state and the initial ρ and g_2 distributions. They are therefore well suited to the description of radial collapse and the formation of horizons and singularities. The simplest model, in which the pressure is set to zero, describes collapsing dust. This situation was first studied by Oppenheimer & Snyder (1939) and has been considered since by many authors (Gautreau & Cohen 1995; Misner & Sharp 1964; Misner *et al.* 1973; Panek 1992). A feature of these studies is the appearance of formulae which have a suggestively Newtonian form. This is usually dismissed as a 'coincidence' (Misner *et al.* 1973, § 32.4). Here we study the collapse process in the Newtonian gauge and show that, far from being coincidental, the Newtonian form of the results is a natural consequence of the equations. The distinguishing feature of the Newtonian gauge approach is that the associated line element is not diagonal. This manifestly breaks time-reversal symmetry, as is appropriate for the description of collapsing matter. Working in this gauge enables us to keep all fields globally defined, so the horizon is easily dealt with and the matching onto an exterior vacuum region is automatically incorporated. This is quite different from previous work (Oppenheimer & Snyder 1939; Misner & Sharp 1964; Misner *et al.* 1973), which usually employs two distinct diagonal metrics, one for the matter region and one

for the vacuum. Finding the correct matching conditions between these metrics is awkward and difficulties are encountered once the horizon has formed.

If $p = 0$ it follows immediately that $G = 0$ and hence $f_1 = 1$. This ensures that the global time coordinate t agrees with the time measured by observers comoving with the fluid. Since $v{\cdot}\mathcal{D}v = 0$ in the absence of pressure, such observers are also freely falling. The function g_2 defines a velocity since, for a particle comoving with the fluid, g_2 is the rate of change of r (which is defined by the Weyl tensor) with proper time t. To emphasize its role as a velocity we replace g_2 with the symbol u for this section. The equations of table 4 now reduce to

$$F = \partial_r u, \tag{6.121}$$

$$M(r,t) = \int_0^r 4\pi r'^2 \rho(r',t)\,\mathrm{d}r', \tag{6.122}$$

which define F and M on a time slice, together with the update equations

$$\partial_t u + u\partial_r u = -M/r^2, \tag{6.123}$$

$$\partial_t M + u\partial_r M = 0. \tag{6.124}$$

Equations (6.123) and (6.124) afford an entirely Newtonian description of the fluid. Equation (6.123) is the Euler equation with an inverse-square gravitational force and (6.124) is the equation for conservation of mass. The L_t derivative plays the role of the 'matter' or 'comoving' derivative for the fluid since, when acting on a scalar, $v{\cdot}\mathcal{D} = L_t$.

The fact that $L_t M = 0$ in the absence of pressure (6.124) is a special case of a more general result. Consider the integral

$$I(r,t) = \int_0^r 4\pi s^2 \rho(s,t) f(s,t)\,\mathrm{d}s, \tag{6.125}$$

where $f(r,t)$ is some arbitrary function which is conserved along fluid streamlines; that is, it obeys

$$L_t f = 0. \tag{6.126}$$

If we now construct $L_t I$ we find that

$$L_t I = u4\pi r^2 \rho f + \int_0^r 4\pi s^2 \big(\rho\partial_t f + f\partial_t \rho\big)\,\mathrm{d}s. \tag{6.127}$$

However, from

$$L_t \rho = -(2u/r + F)\rho \tag{6.128}$$

and equation (6.121), we have

$$\partial_t(r^2\rho) = -\partial_r(ur^2\rho). \tag{6.129}$$

Similarly, from equation (6.126), we see that

$$\partial_t f = -u\partial_r f, \tag{6.130}$$

so it follows that

$$L_t I = u4\pi r^2 \rho f - \int_0^r 4\pi\big(us^2\rho\partial_s f + f\partial_s(us^2\rho)\big)\,\mathrm{d}s = 0. \tag{6.131}$$

Any integral of the type defined by I leads to a quantity which is conserved along the fluid streamlines. The integral for $M(r,t)$ (6.122) is one such example, with f

set to 1. It is clear from its appearance in the Riemann tensor that M represents the 'gravitating energy' of the region enclosed inside r.

Since $G = 0$, we have

$$L_t g_1 = 0 \tag{6.132}$$

and an alternative conserved quantity is therefore defined by

$$\mu(r,t) = \int_0^r 4\pi s^2 \rho(s,t)\,\frac{\mathrm{d}s}{g_1}. \tag{6.133}$$

This is the covariant integral of the density, so is also a covariant scalar quantity; it is simply the total rest-mass energy within r (see box 23.1 of 'Gravitation' (Misner *et al.* 1973) for a discussion of this point in the static case). The relationship between the rest-mass energy μ and the gravitating energy M can be seen more clearly by recalling that

$$g_1{}^2 = 1 - 2M/r + u^2. \tag{6.134}$$

Since

$$M(r,t) - \mu(r,t) = \int_0^r 4\pi s^2 \rho(s,t)(g_1 - 1)\,\frac{\mathrm{d}s}{g_1}, \tag{6.135}$$

the difference between the rest energy μ and the total energy M is governed by $g_1 - 1$. This is then multiplied by the term $4\pi r^2 \rho\,\mathrm{d}r/g_1$, which is the rest mass of a shell of width $\mathrm{d}r$. For $|2M/r - u^2| \ll 1$ we can approximate (6.134) to give

$$g_1 - 1 \approx -M/r + \tfrac{1}{2}u^2, \tag{6.136}$$

which explicitly shows the decomposition of the energy difference into the sum of the Newtonian gravitational potential energy (always negative) and the energy due to the bulk kinetic motion (always positive). It is clear that for a shell of material to escape it must have $g_1 - 1 > 0$ so, with no approximation necessary, we recover the Newtonian escape velocity $u^2 = 2M/r$.

As a further example of the insight provided by the Newtonian gauge, consider the case where the interior of the shell is empty. In this case $M = 0$, so

$$g_1 = (1 + u^2)^{1/2}, \tag{6.137}$$

which shows that g_1 can be interpreted as a relativistic γ factor associated u. This identification is justified if we put $u = \sinh\alpha$, which is reasonable since we know that u can be greater than 1. It is the presence of this additional boost factor in the formula for M compared to μ which, in this case, makes the total gravitating energy greater than the rest mass energy. These results should demonstrate that the physical picture of gravitational collapse in the absence of pressure is really no different from that afforded by Newtonian physics and special relativity. It is therefore no surprise that many of the results agree with those of Newtonian physics. Furthermore, abandoning a description in terms of distorted volume elements and spacetime geometry has allowed us to recover a clear physical picture of the processes involved.

(i) *Analytical solutions*

A useful property of the system of equations obtained when $p = 0$ is that it is easy to construct analytical solutions (Tolman 1934; Bondi 1947). To see this in the Newtonian gauge we write equation (6.124) in the form

$$\left(\frac{\partial M}{\partial t}\right)_r + u\left(\frac{\partial M}{\partial r}\right)_t = 0. \tag{6.138}$$

Since M is a function of r and t only, we can employ the reciprocity relation

$$\left(\frac{\partial M}{\partial t}\right)_r \left(\frac{\partial t}{\partial r}\right)_M \left(\frac{\partial r}{\partial M}\right)_t = -1, \tag{6.139}$$

to deduce that

$$\left(\frac{\partial t}{\partial r}\right)_M = \frac{1}{u}. \tag{6.140}$$

However, we know that u is determined by equation (6.134) and we also know that both M and g_1 are conserved along the fluid streamlines. We can therefore write $g_1 = g_1(M)$ and equation (6.140) can be integrated straightforwardly to give t as a function of r and M.

To perform the integration it is necessary to make a choice for the sign of u. For collapsing matter we clearly require $u < 0$, while for cosmology it turns out that $u > 0$ is the appropriate choice. For this section we can therefore write

$$\left(\frac{\partial t}{\partial r}\right)_M = -(g_1(M)^2 - 1 + 2M/r)^{-1/2}. \tag{6.141}$$

Finally, we need to choose a form for g_1. This amounts to making an initial choice of u, since u and g_1 are related via equation (6.134). For this section we simplify to the case in which the matter is initially at rest. This might provide a reasonable model for a star at the onset of a supernova, in which there is a catastrophic loss of pressure due to vast amounts of neutrino production, and the central core is suddenly left with no supporting pressure. With $u(r, 0) = 0$ we can write

$$g_1{}^2 = 1 - 2M(r_0)/r_0, \tag{6.142}$$

where r_0 labels the initial r coordinate. We can view the value of r_0 as carried along the streamline defined by it at $t = 0$, so can write $r_0 = r_0(t, r)$ and treat M and g_1 as functions of r_0 only. Equation (6.141) now becomes

$$\left(\frac{\partial t}{\partial r}\right)_{r_0} = -\left(\frac{2M}{r} - \frac{2M}{r_0}\right)^{-1/2}, \tag{6.143}$$

which integrates to give

$$t = \left(\frac{{r_0}^3}{2M}\right)^{1/2} (\tfrac{1}{2}\pi - \sin^{-1}(r/r_0)^{1/2} + (r/r_0)^{1/2}(1 - r/r_0)^{1/2}), \tag{6.144}$$

where we have chosen the initial conditions to correspond to $t = 0$.

Equation (6.144) determines a streamline for each initial value r_0 and can therefore be treated as implicitly determining the function $r_0(r, t)$. Since $M(r_0)$ and $g_1(r_0)$ are known, the future evolution of the system is completely determined. Furthermore, quantities such as ρ or $\partial_r u$ can be found directly once $r_0(r, t)$ is known. The above approach is easily extended to deal with initial conditions other than particles starting from rest since, once $M(r_0)$ and $g_1(r_0)$ are known, all one has to do is integrate equation (6.141). The ability to give a global description of the physics in a single gauge allows for simple simulations of a wide range of phenomena (Lasenby *et al.* 1998).

An important restriction on the solution (6.144) is that the streamlines should not cross. Crossed streamlines would imply the formation of shock fronts and in such situations our physical assumption that $p = 0$ will fail. Streamline crossing is avoided if the initial density distribution $\rho(r_0)$ is chosen to be either constant or

a monotonic-decreasing function of r_0. This is physically reasonable and leads to sensible simulations for a collapsing star.

(ii) *Singularity formation*

An immediate consequence of equation (6.144) is that the time taken before a given streamline reaches the origin is given by

$$t_1 = \pi(r_0^3/8M)^{1/2}. \tag{6.145}$$

Since the global time t agrees with the proper time for observers comoving with the fluid, equation (6.145) is also the lapse of proper time from the onset of collapse to termination at the singularity as measured by any particle moving with the dust. As pointed out in § 32.4 of Misner *et al.* (1973), the formula (6.145) agrees with the Newtonian result. For the reasons given above, this should no longer be a surprise.

Since g_1 and M are conserved along a streamline, equation (6.134) shows that $u = g_2$ must become singular as $r \to 0$. Thus the central singularity forms when the first streamline reaches the origin. Near $r_0 = 0$ the initial density distribution must behave as

$$\rho(r_0) \approx \rho_0 - O({r_0}^2), \tag{6.146}$$

so the mass function $M(r_0)$ is

$$M(r_0) = \tfrac{4}{3}\pi {r_0}^3 - O({r_0}^5). \tag{6.147}$$

It follows that

$$\lim_{r_0 \to 0} \pi(r_0^3/8M)^{1/2} = (3\pi/32\rho_0)^{1/2}, \tag{6.148}$$

so the central singularity forms after a time

$$t_f = \left(\frac{3\pi}{32G\rho_0}\right)^{1/2}, \tag{6.149}$$

where ρ_0 is the initial density at the origin and the gravitational constant G has been included.

A simulation of this process is shown in figure 2, which plots the fluid streamlines for the initial density function

$$\rho = \frac{\rho_0}{(1 + (r/a)^2)^2}. \tag{6.150}$$

The streamlines all arrive at the singularity after a finite time and the bunching at t_f can be seen clearly. Solutions can be extended beyond the time at which the singularity first forms by including an appropriate δ function at the origin.

A further point revealed by such simulations is that it is possible to have quite large differences between the total rest mass $\mu_\infty \equiv \mu(r = \infty)$ and the final mass of the singularity, M. For example, for the case plotted in figure 2 we find that $\mu_\infty = 6.26$, whereas $M = 2.17$. In this case nearly three times as many baryons have gone into forming the black hole than is apparent from its mass M. The possibility of large differences between M and μ is usually ignored in discussions of the thermodynamics of black holes (see footnote 14 of Bekenstein (1974)).

(iii) *Horizon formation*

Any particle on a radial path has a covariant velocity of the form

$$v = \cosh u\, e_t + \sinh u\, e_r. \tag{6.151}$$

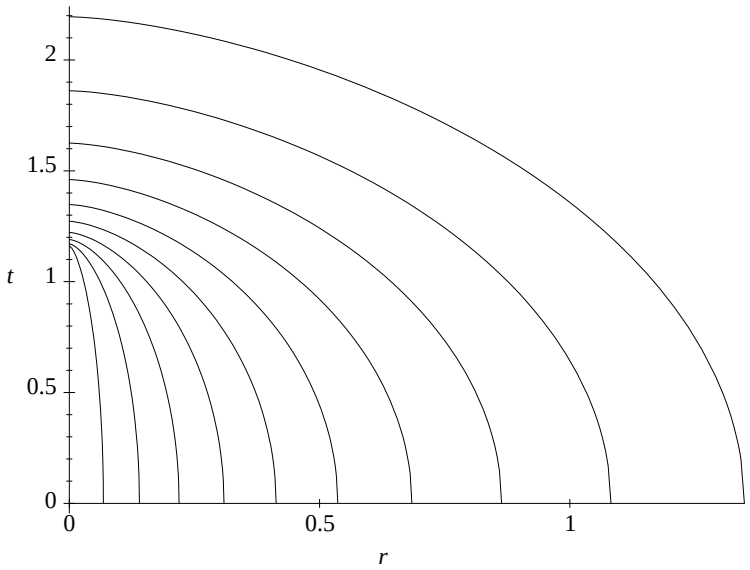

Figure 2. Fluid streamlines for dust collapsing from rest. The initial density is given by equation (6.150) with $\rho_0 = 0.22$ and $a = 1$. The singularity first forms at $t_f = 1.16$.

The underlying trajectory has $\dot{x} = \underline{h}(v)$, so the radial motion is determined by

$$\dot{r} = \cosh u\, g_2 + \sinh u\, g_1. \tag{6.152}$$

Since g_2 is negative for collapsing matter, the particle can only achieve an outward velocity if ${g_1}^2 > {g_2}^2$. A horizon therefore forms at the point where

$$2M(r,t)/r = 1. \tag{6.153}$$

To illustrate the formation of a horizon, we again consider the initial density profile of equation (6.150). By inverting (6.144) at fixed t, r_0 is found as a function of r. From this, $(1 - 2M(r,t)/r)$ is calculated straightforwardly and this quantity is plotted on figure 3 at different time slices. The plots show clearly that the horizon forms at a finite distance from the origin. It is conventional to extend the horizon back in time along the past light cone to the origin ($r = 0$), since any particle inside this surface could not have reached the point at which $1 - 2M/r$ first drops to zero and hence is also trapped (Novikov & Frolov 1989). The ease with which horizon formation is treated again illustrates the advantages of working in a non-diagonal gauge. Such considerations will clearly be important when performing numerical studies of more realistic collapse scenarios.

A final point is that, since u is negative, it follows that $f_1 g_2 - f_2 g_1 = u$ must also be negative. This tells us that the fields that remain after the collapse process has finished are in the class defined by $f_1 g_2 - f_2 g_1 = -1$ at the horizon. This time direction is then frozen into the fields, as discussed in § 6 *d*.

(*f*) *Cosmology*

The equations of table 4 are sufficiently general to deal with cosmology as well as astrophysics. In recent years, however, it has once more become fashionable to include a cosmological constant in the field equations. The derivation of § 6 *b* is largely unaffected by the inclusion of the cosmological term and only a few modifications

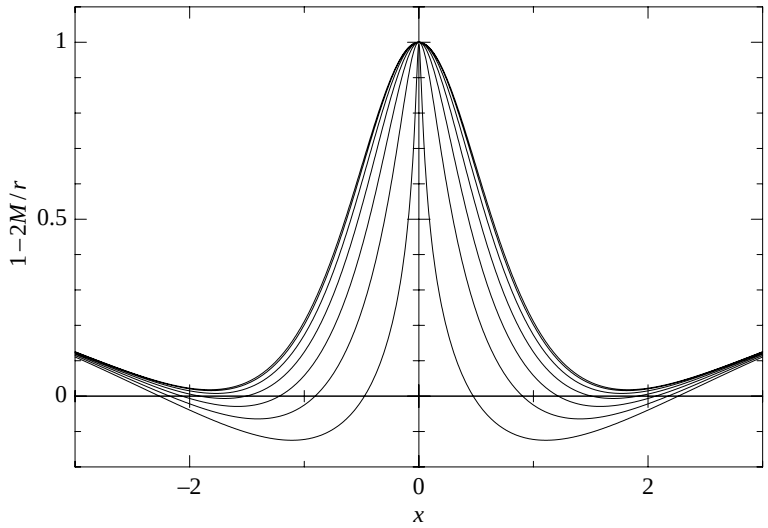

Figure 3. Simulation of collapsing dust in the Newtonian gauge. Successive time slices for the horizon function $(1-2M(r,t)/r)$ versus x are shown, with the top curve corresponding to $t=0$ and lower curves to successively later times. The initial velocity is zero and the initial density profile is given by equation (6.150) with $\rho_0 = 0.22$ and $a = 1$. There is no horizon present initially, but a trapped region quickly forms, since in regions where $1-2M/r < 0$ photons can only move inwards.

to table 4 are required. The full set of equations with a cosmological constant incorporated are summarized in table 5.

In cosmology we are interested in homogeneous solutions to the equations of table 5. Such solutions are found by setting ρ and p to functions of t only and it follows immediately from the $L_r p$ equation that

$$G = 0 \quad \Rightarrow \quad f_1 = 1. \tag{6.154}$$

For homogeneous fields the Weyl component of the Riemann tensor must vanish, since this contains directional information through the e_r vector. The vanishing of this term requires that

$$M(r,t) = \tfrac{4}{3}\pi r^3 \rho, \tag{6.155}$$

which is consistent with the $L_r M$ equation. The $L_t M$ and $L_t \rho$ equations now reduce to

$$F = g_2/r \tag{6.156}$$

and

$$\dot{\rho} = -3g_2(\rho + p)/r. \tag{6.157}$$

But we know that $L_r g_2 = F g_1$, which can only be consistent with (6.156) if

$$F = H(t), \quad g_2(r,t) = rH(t). \tag{6.158}$$

The $L_t g_2$ equation now reduces to a simple equation for $\dot{H}$,

$$\dot{H} + H^2 - \Lambda/3 = -\tfrac{4}{3}\pi(\rho + 3p). \tag{6.159}$$

Finally, we are left with the following pair of equations for g_1:

$$L_t g_1 = 0, \tag{6.160}$$

$$L_r g_1 = ({g_1}^2 - 1)/r. \tag{6.161}$$

Table 5. *Equations governing a radially symmetric perfect fluid—case with a non-zero cosmological constant Λ*

(* indicates equations that differ from those of table 4)

the $\overline{h}$ function	$\overline{h}(e^t) = f_1 e^t, \quad \overline{h}(e^r) = g_1 e^r + g_2 e^t$ $\overline{h}(e^\theta) = e^\theta, \quad \overline{h}(e^\phi) = e^\phi$
the ω function	$\omega(e_t) = G e_r e_t, \quad \omega(e_r) = F e_r e_t$ $\omega(\hat{\theta}) = g_2/r\,\hat{\theta} e_t + (g_1 - 1)/r\, e_r \hat{\theta}$ $\omega(\hat{\phi}) = g_2/r\,\hat{\phi} e_t + (g_1 - 1)/r\, e_r \hat{\phi}$
directional derivatives	$L_t = f_1 \partial_t + g_2 \partial_r, \quad L_r = g_1 \partial_r$
equations relating the $\overline{h}$ and ω functions	$L_t g_1 = G g_2, \quad L_r g_2 = F g_1$ $f_1 = \exp\left\{ \int^r -\frac{G}{g_1}\, \mathrm{d}r \right\}$
definition of M^*	$M \equiv \frac{1}{2} r ({g_2}^2 - {g_1}^2 + 1 - \frac{1}{3}\Lambda r^2)$
remaining derivatives*	$L_t g_2 = G g_1 - M/r^2 + \frac{1}{3} r\Lambda - 4\pi r p$ $L_r g_1 = F g_2 + M/r^2 - \frac{1}{3} r\Lambda - 4\pi r \rho$
matter derivatives	$L_t M = -4\pi g_2 r^2 p, \quad L_t \rho = -(2g_2/r + F)(\rho + p)$ $L_r M = 4\pi g_1 r^2 \rho, \quad L_r p = -G(\rho + p)$
Riemann tensor*	$\mathcal{R}(B) = 4\pi(\rho + p) B \cdot e_t e_t - \frac{1}{3}(8\pi\rho + \Lambda) B$ $-\frac{1}{2}(M/r^3 - \frac{4}{3}\pi\rho)(B + 3\sigma_r B \sigma_r)$
fluid stress-energy tensor	$\mathcal{T}(a) = (\rho + p) a \cdot e_t e_t - p a$

The latter equation yields ${g_1}^2 = 1 + r^2\phi(t)$ and the former reduces to

$$\dot{\phi} = -2H(t)\phi. \tag{6.162}$$

Hence g_1 is given by

$${g_1}^2 = 1 - kr^2 \exp\left\{ -2 \int^t H(t')\, \mathrm{d}t' \right\}, \tag{6.163}$$

where k is an arbitrary constant of integration. It is straightforward to check that (6.163) is consistent with the equations for $\dot{H}$ and $\dot{\rho}$. The full set of equations describing a homogeneous perfect fluid are summarized in table 6.

At first sight, the equations of table 6 do not resemble the usual Friedmann equations. The Friedmann equations are recovered straightforwardly, however, by setting

$$H(t) = \frac{\dot{S}(t)}{S(t)}. \tag{6.164}$$

With this substitution we find that

$${g_1}^2 = 1 - kr^2/S^2 \tag{6.165}$$

and that the $\dot{H}$ and density equations become

$$\ddot{S}/S - \tfrac{1}{3}\Lambda = -\tfrac{4}{3}\pi(\rho + 3p) \tag{6.166}$$

Table 6. *Equations governing a homogeneous perfect fluid*

the $\overline{h}$ function	$\overline{h}(a) = a + a\cdot e_r[(g_1 - 1)e^r + H(t)re^t]$
	$g_1{}^2 = 1 - kr^2 \exp\left\{-2\int^t H(t')\,\mathrm{d}t'\right\}$
the ω function	$\omega(a) = H(t)a\wedge e_t - (g_1 - 1)/r\, a\wedge(e_r e_t)e_t$
the density	$\frac{8}{3}\pi\rho = H(t)^2 - \frac{1}{3}\Lambda + k\exp\left\{-2\int^t H(t')\,\mathrm{d}t'\right\}$
dynamical equations	$\dot{H} + H^2 - \frac{1}{3}\Lambda = -\frac{4}{3}\pi(\rho + 3p)$
	$\dot{\rho} = -3H(t)(\rho + p)$

and

$$(\dot{S}^2 + k)/S^2 - \tfrac{1}{3}\Lambda = \tfrac{8}{3}\pi\rho, \tag{6.167}$$

recovering the Friedmann equations in their standard form (Narlikar 1993). The intrinsic treatment has therefore led us to work directly with the 'Hubble velocity' $H(t)$, rather than the 'distance' scale $S(t)$. There is a good reason for this. Once the Weyl tensor is set to zero, the Riemann tensor reduces to

$$\mathcal{R}(B) = 4\pi(\rho + p)B\cdot e_t e_t - \tfrac{1}{3}(8\pi\rho + \Lambda)B \tag{6.168}$$

and we have now lost contact with an intrinsically defined distance scale. We can therefore rescale the radius variable r with an arbitrary function of t (or r) without altering the Riemann tensor. The Hubble velocity, on the other hand, is intrinsic and it is therefore not surprising that our treatment has led directly to equations for this.

Among the class of radial rescalings a particularly useful one is to rescale r to $r' = S(t)r$. This is achieved with the transformation

$$f(x) = x\cdot e_t e_t + Sx\wedge e_t e_t, \tag{6.169}$$

so that, on applying equation (3.9), the transformed $\overline{h}$ function is

$$\overline{h}'(a) = a\cdot e_t e_t + \frac{1}{S}[(1 - kr^2)^{1/2} a\cdot e_r e^r + a\wedge\sigma_r\sigma_r]. \tag{6.170}$$

The function (6.170) reproduces the standard line element used in cosmology. We can therefore use the transformation (6.169) to move between the 'Newtonian' gauge developed here and the gauge of (6.170). This is useful for later sections, where the Maxwell and Dirac equations are solved in a cosmological background described by (6.170). The differences between these gauges can be understood by considering geodesic motion. A particle at rest with respect to the cosmological frame (defined by the cosmic microwave background) has $v = e_t$. In the gauge of (6.170) such a particle is not moving in the flatspace background (the distance variable r is equated with the comoving coordinate of GR). In the Newtonian gauge, on the other hand, comoving particles are moving outwards radially at a velocity $\dot{r} = H(t)r$, though this expansion centre is not an intrinsic feature. Of course, attempting to distinguish these pictures is a pointless exercise, since all observables must be gauge invariant. All that is of physical relevance is that, if two particles are at rest with respect to the cosmological frame (defined by the cosmic microwave background), then the

light-travel time between these particles is an increasing function of time and light is redshifted as it travels between them.

(i) *Dust models*

The utility of the Newtonian gauge in cosmology has been independently discovered by other authors (Gautreau 1984; Ellis & Rothman 1993). Here we illustrate its advantages for dust models ($p = 0$). Setting p to zero implies that

$$H(t) = -\dot{\rho}/3\rho, \tag{6.171}$$

so

$$g_1{}^2 = 1 - kr^2\rho^{2/3} \tag{6.172}$$

and

$$H(t) = (\tfrac{8}{3}\pi\rho - k\rho^{2/3} + \tfrac{1}{3}\Lambda)^{1/2}. \tag{6.173}$$

We are therefore left with a single first-order differential equation for ρ. Explicit solutions of this equations are often not needed, as we can usually parametrize time by the density $\rho(t)$.

If we now look at the trajectories defined by the fluid, we find that

$$v = e_t, \tag{6.174}$$

$$\Rightarrow \quad \dot{x} = e_t + rH(t)e_r. \tag{6.175}$$

It follows that

$$\dot{r}/r = H(t) = -\dot{\rho}/3\rho \tag{6.176}$$

and hence that

$$r/r_0 = (\rho/\rho_0)^{-1/3}. \tag{6.177}$$

The fluid streamlines form a family of spacetime curves spreading out from the origin at the initial singularity (when $\rho = \infty$). The Newtonian gauge therefore describes an expanding universe in a very simple, almost naive, manner. Since all points in a homogeneous cosmology are equivalent, we can consider ourselves to be located at $r = 0$. The Newtonian gauge then pictures our observable universe as a ball of dust expanding outwards radially from us.

Whilst the picture provided by the Newtonian gauge has no physical reality of its own, it does have some heuristic merit and can provide a useful aid to one's physical intuition. For example, consider the familiar relationship between angular size and redshift. An initially surprising feature of this relationship is that, beyond a certain redshift, angular sizes stop decreasing and start increasing. This result is easily understood in the Newtonian gauge. Consider the photon paths shown in figure 4. (These paths are for a $k = 0$ and $\Lambda = 0$ universe, though the comments are applicable more generally.) Suppose that at some finite time t_0 we receive a photon from the distant past. This photon must have followed part of a path which begins on the origin at $t = 0$. Before a certain time in the past, therefore, photons received by us must have initially travelled outwards before turning round and reaching us. The angular size of an observed object is then that appropriate to the actual r coordinate of the object when it emitted the photons. Since the value of r decreases before a certain time, angular sizes appear larger for objects which emitted photons before this time.

A radial null geodesic has a trajectory

$$x(\tau) = t(\tau)e_t + r(\tau)e_r. \tag{6.178}$$

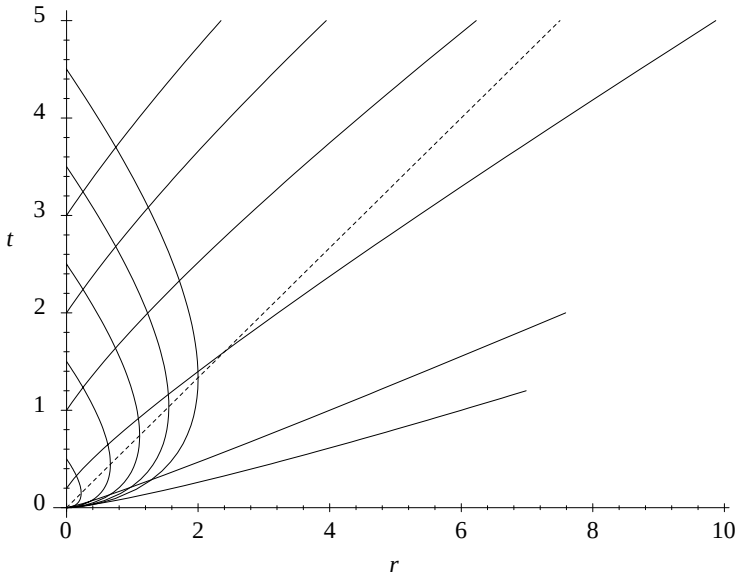

Figure 4. Null geodesics for a dust-filled $k = \Lambda = 0$ universe in the Newtonian gauge. All geodesics start from the origin at the initial singularity (here set at $t = 0$). The dashed line gives the critical distance where photons 'turn round' in this gauge.

For these trajectories,

$$v = \underline{h}^{-1}(\dot{x}) = \dot{t}\left(e_t - \frac{rH(t)}{g_1}e_r\right) + \frac{\dot{r}}{g_1}e_r \tag{6.179}$$

and the condition that $v^2 = 0$ reduces to

$$\dot{r}/g_1 = \dot{t}(rH(t)/g_1 \pm 1), \tag{6.180}$$

$$\Rightarrow \quad \frac{\mathrm{d}r}{\mathrm{d}t} = rH(t) \pm g_1. \tag{6.181}$$

Since $H(t)$ is positive in a cooling universe, the distance at which photons 'turn round' is defined by

$$r_c = \frac{g_1}{H(t)} = (\tfrac{8}{3}\pi\rho + \tfrac{1}{3}\Lambda)^{-1/2}. \tag{6.182}$$

(In Lasenby *et al.* (1995), we mistakenly referred to this distance as a particle horizon.) For $k = \Lambda = 0$ it is simple to show that this distance corresponds to a redshift of 1.25. This result relates physically measurable quantities, so is gauge invariant.

(ii) *Closed universe models*

The Newtonian gauge presents a particularly simple picture for a closed universe ($k > 0$). For $k > 0$ the requirement that $1 - kr^2\rho^{2/3}$ is positive means that

$$r^3\rho > k^{-3/2}. \tag{6.183}$$

This places a limit on the speed with which the dust can expand, so the solution describes a finite ball of dust expanding into a vacuum. This ball expands out to some fixed radius before turning round and contracting back to the origin. The turning point is achieved where, for $\Lambda = 0$,

$$H(t) = 0, \tag{6.184}$$

$$\Rightarrow \quad \rho = (3k/8\pi)^3, \tag{6.185}$$

so the maximum radius is

$$r_{\max} = \frac{8\pi}{3k^{3/2}}. \tag{6.186}$$

The time taken to reach the future singularity is given by (6.149), since this cosmological model is a special case of spherical collapse in which the density is uniform. This picture of a model for a closed universe is both simple and appealing. It allows us to apply Newtonian reasoning while ensuring consistency with the full relativistic theory.

The finite ball model for a $k > 0$ cosmology is clearly useful when considering experiments with particles carried out near the origin, but globally one must consider the boundary properties of the ball. A crucial question is whether the particle horizon (the largest region of the universe with which an observer at the origin is in causal contact) extends past the edge of the ball or not. It can be shown that this horizon always lies inside the radius at which g_1 becomes imaginary, except at the turnaround point (the point at which the ball reaches its maximum radius), where the two radii coincide. A suitable choice of cutoff radius is therefore available in either the expanding or contracting phase separately, but what happens at the turnaround point is potentially ambiguous. When discussing field theory, however, the finite ball model is inadequate. One must instead use a global gauge so, in § 7 *d*, we introduce the 'stereographic projection gauge'. This provides a global solution which can be used in the study of electromagnetism (§ 7 *d*) and the Dirac equation (§ 8 *c*) in a cosmological background. It is possible to treat the stereographic projection gauge solution in a form of the Newtonian gauge, though this possibility is not explored here.

7. Electromagnetism in a gravitational background

In § 4 we derived field equations for the gravitational and Dirac fields. We now turn to the derivation of the Maxwell equations in a gravitational background. A number of applications of these equations are then discussed, including a simple derivation of the characteristic surfaces for both the Maxwell and Dirac equations.

The basic dynamical variable is the electromagnetic vector potential A, for which the coupling to spinor fields was derived in § 3 *c*. Under phase rotations of the spinor field, A transforms as

$$eA \mapsto eA - \nabla\phi. \tag{7.1}$$

It follows that, under a displacement, A must transform in the same way as $\nabla\phi$; that is,

$$A \mapsto \overline{f}(A(x')). \tag{7.2}$$

The covariant form of the vector potential is therefore

$$\mathcal{A} = \overline{h}(A), \tag{7.3}$$

which is the term that appeared in the Dirac equation (3.62).

From A, the Faraday bivector F is defined by

$$F \equiv \nabla \wedge A. \tag{7.4}$$

This definition implies that, under displacements, $F(x)$ is transformed to $F'(x)$, where

$$F'(x) = \nabla \wedge \overline{f}A(x') = \overline{f}(\nabla_{x'} \wedge A(x')) = \overline{f}F(x'). \tag{7.5}$$

It follows that the covariant analogue of F is defined by

$$\mathcal{F} = \overline{h}(F), \tag{7.6}$$

which is covariant under position and rotation-gauge transformations and is invariant under phase changes.

The same covariant quantity $\mathcal{F}$ is obtained if one follows the route used for the construction of $R(a\wedge b)$ at (4.2). In particular, the contracted commutator of two covariant derivatives gives (in the absence of torsion)

$$\overline{h}(\partial_b)\wedge\overline{h}(\partial_a)[\mathrm{D}_a, \mathrm{D}_b]\psi = \tfrac{1}{2}\overline{h}(\partial_b)\overline{h}(\partial_a)R(a\wedge b)\psi = \tfrac{1}{2}\mathcal{R}\psi. \tag{7.7}$$

The analogous construction for the 'internal' covariant derivative

$$\mathrm{D}^{\mathrm{I}}_a\psi = a\cdot\nabla\psi - ea\cdot A\,\psi\mathrm{i}\sigma_3 \tag{7.8}$$

gives

$$\overline{h}(\partial_b)\wedge\overline{h}(\partial_a)[\mathrm{D}^{\mathrm{I}}_a, \mathrm{D}^{\mathrm{I}}_b]\psi = e\overline{h}(\partial_b)\wedge\overline{h}(\partial_a)(a\wedge b)\cdot F\psi\mathrm{i}\sigma_3 = 2e\mathcal{F}\psi\mathrm{i}\sigma_3, \tag{7.9}$$

which clearly identifies $\mathcal{F}$ as a covariant quantity. Unlike $\mathcal{R}$, however, $\mathcal{F}$ is a bivector and equation (7.9) exhibits a curious interaction between this bivector on the left of ψ and the fixed bivector $\mathrm{i}\sigma_3$ on the right.

Having defined the covariant bivector $\mathcal{F}$, it is clear that the appropriate generalization of the electromagnetic action to include gravitational interactions is

$$S = \int |\mathrm{d}^4x|\det(\underline{h})^{-1}(\tfrac{1}{2}\mathcal{F}\cdot\mathcal{F} - \mathcal{A}\cdot\mathcal{J}), \tag{7.10}$$

where here $\mathcal{J}$ is the covariant charge current. Unlike the Dirac action, the rotation-gauge field $\Omega(a)$ does not appear in this action. It follows that the electromagnetic field does not act as a source of spin. The action (7.10) is varied with respect to A, with $\overline{h}(a)$ and $\mathcal{J}$ treated as external fields. The result of this is the equation

$$\nabla\cdot G = J, \tag{7.11}$$

where

$$G \equiv \underline{h}\overline{h}(F)\det(\underline{h})^{-1} \tag{7.12}$$

and

$$J \equiv \underline{h}(\mathcal{J})\det(\underline{h})^{-1}. \tag{7.13}$$

Equation (7.11) combines with the identity

$$\nabla\wedge F = 0 \tag{7.14}$$

to form the full set of Maxwell equations in a gravitational background. We again see that the classical field equations can be expressed in a form from which all reference to the rotation gauge has been removed.

Some insight into the equations (7.11) and (7.14) is obtained by performing a space-time split (see § 2 *a*) and writing

$$F = \boldsymbol{E} + \mathrm{i}\boldsymbol{B}, \tag{7.15}$$

$$G = \boldsymbol{D} + \mathrm{i}\boldsymbol{H}, \tag{7.16}$$

$$J\gamma_0 = \rho + \boldsymbol{J}. \tag{7.17}$$

In terms of these variables Maxwell's equations in a gravitational background take the familiar form

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0, \quad \boldsymbol{\nabla}\cdot\boldsymbol{D} = \rho, \quad \boldsymbol{\nabla}\times\boldsymbol{E} = -\frac{\partial\boldsymbol{B}}{\partial t}, \quad \boldsymbol{\nabla}\times\boldsymbol{H} = \boldsymbol{J} + \frac{\partial\boldsymbol{D}}{\partial t}, \tag{7.18}$$

where $\boldsymbol{\nabla} = \gamma_0 \wedge \nabla = \sigma_i \partial_{x^i}$ is the 3D vector derivative and the bold cross $\times$ is the traditional vector cross product:

$$\boldsymbol{a} \times \boldsymbol{b} = -\mathrm{i}\boldsymbol{a} \times \boldsymbol{b}. \tag{7.19}$$

Equation (7.18) shows that the $\overline{h}$ field defines the dielectric properties of the space through which the electromagnetic fields propagate, with $\det(\underline{h})^{-1}\underline{h}\overline{h}$ determining the generalized permittivity–permeability tensor. Many phenomena, including the bending of light, can be understood easily in terms of the properties of the dielectric defined by the $\overline{h}$ field.

(iii) *Covariant form of the Maxwell equations*

So far we have failed to achieve a manifestly covariant form of the Maxwell equations. We have, furthermore, failed to unite the separate equations into a single equation. In the absence of gravitational effects, the equations $\nabla \cdot F = J$ and $\nabla \wedge F = 0$ combine into the single equation

$$\nabla F = J. \tag{7.20}$$

The significance of this equation is that the ∇ operator is invertible, whereas the separate $\nabla\cdot$ and $\nabla\wedge$ operators are not (Hestenes & Sobczyk 1984; Gull *et al.* 1993*b*). Clearly, we expect that such a unification should remain possible after the gravitational gauge fields have been introduced. To find a covariant equation, we first extend the 'wedge' equation (4.36) to include higher-grade terms. To make the derivation general we include the spin term, in which case we find that

$$\begin{aligned}\dot{\mathcal{D}}\wedge\dot{\overline{h}}(a\wedge b) &= [\dot{\mathcal{D}}\wedge\dot{\overline{h}}(a)]\wedge\overline{h}(b) - \overline{h}(a)\wedge[\dot{\mathcal{D}}\wedge\dot{\overline{h}}(b)] \\ &= \kappa[\overline{h}(a)\cdot\mathcal{S}]\wedge\overline{h}(b) - \kappa\overline{h}(a)\wedge[\overline{h}(b)\cdot\mathcal{S}] = \kappa\mathcal{S}\times\overline{h}(a\wedge b)\end{aligned} \tag{7.21}$$

and, more generally, we can write

$$\mathcal{D}\wedge\overline{h}(A_r) = \overline{h}(\nabla\wedge A_r) + \kappa\langle\mathcal{S}\overline{h}(A_r)\rangle_{r+1}. \tag{7.22}$$

We can therefore replace equation (7.14) by

$$\mathcal{D}\wedge\mathcal{F} - \kappa\mathcal{S}\times\mathcal{F} = 0. \tag{7.23}$$

We next use the rearrangement

$$\begin{aligned}\nabla\cdot(\underline{h}(\mathcal{F})\det(\underline{h})^{-1}) &= \mathrm{i}\nabla\wedge(\mathrm{i}\underline{h}(\mathcal{F})\det(\underline{h})^{-1}) = \mathrm{i}\nabla\wedge(\overline{h}^{-1}(\mathrm{i}\mathcal{F})) \\ &= \mathrm{i}\overline{h}^{-1}[\mathcal{D}\wedge(\mathrm{i}\mathcal{F}) + \kappa(\mathrm{i}\mathcal{F})\times\mathcal{S}]\end{aligned} \tag{7.24}$$

to write equation (7.11) as

$$\mathcal{D}\cdot\mathcal{F} - \kappa\mathcal{S}\cdot\mathcal{F} = \mathrm{i}\overline{h}(J\mathrm{i}) = \mathcal{J}. \tag{7.25}$$

Equations (7.23) and (7.25) now combine into the single equation

$$\mathcal{D}\mathcal{F} - \kappa\mathcal{S}\mathcal{F} = \mathcal{J}, \tag{7.26}$$

which achieves our objective. Equation (7.26) is manifestly covariant and the appearance of the $\mathcal{D}\mathcal{F}$ term is precisely what one might expect on 'minimal-coupling' grounds. The appearance of the spin term is a surprise, however. Gauge arguments alone would not have discovered this term and it is only through the construction of a gauge-invariant action integral that the term is found. Equation (7.26) should be particularly useful when considering electromagnetic effects in regions of high spin density, such as neutron stars.

To complete the description of electromagnetism in a gravitational background we

need a formula for the free-field stress-energy tensor. Applying the definition (4.18), we construct

$$\mathcal{T}_{\rm em}\underline{h}^{-1}(a) = \tfrac{1}{2}\det(\underline{h})\partial_{\overline{h}(a)}\langle\overline{h}(F)\overline{h}(F)\det(\underline{h})^{-1}\rangle = \overline{h}(a\cdot F)\cdot\mathcal{F} - \tfrac{1}{2}\underline{h}^{-1}(a)\mathcal{F}\cdot\mathcal{F}. \quad (7.27)$$

Hence,

$$\mathcal{T}_{\rm em}(a) = \overline{h}(\underline{h}(a)\cdot F)\cdot\mathcal{F} - \tfrac{1}{2}a\mathcal{F}\cdot\mathcal{F} = (a\cdot\mathcal{F})\cdot\mathcal{F} - \tfrac{1}{2}a\mathcal{F}\cdot\mathcal{F} = -\tfrac{1}{2}\mathcal{F}a\mathcal{F}, \quad (7.28)$$

which is the natural covariant extension of the gravitation-free form $-\frac{1}{2}FaF$. The tensor (7.28) is symmetric, as one expects for fields with vanishing spin density.

(*a*) *Characteristic surfaces*

In the STA the Maxwell and Dirac equations are both first-order differential equations involving the vector derivative ∇. For electromagnetism, this first-order form of the equations offers many advantages over the equivalent second-order theory (Hestenes 1966; Gull *et al.* 1993*b*). We have now seen that gravitational interactions modify both these equations in such a way that the vector derivative ∇ is replaced by the position-gauge covariant derivative $\overline{h}(\nabla)$. As an illustration of the utility of first-order equations, both without and with gravitational effects, we now give a simple derivation of the properties of characteristic surfaces.

Consider, initially, a generic equation of the type

$$\nabla\psi = f(\psi, x), \quad (7.29)$$

where ψ is an arbitrary field and f is a known function. Suppose that an initial set of data is given over some three-dimensional surface in spacetime and we wish to propagate this information off the surface into the adjoining region. We pick three vectors, a, b and c, which are tangent to the surface. From our initial data we can construct $a\cdot\nabla\psi$, $b\cdot\nabla\psi$ and $c\cdot\nabla\psi$. Now define

$$n \equiv {\rm i}a\wedge b\wedge c \quad (7.30)$$

and use $n\nabla = n\cdot\nabla + n\wedge\nabla$ to decompose (7.29) into

$$n\cdot\nabla\psi = -n\wedge\nabla\psi + nf(\psi). \quad (7.31)$$

The right-hand side of (7.31) contains the term

$$n\wedge\nabla\psi = {\rm i}(a\wedge b\wedge c)\cdot\nabla\psi = {\rm i}(a\wedge b\,c\cdot\nabla\psi - a\wedge c\,b\cdot\nabla\psi + b\wedge c\,a\cdot\nabla\psi), \quad (7.32)$$

which is therefore known. It follows that we know all the terms on the right-hand side of equation (7.31) and can therefore construct $n\cdot\nabla\psi$. This gives us the derivative required to propagate off the surface. The only situation for which propagation is impossible is when n remains in the surface. This occurs when

$$n\wedge(a\wedge b\wedge c) = 0 \quad \Rightarrow \quad n\wedge(n{\rm i}) = 0 \quad \Rightarrow \quad n\cdot n = 0. \quad (7.33)$$

It follows that the characteristic surfaces for any first-order equation of the type defined by (7.29) are null surfaces. These considerations automatically include the Maxwell and Dirac equations. It is notable how this result follows from purely algebraic considerations.

The generalization to a gravitational background is straightforward. Equation (7.26) is generalized to

$$\overline{h}(\nabla)\psi = f(\psi) \quad (7.34)$$

and we assume that a gauge choice has been made so that all the fields (apart from

ψ) are known functions of x. Again, we assume that the initial data consist of values for ψ over some three-dimensional surface, so we can still determine $a\cdot\nabla\psi$, etc. Since

$$a\cdot\nabla = \underline{h}^{-1}(a)\cdot\overline{h}(\nabla), \tag{7.35}$$

it follows that the vector of interest is now

$$\mathrm{i}\underline{h}^{-1}(a)\wedge\underline{h}^{-1}(b)\wedge\underline{h}^{-1}(c) = \mathrm{i}\underline{h}^{-1}(n\mathrm{i}) = \overline{h}(n)\det(\underline{h})^{-1}. \tag{7.36}$$

This time we multiply equation (7.34) by $\overline{h}(n)$ and find that $\overline{h}(n)\cdot\overline{h}(\nabla)\psi$ can be constructed entirely from known quantities. We can therefore propagate in the $\underline{h}\overline{h}(n)$ direction, so we now require that this vector does not lie in the initial surface. The analogue of (7.33) is therefore

$$\underline{h}\overline{h}(n)\cdot n = 0 \quad \text{or} \quad \overline{h}(n)^2 = 0 \tag{7.37}$$

and the characteristic surfaces are now those for which $\overline{h}(n)$ is null. This is the obvious covariant extension of n being a null vector.

(*b*) *Point charge in a black hole background*

The problem of interest here is to find the fields generated by a point source held at rest outside the horizon of a radially symmetric black hole. The $\overline{h}$ function in this case can be taken as that of equation (6.79). The solution to this problem can be found by adapting the work of Copson (1928) and Linet (1976) to the present gauge choices. Assuming, for simplicity, that the charge is placed on the z-axis a distance a from the origin ($a > 2M$), the vector potential can be written in terms of a single scalar potential $V(r,\theta)$ as

$$A = V(r,\theta)\left(e_t + \frac{\sqrt{2Mr}}{r-2M}e_r\right). \tag{7.38}$$

It follows that

$$\boldsymbol{E} = -\boldsymbol{\nabla}V, \quad \boldsymbol{B} = -\frac{\sqrt{2Mr}}{r(r-2M)}\frac{\partial V}{\partial\theta}\sigma_\phi, \quad \boldsymbol{H} = 0, \quad \boldsymbol{D} = -\frac{\partial V}{\partial r}\sigma_r - \frac{1}{r-2M}\frac{\partial V}{\partial\theta}\sigma_\theta \tag{7.39}$$

and

$$\mathcal{F} = -\frac{\partial V}{\partial r}\sigma_r - \frac{1}{r-2M}\frac{\partial V}{\partial\theta}(\sigma_\theta + \sqrt{2M/r}\mathrm{i}\sigma_\phi). \tag{7.40}$$

The Maxwell equations now reduce to the single partial differential equation

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right) + \frac{1}{r(r-2M)}\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial V}{\partial\theta}\right) = -\rho, \tag{7.41}$$

where $\rho = q\delta(\boldsymbol{x}-\boldsymbol{a})$ is a δ function at $z = a$. This was the problem originally tackled by Copson (1928), who obtained a solution that was valid locally in the vicinity of the charge, but contained an additional pole at the origin. Linet (1976) modified Copson's solution by removing the singularity at the origin to produce a potential $V(r,\theta)$ whose only pole is on the z-axis at $z = a$. Linet's solution is

$$V(r,\theta) = \frac{q}{ar}\frac{(r-M)(a-M) - M^2\cos^2\theta}{D} + \frac{qM}{ar}, \tag{7.42}$$

where

$$D = [r(r-2M) + (a-M)^2 - 2(r-M)(a-M)\cos\theta + M^2\cos^2\theta]^{1/2}. \tag{7.43}$$

An important feature of of this solution is that once (7.42) is inserted back

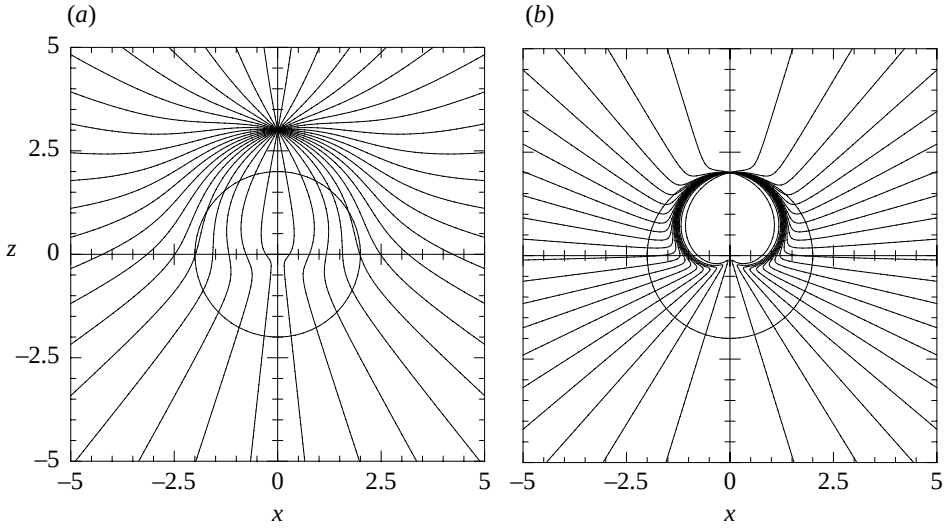

Figure 5. Streamlines of the $\boldsymbol{D}$ field. The horizon is at $r = 2$ and the charge is placed on the z-axis. The charge is at $z = 3$ and $z = 2.01$ for the top and bottom diagrams, respectively. The streamlines are seeded so as to reflect the magnitude of $\boldsymbol{D}$. The streamlines are attracted towards the origin but never actually meet it. Note the appearance of a 'cardioid of avoidance' as the charge gets very close to the horizon. The equation for this cardioid is $r = M(1 + \cos\theta)$, which is found by setting $D = 0$ when $a = 2M$.

into (7.40) the resultant $\mathcal{F}$ is both finite and continuous at the horizon. Furthermore, since $\overline{h}$ is well-defined at the horizon, both F and G must also be finite and continuous there. Working in the Newtonian gauge has enabled us to construct a global solution and we can therefore study its global properties. One simple way to illustrate the global properties of the solution is to plot the streamlines of $\boldsymbol{D}$ which, from equation (7.18), is divergenceless away from the source. The streamlines should therefore spread out from the charge and cover all space. Since the distance scale r is fixed to the gravitationally defined distance, the streamlines of $\boldsymbol{D}$ convey genuine intrinsic information. Hence the plots are completely unaffected by our choice for the g_1 or g_2 functions, or indeed our choice of t coordinate. Figure 5 shows streamline plots for charges held at different distances above the horizon. Similar plots were first obtained by Hanni & Ruffini (1973) although, as they worked with the Schwarzschild metric, they were unable to extend their plots through the horizon. The plots reveal an effective contraction in the radial direction. It is not hard to show that the contraction is precisely that of a particle moving with the free-fall velocity $(2M/a)^{1/2}$ relative to a fixed observer.

The description presented here of the fields due to a point charge is very different from that advocated by the 'Membrane Paradigm' (Thorne *et al.* 1986). The membrane paradigm was an attempt to develop the theory of black holes in a way that, as far as possible, employed only familiar physical concepts. In this way, gravitational effects could be incorporated correctly without requiring an understanding of the full GR treatment. The hope was that astrophysicists would adopt this paradigm when modelling regions where black hole physics could be significant, such as at the heart of a quasar. The paradigm works by drawing a veil over the horizon (the membrane) and concentrating on the physics outside the horizon as seen by observers remaining at a fixed distance (fiducial observers). Our view, however, is that it is the Newtonian

gauge which provides the clearest understanding of the physics of black holes whilst requiring minimal modification to Newtonian and special-relativistic ideas. Furthermore, writing the Maxwell equations in the form (7.18) removes any difficulties in applying conventional reasoning to the study of electromagnetism in a gravitational background.

There are other ways that our approach offers advantages over the membrane paradigm. When applying the membrane paradigm one has to work with quantities which are singular at the horizon and this is hardly a recipe for applying traditional intuition! As we have seen, once formulated in the Newtonian gauge (or any other gauge admissible in GTG), all physical quantities are finite. The membrane paradigm also warns physicists against producing plots such as figure 5, because such plots depend on the choice of radial coordinate. However, our intrinsic approach makes it clear that such plots *are* meaningful, because r is determined uniquely by the Riemann tensor. Presenting the plots in the form of figure 5 enables direct physical information to be read off. In short, the simple physical picture provided by our intrinsic method and Newtonian-gauge solution disposes of any need to adopt the artificial ideas advocated by the membrane paradigm.

(*c*) *Polarization repulsion*

An interesting feature of the above solution (7.42) is the existence of a repulsive 'polarization' force (Smith & Will 1980), one effect of which is that a smaller force is needed to keep a charged particle at rest outside a black hole than an uncharged one. In their derivation of this force, Smith & Will (1980) employed a complicated energy argument which involved renormalising various divergent integrals. Here we show that the same force can be derived from a simple argument based on the polarization effects of the dielectric described by a black hole. First, however, we must be clear how force is defined in GTG. In the presence of an electromagnetic field, the equation of motion for a point particle (4.61) is modified to

$$m\dot{v} = [q\mathcal{F} - m\omega(v)]\cdot v. \tag{7.44}$$

We therefore expect that any additional force should also be described by a covariant bivector which couples to the velocity the same way that $\mathcal{F}$ does. So, if we denote the externally applied force as W, the equation of motion for a neutral test particle becomes

$$m\dot{v} = [W - m\omega(v)]\cdot v. \tag{7.45}$$

Now, suppose that W is chosen so that the particle remains at a fixed distance a outside the horizon of a black hole. The equation for the trajectory is

$$x(\tau) = t(\tau)e_t + ae_r(\theta_0, \phi_0), \tag{7.46}$$

where the constants θ_0 and ϕ_0 specify the angular position of the particle. The covariant velocity is

$$v = \dot{t}\underline{h}^{-1}(e_t) \tag{7.47}$$

and the condition that $v^2 = 1$ forces

$$\dot{t}^2(1 - 2M/a) = 1, \tag{7.48}$$

so that $\dot{t}$, and hence v, are constant. Equation (7.45) now reduces to

$$W - m\omega(v) = 0, \tag{7.49}$$

$$\Rightarrow \quad W = m\dot{\Omega}(e_t) = \frac{Mm}{a^2}\left(1 - \frac{2M}{a}\right)^{-1/2}\sigma_a, \tag{7.50}$$

where σ_a is the unit outward spatial vector from the source to the particle. The magnitude of the force is therefore

$$|W| = \frac{Mm}{a^2}\left(1 - \frac{2M}{a}\right)^{-1/2} \tag{7.51}$$

and it is not hard to check that this result is gauge invariant. Equation (7.51) reduces to the Newtonian formula at large distances and becomes singular as the horizon is approached, where an infinite force is required to remain at rest.

We now want to see how this expression for the force is modified if the particle is charged and feels a force due to its own polarization field. From (7.44) the extra term in the force is simply $\mathcal{F}$ and only the radial term contributes. From equation (7.40) this is just $-\partial_r V\sigma_r$. Since the charge lies on the z-axis, we need only look at V along this axis, for which

$$V(z) = \frac{q}{|z-a|} - \frac{qM}{|z-a|}\left(\frac{1}{a} + \frac{1}{z}\right) + \frac{qM}{az}. \tag{7.52}$$

The singular terms must be due to the particle's own Coulomb field and so do not generate a polarization force. The only term which generates a force is therefore the final one, which is precisely the term that Linet added to Copson's original formula! This term produces an outward-directed force on the charge, of magnitude q^2M/a^3. The applied force is therefore now

$$W = \left(\frac{Mm}{a^2}\left(1 - \frac{2M}{a}\right)^{-1/2} - \frac{Mq^2}{a^3}\right)\sigma_a, \tag{7.53}$$

so a smaller force is needed to keep the particle at rest outside the horizon. This result agrees with that found in Smith & Will (1980), though our derivation avoids the need for infinite mass renormalization and is considerably simpler. This result is a good example of the importance of finding global solutions. The polarization force is felt outside the horizon, yet the correction term that led to it was motivated by the properties of the field at the origin.

(*d*) *Point charge in a* $k > 0$ *cosmology*

We saw in § 6 *f* that one form of the $\overline{h}$ function for a homogeneous cosmology is defined by (6.170)

$$\overline{h}(a) = a\cdot e_t e_t + \alpha[(1-kr^2)^{1/2}a\cdot e_r e^r + a\wedge\sigma_r\,\sigma_r], \tag{7.54}$$

where $\alpha = 1/S$. When k is positive, however, the function (7.54) is undefined for $r > k^{-1/2}$ and so fails to define a global solution. A globally valid solution is obtained with the displacement

$$f(x) = x\cdot e_t e_t + \frac{r}{1 + \frac{1}{4}kr^2}e_r, \tag{7.55}$$

which results in the particularly simple solution

$$\overline{h}'(a) = a\cdot e_t e_t + \alpha(1 + \tfrac{1}{4}kr^2)a\wedge e_t e_t. \tag{7.56}$$

This solution is well defined for all x and generates an 'isotropic' line element. (The solution can also be viewed as resulting from a stereographic projection of a 3-sphere.)

We want to find the fields due to a point charge in the background defined by (7.56). Since the $\overline{h}$ function is diagonal, we start with the obvious ansatz

$$A = \alpha V(\boldsymbol{x}) e_t, \tag{7.57}$$

so that

$$\boldsymbol{E} = -\alpha \boldsymbol{\nabla} V \tag{7.58}$$

and

$$\boldsymbol{D} = -[1 + kr^2/4]^{-1} \boldsymbol{\nabla} V. \tag{7.59}$$

It follows that the equation we need to solve is simply

$$-\boldsymbol{\nabla}\cdot([1 + \tfrac{1}{4}kr^2]^{-1}\boldsymbol{\nabla} V) = q\delta(\boldsymbol{x} - \boldsymbol{a}), \tag{7.60}$$

where the charge q is located at $\boldsymbol{x} = \boldsymbol{a}$. This equation can be solved using the general technique described by Hadamard (see also Copson (1928) for a discussion of a similar problem.) The solution to (7.60) turns out to be

$$V = \frac{1 + \frac{1}{4}ka^2 - \frac{1}{2}ks}{[s(1 + \frac{1}{4}ka^2 - \frac{1}{4}ks)]^{1/2}}, \tag{7.61}$$

where

$$s = \frac{(\boldsymbol{x} - \boldsymbol{a})^2}{1 + \frac{1}{4}kr^2}, \quad a = \sqrt{\boldsymbol{a}^2}. \tag{7.62}$$

As a simple check, V reduces to the usual Coulomb potential when $k = 0$.

Fieldline plots of the $\boldsymbol{D}$ field defined from (7.61) are shown in figure 6. The fieldlines follow null geodesics and clearly reveal the existence of an image charge. The reason for this can be seen in the denominator of V. Not only is V singular at $s = 0$, it is also singular at

$$1 + \tfrac{1}{4}ka^2 - \tfrac{1}{4}ks = 0. \tag{7.63}$$

So, if the charge lies on the z-axis, the position of its image is found by solving

$$\frac{k(z - a)^2}{4 + kz^2} = 1 + \tfrac{1}{4}ka^2, \tag{7.64}$$

from which we find that the image charge is located at

$$z = -4/(ka). \tag{7.65}$$

(This is the stereographic projection of the opposite point to the charge on a 3-sphere.) However, if we try to remove this image charge by adding a point source at its position, we find that the fields vanish everywhere, since the new source has its own image which cancels the original charge. The image charge is therefore an unavoidable feature of $k > 1$ cosmologies. We comment on this further in § 9.

8. The Dirac equation in a gravitational background

In this paper we began by introducing the gravitational gauge fields and minimally coupling these to the Dirac action. For our final major application we return to the Dirac equation in a gravitational background. In the absence of an $\mathcal{A}$ field, the minimally coupled equation (3.62) is

$$D\psi i\sigma_3 = m\psi\gamma_0. \tag{8.1}$$

Here we consider two applications: the case of a black hole background; and cosmological models in which the universe is not at critical density.

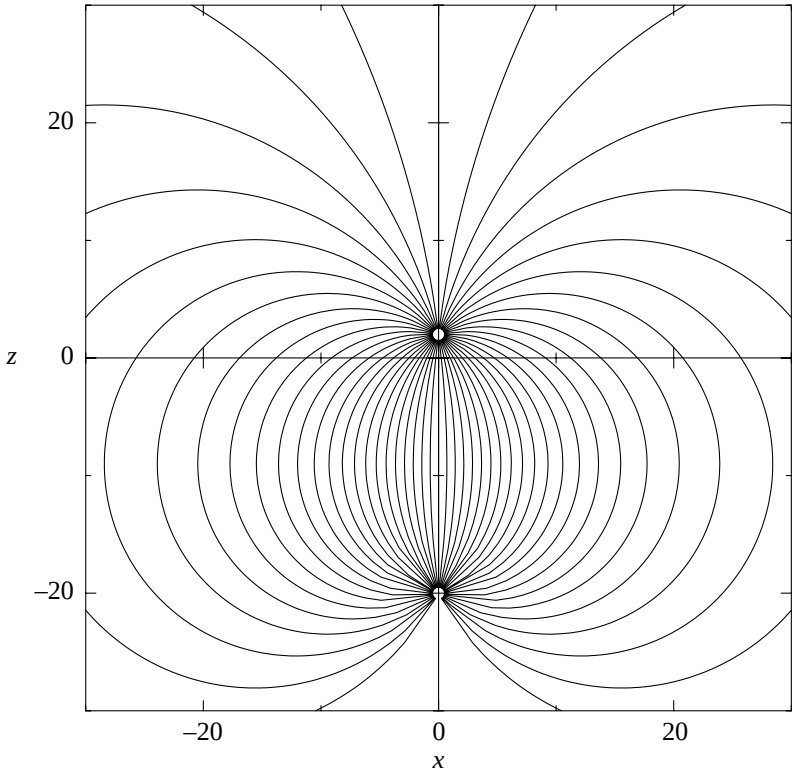

Figure 6. Fieldlines of the $\boldsymbol{D}$ field for a point charge in a $k > 0$ universe. The fieldlines follow null geodesics, which are circles in this gauge. The existence of an image charge is clear.

(*a*) *Black hole background*

We have already demonstrated that the Newtonian gauge solution dramatically simplifies the study of black hole physics, so this is the natural gauge in which to study the Dirac equation. We therefore start with an analysis in this gauge and then consider the gauge invariance of the predictions made. With the gravitational fields as described in § 6 *d*, the Dirac equation (8.1) becomes

$$\nabla\psi\mathrm{i}\sigma_3 - (2M/r)^{1/2}\gamma_0(\partial_r\psi + 3/(4r)\psi)\mathrm{i}\sigma_3 = m\psi\gamma_0. \tag{8.2}$$

If we premultiply by γ_0 and introduce the symbol j to represent right-sided multiplication by $\mathrm{i}\sigma_3$, so that $\mathrm{j}\psi \equiv \psi\mathrm{i}\sigma_3$, then equation (8.2) becomes

$$\mathrm{j}\partial_t\psi = -\mathrm{j}\boldsymbol{\nabla}\psi + \mathrm{j}(2M/r)^{1/2}r^{-3/4}\partial_r(r^{3/4}\psi) + m\bar{\psi}, \tag{8.3}$$

where $\bar{\psi} \equiv \gamma_0\psi\gamma_0$. One feature that emerges immediately is that the Newtonian gauge has recovered a Hamiltonian form of the Dirac equation (see Doran *et al.* (1996*b*) for a discussion of operators, Hamiltonians and Hermiticity in the STA approach to Dirac theory). Since the Newtonian gauge involves the notion of a global time, it might have been expected that this gauge would lend itself naturally to a Hamiltonian formulation. A question of some interest is whether it is always possible to make such a gauge choice and we hope to address this in the near future.

The Hamiltonian (8.3) contains a subtlety: it is Hermitian only away from the origin. To see why, consider the interaction term

$$H_{\mathrm{I}}(\psi) = \mathrm{j}(2M/r)^{1/2}r^{-3/4}\partial_r(r^{3/4}\psi). \tag{8.4}$$

For this we find that

$$\int \mathrm{d}^3x\,\langle\phi^\dagger H_{\mathrm{I}}(\psi)\rangle_S = \sqrt{2M}\int \mathrm{d}\Omega\int_0^\infty r^2\,\mathrm{d}r\,r^{-5/4}\langle\phi^\dagger\partial_r(r^{3/4}\psi)\mathrm{i}\sigma_3\rangle_S$$
$$= \sqrt{2M}\int \mathrm{d}\Omega\int_0^\infty \mathrm{d}r\,\langle r^{3/4}\phi^\dagger\partial_r(r^{3/4}\psi)\mathrm{i}\sigma_3\rangle_S$$
$$= \int \mathrm{d}^3x\,\langle(H_{\mathrm{I}}(\phi)^\dagger\psi\rangle_S + \sqrt{2M}\int \mathrm{d}\Omega\,[r^{3/2}\langle\phi^\dagger\psi\mathrm{i}\sigma_3\rangle_S]_0^\infty, \quad (8.5)$$

where $\langle\,\rangle_S$ denotes the projection onto the 'complex' 1 and $\mathrm{i}\sigma_3$ terms and $\phi^\dagger \equiv \gamma_0\tilde{\phi}\gamma_0$ (see Appendix A and Doran *et al.* 1996*b*). For all normalized states the final term in (8.5) tends to zero as $r\to\infty$. But it can be shown from (8.3) that wave functions tend to the origin as $r^{-3/4}$, so the lower limit is finite and H_{I} is therefore not (quite) a Hermitian operator. This immediately rules out the existence of normalizable stationary states with constant real energy.

Equation (8.3) can be used to propagate a spinor defined over some initial spatial surface and numerical simulations based on this equation give a good picture of the scattering induced by a black hole. Here, however, we wish to focus on an analytical approach. A result that follows immediately from the Hamiltonian form of the Dirac equation is that (8.3) is manifestly separable, so that we can write

$$\psi(x) = \alpha(t)\psi(\boldsymbol{x}). \quad (8.6)$$

As usual, the solution of the t equation is

$$\alpha(t) = \exp(-\mathrm{i}\sigma_3 Et), \quad (8.7)$$

where E is the separation constant. The non-Hermiticity of H_{I} means that E cannot be purely real if ψ is normalizable. The imaginary part of E is determined by equation (8.5) and, for suitably normalized states, we find that

$$\mathrm{Im}(E) = -\lim_{r\to 0} 2\pi\sqrt{2M}\langle\psi^\dagger\psi\rangle r^{3/2}. \quad (8.8)$$

This equation shows that the imaginary part of E is necessarily negative, so the wave function decays with time. This is consistent with the fact that the streamlines generated by the conserved current $\psi\gamma_0\tilde{\psi}$ are timelike curves and, once inside the horizon, must ultimately terminate on the origin.

With the t dependence separated out, equation (8.3) reduces to

$$\boldsymbol{\nabla}\psi - (2M/r)^{1/2}r^{-3/4}\partial_r(r^{3/4}\psi) = \mathrm{j}E\psi - \mathrm{j}m\bar{\psi}. \quad (8.9)$$

To solve this equation we next separate out the angular dependence. This is achieved using the spherical monogenics, which are the STA equivalent of the χ_{jlm} Pauli spinors. These are described in detail in Doran *et al.* (1996*b*) and here we quote the necessary formulae. The unnormalized monogenic ψ_l^m is defined by

$$\psi_l^m = [(l+m+1)P_l^m(\cos\theta) - P_l^{m+1}(\cos\theta)\mathrm{i}\sigma_\phi]\mathrm{e}^{m\phi\mathrm{i}\sigma_3}, \quad (8.10)$$

where $l\geqslant 0$, $-(l+1)\leqslant m\leqslant l$, and P_l^m are the associated Legendre polynomials. The two properties of the ψ_l^m relevant here are

$$\boldsymbol{\nabla}\psi_l^m = -(l/r)\sigma_r\psi_l^m \quad (8.11)$$

and

$$\boldsymbol{\nabla}(\sigma_r\psi_l^m) = (l+2)/r\psi_l^m. \quad (8.12)$$

Now, the operator $K(\psi) = \gamma_0(1-\boldsymbol{x}\wedge\boldsymbol{\nabla})\psi\gamma_0$ commutes with the Hamiltonian defined

by (8.3). Constructing eigenstates of this operator with eigenvalue κ, leads to solutions of the form

$$\psi(\boldsymbol{x},\kappa)=\begin{cases}\psi_l^m u(r)+\sigma_r\psi_l^m v(r)\mathrm{i}\sigma_3, & \kappa=l+1,\\ \sigma_r\psi_l^m u(r)\sigma_3+\psi_l^m \mathrm{i}v(r), & \kappa=-(l+1),\end{cases} \tag{8.13}$$

where κ is a non-zero integer and $u(r)$ and $v(r)$ are complex functions of r (i.e. sums of a scalar and an $\mathrm{i}\sigma_3$ term). Substituting (8.13) into equation (8.9), and using the properties of the spherical monogenics, we arrive at the coupled radial equations

$$\begin{pmatrix}1 & -(2M/r)^{1/2}\\ -(2M/r)^{1/2} & 1\end{pmatrix}\begin{pmatrix}u_1'\\ u_2'\end{pmatrix}=\boldsymbol{A}\begin{pmatrix}u_1\\ u_2\end{pmatrix}, \tag{8.14}$$

where

$$\boldsymbol{A}\equiv\begin{pmatrix}\kappa/r & \mathrm{j}(E+m)-(2M/r)^{1/2}(4r)^{-1}\\ \mathrm{j}(E-m)-(2M/r)^{1/2}(4r)^{-1} & -\kappa/r\end{pmatrix}, \tag{8.15}$$

u_1 and u_2 are the reduced functions defined by

$$u_1=ru,\quad u_2=\mathrm{j}rv \tag{8.16}$$

and the primes denote differentiation with respect to r. (We continue to employ the abbreviation j for $i\sigma_3$.)

To analyse (8.14) we first rewrite it in the equivalent form

$$(1-2M/r)\begin{pmatrix}u_1'\\ u_2'\end{pmatrix}=\begin{pmatrix}1 & (2M/r)^{1/2}\\ (2M/r)^{1/2} & 1\end{pmatrix}\boldsymbol{A}\begin{pmatrix}u_1\\ u_2\end{pmatrix}. \tag{8.17}$$

This makes it clear that the equations have regular singular points at the origin and horizon ($r=2M$), as well as an irregular singular point at $r=\infty$. To our knowledge, the special function theory required to deal with such equations has not been developed. Without it we either attempt a numerical solution, or look for power series with a limited radius of convergence. Here we consider the latter approach and look for power-series solutions around the horizon. To this end we introduce the series

$$u_1=\eta^s\sum_{k=0}^{\infty}a_k\eta^k,\quad u_2=\eta^s\sum_{k=0}^{\infty}b_k\eta^k, \tag{8.18}$$

where $\eta=r-2M$. The index s controls the radial dependence of ψ at the horizon, so represents a physical quantity. To find the values that s can take, we substitute (8.18) into (8.17) and set $\eta=0$. This results in the equation

$$\frac{s}{2M}\begin{pmatrix}a_0\\ b_0\end{pmatrix}=\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}\begin{pmatrix}\kappa/(2M) & \mathrm{j}(E+m)-(8M)^{-1}\\ \mathrm{j}(E-m)-(8M)^{-1} & -\kappa/(2M)\end{pmatrix}\begin{pmatrix}a_0\\ b_0\end{pmatrix}. \tag{8.19}$$

Rewriting this in terms of a single matrix and setting its determinant to zero yields the two indicial roots

$$s=0\quad\text{and}\quad s=-\tfrac{1}{2}+4\mathrm{j}ME. \tag{8.20}$$

The $s=0$ solution is entirely sensible—the power series is analytic and nothing peculiar happens at the horizon. The existence of this root agrees with our earlier observation that one can evolve the time-dependent equations without encountering

any difficulties at the horizon. The second root is more problematic, as it leads to solutions which are ill defined at the horizon. Before assessing the physical content of these roots, however, we must first check that they are gauge invariant.

If, instead of working in the Newtonian gauge, we keep the gauge unspecified then, after separating out the angular dependence, the equations reduce to

$$\begin{pmatrix} L_r & L_t \\ L_t & L_r \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \kappa/r - \frac{1}{2}G & \mathrm{j}m - \frac{1}{2}F \\ -\mathrm{j}m - \frac{1}{2}F & -\kappa/r - \frac{1}{2}G \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{8.21}$$

We can still assume that the t dependence is of the form $\exp(-\mathrm{j}Et)$, so that equation (8.21) becomes

$$\begin{pmatrix} g_1 & g_2 \\ g_2 & g_1 \end{pmatrix} \begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \boldsymbol{B} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \tag{8.22}$$

where

$$\boldsymbol{B} \equiv \begin{pmatrix} \kappa/r - \frac{1}{2}G + \mathrm{j}f_2E & \mathrm{j}(m + f_1E) - \frac{1}{2}F \\ -\mathrm{j}(m - f_1E) - \frac{1}{2}F & -\kappa/r - \frac{1}{2}G + \mathrm{j}f_2E \end{pmatrix}. \tag{8.23}$$

Now, since ${g_1}^2 - {g_2}^2 = 1 - 2M/r$ holds in all gauges, we obtain

$$(1 - 2M/r) \begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} g_1 & -g_2 \\ -g_2 & g_1 \end{pmatrix} \boldsymbol{B} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \tag{8.24}$$

and substituting in the power series (8.18) and setting $\eta = 0$ produces the indicial equation

$$\det \left[\begin{pmatrix} g_1 & -g_2 \\ -g_2 & g_1 \end{pmatrix} \boldsymbol{B} - \frac{s}{r} I \right]_{r=2M} = 0, \tag{8.25}$$

where I is the identity matrix. Employing the result that

$$g_1G - g_2F = \tfrac{1}{2}\partial_r({g_1}^2 - {g_2}^2) = M/r^2, \tag{8.26}$$

we find that the solutions to the indicial equation are now

$$s = 0 \quad \text{and} \quad s = -\tfrac{1}{2} + 4\mathrm{j}ME(g_1f_2 - g_2f_1). \tag{8.27}$$

However, in § 6 *d* we established that $g_1f_2 - g_2f_1 = +1$ at the horizon for all solutions with a forward time direction. This demonstrates that the indices are indeed intrinsic, with the sign of the imaginary term for the singular root picking up information about the time direction implicit in the presence of the horizon.

The fact that $s = 0$ is always a solution of the indicial equation means that solutions always exist which are analytic at the horizon. Determining the split between ingoing and outgoing states of these solutions enables one to calculate reflection coefficients and scattering amplitudes. The question we wish to consider is whether the second singular root can be physically significant. To address this we look at the current. The covariant current $\mathcal{J}$ is given by $\psi\gamma_0\tilde{\psi}$ and satisfies $\mathcal{D}\cdot\mathcal{J} = 0$. The corresponding non-covariant quantity is therefore

$$J = \underline{h}(\psi\gamma_0\tilde{\psi}) \det(\underline{h})^{-1}, \tag{8.28}$$

which satisfies the flatspace conservation equation $\nabla\cdot J = 0$. It is the streamlines of J that are plotted as functions of x and determine the flow of density. The crucial

terms in J are the time component and radial component, which (ignoring the overall exponential decay term) are given by

$$\gamma_0 \cdot J = \frac{1}{r^2}[f_1(u_1\tilde{u}_1 + u_2\tilde{u}_2) + f_2(u_1\tilde{u}_2 + \tilde{u}_1 u_2)]\psi_l^m \tilde{\psi}_l^m \tag{8.29}$$

and

$$e^r \cdot J = \frac{1}{r^2}[g_1(u_1\tilde{u}_2 + \tilde{u}_1 u_2) + g_2(u_1\tilde{u}_1 + u_2\tilde{u}_2)]\psi_l^m \tilde{\psi}_l^m. \tag{8.30}$$

The $\{f_i\}$ and $\{g_i\}$ are finite for all admissible solutions so, for the $s = 0$ solution, the components of J are well defined at the horizon. Furthermore, it is easily shown that for $s = 0$ the radial flux at the horizon always points inwards. The $s = 0$ root therefore describes the case where the flux crosses the horizon and continues onto the singularity.

For the singular root we must first decide on a branch for the solution so that ψ is fully specified on both sides of the horizon. We can then assess whether the discontinuity in ψ, and the discontinuity in the current generated by it, are physically acceptable. To do this, we first write η^s as

$$\eta^s = \exp\{(-\tfrac{1}{2} + 4\mathrm{j}ME)\ln(r - 2M)\}. \tag{8.31}$$

We can now write

$$\ln(r - 2M) = \ln|r - 2M| + \mathrm{j}\arg(r - 2M) \tag{8.32}$$

and for the choice of argument we take

$$\arg(r - 2M) = \begin{cases} 0, & r > 2M, \\ -\pi, & r < 2M. \end{cases} \tag{8.33}$$

(The choice of sign will be discussed further below.) If we now take the limit $r \to 2M$ from above and below we find that the γ_0 component of J is given by

$$\gamma_0 \cdot J = A(\theta, \phi)e^{-2\epsilon t}|r - 2M|^{-1+8M\epsilon} \times \begin{cases} 1, & r > 2M, \\ \exp\{8\pi M E_r\}, & r < 2M, \end{cases} \tag{8.34}$$

where $A(\theta, \phi)$ is a positive-definite finite term and we have split E into real and imaginary parts as

$$E = E_r - \mathrm{j}\epsilon. \tag{8.35}$$

Equation (8.34) is valid in all gauges for which $g_1 f_2 - g_2 f_1 = +1$ at the horizon. While the density $\gamma_0 \cdot J$ is singular at the horizon, the presence of the positive term $8M\epsilon$ ensures that any integral over the horizon is finite and the solution is therefore normalizable. This link between the properties of ψ at the horizon and at the origin (where ϵ is determined) provides another example of the importance of finding global solutions to the field equations. The radial current now turns out to be

$$e^r \cdot J = \frac{A(\theta, \phi)}{4M} e^{-2\epsilon t}|r - 2M|^{8M\epsilon} \times \begin{cases} 1, & r > 2M, \\ -\exp\{8\pi M E_r\}, & r < 2M \end{cases} \tag{8.36}$$

and is therefore zero at the horizon and inward pointing everywhere inside the horizon. It appears that the existence of the imaginary contribution to E does ensure that that the singular solutions have sensible physical properties and the singularity in ψ at the horizon is no worse than that encountered in the ground state of the hydrogen

atom (Martellini & Treves 1977). What is less clear, however, is the extent to which the properties of ψ at the horizon are compatible with the original equation (8.1). In particular, since ψ is both singular and non-differentiable at the horizon, it does not appear that the singular root can be viewed as defining a solution of (8.1) over all space.

(*b*) *The Hawking temperature*

A number of authors have attempted to give derivations of the Hawking temperature and particle flux due to a black hole from an analysis of first-quantized theory, i.e. from the properties of wave equations alone (Damour & Ruffini 1976; Zhao Zheng *et al.* 1981, 1982). This work has generated some controversy (Martellini & Treves 1977), so it is interesting to assess how the ideas stand up in GTG. These derivations focus on the singular solutions to the wave equation (either Klein–Gordon or Dirac) and study the properties of these solutions under the assumption that the energy is real. If one ignores the problems that $\epsilon = 0$ introduces for the normalizability of ψ and presses ahead, then from (8.36) there is now a non-zero current at the horizon and, furthermore, there is a net creation of flux there. The ratio of the outward flux to the total flux is simply

$$\frac{e^r \cdot J_+}{e^r \cdot J_+ - e^r \cdot J_-} = \frac{1}{e^{8\pi ME} + 1}, \tag{8.37}$$

which defines a Fermi–Dirac distribution with temperature

$$T = \frac{1}{8\pi M k_B}. \tag{8.38}$$

Remarkably, this is the temperature found by Hawking (1974). The fact that both the correct Fermi–Dirac statistics and Hawking temperature can be derived in this manner is astonishing, since both are thought to be the result of quantum field theory. But what can we really make of this derivation? The first problem is that setting ϵ to zero means that the density is no longer normalizable at the horizon—any integral of the density over the horizon region diverges logarithmically, which is clearly unphysical. A further problem relates to the choice of branch (8.33). Had the opposite branch been chosen we would not have obtained (8.37) and, as pointed out in Martellini & Treves (1977), there is no *a priori* justification for the choice made in (8.33).

For the above reasons, the 'derivation' of (8.37) cannot be viewed as being sound. The remarkable thing is that the same techniques can be used to 'derive' the correct temperatures for the Reissner–Nordström and Rindler cases, as well as the Schwinger production rate in a constant electric field. This will be discussed elsewhere. These further analyses contain another surprise: the temperature at the interior horizon of a Reissner–Nordström black hole is necessarily negative! However, while these analyses are both interesting and suggestive, it is only through a study of the full quantum field theory in a black hole background that one can be sure about particle production rates. This too will be treated elsewhere.

A final point in this section is that all our analyses have been based on working with the correct time-asymmetric solutions admitted in GTG. On attempting to force through the analysis in the 'Schwarzschild' gauge ($g_2 = f_2 = 0$), one discovers that the indices are now given by

$$s = -\tfrac{1}{2} \pm 4\mathrm{j}ME. \tag{8.39}$$

In this case, no analytic solution is possible and even the presence of an exponential damping factor does not produce a normalizable current at the horizon. This only serves to reinforce the importance of working with global solutions, since there is no doubt that the presence of non-singular normalizable solutions is an intrinsic feature of horizons.

(*c*) *The Dirac equation in a cosmological background*

As a second application we consider the Dirac equation in a cosmological background. We have a choice of form of $\overline{h}$ function to use, of which the simplest is that defined by (6.170),

$$\overline{h}'(a) = a\cdot e_t e_t + (1/S)[(1-kr^2)^{1/2} a\cdot e_r e^r + a\wedge\sigma_r\sigma_r]. \tag{8.40}$$

The Dirac equation in this background takes the form

$$(e^t\partial_t + (1/S)[(1-kr^2)^{1/2}e^r\partial_r + e^\theta\partial_\theta + e^\phi\partial_\phi])\psi \mathrm{i}\sigma_3 + \tfrac{1}{2}\left(3H(t)e_t - \frac{2}{rS}[(1-kr^2)^{1/2}-1]e_r\right)\psi\mathrm{i}\sigma_3 = m\psi\gamma_0, \tag{8.41}$$

where the various functions are as defined in § 6 *f*. Our question is this: can we find solutions to (8.41) such that the observables are homogeneous? There is clearly no difficulty if $k=0$ since, with $p=0$, equation (8.41) is solved by

$$\psi = \rho^{1/2}\mathrm{e}^{-\mathrm{i}\sigma_3 mt} \tag{8.42}$$

and the observables are fixed vectors which scale as $\rho(t)$ in magnitude (Lasenby *et al.* 1993*d*). But what happens when $k\neq 0$? It turns out that the solution (8.42) must be modified to (Lasenby *et al.* 1993*d*)

$$\psi = \frac{\rho^{1/2}}{1+\sqrt{1-kr^2}}\mathrm{e}^{-\mathrm{i}\sigma_3 mt}. \tag{8.43}$$

For the case of $k>0$, both $\overline{h}(a)$ and ψ are only defined for $r<k^{-1/2}$. This problem is overcome by using the displacement (7.55) to transform to the global solution of equation (7.56). In this case ψ is given by

$$\psi = \tfrac{1}{2}(1+\tfrac{1}{4}kr^2)\rho^{1/2}\mathrm{e}^{-\mathrm{i}\sigma_3 mt}, \tag{8.44}$$

which diverges as $r\to\infty$.

In both the $k>0$ and $k<0$ cases, ψ contains additional r dependence and so is not homogeneous. Furthermore, the observables obtained from ψ are also inhomogeneous. Unlike the case of a constant magnetic field (Itzykson & Zuber 1980), the centre of inhomogeneity is fixed at $r=0$ and no stationary state solutions to (8.41) exist centred on any other point. In principle one could therefore determine the origin of this space from measurements of the current density. This clearly violates the principle of homogeneity, though it is not necessarily inconsistent with experiment. The implications for cosmology of this fact are discussed in the following section. (Some consequences for self-consistent solutions of the Einstein–Dirac equations are discussed in Lasenby *et al.* (1993*d*), Challinor *et al.* (1997) and, in the context of GR, in Isham & Nelson (1974).)

The fact that quantum fields see this 'preferred' direction in $k\neq 0$ models, whereas classical phenomena do not, reflects the gauge structure of the theory. Dirac spinors are the only fields whose action couples them directly to the $\omega(a)$ function. All other matter fields couple to the gravitational field through the $\overline{h}$ field only. Dirac spinors

therefore probe the structure of the gravitational fields directly at level of the ω field, which is inhomogeneous for $k \neq 0$ models. On the other hand, classical fields only interact via the covariant quantities obtained from the gravitational fields, which are homogeneous for all values of k. This conclusion is reinforced by the fact that the Klein–Gordon equation, for which the action does not contain the $\omega(a)$ field, *does* have homogenous solutions in a $k \neq 0$ universe.

9. Implications for cosmology

In § 6 *f*, we discussed some aspects of cosmology as examples of the general treatment of time-varying spherically symmetric systems. There we drew attention to the utility of the Newtonian gauge as a tool for tackling problems in cosmology. In addition, in §§ 7 *d* and 8 *c* we studied the Maxwell and Dirac equations in various cosmological backgrounds. In this section we draw together some of our conclusions from these sections. Specifically, we discuss redshifts, difficulties with $k \neq 0$ models and the definitions of mass and energy for cosmological models.

(*a*) *Cosmological redshifts*

As a final demonstration of the use of the Newtonian gauge, consider a photon following a null path in the $\theta = \frac{1}{2}\pi$ plane. In this case the photon's momentum can be written as

$$P = \Phi R(\gamma_0 + \gamma_1)\tilde{R}, \tag{9.1}$$

where

$$R = \mathrm{e}^{\alpha \mathrm{i}\sigma_3/2} \tag{9.2}$$

and Φ is the frequency measured by observers comoving with the fluid. We restrict to the pressureless case, so $G = 0$ and $f_1 = 1$, but will allow ρ to be r dependent. A simple application of equation (4.61) produces

$$\partial_\tau \Phi = -\Phi^2 \left(\frac{g_2}{r} \sin^2 \chi + \partial_r g_2 \cos^2 \chi \right), \tag{9.3}$$

where $\chi = \phi + \alpha$. However, since $f_1 = 1$, we find that $\partial_\tau t = \Phi$, so

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -\Phi \left(\frac{g_2}{r} \sin^2 \chi + \partial_r g_2 \cos^2 \chi \right), \tag{9.4}$$

which holds in any spherically symmetric pressureless fluid.

For the case of a cosmological background we have $g_2 = H(t)r$, so the angular terms drop out of equation (9.4) and we are left with the simple equation

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -H(t)\,\Phi = \frac{\dot{\rho}}{3\rho}, \tag{9.5}$$

which integrates to give the familiar redshift versus density relation

$$1 + z = (\rho_1/\rho_0)^{1/3}. \tag{9.6}$$

Other standard cosmological relations, such as the luminosity distance and angular diameter versus redshift formulae, can be easily derived in this gauge (see also § 6 *f*).

In Lasenby *et al.* (1993*d*), equation (9.6) was derived in the gauge of equation (6.170), in which all particles comoving with the cosmological fluid are at rest in the background spacetime. In this gauge the redshift can be attributed to a loss of energy to the gravitational field, although this is a gauge-dependent viewpoint—the

only physical statement that one can make is embodied in equation (9.6). The explanation of cosmological redshifts in our theory has nothing to do with 'tired light', or spacetime playing a dynamic role by expanding, or even anything to do with Doppler shifts. The redshift is simply a consequence of the assumption of homogeneity. Ultimately, all physical predictions are independent of the gauge in which they are made, although certain gauges may have useful computational or heuristic value.

(*b*) $k \neq 0$ *cosmologies*

At the level of classical (i.e. non-quantum) physics, there is no doubt that $k \neq 0$ cosmologies are homogeneous. This is true in both GR and GTG. No prediction derived for classical systems of point particles or for electromagnetic fields can reveal a preferred spatial direction in these models. However, we saw in § 8 *c* that it is impossible to find homogeneous solutions of the Dirac equation in a $k \neq 0$ universe. The consequences of this are, in principle, observable, since the local density gradient will reveal a preferred radial direction. It has already been pointed out that it is impossible to find a *self-consistent* solution of the combined system of Dirac–Einstein equations for any case other than a spatially flat cosmology (Lasenby *et al.* 1993*d*; Isham & Nelson 1974). We believe that this is the first time that it has been pointed out that even a *non*-self-consistent Dirac field would be observably inhomogeneous. This is a more damaging result for $k \neq 0$ cosmologies, since it reveals inhomogeneity without assuming that spin–torsion effects have anything to do with the dynamics of the universe.

While the properties of Dirac fields pose theoretical difficulties for $k \neq 0$ models, there is no contradiction with present observations. Furthermore, one could question the validity of inferences drawn in extrapolating the Dirac equation to cosmological scales. There is, however, a purely classical effect which does lead one to question the validity of $k > 0$ models. As we have seen, when looking at the properties of fields in a $k > 0$ background, it is necessary to work in a globally defined gauge, such as that of equation (7.56). In this case the Maxwell equations show that each point charge must have an image charge present in a remote region of the universe. This is a consequence of a closed universe that we have not seen discussed, although it has doubtless been pointed out before. The necessity for this image charge raises many problems in attempting to take such a universe seriously.

(*c*) *Mass and energy in cosmological models*

Setting aside the problems with $k \neq 0$ models, a further issue on which our theory sheds some light is discussions of the total matter and energy content of the universe. In § 6 *e* we discussed the distinction between the rest-mass energy and the total gravitating energy inside a sphere of radius r. Since cosmological models are a special case of the general theory outlined in § 6, this same distinction should be significant in cosmology.

In § 6 *e* we identified the total gravitating energy of a sphere of radius r with the function $M(r,t)$. For all cosmological models, this is given by (6.155)

$$M(r,t) = \tfrac{4}{3}\pi r^3 \rho(t). \tag{9.7}$$

In strictly homogeneous models, the Weyl tensor vanishes and we lose an intrinsically defined distance scale. But, if we consider cosmological models as the limiting case of spherically symmetric systems, then there seems little doubt that (9.7) is still the correct expression for the gravitating energy within a sphere of radius r surrounding

the origin. Moreover, attempts to discover the gravitating content of a region rely on perturbations away from ideal uniformity. In these cases an intrinsic distance scale is well defined, since a Weyl tensor is again present. (Determining the gravitating content of a region is important in, for example, determinations of Ω—the ratio of the actual density of the universe to the critical density.) On the other hand, the total rest mass energy within a sphere of radius r centred on the origin must still be given by (6.133)

$$\mu(r,t) = \int_0^r 4\pi s^2 \rho(t)\, \frac{\mathrm{d}s}{g_1}. \tag{9.8}$$

This remains a covariant scalar quantity and is just $\rho(t)$ multiplied by the covariant volume integral (this is the volume one would measure locally using light paths or fixed rods).

We have now defined two covariant scalar quantities, $M(r,t)$ and $\mu(r,t)$, both of which are conserved along fluid streamlines in the absence of pressure. If the identifications made in the spherically symmetric case remain valid in the homogeneous case, then the difference between these should give the additional contribution to the total energy beyond the rest-mass energy. For spatially flat universes we have $g_1 = 1$ so there is no difference. (In terms of the Newtonian-gauge description of § 6 *e*, the gravitational potential energy cancels the kinetic energy.) However, for $k \neq 0$ models, there is a difference because $\mu(r,t)$ is now given by

$$\mu(r) = \int_0^r \frac{4\pi s^2 \rho(s)}{(1 - k s^2 \rho^{2/3})^{1/2}}\, \mathrm{d}s. \tag{9.9}$$

An interesting place to study the difference between M and μ is in a $k > 0$ universe at its 'turnaround' point, as described in § 6 *f*. There one finds that, to lowest order in r, the difference is given by

$$M(r) - \mu(r) \approx -\frac{3M(r)^2}{5r}, \tag{9.10}$$

which is precisely the Newtonian formula for the self-potential of a uniform sphere of mass $M(r)$. This is what we would have expected since, at the turnaround point, the kinetic energy vanishes.

The above should only be viewed as suggestive, but one idea that it appears to rule out is the popular suggestion that the total energy density of the universe should be zero (Tryon 1973, 1984). If the above analysis is correct then there is no possibility of the total energy density $M(r,t)$ ever being zero. Furthermore, for spatially flat models—which we consider the most likely—the total energy density resides entirely in the rest-mass energies of the particles in the universe and cannot be cancelled by a negative gravitational contribution.

10. Conclusions

In this paper we developed a theory of gravity consisting of gauge fields defined in a flat background spacetime. The theory is conceptually simple and the role of the gauge fields is clearly understood—they ensure invariance under arbitrary displacements and rotations. Whilst it is possible to maintain a classical picture of the rotation gauge group, a full understanding of its role is only achieved once the Dirac action is considered. The result is a theory which offers a radically different interpretation of gravitational interactions from that provided by GR. Despite this,

the two theories agree in their predictions for a wide range of phenomena. Differences only begin to emerge over issues such as the role of topology, our insistence on the use of global solutions and in the interaction with quantum theory. Furthermore, the separation of the gauge fields into one for displacements and one for local rotations is suggestive of physical effects being separated into an inertia field and a force field. Indeed, there is good reason to believe that mass should enter relativistic multiparticle wave equations in the manner of the $\overline{h}$ field (Doran *et al.* 1996*b*). It is possible that, in the development of a multiparticle theory, the $\overline{h}(a)$ and $\Omega(a)$ fields will be extended in quite distinct ways. Such possibilities do not appear to be open to a GR-type theory, with its reliance on the metric as the 'foundation of all' (Misner *et al.* 1973). Probably the closest approach to the theory developed here is the spin-two field theory discussed by many authors (see box 17.2 of Misner *et al.* (1973) and Feynman *et al.* (1995)). This theory is usually viewed as reproducing GR exactly, albeit in a somewhat ugly form due to the existence of a background spacetime and the reliance on infinite series of the field variable. By contrast, we hope to have demonstrated that GTG has an internal attractiveness of its own, as well as simplicity due to its first-order nature.

A crucial question to address is whether any experimental tests are likely to distinguish between GR and GTG in the immediate future. The biggest differences between GR and GTG to emerge to date lie in the treatment of black hole singularities (Doran *et al.* 1998*a*; Doran 1998; Lasenby *et al.* 1996), but these are unlikely to be testable for some considerable time! A more promising area is the link between gravity and quantum spin. GTG makes a clear prediction for the type and magnitude of this interaction, whereas it is not uniquely picked out in ECKS theory (the extension of GR to incorporate torsion) or in more general Poincaré gauge theory. Any experiment measuring this interaction would therefore provide a clear test of GTG. A partial exploration of the effects of spin interactions in GTG is contained in Doran *et al.* (1998*b*).

The techniques developed here reveal some remarkable properties of spherically symmetric systems. It has been known since the 1930s (Milne & McCrea 1934) that, in the absence of pressure, the dynamical equations of cosmology can be cast in a Newtonian form. We have now shown that a single, unified Newtonian treatment can be given for all spherically symmetric pressureless fluids, whether homogeneous or not. Furthermore, effects which have hitherto been viewed as the result of spacetime curvature can now be understood in a simple alternative fashion. The result is a physical picture in which the background spacetime has no effect on either dynamics or kinematics. This, we believe, is both new and potentially very useful. For example, simulations of black hole formation, and studies of the behaviour of the universe as a whole, can be carried out in exactly the same framework. All previous studies have relied on cutting and pasting various metrics together, with the result that no clear, global view of the underlying physics can be achieved. These advantages are exploited in Lasenby *et al.* (1998) to model the growth of a spherically symmetric perturbation in a homogeneous background cosmology and to study the effect of the perturbation on the cosmic microwave background.

The intrinsic method described here, and used to study spherically symmetric systems, is quite general and can be applied to a wide range of problems. In Doran *et al.* (1996*a*) the method is applied to a restricted class of cylindrically symmetric systems. Further publications will present treatments of more general cylindrically symmetric systems and of axisymmetric systems. In all cases studied to date, the

intrinsic method has brought considerable clarity to what would otherwise be a largely mysterious mess of algebra. This is achieved by removing the dependence on an arbitrary coordinate system, and instead working directly with physical quantities. The same technique also looks well suited to the study of cosmological perturbation theory, about which there has been considerable recent debate (Ellis 1995).

The interaction between Dirac theory and the gauge theory developed here revealed a number of surprises. The first was that consistency of the action principle with the minimal-coupling procedure restricted us to a theory which is unique up to the possible inclusion of a cosmological constant. The second was that spatially flat cosmologies are the only ones that are consistent with homogeneity at the level of the single-particle Dirac equation. The final surprise was provided by a study of the Dirac equation in a black hole background, which revealed a remarkable link with the Hawking temperature and quantum field theory. Much work remains to settle the issues raised by this final point, however.

As a final remark, we also hope to have demonstrated the power of geometric algebra in analysing many physical problems. Many of the derivations performed in this paper would have been far more cumbersome in any other language and none are capable of the compact expressions provided by geometric algebra for, say, the Riemann tensor. In addition, use of geometric algebra enabled us to remove all reference to coordinate frames from the fundamental equations. This is a real aid to providing a clear physical understanding of the mathematics involved. We would encourage anyone interested in studying the consequences of our theory to take time to master the techniques of geometric algebra.

We thank Anton Garrett for his careful reading of this manuscript and his many suggestions for improvements. We also thank David Hestenes for many enjoyable and thought-provoking discussions and George Ellis for a number of helpful suggestions.

Appendix A. The Dirac operator algebra

In the Dirac–Pauli representation, the γ matrices are defined as

$$\hat{\gamma}_0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \quad \hat{\gamma}_k = \begin{pmatrix} 0 & -\hat{\sigma}_k \\ \hat{\sigma}_k & 0 \end{pmatrix}, \tag{A 1}$$

where the $\hat{\sigma}_k$ are the standard Pauli matrices (Bjorken & Drell 1964; Itzykson & Zuber 1980). The Dirac γ matrices act on spinors, which are four-component complex column vectors. A spinor $|\psi\rangle$ is placed in one-to-one correspondence with an eight-component even element of the STA via (Doran *et al.* 1993*c*)

$$|\psi\rangle = \begin{pmatrix} a^0 + \mathrm{j}a^3 \\ -a^2 + \mathrm{j}a^1 \\ -b^3 + \mathrm{j}b^0 \\ -b^1 - \mathrm{j}b^2 \end{pmatrix} \leftrightarrow \psi = a^0 + a^k \mathrm{i}\sigma_k + \mathrm{i}(b^0 + b^k \mathrm{i}\sigma_k), \tag{A 2}$$

where j here denotes the unit scalar imaginary of conventional quantum mechanics. The action of the $\{\hat{\gamma}_\mu\}$, j and $\hat{\gamma}_5 = -\mathrm{j}\hat{\gamma}_0\hat{\gamma}_1\hat{\gamma}_2\hat{\gamma}_3$ operators maps to

$$\hat{\gamma}_\mu|\psi\rangle \leftrightarrow \gamma_\mu\psi\gamma_0, \quad \mathrm{j}|\psi\rangle \leftrightarrow \psi\mathrm{i}\sigma_3, \quad \hat{\gamma}_5|\psi\rangle \leftrightarrow \psi\sigma_3. \tag{A 3}$$

The Dirac equation,

$$\hat{\gamma}_\mu(\mathrm{j}\partial_\mu - eA_\mu)|\psi\rangle = m|\psi\rangle, \tag{A 4}$$

now takes the STA form

$$\gamma^\mu(\partial_\mu\psi\mathrm{i}\sigma_3 - eA_\mu\psi)\gamma_0 = m\psi. \tag{A 5}$$

Recombining to form the vectors $\nabla = \gamma^\mu\partial_\mu$ and $A = \gamma^\mu A_\mu$, and postmultiplying by γ_0, we arrive at the Dirac equation in the form

$$\nabla\psi\mathrm{i}\sigma_3 - eA\psi = m\psi\gamma_0. \tag{A 6}$$

Under Lorentz transformations, the spinor ψ transforms single sidedly to $R\psi$, hence the presence of the fixed γ_0 and γ_3 vectors on the right-hand side of ψ does not break Lorentz invariance.

The role of the Dirac adjoint is played by the geometric operation of reversion and the quantum inner product projects out the $\{1, \mathrm{i}\sigma_3\}$ components from a general multivector. So, for example, the real part of the inner product $\langle\bar{\psi}_1|\psi_2\rangle$ is given in the STA by $\langle\tilde{\psi}_1\psi_2\rangle$ and the imaginary part by $-\langle\tilde{\psi}_1\psi_2\mathrm{i}\sigma_3\rangle$. The Dirac current $J^\mu = \langle\bar{\psi}|\hat{\gamma}^\mu|\psi\rangle$ is now replaced by the set of components

$$\langle\tilde{\psi}\gamma^\mu\psi\gamma_0\rangle = \gamma^\mu\cdot(\psi\gamma_0\tilde{\psi}). \tag{A 7}$$

These are simply the components of the vector $\psi\gamma_0\tilde{\psi}$, decomposed in the $\{\gamma^\mu\}$ frame. Reference to the frame is removed from the vector by defining the current as

$$\mathcal{J} = \psi\gamma_0\tilde{\psi}. \tag{A 8}$$

Similarly, the role of the spin current is played by the vector

$$s = \psi\gamma_3\tilde{\psi} \tag{A 9}$$

and the spin trivector is simply is. The Dirac Lagrangian has the equivalent STA form

$$\langle\bar{\psi}|(\hat{\gamma}^\mu(\mathrm{j}\partial_\mu - eA_\mu) - m)|\psi\rangle \leftrightarrow \langle\nabla\psi\mathrm{i}\gamma_3\tilde{\psi} - eA\psi\gamma_0\tilde{\psi} - m\psi\tilde{\psi}\rangle, \tag{A 10}$$

which is the form used in the main text. A more detailed discussion of the STA formulation of Dirac theory is contained in Doran *et al.* (1996*b*).

Appendix B. Some results in multivector calculus

We begin with a set of results for the derivative with respect to the vector a in an n-dimensional space (Hestenes & Sobczyk 1984)

$$\left.\begin{aligned} &\partial_a a\cdot b = b, && \partial_a a^2 = 2a,\\ &\partial_a\cdot a = n, && \partial_a a\cdot A_r = rA_r,\\ &\partial_a\wedge a = 0, && \partial_a a\wedge A_r = (n-r)A_r,\\ &\partial_a a = n, && \dot{\partial}_a A_r\dot{a} = (-1)^r(n-2r)A_r. \end{aligned}\right\} \tag{B 1}$$

The results needed for the multivector derivative in this paper are

$$\partial_X\langle XA\rangle = P_X(A), \quad \partial_X\langle\tilde{X}A\rangle = P_X(\tilde{A}), \tag{B 2}$$

where $P_X(A)$ is the projection of A onto the grades contained in X. These results are combined using Leibniz's rule; for example,

$$\partial_\psi\langle\psi\tilde{\psi}\rangle = \dot{\partial}_\psi\langle\dot{\psi}\tilde{\psi}\rangle + \dot{\partial}_\psi\langle\psi\dot{\tilde{\psi}}\rangle = 2\tilde{\psi}. \tag{B 3}$$

For the action principle we also require results for the multivector derivative with respect to the directional derivatives of a field ψ. The aim is again to refine the

calculus so that it becomes possible to work in a frame-free manner. (The derivations presented here supersede those given previously in Lasenby *et al.* (1993*b*).) We first introduce the fixed frame $\{e^j\}$, with reciprocal $\{e_k\}$, so that $e^j \cdot e_k = \delta^j_k$. The partial derivative of ψ with respect to the coordinate $x^j = e^j \cdot x$ is abbreviated to $\psi_{,j}$ so that

$$\psi_{,j} \equiv e_j \cdot \nabla \psi. \tag{B 4}$$

We can now define the frame-free derivative

$$\partial_{\psi_{,a}} \equiv a \cdot e_j \partial_{\psi_{,j}}. \tag{B 5}$$

The operator $\partial_{\psi_{,a}}$ is the multivector derivative with respect to the a derivative of ψ. The fundamental property of $\partial_{\psi_{,a}}$ is that

$$\partial_{\psi_{,a}} \langle b \cdot \nabla \psi M \rangle = a \cdot b P_\psi(M). \tag{B 6}$$

Again, more complicated results are built up by applying Leibniz's rule. The Euler–Lagrange equations for the Lagrangian density $\mathcal{L} = \mathcal{L}(\psi, a \cdot \nabla \psi)$ can now be given in the form

$$\partial_\psi \mathcal{L} = \partial_a \cdot \nabla (\partial_{\psi_{,a}} \mathcal{L}), \tag{B 7}$$

which is the form applied in the main text of this paper.

We also need a formalism for the derivative with respect to a linear function. Given the linear function $\underline{h}(a)$ and the fixed frame $\{e_i\}$, we define the scalar coefficients

$$h_{ij} \equiv e_i \cdot \underline{h}(e_j). \tag{B 8}$$

The individual partial derivatives $\partial_{h_{ij}}$ are assembled into a frame-free derivative by defining

$$\partial_{\underline{h}(a)} \equiv a \cdot e_j e_i \partial_{h_{ij}}. \tag{B 9}$$

The fundamental property of $\partial_{\underline{h}(a)}$ is that

$$\partial_{\underline{h}(a)} \underline{h}(b) \cdot c = a \cdot e_j e_i \partial_{h_{ij}} (h_{lk} b^k c^l) = a \cdot e_j e_i c^i b^j = a \cdot bc, \tag{B 10}$$

which, together with Leibniz's rule, is sufficient to derive all the required properties of the $\partial_{\underline{h}(a)}$ operator. For example, if B is a fixed bivector,

$$\begin{aligned} \partial_{\underline{h}(a)} \langle \underline{h}(b \wedge c) B \rangle &= \dot{\partial}_{\underline{h}(a)} \langle \dot{\underline{h}}(b) \underline{h}(c) B \rangle - \dot{\partial}_{\underline{h}(a)} \langle \dot{\underline{h}}(c) \underline{h}(b) B \rangle \\ &= a \cdot b \underline{h}(c) \cdot B - a \cdot c \underline{h}(b) \cdot B = \underline{h}[a \cdot (b \wedge c)] \cdot B. \end{aligned} \tag{B 11}$$

This result extends immediately to give

$$\partial_{\underline{h}(a)} \langle \underline{h}(A_r) B_r \rangle = \langle \underline{h}(a \cdot A_r) B_r \rangle_1. \tag{B 12}$$

In particular,

$$\partial_{\underline{h}(a)} \det(\underline{h}) = \partial_{\underline{h}(a)} \langle \underline{h}(I) I^{-1} \rangle = \underline{h}(a \cdot I) I^{-1} = \det(\underline{h}) \overline{h}^{-1}(a), \tag{B 13}$$

where the definition of the inverse (2.53) has been employed. The derivation of (B 13) affords a remarkably direct proof of the formula for the derivative of the determinant of a linear function.

The above results hold equally if $\underline{h}$ is replaced throughout by its adjoint $\overline{h}$, which is the form of the derivative used throughout the main text. Note, however, that

$$\partial_{\underline{h}(a)} \overline{h}(b) = \partial_{\underline{h}(a)} \langle \underline{h}(c) b \rangle \partial_c = a \cdot cb \partial_c = ba. \tag{B 14}$$

Thus the derivatives of $\underline{h}(b)$ and $\overline{h}(b)$ give different results, regardless of any symmetry properties of $\underline{h}$. This has immediate implications for the symmetry (or lack of symmetry) of the functional stress-energy tensors for certain fields.

We finally need some results for derivatives with respect to the bivector-valued linear function $\Omega(a)$. The extensions are straightforward and we just give the required results:

$$\partial_{\Omega(a)}\langle\Omega(b)M\rangle = a\cdot b\langle M\rangle_2 \tag{B 15}$$

$$\partial_{\Omega(b),a}\langle c\cdot\nabla\Omega(d)M\rangle = a\cdot cb\cdot d\langle M\rangle_2. \tag{B 16}$$

Appendix C. The translation of tensor calculus

The reformulation of the gauge theory presented in this paper in terms of conventional tensor calculus proceeds as follows. A choice of gauge is made and a set of scalar coordinates $\{x^\mu\}$ is introduced. The coordinate frame $\{e_\mu\}$,

$$e_\mu \equiv \frac{\partial x}{\partial x^\mu}, \tag{C 1}$$

and reciprocal frame $\{e^\mu\}$,

$$e^\mu \equiv \nabla x^\mu, \tag{C 2}$$

are then constructed. From these one constructs the vectors

$$g_\mu \equiv \underline{h}^{-1}(e_\mu), \quad g^\mu \equiv \overline{h}(e^\mu). \tag{C 3}$$

The metric is then given by the 4×4 matrix

$$g_{\mu\nu} \equiv g_\mu\cdot g_\nu. \tag{C 4}$$

If the x dependence in $g_{\mu\nu}$ is replaced by dependence solely on the coordinates $\{x^\mu\}$ then we recover Riemann–Cartan geometry, where all relations are between coordinates and the concept of a point as a vector is lost.

The connection is defined by (following the conventions of Hehl *et al.* (1976) and Nakahara (1990))

$$\mathcal{D}_\mu g_\nu = \Gamma^\alpha_{\mu\nu} g_\alpha, \tag{C 5}$$

where $\mathcal{D}_\mu = g_\mu\cdot\mathcal{D} = \partial_\mu + \omega(g_\mu)\times$. We can therefore write

$$\Gamma^\lambda_{\mu\nu} = g^\lambda\cdot(\mathcal{D}_\mu g_\nu). \tag{C 6}$$

Since

$$\partial_\mu g_{\nu\lambda} = (\mathcal{D}_\mu g_\nu)\cdot g_\lambda + g_\nu\cdot(\mathcal{D}_\mu g_\lambda), \tag{C 7}$$

we find that

$$\partial_\mu g_{\nu\lambda} = \Gamma^\alpha_{\mu\nu} g_{\alpha\lambda} + \Gamma^\alpha_{\mu\lambda} g_{\alpha\nu}, \tag{C 8}$$

which recovers 'metric compatibility' of the connection. This is nothing more than the statement that the $a\cdot\mathcal{D}$ operator satisfies Leibniz's rule. Equation (C 8) can be inverted to show that the connection contains a component given by the standard Christoffel symbol. The connection can then be written

$$\Gamma^\nu_{\lambda\mu} = \left\{{\nu \atop \lambda\mu}\right\} - K_{\lambda\mu}{}^\nu, \tag{C 9}$$

where $K_{\lambda\mu}{}^\nu$ is the contorsion tensor and is given by

$$K_{\lambda\mu}{}^\nu = -S_{\lambda\mu}{}^\nu + S_\lambda{}^\nu{}_\mu - S^\nu{}_{\lambda\mu}. \tag{C 10}$$

Here $S_{\lambda\mu}{}^\nu$ is the torsion tensor, equal to the antisymmetric part of the connection,

$$\begin{aligned} S_{\lambda\mu}{}^\nu &= \tfrac{1}{2}(\Gamma^\nu_{\lambda\mu} - \Gamma^\nu_{\mu\lambda}) = \tfrac{1}{2}g^\nu\cdot(\mathcal{D}_\lambda g_\mu - \mathcal{D}_\mu g_\lambda) \\ &= -\tfrac{1}{2}[g_\mu\cdot(\mathcal{D}_\lambda g^\nu) - g_\lambda\cdot(\mathcal{D}_\mu g^\nu)] = -\tfrac{1}{2}(g_\mu\wedge g_\lambda)\cdot(\mathcal{D}\wedge g^\nu). \end{aligned} \tag{C 11}$$

The modified torsion tensor $T_{\lambda\mu}{}^{\nu}$ is defined by

$$T_{\lambda\mu}{}^{\nu} = S_{\lambda\mu}{}^{\nu} + 2\delta^{\nu}_{[\lambda}S_{\mu]\alpha}{}^{\alpha} = \tfrac{1}{2}\kappa(g_\lambda \wedge g_\mu)\cdot\mathcal{S}(g^\nu), \tag{C 12}$$

where $\mathcal{S}(a)$ is the torsion bivector as expected. When the spin is produced entirely by spin-$\frac{1}{2}$ particles we can write $\mathcal{S}(a) = a\cdot\mathcal{S}$, where $\mathcal{S}$ is the spin trivector. For this case

$$K_{\lambda\mu\nu} = -S_{\lambda\mu\nu} = -\tfrac{1}{2}(g_\lambda \wedge g_\mu \wedge g_\nu)\cdot\mathcal{S}. \tag{C 13}$$

If we now consider the covariant derivative of a covariant vector $\mathcal{A} = A^\alpha g_\alpha = A_\alpha g^\alpha$, we find that

$$\mathcal{D}_\mu\mathcal{A} = \mathcal{D}_\mu(A^\alpha g_\alpha) = (\partial_\mu A^\alpha)g_\alpha + A^\alpha \Gamma^\beta_{\mu\alpha} g_\beta = (\partial_\mu A^\alpha + \Gamma^\alpha_{\mu\beta}A^\beta)g_\alpha, \tag{C 14}$$

so that the components of the vector $\mathcal{D}_\mu\mathcal{A}$ are those expected for tensor calculus. Obviously, the fact that $A^\alpha g_\alpha = A_\alpha g^\alpha$ implies that $A_\mu = A^\alpha g_{\alpha\mu}$, so indices are raised and lowered in the expected manner.

For covariant quantities such as the Riemann tensor the translation to tensor calculus is straightforward:

$$R^{\mu}{}_{\nu\rho\sigma} = (g^\mu \wedge g_\nu)\cdot\mathcal{R}(g_\sigma \wedge g_\rho). \tag{C 15}$$

The general scheme is that any covariant quantity in GTG can be decomposed into tensor components by applying either the $\{g_\mu\}$ or $\{g^\mu\}$, or a mixture of both, to yield a tensor with the appropriate number of upstairs and downstairs indices. So, for example, $\mathcal{F}$ can be decomposed to $F_{\mu\nu} = \mathcal{F}\cdot(g_\mu \wedge g_\nu)$, $F_\mu{}^\nu = \mathcal{F}\cdot(g_\mu \wedge g^\nu)$ or $F^{\mu\nu} = \mathcal{F}\cdot(g^\mu \wedge g^\nu)$. Tensor calculus is poor at revealing which, if any, of the components represent a physical observable. Such issues are much clearer in GTG, which focuses attention on the single entity $\mathcal{F}$.

A vierbein $e_\mu{}^i$ (essentially an orthonormal tetrad) is given by

$$e_\mu{}^i = g_\mu\cdot\gamma^i, \tag{C 16}$$

$$e^\mu{}_i = g^\mu\cdot\gamma_i, \tag{C 17}$$

where $\{\gamma^i\}$ is a fixed orthonormal frame. Any position dependence in the $\{\gamma^i\}$ is eliminated with a suitable rotor transformation. When matrix operators $\{\hat{\gamma}^i\}$ are required these are replaced by the $\{\gamma^i\}$ frame vectors using the method described in Appendix A. In this way frame-free vectors can be assembled. For example, the Dirac operator (Gockeler & Schucker 1987, ch. 11)

$$\not{D}|\psi\rangle \equiv e^\mu{}_i\hat{\gamma}^i\left(\frac{\partial}{\partial x^\mu} + \tfrac{1}{4}\omega_{jk\mu}\hat{\gamma}^j\hat{\gamma}^k\right)|\psi\rangle \tag{C 18}$$

has the STA equivalent

$$g^\mu\cdot\gamma_i\gamma^i\left[\frac{\partial}{\partial x^\mu} + \tfrac{1}{4}\omega(g_\mu)\cdot(\gamma_k\wedge\gamma_j)\gamma^j\gamma^k\psi\gamma_0\right] = \overline{h}(\nabla)\psi\gamma_0 + \tfrac{1}{2}g^\mu\omega(g_\mu)\psi\gamma_0 = D\psi\gamma_0. \tag{C 19}$$

The above relations enable many results from Riemann–Cartan geometry to be carried over into our formalism.

A similar translation scheme is easily set up for the language of differential forms, which is much closer to the spirit of geometric algebra than tensor calculus. Differential forms are scalar-valued functions of an antisymmetrized set of vectors. They can easily be mapped to an equivalent multivector and a full translation into geometric algebra is quite straightforward. Here we note in passing the geometric algebra

equivalent of the Hodge dual of a differential form, which is

$$^{*}\alpha_r \mapsto -\det(\underline{h})^{-1}\underline{h}\overline{h}(\tilde{A}_r)\mathrm{i}, \tag{C 20}$$

where A_r is the multivector equivalent of α_r.

References

Abrahams, A. M. & Evans, C. R. 1993 Critical behaviour and scaling in vacuum axisymmetric gravitational collapse. *Phys. Rev. Lett.* **70**, 2980.

Bekenstein, J. D. 1974 Generalized second law of thermodynamics in black-hole physics. *Phys. Rev.* D **9**, 3292.

Bjorken, J. D. & Drell, S. D. 1964 *Relativistic quantum mechanics*, vol. 1. New York: McGraw-Hill.

Bondi, H. 1947 Spherically symmetrical models in general relativity. *Mon. Not. R. Astron. Soc.* **107**, 410.

Brackx, F., Delanghe, R. & Serras, H. (eds) 1993 *Clifford algebras and their applications in mathematical physics*. Dordrecht: Kluwer.

Cartan, E. 1922 Sur un généralisation de la notion de courbure de Riemann et les espaces à torsion. *C. R. Acad. Sci., Paris* **174**, 593.

Challinor, A. D., Lasenby, A. N., Doran, C. J. L. & Gull, S. F. 1997 Massive, non-ghost solutions for the Dirac field coupled self-consistently to gravity. *Gen. Rel. Grav.* **29**, 1527.

Chisholm, J. S. R. & Common, A. K. (eds) 1986 *Clifford algebras and their applications in mathematical physics*. Dordrecht: Reidel.

Choptuik, M. W. 1993 Universality and scaling in gravitational collapse of a massless scalar field. *Phys. Rev. Lett.* **70**, 9.

Clifford, W. K. 1878 Applications of Grassmann's extensive algebra. *Am. J. Math.* **1**, 350.

Copson, E. T. 1928 On electrostatics in a gravitational field. *Proc. R. Soc. Lond.* A **118**, 184.

Costa de Beauregard, O. 1963 Translational inertial spin effect. *Phys. Rev.* **129**, 466.

Damour, T. & Ruffini, R. 1976 Black-hole evaporation in the Klein–Sauter–Heisenberg–Euler formalism. *Phys. Rev.* D **14**, 332.

d'Inverno, R. 1992 *Introducing Einstein's relativity*. Oxford University Press.

Doran, C. J. L. 1994 Geometric algebra and its application to mathematical physics. Ph.D. thesis, Cambridge University.

Doran, C. J. L. 1998 Integral equations and Kerr–Schild fields. I. Spherically-symmetric fields. *Class. Quantum Grav.* (Submitted.)

Doran, C. J. L., Lasenby, A. N. & Gull, S. F. 1993*a* Gravity as a gauge theory in the spacetime algebra. In *Clifford algebras and their applications in mathematical physics* (ed. F. Brackx, R. Delanghe & H. Serras), p. 375. Dordrecht: Kluwer.

Doran, C. J. L., Lasenby, A. N. & Gull, S. F. 1993*b* Grassmann mechanics, multivector derivatives and geometric algebra. In *Spinors, twistors, Clifford algebras and quantum deformations* (ed. Z. Oziewicz, B. Jancewicz & A. Borowiec), p. 215. Dordrecht: Kluwer.

Doran, C, J. L., Lasenby, A. N. & Gull, S. F. 1993*c* States and operators in the spacetime algebra. *Found. Phys.* **23**, 1239.

Doran, C. J. L., Hestenes, D., Sommen, F. & van Acker, N. 1993*d* Lie groups as spin groups. *J. Math. Phys.* **34**, 3642.

Doran, C. J. L., Lasenby, A. N. & Gull, S. F. 1996*a* Physics of rotating cylindrical strings. *Phys. Rev.* D **54**, 6021.

Doran, C. J. L., Lasenby, A. N., Gull, S. F., Somaroo, S. S. & Challinor, A. D. 1996*b* Spacetime algebra and electron physics. *Adv. Imag. Elect. Phys.* **95**, 271.

Doran, C. J. L., Lasenby, A. N. & Gull, S. F. 1998*a* Integral equations and Kerr–Schild fields. II. The Kerr solution. *Class. Quantum Grav.* (Submitted.)

Doran, C. J. L., Lasenby, A. N., Challinor, A. D. & Gull, S. F. 1998*b* Effects of spin–torsion in gauge theory gravity. *J. Math. Phys.* (In the press.)

Eguchi, T., Gilkey, P. B. & Hanson, A. J. 1980 Gravitation, gauge theories and differential geometry. *Phys. Rep.* **66**, 213.

Ellis, G. F. R. & Rothman, T. 1993 Lost horizons. *Am. J. Phys.* **61**, 883.

Ellis, G. F. R. 1995 The covariant and gauge invariant approach to perturbations in cosmology. In *Current topics in astrofundamental physics: the early universe* (ed. N. Sánchez & A. Zichichi), p. 1. Dordrecht: Kluwer.

Estabrook, F. B. & Wahlquist, H. D. 1965 Dyadic analysis of spacetime congruences. *J. Math. Phys.* **5**, 1629.

Feynman, R. P., Morningo, F. B. & Wagner, W. G. 1995 *Feynman lectures on gravitation.* Reading, MA: Addison-Wesley.

Gautreau, R. 1984 Curvature coordinates in cosmology. *Phys. Rev.* D **29**, 186.

Gautreau, R. 1995 Light cones inside the Schwarzschild radius. *Am. J. Phys.* **63**, 431.

Gautreau, R. & Cohen, J. M. 1995 Gravitational collapse in a single coordinate system. *Am. J. Phys.* **63**, 991.

Gockeler, M. & Schucker, T. 1987 *Differential geometry, gauge theories, and gravity.* Cambridge University Press.

Grassmann, H. 1877 Der ort der Hamilton'schen quaternionen in der ausdehnungslehre. *Math. Ann.* **12**, 375.

Gull, S. F. 1991 Charged particles at potential steps. In *The electron* (ed. A. Weingartshofer & D. Hestenes), p. 37. Dordrecht: Kluwer.

Gull, S. F., Lasenby, A. N. & Doran, C. J. L. 1993*a* Imaginary numbers are not real—the geometric algebra of spacetime. *Found. Phys.* **23**, 1175.

Gull, S. F., Lasenby, A. N. & Doran, C. J. L. 1993*b* Electron paths, tunnelling and diffraction in the spacetime algebra. *Found. Phys.* **23**, 1329.

Hanni, R. S. & Ruffini, R. 1973 Lines of force of a point charge near a Schwarzschild black hole. *Phys. Rev.*, D **8**, 3259.

Hawking, S. W. 1974 Black hole explosion? *Nature* **248**, 30.

Hawking, S. W. 1993 *Black holes and baby universes and other essays.* London: Bantam.

Hawking, S. W. & Ellis, G. F. R. 1973 *The large scale structure of spacetime.* Cambridge University Press.

Hecht, R. D., Lemke, J. & Wallner, R. P. 1991 Can Poincaré gauge theory be saved? *Phys. Rev.* D **44**, 2442.

Hehl, F. W., von der Heyde, P., Kerlick, G. D. & Nester, J. M. 1976 General relativity with spin and torsion: foundations and prospects. *Rev. Mod. Phys.* **48**, 393.

Hestenes, D. 1966 *Spacetime algebra.* New York: Gordon and Breach.

Hestenes, D. 1974*a* Proper particle mechanics. *J. Math. Phys.* **15**, 1768.

Hestenes, D. 1974*b* Proper dynamics of a rigid point particle. *J. Math. Phys.* **15**, 1778.

Hestenes, D. 1975 Observables, operators, and complex numbers in the Dirac theory. *J. Math. Phys.* **16**, 556.

Hestenes, D. 1985 *New foundations for classical mechanics.* Dordrecht: Reidel.

Hestenes, D. 1986*a* A unified language for mathematics and physics. In *Clifford algebras and their applications in mathematical physics* (ed. J. S. R. Chisholm & A. K. Common), p. 1. Dordrecht: Reidel.

Hestenes, D. 1986*b* Curvature calculations with spacetime algebra. *Int. J. Theor. Phys.* **25**, 581.

Hestenes, D. 1991 The design of linear algebra and geometry. *Acta Appl. Math.* **23**, 65.

Hestenes, D. & Sobczyk, G. 1984 *Clifford algebra to geometric calculus.* Dordrecht: Reidel.

Isham, C. J. & Nelson, J. E. 1974 Quantization of a coupled Fermi field and Robertson–Walker metric. *Phys. Rev.* D **10**, 3226.

Itzykson, C. & Zuber, J.-B. 1980 *Quantum field theory.* New York: McGraw-Hill.

Ivanenko, D. & Sardanashvily, G. 1983 The gauge treatment of gravity. *Phys. Rep.* **94**, 1.

Kaufmann, W. J. 1979 *The cosmic frontiers of general relativity.* New York: Penguin.

Khodunov, A. V. & Zhytnikov, V. V. 1992 Gravitational equations in spacetime with torsion. *J. Math. Phys.* **33**, 3509.

Kibble, T. W. B. 1961 Lorentz invariance and the gravitational field. *J. Math. Phys.* **2**, 212.

Kramer, D., Stephani, H., MacCallum, M. & Herlt, E. 1980 *Exact solutions of Einstein's field equations*. Cambridge University Press.

Kruskal, M. D. 1960 Maximal extension of the Schwarzschild metric. *Phys. Rev.* **119**, 1743.

Lasenby, A. N., Doran, C. J. L. & Gull, S. F. 1993*a* Grassmann calculus, pseudoclassical mechanics and geometric algebra. *J. Math. Phys.* **34**, 3683.

Lasenby, A. N., Doran, C. J. L. & Gull, S. F. 1993*b* A multivector derivative approach to Lagrangian field theory. *Found. Phys.* **23**, 1295.

Lasenby, A. N., Doran, C. J. L. & Gull, S. F. 1993*c* 2-spinors, twistors and supersymmetry in the spacetime algebra. In *Spinors, twistors, Clifford algebras and quantum deformations* (ed. Z. Oziewicz, B. Jancewicz & A. Borowiec), p. 233. Dordrecht: Kluwer.

Lasenby, A. N., Doran, C. J. L. & Gull, S. F. 1993*d* Cosmological consequences of a flat-space theory of gravity. In *Clifford algebras and their applications in mathematical physics* (ed. F. Brackx, R. Delanghe & H. Serras), p. 387. Dordrecht: Kluwer.

Lasenby, A. N., Doran, C. J. L. & Gull, S. F. 1995 Astrophysical and cosmological consequences of a gauge theory of gravity. In *Current topics in astrofundamental physics* (ed. N. Sánchez & A. Zichichi), p. 359. Singapore: World Scientific.

Lasenby, A. N., Doran, C. J. L., Dabrowski, Y. & Challinor, A. D. 1996 Rotating astrophysical systems and a gauge theory approach to gravity. In *Current topics in astrofundamental physics* (ed. N. Sánchez & A. Zichichi), p. 380. Singapore: World Scientific.

Lasenby, A. N., Doran, C. J. L., Hobson, M. P., Dabrowski, Y. & Gull, S. F. 1998 Microwave background anisotropies due to nonlinear structures. I. Improved theoretical models. *Mon. Not. R. Astron. Soc.* (Submitted.)

Lemaître, G. 1933 Spherical condensations in the expanding universe. *C. R. Acad. Sci., Paris* **196**, 903.

Linet, B. 1976 Electrostatics and magnetostatics in the Schwarzschild metric. *J. Phys.*, A **9**, 1081.

Logunov, A. A. & Loskutov, Yu. M. 1988 Once more on the nonuniqueness of the predictions of the general theory of relativity. *Theor. Math. Phys.* **76**, 779.

Martellini, M. & Treves, A. 1977 Comment on the Damour–Ruffini treatment of black-hole evaporation. *Phys. Rev.* D **15**, 2415.

Micali, A., Boudet, R. & Helmstetter, J. (eds) 1991 *Clifford algebras and their applications in mathematical physics*. Dordrecht: Kluwer.

Milne, E. A. & McCrea, W. H. 1934 Newtonian universes and the curvature of space. *Q. Jl Maths.* **5**, 73.

Misner, C. W. & Sharp, D. H. 1964 Relativistic equations for adiabatic, spherically symmetric gravitational collapse. *Phys. Rev.* **136**, 571.

Misner, C. W., Thorne, K. S. & Wheeler, J. A. 1973 *Gravitation*. San Francisco, CA: Freeman.

Nakahara, M. 1990 *Geometry, topology and physics*. Bristol: Adam Hilger.

Narlikar, J. V. 1993 *Introduction to cosmology*. Cambridge University Press.

Newman, E. T. & Janis, A. I. 1965 Note on the Kerr spinning-particle metric. *J. Math. Phys.* **6**, 915.

Novikov, I. D. & Frolov, V. P. 1989 *Physics of black holes*. Dordrecht: Kluwer.

Oppenheimer, J. R. & Snyder, H. 1939 On continued gravitational contraction. *Phys. Rev.* **56**, 455.

Panek, M. 1992 Cosmic background radiation anisotropies from cosmic structures: models based on the Tolman solution. *Ap. J.* **388**, 225.

Penrose, R. & Rindler, W. 1984 *Spinors and spacetime. Two-spinor calculus and relativistic fields*, vol. I, Cambridge University Press.

Rauch, R. T. 1982 Equivalence of an $R+R^2$ theory of gravity to Einstein–Cartan–Sciama–Kibble theory in the presence of matter. *Phys. Rev.* D **26**, 931.

Sciama, D. 1964 The physical structure of general relativity. *Rev. Mod. Phys.* **36**, 463, 1103.

Smith, A. G. & Will, C. M. 1980 Force on a static charge outside a Schwarzschild black hole. *Phys. Rev.* D **22**, 1276.

Sobczyk, G. 1981 Spacetime approach to curvature. *J. Math. Phys.* **22**, 333.

Stephani, H. 1982 *General relativity.* Cambridge University Press.

Thorne, K. S., Price, R. H. & Macdonald, D. A. 1986 *Black holes: the membrane paradigm.* Yale University Press.

Tolman, R. C. 1934 Effect of inhomogeneity on cosmological models. *Proc. Natn. Acad. Sci. USA* **20**, 169.

Tryon, E. P. 1973 Is the Universe a vacuum fluctuation? *Nature* **246**, 396.

Tryon, E. P. 1984 What made the world? *New Scientist* **8**, 14.

Vold, T. G. 1993*a* An introduction to geometric algebra with an application to rigid body mechanics. *Am. J. Phys.* **61**, 491.

Vold, T. G. 1993*b* An introduction to geometric calculus and its application to electrodynamics. *Am. J. Phys.* **61**, 505.

Wahlquist, H. D. 1992 The problem of exact interior solutions for rotating rigid bodies in general relativity. *J. Math. Phys.* **33**, 304.

Weyl, H. 1950 A remark on the coupling of gravitation and electron. *Phys. Rev.* **77**, 699.

Weyssenhoff, J. & Raabe, A. 1947 Relativistic dynamics of spin-fluids and spin-particles. *Acta Phys. Pol.* **9**, 7.

Utiyama, R. 1956 Invariant theoretical interpretation of interaction. *Phys. Rev.* **101**, 1597.

Zheng, Z., Yuan-xing, G. & Liao, L. 1981 Hawking evaporation of Dirac particles. *Chin. Phys.* **1**, 934.

Zheng, Z., Yuan-xing, G. & Liao, L. 1982 On the Hawking evaporation of Dirac particles in Kerr–Newman spacetime. *Chin. Phys.* **2**, 386.