International Journal of Theoretical Physics, Vol. 30, No. 9, 1991

Holonomy and Path Structures in General Relativity and Yang–Mills Theory

J. W. Barrett[1]

Received November 12, 1990

This article is about a different representation of the geometry of the gravitational field, one in which the paths of test bodies play a crucial role. The primary concept is the geometry of the motion of a test body, and the relation between different such possible motions. Space-time as a Lorentzian manifold is regarded as a secondary construct, and it is shown how to construct it from the primary data. Some technical problems remain. Yang-Mills fields are defined by their holonomy in an analogous construction. I detail the development of this idea in the literature, and give a new version of the construction of a bundle and connection from holonomy data. The field equations of general relativity are discussed briefly in this context.

PROLOGUE

The following text was written in 1985, forming the author's Ph.D. thesis. In preparing it for publication I have edited and amended it in a number of places, particularly updating the citations of work which has appeared since 1985, or has been brought to my attention since then. But I have tried to keep to my original intentions regarding the particular point of view about fundamental physics which I tried to express in 1985.

The work in Section 2, on Yang–Mills fields (connections), has been developed in innumerable ways over the years from Dirac onward, with many of the developers working separately, it seems, particularly those in the physics community. The introduction provides a guide to the literature which I know about on this subject—but is surely not complete, and I expect that more of this fragmented literature will come to light as the years pass. The material in Section 2 provides the necessary synthesis for the two sections which follow.

[1]Department of Physics, The University, Newcastle upon Tyne, NE1 7RU England.

0020-7748/91/0900-1171$06.50/0

In Sections 3 and 4 I apply the same philosophy and method to gravitational fields (Riemannian metrics), following the physical reasoning I outline in the Introduction. This departure appeared to be novel at the time, and still does. There are some loose parallels with the physical motivation of the twistor program in general relativity, as regards recasting the gravitational field in terms of the incidence of lines. Still looser is any similarity to string theories, where lines and loops wander on manifolds. If there is a connection at any technical level, it remains to be made.

I paid attention to the intuitive physical ideas (and the Yang–Mills analogy), sacrificing the ability to calculate and, probably, the mathematical precision and comprehensiveness which would allow the theory to develop further. However, from the point of view of theoretical physics, the next input must be quantum theory, and in quantum theory points and lines on manifolds as physical objects cannot be defended. Thus, I do not believe that a tighter mathematical formulation of the ideas I present here will become a part of fundamental physics in a direct way. Rather, my hope is that the ideas will suggest a physical reformulation, involving perhaps discreteness or quantum theory, or both.

Terms Assumed Throughout. Suppose B and C are two topological spaces, and M a differentiable manifold with a singled-out basepoint $*$. Then I use the following notation.

$\mathrm{Map}(B, C)$	denotes the set of continuous maps $B \to C$
I	the unit interval $[0, 1]$
TM	the tangent bundle of M
T_xM	the tangent space at x
M^I	the unpointed path space: the set of piecewise smooth maps $I \to M$
PM	the pointed path space: the subset of M^I with $p(0) = *$, $p \in PM$; note that, in places indicated, when E is a bundle over M, PE means the space of paths with $p(0)$ anywhere in the fiber over $*$
ΩM	the loop space: the subset of PM with $p(1) = *$
p^{-1}	for $p \in M^I$, p^{-1} is the *reverse* path: $p^{-1}(i) = p(1-i)$
$\circ$	denotes the composition of paths: for $p_2, p_1 \in M^I$ with $p_1(1) = p_2(0)$, $p_2 \circ p_1$ is the path

$$i \to \begin{cases} p_1(2i) & i \leq 1/2 \\ p_2(2i-1) & i \geq 1/2 \end{cases}$$

$\sim$ the tilde convention: for any map $\tilde{\psi}: A \to \mathrm{Map}(I, B)$, the same symbol ψ without the tilde represents the associated map

$$\psi; A \times I \to B$$
$$(u, i) \to \tilde{\psi}(u)[i]$$

■ the end-of-proof symbol

1. INTRODUCTION

The idea of "gravitational field" is more or less accepted as the basic notion of Einstein's general relativity. One imagines, in the absence of matter, a space-time manifold obeying the vacuum Einstein equations. Observers, or test particles, may venture into parts of this manifold to make such measurements as they please, and, so long as their mass is negligibly small, they may do so without disturbing the gravitational field. The field has an existence as a physical quantity independent of the test particles, whose proper lengths the field purports to indicate.

The mathematics of gravitation has changed very little since Einstein's original foundation of general relativity—one might mention the later introduction of spin density—but the conception of the gravitational field has altered significantly, if we can judge from Einstein's book on relativity theory (Einstein, 1922):

> For the concept of space the following seems essential. We can form new bodies by bringing bodies B, C, ... up to body A; we say that we *continue* body A. We can continue body A in such a way that it comes into contact with any other body, X. The ensemble of all continuations of body A we can designate as the "space of the body A". Then it is true that all bodies are in the "space of the (arbitrarily chosen) body A". In this sense we cannot speak of space in the abstract, but only of the "space belonging to a body A". The Earth's crust plays such a dominant role in our daily life in judging the relative positions of bodies that it has led to an abstract conception of space which cannot be defended. In order to free ourselves from this fatal error we shall speak only of "bodies of reference" or "space of reference". It was only through the theory of general relativity that the refinement of these concepts became necessary, as we shall see later ... [Einstein (1922, p. 2)]

In this passage, Einstein is against the notion of an abstract space or space-time having an existence independent of the measuring bodies. In other places, it is true, he did support the analogy between the electromagnetic field and the gravitational field, an analogy which probably did most to establish the now-accepted notion of the gravitational field. However, a bit of selective quotation serves to make this particular point.

Quantum gravity has very much taken its cue from the conventional field point of view. One considers the space of all metrics on a given manifold and tries to form a statistical, probabilistic, theory on this space, which satisfies the conventional ideals of a quantum theory. In other words, the attempted theory purports to deal only with the pure field, and does not consider as part of its scope the measuring bodies which Einstein considered to be basic to his conception of "space." We know now that this conception of quantum gravity has all sorts of objectional features (which, of course, may be due to other aspects of the theory) (see, e.g., Isham, 1981). One of the unpleasant aspects is the necessity of considering the diffeomorphism symmetry of the configuration space. This arises because quantum field theory is based on the local vector space structure of the configuration space (Isham, 1984), and so is based on the space of all metrics. However, diffeomorphism-related metrics are considered to describe the same physical object, and so one is forced to "factor through" the diffeomorphism symmetry. Now this symmetry is an essentially unphysical one, and arises because the space of all metrics is really the wrong concept, it does not have a one-to-one correspondence with real physical quantities. What one should really work with are the set of distinct geometries. The diffeomorphism symmetry is a symmetry of a representation, via a particular differentiable manifold.

This article is about a different representation of the geometry of the gravitational field, one in which the test bodies play a crucial role. It is an analysis of what the relationships between the measured quantities of the motion of the test bodies are. It describes the effect of the geometry of space-time on the geometries of the test-particle motions. The key idea is suggested by the quote of Einstein above: a point on a manifold is defined as the set of all particles which arrive "there." In other words, we have to say what a particle path is, and what it means for particle paths to be "in coincidence."

Particle paths shall be defined so that they all start at the same place. A particle path is defined by its geometry—the specification of the angles through which the particle bends, in which directions, and at what proper distances along the path. In other words, the intrinsic geometry (proper distances along the path) and extrinsic geometry (parallel transport of vectors along the path) are specified.

The information in the gravitational field is involved in grouping together, in a particular way which will be described more fully below, all the possible particle paths into sets which represent the particles whose paths are "in coincidence." These sets form the points of the space-time manifold. This is actually all that is required! What one finds is that the geometry of the particle paths is sufficient to specify the geometry of the resulting space-time. The result is a theory of gravity in which the test particles are firmly

mixed up with the phenomenon of the field itself. In fact, the field is "made out of" the motions of the test bodies. There is no conception of an independent gravitational field divorced from the theory of the propagation of matter.

So the presence of gravity is felt by the fact that it alters the sets of coincident particle motions. The coincidence is specified by an equivalence relation on the set of all possible particle geometries. I have termed the information that goes into this equivalence relation the "holonomy" of the gravitational field, the word holonomy being using in rather a loose sense. The point is that two particle paths end at the same point if they form a closed loop on the space-time manifold, which starts and ends at their mutual starting points. So to specify the equivalence relation one needs to know the geometry of the particles which move in closed loops only, and in addition the Lorentz group holonomy element (in the strict sense of the word holonomy) of the closed loop. This is the way in which the gravitational "field" is specified. The mathematical details of this are best left to the following sections.

So, with the motivation the analysis of gravity, why is a large section of the article about classical Yang–Mills fields? The reason is that Yang–Mills theory is very similar to general relativity, but, as is usually the case, it is a simpler theory. The Yang–Mills result (the representation theorem of Section 2) both provides a comparison with general relativity, and also a technical result which is of use in building the linear connection of the gravitational field. Moreover, the technical tools used on the way in Section 2 bear a strong resemblance to, and in fact motivate, the techniques of Section 3.

Table I contains a comparison of the main features of Yang–Mills theory in Section 2 and gravity in Section 3; the notation has been deliberately chosen to enhance the similarity.

The strong resemblances displayed in the comparison throw fresh light on the debate about the status of gravity as a gauge theory. From the point of view of holonomy, gravity has no diffeomorphism symmetry: one just specifies the set P and the map h. The "gauge freedom" of gravity is still present, however, in a rather subtle form. The set P (see Section 3) contains the geometries of the paths which form closed loops. Suppose, for the sake of argument, that the paths are composed of n linear sections (straight lines). Then, to specify an element of P (and hence give information about the gravitational field) one can arbitrarily specify the first $n-1$ pieces. Then the geometry of the last section of the path is entirely fixed by the requirement that it be in P. We can say that the gauge freedom of gravity is the freedom to specify the first $n-1$ sections at will, and the real information (the "holonomy") is the fixed value of the geometry of the last section. This notion of

Table I

Construct	YM	Gr
G	Structure group	Lorentz group
M	Base space	Tangent space
Geometry of loops	ΩM	P
Holonomy mapping	$H: \Omega M \to G$	$h: P \to G$
Equivalence relation on $PM \times G$, $(p, g) \sim (p', g')$	$p(1) = p'(1)$, $g' = H(p^{-1} \circ p')g$	There exists $\pi \in P$: $\pi = (H(\pi)p^{-1}) \circ p'$, $g' = h(\pi)g$
Constructed set $PM \times G/\sim$	Bundle	Manifold and frame bundle
Lifting	Lifting function l_*	Inverse development map Δ, lifting in frame bundle
whose smoothness gives	Charts on the bundle	Charts on the manifold
and which defines	Connection	Metric and connection

gauge freedom is of a different character from the rather empty notion of "diffeomorphism symmetry." It clearly relates to choices made by the experimenter about the information which is to be gathered. Thus, it is a symmetry with an explicit physical significance. This interpretation of the "gauge freedom" of gravity is only possible if one has in mind the physical identification of the paths of the holonomy description (which might be regarded as mere mathematical artifacts) with particle path geometries.

The last section is concerned with the field equations of gravity, which have so far not been introduced. The equations are presented in a form which relates the (infinitesimal) holonomy of the field to the matter momentum and angular momentum. The details are best left to that section. It suffices to remark here that by taking a very specific point of view on the construction of the gravitational field, as is described in this paper, one gets a very specific point of view about the field equations: the metric-compatible aspect of the connection is a matter of definition, but the torsion-free aspect is a field equation, with the same status as the Einstein field equation. In fact, the torsion equation is naturally paired with the Einstein equation: it is part of the "angular momentum" field equation, the Einstein equation being the "linear momentum" field equation.

2. THE HOLONOMY REPRESENTATION OF GAUGE FIELDS

2.1. Introduction

Classical Yang–Mills fields are usually described by potentials, or connections on a principal fiber bundle. A physical field configuration is really a set of potentials related by gauge transformations, so that in other words

the *configuration space* is the space of orbits of the action of the gauge group $\mathcal{G}$ in the space $\mathcal{A}$ of all connections on some given bundle. This orbit space has been much studied because one can view Yang–Mills quantum field theory as a theory of integration on the orbit space, albeit in an imprecise fashion (Atiyah, 1980; Singer, 1981; Babelon and Viallet, 1981). There is also a canonical version, where geodesics on orbit space correspond to the Hamiltonian flow (see also Narasimhan and Ramadas, 1979).

For good physical reasons, then, one would like to have more information about the configuration space $\mathcal{A}/\mathcal{G}$. To achieve a different characterization of the configuration space, it is possible to consider the *holonomy mapping* of the connection as the fundamental object. This mapping takes the loop space ΩM of the base manifold M into the structure group G of the bundle by mapping a loop into its holonomy element. The loop space is the set of all piecewise smooth paths in M which start and end at an arbitrarily singled-out point, denoted $*$, in M, with a topology which will be discussed later. The holonomy element of a loop is defined in terms of the horizontal lifting of the loop into the total space of the bundle. The two endpoints of the lifting define a G-element which translates one of the endpoints to the other, using the G-action on the bundle. This is the holonomy element. Section 2.2 describes how a connection gives rise to the concepts of horizontal lifting and holonomy, and why the machinery of bundles is the most appropriate.

Discussion of the holonomy in electromagnetism goes back to Dirac (1931) in the context of magnetic monopoles. Aharonov and Bohm (1959) discussed parallel transport as a phase shift of a Schrödinger wavefunction, and recognized the physical significance of the holonomy operator for a closed loop, being more directly involved in the behavior of the Schrödinger wavefunction than magnetic forces. Some of these ideas were used in an attempt to find gauge-invariant quantizations of electromagnetic and other gauge fields (Mandelstam 1962*a*,*b*, 1968*a*,*b*; Bialynicki-Birula, 1963).

There are two important facts about the holonomy mapping. First, different physical configurations give rise to different holonomy mappings, and second, the set of *restricted* gauge transformations $\mathcal{G}_* \subset \mathcal{G}$, that is, ones which act trivially at the basepoint, leave the holonomy mapping invariant. This means that, if we consider gauge equivalence to be restricted in this way, the physical configurations are faithfully and uniquely represented by their holonomy mappings. The residual gauge freedom of field rotations at the basepoint is parametrized by $G \sim \mathcal{G}/\mathcal{G}_*$, a finite-dimensional Lie group, in contrast to the original infinite-dimensional gauge group $\mathcal{G}$. Thus, the restricted, or pointed, configuration space $\mathcal{A}/\mathcal{G}_*$ is equivalent to a set of certain types of mappings $\Omega M \to G$, namely those which arise as holonomy mappings.

The results demonstrated in this section show that there is a simple set of conditions H1–H3, given below, defining the relevant subset $\mathscr{H} \subset \mathrm{Map}(\Omega M, G)$ and a simple reconstruction of both the principal bundle over M and the connection on it.

This works equally well for any principal G-bundle over M, so that it is more natural to consider as the configuration space the disjoint union over inequivalent bundles of the orbit spaces for each bundle. This can be much more elegantly stated as the set $\mathscr{F}_*$ of triples (B, Γ, b), where B is any principal G-bundle, Γ a connection on it, and b a point in the fiber over the basepoint $*$, with equivalence

$$(B, \Gamma, b) \sim (B', \Gamma', b')$$

if there is a bundle isomorphism $B \to B'$ taking Γ to Γ' and b to b', and such that it is the identity on M. This definition is more useful here, since the holonomy mapping involves only the base manifold M in its definition, the reconstruction process manufactures "new" bundles as well as connections. Note that since a preferred point in the fiber over $*$ is preserved, these triples are equivalent to orbits of the restricted gauge group in the space of potentials. To be precise, if C is a particular bundle with basepoint c, and $\mathscr{F}_*^C \subset \mathscr{F}_*$ is defined as the subset of triples (B, Γ, b) where B is isomorphic to C, and $\mathscr{A}$ and $\mathscr{G}_*$ are the potentials and restricted gauge group of (C, c), then $\mathscr{F}_*^C$ is in bijective correspondence with $\mathscr{A}/\mathscr{G}_*$, in the obvious way.

The two main results are as follows.

Reconstruction Theorem. Suppose M is a connected manifold with basepoint $*$, and $H: \Omega M \to G$ satisfies conditions H1–H3 below; then there exists a differentiable principal fiber bundle $B = (E, \pi, M, G)$, a point $b \in \pi^{-1}(*)$ and a connection Γ on B such that H is the holonomy mapping of (B, Γ, b).

Representation Theorem. If M is a connected, Hausdorff manifold, the correspondence of the reconstruction theorem between the Yang–Mills configuration space $\mathscr{F}_*$, defined as the set of triples (B, Γ, b) as above, and the set of holonomy maps $\mathscr{H} \subset \mathrm{Map}(\Omega M, G)$, defined by the conditions H1–H3, is a bijection.

Results of this nature can be proved in a variety of settings, for example, for differentiable or topological bundles. The infinitesimal connection (as described above) might be replaced by some more general gadget, e.g., a lifting function. In the literature, generalizations of infinitesimal connections are rather vaguely called "connections."

The first mention of a result such as the reconstruction theorem occurs in Kobayashi (1954) in a short note without proofs. The setting is rather

similar to the one used here, differentiable bundles and infinitesimal connections, but Kobayashi's axioms do not seem to include any mention of differentiability (see H3 below). The reconstruction theorem appears in a more general guise in papers on bundle topology which appeared shortly afterward. Milnor (1956) considered locally trivial bundles ("fiber" bundles) over a simplicial complex, with a bundle slicing function playing the role of a connection. Milnor's analogue of H2 is particularly elegant. The constructions are all topological, rather than differentiable.

The most general setting is that of Lashof (1956), who considered topological principal bundles (not even necessarily locally trivial) with a general lifting function playing the role of the connection. The lifting function is much more general than an infinitesimal connection because the lifting may not commute with the structure group action. This is also connected with the fact that in this formulation, paths which differ only by reparametrizations and other similar operations (cf. H2) are *not* considered to be equivalent. Lashof has a notion of equivalence of maps $\Omega M \to G$ which renders the equivalence classes in a bijective correspondence with the set of inequivalent G-bundles over M. This is thus a coarse version of the representation theorem, which does not distinguish between different connections on the same bundle. Interestingly, Lashof's constructions are rather close to the gravitational constructions presented here.

The differential aspect of the subject was explored by Teleman (1960, 1963) shortly afterward. He later wrote two papers which are probably the most comprehensive overview of all of these sorts of results. In the first (Teleman, 1969*a*), the reconstruction theorem appears (Theorem 3) in a general guise, a connection being a lifting function which satisfies analogues of H1 and H2. Various special cases (simplicial, differential, complex analytic) are discussed in the second paper (Teleman, 1969*b*).

The constructions were later rediscovered by physicists (Giles, 1981; Anandan, 1983; Barrett, 1985, 1989; Fischer, 1986). For the most part, they confined themselves to the algebraic part of the constructions rather than the topological aspect. Chan and Tsou (1986) and Chan *et al.* (1986) found an interesting extension of the reconstruction to presenting the data in the form of a connection on loop space ΩM, with applications to monopoles. The last paper also contains references to many other associated papers in the physics literature. It is interesting to note that Dirac's (1931) paper on monopoles discussed a homomorphism $\Omega M \to U1$, with M three-dimensional Euclidean space minus a point, which by the reconstruction theorem immediately translates into a connection on a bundle, the modern understanding of monopoles. The inequivalent bundles correspond to the different monopole charges which Dirac found.

Many physicists were interested in the character of the homomorphism $\Omega M \to G$ in the case that G acts on a vector space, which goes under the name "Wilson loop" in the physics literature (Wilson, 1974). See, e.g., Durhuus (1980) for a lattice version of a result which goes back to Teleman (1969*a*, Theorem 5) at least. Polyakov (1979) discusses some of the uses of Wilson loops in quantum field theory; it is a subject that was taken up by many other authors.

2.1.1. Conditions H1–H3

The first condition on a holonomy map H is straightforward:

H1. H is a homomorphism of the composition law of loops

$$H(\omega_2 \circ \omega_1) = H(\omega_1)H(\omega_2)$$

H1 is consistent with the holonomy element being defined as the G-element which translates the point b to its image point under parallel transport *backward* around the loop. This will be taken as the definition of the holonomy element. To be more explicit, if ω' is a horizontal lift of ω with $\omega'(1) = b$, then $bH(\omega) = \omega'(0)$ (Kobayashi and Nomizu, 1963).

The second condition axiomatizes several related features of H. H takes the same value on loops which differ by a reparametrization, or by the addition or removal of path sections which "double back" on themselves. This is formalized by the definition of *thin loops*: θ is a thin loop if there exists a homotopy of θ to the trivial loop, with the image of the homotopy lying entirely within the image of θ. This makes precise the notion that a thin loop in M does not enclose any area of M.

H2. H takes the same value on *thinly equivalent* loops: $\omega_1 \sim \omega_2$ if $\omega_1 \circ \omega_2^{-1}$ is thin.

The path ω^{-1} is the reverse path of ω: $\omega^{-1}(i) = \omega(1-i)$.

This property, together with H1, implies that $H(\omega^{-1}) = H^{-1}(\omega)$. In fact we can consider the whole H-group structure of ΩM (Spanier, 1966), and since the homotopies which make loop composition homotopy-associative and homotopy-invertible are actually thin homotopies describing thin equivalences, ΩM factored by thin equivalence is a group, and H defines a homomorphism of groups. This is explained further in Section 2.3. We can think of thin equivalence as a restricted notion of homotopy equivalence, intuitively similar to homotopy theory on a very fine mesh or sieve. The holes of the mesh stop the homotopies sweeping out across areas of M.

H1 and H2 are all the algebraic properties required of H. It remains to formulate a notion of smoothness for the holonomy mapping for axiom H3. One should not arbitrarily impose conditions of continuity and differentiability for purely technical reasons, but in an ideal world, these things should relate to questions of physical importance, about the way in which the measurements that the field represents are made. This point was made to me by Chris Isham. In Yang–Mills theory this is a little difficult to interpret, since except for the case of electromagnetism, one does not measure the classical field directly, it just appears as a construct in the quantum field theory. Nevertheless, the question is certainly important for classical gravity and electromagnetism.

Any classical field configuration represents an infinite amount of information, and since one can only ever measure a finite set of values, it is an idealization of the true situation. Any continuum of quantities represents a sequence (in the mathematical sense of a countable set in an order labeled by the integers) of measurements performed to greater accuracy as the number of measurements increases. Perhaps it would be more accurate to say that the field configuration represents a measurement algorithm, rather than any set of measurements themselves. Now in any theory envisaging an infinite set of possible outcomes this set should at least be a topological space, so that one can meaningfully discuss the convergence of a set of approximations representing the process of measurement. Clearly, the topology involved should relate closely to the way in which the theory envisages the measurements being made. Conventionally, when discussing theories involving an infinite amount of information, such as a classical field theory, there seem to be two distinct possible "sources" of the infinity. On one hand, there may be an infinite amount of information contained in a given field configuration, and on the other, there may be an infinite range of possibilities for that field configuration. So for a field theory, say involving real-valued functions on space-time, one needs both a topology on space-time to define a field as a continuous function, and a topology on the configuration space to determine which field configurations neighbor each other. Clearly, when one relates these topologies to the measurement process, one can see that they should be intimately related to each other. Similar sorts of remarks apply to the smooth structures of these spaces.

The most immediate problem is to find the structure on ΩM which makes holonomy mappings continuous, at least, and hopefully differentiable. Since much of the theory has a strong algebraic topological flavor to it, the first thing to try, and in fact discard, is the compact-open topology. The problem with this is simple: holonomy mappings are not in general continuous with this topology. A simple example will serve to demonstrate this. Let space-time be R^2 and consider an electromagnetic field on it with constant

field strength F. For any closed loop the holonomy element is

$$H(\omega) = \exp \int_V iF\, dx^1 \wedge dx^2$$

where V is the volume bounded by ω. Consider the family of loops $\tilde{\psi}: I \to \Omega\mathbb{R}^2$ given by

$$\psi(s, t) = \begin{cases} f(s)t(1-t)\exp(it/s), & s \neq 0 \\ 0, & s = 0 \end{cases}$$

where $\mathbb{R}^2$ has been identified with the complex numbers. The parameter t is the time along the path, and s is the family parameter. If f is a continuous real function with $f(0) = 0$, then ψ is continuous, and so $\tilde{\psi}$ is continuous in the compact-open topology. The family describes a continuous set of loops which, as $s \to 0$, simultaneously wind around faster and faster, and shrink to zero radius. The area integral is easily computed, putting $r(t) = f(s)t(1-t)$ and $\theta(t) = t/s$, then the area enclosed by loop s is

$$\int \tfrac{1}{2} r^2\, d\theta = \begin{cases} f^2(s)/60s, & s \neq 0 \\ 0, & s = 0 \end{cases}$$

Clearly, the choice $f(s) = s^{1/2}$ renders the holonomy $H\tilde{\psi}: I \to U1$ a discontinuous function. The consequence of this is that although many of the constructions to be used further on are very strongly linked with algebraic topology, the standard topology used in algebraic topology is not relevant. On a technical level, the reason is that essentially the parallel transport operator is the solution of an ordinary differential equation, and so it is important that the topology controls the derivatives of the paths. In the above example, the derivative of ψ is not continuous at $s = 0$. On a physical level, this failure stems from the fact that there were no good physical reasons for supposing that the compact-open topology is relevant.

To return to the physical motivation, what is needed is a topology on ΩM which relates more closely to the physics of the measurements. The physics is based on entirely classical (nonquantum) ideas. Suppose we just limit attention to the subspace of loops which are piecewise linear, i.e., composed of a finite number of straight-line segments. Then if space-time is again $\mathbb{R}^4$, and we consider the loops with $n+1$ sections, these can be parametrized b the positions of the "corners," i.e., $(\mathbb{R}^4)^n$. Thus, measuring the shape of the loop is reduced to measuring a finite number of particle positions,

and this is exactly where the topology and differentiable structure of space-time come into use.[2] Thus, we want to postulate that the holonomy is a smooth map $(\mathbb{R}^4)^n \to G$. So, to consider the space of piecewise linear loops, the smoothness requirement is:

For any integer $n > 0$, if ψ: $(\mathbb{R}^4)^n \to \Omega M$ is the family of piecewise linear loops with $n+1$ linear pieces, then the composite map $H\psi$: $\mathbb{R}^{4n} \to \Omega M \to G$ is a smooth map.

It would be quite possible, and physically well motivated, to work with the space of piecewise linear loops. However, the linear structure of the base space (here $\mathbb{R}^4$) is not really relevant, and we want the results to hold equally well on any differentiable manifold. It is also technically rather inconvenient. So the remedy is to allow the above situation to relax under arbitrary diffeomorphisms of the base space.

The final form of the axiom H3 encapsulates the notion that a smooth finite-dimensional family of loops has a smoothly-varying holonomy image in G. A *smooth finite-dimensional family of loops* is a map $\tilde{\psi}$: $U \to \Omega M$ with U an open subset of $\mathbb{R}^n$ for any n, which is smooth in the sense that the associated map

$$\psi: \quad U \times I \to M$$

$$(u, i) \to \tilde{\psi}(u)[i]$$

is continuous, and smooth (C^∞) on the subintervals

$$U \times [i_n, i_{n+1}] \quad \text{for} \quad i_0 = 0 < i_1 < \cdots i_k < 1 = i_{k+1}, \qquad n = 0, 1, \ldots, k$$

H3. For any smooth finite-dimensional family of loops $\tilde{\psi}$: $U \to \Omega M$, the composite map $H\tilde{\psi}$: $U \to \Omega M \to G$ is smooth.

The collection of all such maps $\tilde{\psi}$ from any such U into ΩM defines the induced topology on ΩM. This is the finest topology which makes all these maps continuous. Any map which obeys the axiom H3 is continuous (Spanier, 1966; Dugundji, 1966).

2.1.2. The Representation Map

If H is the holonomy mapping of $(B, \Gamma, b) \in \mathscr{F}_*$, then it has to be shown that properties H1–H3 hold. H1 is straightforward. To prove H2, we have to show first that $H(\theta) =$ identity when θ is a thin loop.

[2] I am indebted to Dr. R. W. Tucker for bringing to my attention the fact that if parts of the loop considered are spacelike, then it does not have the direct interpretation of being composed of particle paths. One just has to consider it as a figure in space-time, perhaps as marked out by, and measured with, the aid of auxiliary particles.

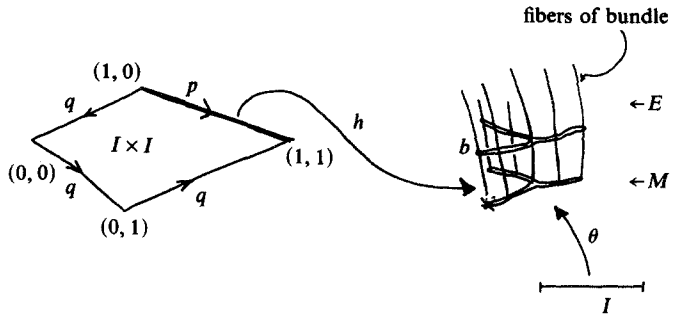


Fig. 1

Let $\tilde{h}: I \to \Omega M$ be the smooth homotopy of θ which makes it a thin loop, so that $\theta = \tilde{h}(1)$, $\tilde{h}(0)$ is the trivial loop, and the map $h: I \times I \to M$ has $\mathrm{Im}(h) = \mathrm{Im}(\theta)$ (using the tilde convention).

We can study the lifting in the pullback bundle $B_h = (E_h, \pi_h, I \times I, G)$ of B by the homotopy h. Since the image of h is one-dimensional, the curvature form in B_h is zero, and B_h has the canonical flat connection. It follows that the path $p: i \to (1, i)$ in $I \times I$ can be deformed to the path

$$q: \quad i \to \begin{cases} (1-3i, 0), & 0 \le i \le 1/3 \\ (0, 3i-1), & 1/3 \le i \le 2/3 \\ (3i-2, 1), & 2/3 \le i \le 1 \end{cases}$$

without altering the endpoints of the lift in B_h. The map h carries p to the path θ in M and q to the trivial path in M (Figure 1). Since the canonical homomorphism $B_h \to B$ carries horizontal lifts into horizontal lifts, it follows that there is a horizontal lift of θ in E which has the same endpoints as the trivial lift of the trivial loop in M, and so the endpoints of the θ lift are the same point. Hence the holonomy of θ is the identity element of G. It should be noted that the homotopy h may be smooth only on subintervals $I \times [i_n, i_{n+1}]$, but that the conclusion is unaffected.

If $\omega_1 \sim \omega_2$, then $H(\omega_1 \circ \omega_2^{-1}) = \mathrm{id}$, and since $H(\omega_2^{-1}) = H(\omega_2)^{-1}$, it follows, using H1, that $H(\omega_2) = H(\omega_1)$. Property H3 follows from a slight modification of the proof in Kobayashi and Nomizu (1963, p. 74).

Since these conditions are satisfied, the *representation map* $\mathscr{R}: \mathscr{F}_* \to \mathscr{H}$, which takes (B, Γ, b) to its holonomy map, is defined.

2.1.3. Bundle Construction

In this section it is supposed that $H \in \mathscr{H}$, i.e., it is a map $\Omega M \to G$ satisfying H1–H3. Using H, we can construct the total space E of a principal

bundle B as the set

$$E = PM \times G/R$$

where PM is the path space of M, piecewise smooth paths with $p(0) = *$, and R is an equivalence relation quotienting $PM \times G$:

$$(p, g) \sim (p', g') \quad \text{if} \quad p(1) = p'(1) \quad \text{and} \quad g' = H(p^{-1} \circ p')g$$

The loop $p^{-1} \circ p'$ is formed by composing p' with p parametrized in the inverse direction. By virtue of properties H1 and H2, R is an equivalence relation. In the following the equivalence class of (p, g) will be denoted $\{p, g\}$.

It is interesting to note that this construction is analogous to the associated bundle construction, where $\Omega M \to PM \to M$ is analogous to the principal bundle with fiber ΩM and H defines the action of ΩM on G. This analogy is made precise by turning ΩM quotiented by thin equivalence into a topological group (Teleman, 1960; Milnor, 1956).

The projection map and G-action are

$$\begin{aligned} \pi: \quad & E \to M \\ & \{p, g\} \to p(1) \\ \mu: \quad & E \times G \to G \\ & \{p, g\}h \to \{p, gh\} \end{aligned}$$

The preferred point in the fiber over $* \in M$ is

$$b = \{t, \mathrm{id}\}, \qquad \text{where } t \text{ is the trivial path}$$

The lifting function is also implicit in the construction

$$\begin{aligned} l_*: \quad & PM \times G \to PE \\ & (p, g) \to q \end{aligned}$$

with PE the space of paths in E starting over $*$, and q is the path

$$\begin{aligned} q: \quad & I \to E \\ & i \to \{K(p, i), g\} \end{aligned}$$

the contraction $K(p, i)$ representing the section of path p from 0 to i:

$$\begin{aligned} K: \quad & PM \times I \to PM \\ & K(p, i)[j] = p(ij) \end{aligned}$$

Now having a preferred point b in the fiber $\pi^{-1}(*)$ provides an isomorphism between this fiber and the group G. So the lifting function l_* gives

for each path p in the base starting at $*$ and a point g in the fiber over $*$ a path in E starting at g which is carried back onto p by the projection π. For a differentiable bundle with a connection a lifting function is given by assigning $l_*(p, g)$ the unique horizontal curve starting at g which covers p. The lifting described here may, *a priori*, not be of this type. But with the aid of axioms H1–H3 we will show that it is.

As yet E has no differentiable structure. Suppose U is a contractible open set of M. Then there exists a smooth family of paths $\tilde{\psi}$: $U \to PM$ such that $\tilde{\psi}(u)$ ends at u. Using the endpoint map e: $PM \to M$, $e(p) = p(1)$, this condition can be restated as $e\tilde{\psi} = \mathrm{id}$.

The chart C_ψ for $\pi^{-1}(U) \subset E$ is

$$C_\psi: \quad U \times G \xrightarrow{(\tilde{\psi},\mathrm{id})} PM \times G \xrightarrow{\alpha} PM \times G/R = E$$

where α is the canonical projection. C_ψ is a bijection as a map $U \times G \to \pi^{-1}(U)$. It is interesting to note, and will be useful later, that $\alpha = e' l_*$, where e': $PE \to E$ is the endpoint map of PE, so that the chart C_ψ is the map

$$C_\psi: \quad U \times G \xrightarrow{(\tilde{\psi},\mathrm{id})} PM \times G \xrightarrow{l_*} PE \xrightarrow{e'} E$$

Hence we can say that the lifting function is the relevant structure of E which is used to define the charts.

Lemma 1. The charts have smooth transition functions.

Proof. Suppose there is a second chart $C'_{\psi'}$ with $\tilde{\psi}'$: $V \to PM$; then, putting $W = U \cap V$ and defining $\tilde{\chi}$: $W \to \Omega M$

$$w \to \tilde{\psi}^{-1}(W) \circ \tilde{\psi}(w)$$

$$C'(w, g) = \{\tilde{\psi}'(w), g\} = \{\tilde{\psi}(w), H(\tilde{\chi}(w))g\} = C(w, H(\tilde{\chi}(w))g)$$

So $C^{-1}C'$ is the map

$$W \times G \xrightarrow{(\mathrm{diag},\mathrm{id})} W \times W \times G \xrightarrow{(\mathrm{id},H\tilde{\chi},\mathrm{id})} W \times G \times G \xrightarrow{(\mathrm{id},\mathrm{compose})} W \times G$$

which is smooth by property H3. ■

Since C_ψ commutes with the G-action on E, we have the following result.

Proposition. $E \xrightarrow{\pi} M$ is a principal fiber bundle with group G. ■

In a similar way, it is possible to show that the lifting function l_* is smooth, in the sense that a smooth family of paths in M lifts to a smooth family of paths in E.

For the present, we shall assume the proof of Lemma 3, which states that l_* is the horizontal lifting of a connection on B. It is clear that the connection is uniquely specified by the lifting function. Thus, the construction of a triple (B, Γ, b) is complete. This defines the *construction map* $\mathscr{C}: \mathscr{H} \to \mathscr{F}_*$.

Proof of the Reconstruction Theorem. To complete this proof, it remains to show that $\mathscr{R}\mathscr{C} = \mathrm{id}$, or, in other words, that if (B, Γ, b) is constructed from $H \in \mathscr{H}$ and has holonomy mapping H', then $H = H'$. This is straightforward. ■

Proof of the Representation Theorem. The reconstruction theorem shows that $\mathscr{R}\mathscr{C} = \mathrm{id}$, so it remains to show that $\mathscr{C}\mathscr{R} = \mathrm{id}$. Suppose that $\mathscr{R}(B, \Gamma, b) = H$ and $\mathscr{C}(H) = (\bar{B}, \bar{\Gamma}, \bar{b})$, $b = (e, \pi, m, g)$, and $\bar{B} = (\bar{E}, \bar{\pi}, M, G)$, and that l_* is the lifting function of (B, Γ, b). Consider the map

$$PM \times G \xrightarrow{l_*} PE \xrightarrow{e'} E$$

This factors through the relation R_H on $PM \times G$ to give a map $\phi: \bar{E} \to E$. The map ϕ is a bijection and commutes with the G-actions on $\bar{E}$ and E.

Lemma 2. ϕ is a diffeomorphism.

Proof. To show that ϕ is a smooth bijection, it is enough to show that the charts C_Ψ defined for $\bar{E}$ map smoothly to E. The map ϕC_Ψ is

$$U \times G \xrightarrow{(\Psi, \mathrm{id})} PM \times G \xrightarrow{l_*} PE \xrightarrow{e'} E$$

which is smooth due to the fact that a smooth family of paths lifts to a smooth family. Since ϕ is the identity on M and an isomorphism of the fibers, it is clear that ϕC_ψ is a chart for E, and so ϕ is a diffeomorphism. ■

It is also easy to check that the induced mapping of paths takes the lifting function of (B, Γ, b) into the lifting function of $(\bar{B}, \bar{\Gamma}, \bar{b})$. Hence ϕ is a bundle isomorphism preserving M, Γ, and b.

2.1.4. The Reconstructed Lifting Function Defines a Connection

Now the missing lemma in the proof of the reconstruction theorem will be dealt with.

Lemma 3. Let $B = (E, \pi, M, G)$ with lifting function l_* be constructed as in Section 2.1.3. Suppose q is the lift of any path $p \in PM$, and q passes through point $c \in E$, $q(i) = c$. Then $(dq/di)(i)$ depends only on $(dp/di)(i)$. In fact there is a linear injection $\Gamma_c: T_{\pi(c)}M \to T_cE$ such that $\Gamma_c(dp/di) = dq/di$. The images of Γ_c for $c \in E$ define a smooth distribution (Kobayashi and Nomizu, 1963) on E.

Since the lifting property of l_* guarantees that $\pi_* \Gamma_c = \mathrm{id}$, it follows from Lemma 3 that Γ defines a connection.

Proof of Lemma 3. The first point is that the lifting defined by l_* is *local* in the sense that the lift of a section $J \subset I$ of a path $p \in PM$ depends only on the section $p_{\|}J$ of p. More precisely, if $p' \in PM$ agrees with p on J, then on J the lifts of p and p' are related by right translation with a fixed G-element. So the tangent vector of a lift at a point $c \in E$ can depend only on the classes of paths in M which agree in some open neighborhood of $\pi(c)$. For example, it might depend on any of the derivatives of the path p at the point $\pi(c)$. It shall be shown, however, through Lemmas 4–8, that the three conditions H1–H3 in combination are sufficiently restrictive for the tangent to the lift to be determined only by a linear mapping of the first derivative of p at $\pi(c)$.

Lemma 4 eliminates a possible pathology of the holonomy mapping (see also Section 2.5).

Lemma 4. Suppose $\tilde{\psi}: I \to \Omega M$ is a smooth one-parameter family of loops with $\tilde{\psi}(0) = t$, the trivial loop; then $(d/di)(H\tilde{\psi})(0) = 0$.

Proof. Since only the behavior of ψ near the point $* \in M$ is important, we can assume without loss of generality that M is the m-dimensional vector space $\mathbb{R}^m$ with $*$ as the origin. The strategy is to embed $\tilde{\psi}$ in an m-dimensional family of loops $\tilde{\omega}: I^m \to \Omega\mathbb{R}^m$. This is defined by its associated map ω

$$\omega: I^m \times I \to \mathbb{R}^m$$

$$(s_1, s_2, \ldots, s_m, t) \to (\psi_1(s_1, t), \psi_2(s_2, t), \ldots, \psi_m(s_m, t))$$

In this formula ψ_n is the nth coordinate of the function ψ, t denotes the time along the paths, and the s's are the shrinking parameters. The original family $\tilde{\psi}$ is the diagonal of the parameter space I^m: $\tilde{\psi} = \tilde{\omega}\Delta$ with $\Delta: I \to I^m$, $s \to (s, s, \ldots, s)$. Starting from a point d on the diagonal of I^m and moving toward a face of I^m by keeping all coordinates of I^m except the nth fixed, we see that the function $\tilde{\omega}$ smoothly collapses the nth coordinate of the loop so that on the face $s_n = 0$ of I^m the loop has no variation in its nth coordinate. It is the projection onto the plane $x_n = 0$ of $\mathbb{R}^m$ of the original loop $\tilde{\omega}(d)$ (Figure 2).

Now $\tilde{\omega}$ is a smooth family of loops and so $H\tilde{\omega}$ is a smooth mapping. Consequently,

$$\frac{d}{di}(H\tilde{\psi}) = \frac{d}{di}(H\tilde{\omega}\Delta) = \sum_n \frac{\partial}{\partial s_n}(H\tilde{\omega})$$

Evaluating these derivatives at $i = 0$, each term on the right is just a derivative along a coordinate axis $(0, 0, \ldots, s_n, \ldots, 0)$ of I^m. But the loop

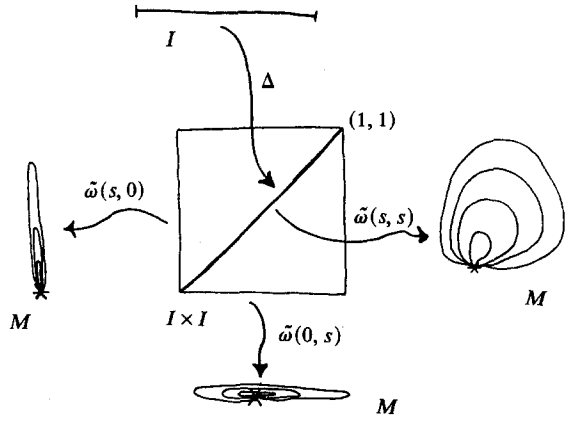


Fig. 2

$t \to (0, 0, \ldots, \psi(s_n, t), \ldots, 0)$ is a thin loop, and the holonomy element is the identity for all values of s_n. Hence $(\partial/\partial s_n)(H\tilde{\omega})(0) = 0$. Then the conclusion of Lemma 4 follows. ■

Returning to the proof of Lemma 3, the local property of the lifting means that it is possible to define the lifting of all paths in M, not only the ones which start at $*$. This lifting has the *product* property: If q lifts p, q' lifts p', and $q(1) = q'(0)$, then $q' \circ q$ lifts $p' \circ p$. It also enjoys the obvious G-invariance and reparametrization invariances. There is a smoothness property which is discussed more below, and a thin equivalence property: If q lifts p, q' lifts p', p is thinly equivalent to p' [$p^{-1} \circ p'$ is a thin loop based at $p(0)$], and $q(0) = q'(0)$, then $q(1) = q'(1)$.

Instead of considering directly the lifting of tangent vectors, it is useful to consider, as a generalization, the lifting of one-parameter families of paths $\tilde{\psi}: I \to M^I$ [$M^I \subset \mathrm{Map}(I, M)$ denoting the unpointed piecewise smooth path space] which shrink at parameter zero to the constant path at some point in M (Figure 3). Such a family is specified by the smooth map $\psi: I \times I \to M$, with $\psi(0, t) = x$. The first factor will be denoted s, the shrinking parameter, the second, t, is the time along the path. The quantity

$$v^1 - v^0 = \frac{\partial \psi(0, 1)}{\partial s} - \frac{\partial \psi(0, 0)}{\partial s}$$

generalizes the tangent vector of a path in M. For if $p: I \to M$ is a path, then putting $\psi(s, t) = p(st)$, it follows that $v^1 = (dp/di)(0)$ and $v^0 = 0$. The vectors v^0, v^1 will be called the boundary vectors of ψ.

Now suppose that $\lambda: I \times I \to E$ lifts ψ, i.e., for fixed s, the path $\lambda(s, t)$ is a lift of $\psi(s, t)$. The function $s \to \lambda(s, 0)$ determining the starting point of the

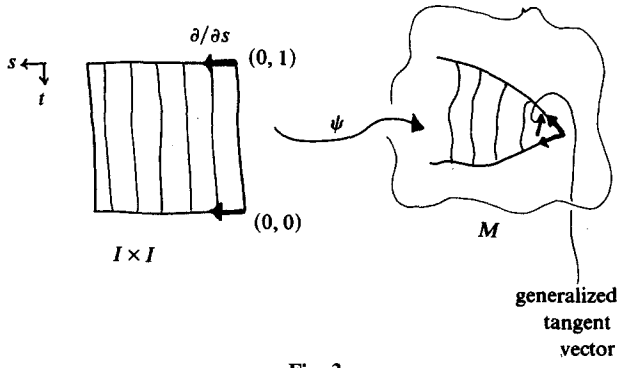


Fig. 3

lifts is arbitrary. If this is chosen to be smooth, then λ is also smooth. Consider the boundary vectors of λ,

$$w^1 = \frac{\partial \lambda\,(0, 1)}{\partial s}, \qquad w^0 = \frac{\partial \lambda\,(0, 0)}{\partial s}$$

Since $t \to \psi(0, t)$ is the trivial path, putting $c = \lambda\,(0, t)$, both w^1 and w^0 lie in T_cE.

The first part of Lemma 3 is a consequence of the following result.

Lemma 5. There exists a linear injection Γ_c: $T_{\pi(c)}M \to T_cE$ such that for any functions ψ and λ defined as above

$$w^1 - w^0 = \Gamma_c(v^1 - v^0)$$

Proof. The proof spans Lemmas 6–8.

Lemma 6. Suppose we have a number of smooth one-parameter families of paths $\tilde{\psi}_k$ in M, $k = 1, 2, \ldots, n$, shrinking to the point $*$, with lifts $\tilde{\lambda}_k$ shrinking to the point $b \in \pi^{-1}(*)$, defined as above, and such that

$$\psi_k(s, 1) = \psi_{k+1}(s, 0) \qquad \text{and} \qquad \psi_n(s, 1) = \psi_1(s, 0) = *$$

so that the composition γ: $i \to \tilde{\psi}_n(i) \circ \tilde{\psi}_{n-1}(i) \circ \cdots \tilde{\psi}_1(i)$ is a map $I \to \Omega M$ with $0 \to t$; then $\sum_{k=1}^{n} w_k^1 - w_k^0 = 0$.

Proof. Define $g_k: I \to G$ to be the maps such that

$$\lambda_{k+1}(s, 0) = \lambda_k(s, 1) g_k(s), \qquad k = 0, 1, \ldots, n$$

putting $\lambda_0(s, t) = \lambda_{n+1}(s, t) = b$ (Figure 4). Then $g_k(0) = \mathrm{id}$, and $H(\gamma(s)) = g_0(s) g_1(s) g_2(s) \cdots g_n(s)$. Lemma 4 then gives

$$\sum_{k=0}^{n} \frac{dg_k(0)}{ds} = 0$$

Using the Leibnitz rule on the defining equation of the g's gives

$$w^0_{k+1} = w^1_k + b \frac{dg_k(0)}{ds}$$

so that $\sum_{k=0}^{n} w^0_{k+1} - w^1_k = 0$, and the result follows since $w^1_0 = w^0_{n+1} = 0$. ■

Lemma 7. Suppose we have two smooth one-parameter families of paths $\tilde{\psi}$ and $\tilde{\psi}'$ in M shrinking to the point $*$, with lifts $\tilde{\lambda}$ and $\tilde{\lambda}'$ shrinking to the point $b \in \pi^{-1}(*)$, and the boundary vectors of ψ, ψ' satisfy $v'^1 = v^1$, $v'^0 = v^0$; then it follows that the boundary vectors of λ, λ' satisfy $w'^1 - w'^0 = w^1 - w^0$.

Proof. Families of paths $\tilde{\eta}$, $\tilde{\alpha}$, $\tilde{\beta}: I \to M^I$ will be defined in such a way that the composition

$$\gamma(i) = \tilde{\eta}^{-1}(i) \circ \tilde{\beta}(i) \circ \tilde{\psi}^{-1}(i) \circ \tilde{\alpha}(i) \circ \tilde{\psi}'(i) \circ \tilde{\eta}(i)$$

is a map $I \to \Omega M$. The β and α will be defined so that $w^1_\beta - w^0_\beta = 0$ and $w^1_\alpha - w^0_\alpha = 0$. Then Lemma 6 will give the desired result since the contributions to $\sum w^1 - w^0$ will cancel from $\tilde{\eta}$ and $\tilde{\eta}^{-1}$.

Since only the behavior of the functions in a neighborhood of $*$ is required, it can be assumed without loss of generality that $M = \mathbb{R}^m$, with $*$

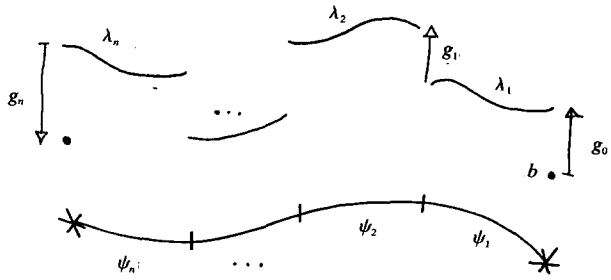


Fig. 4. Picture at a fixed value of s.

as the origin. The functions are

$$\eta(s, t) = t\psi'(s, 0)$$
$$\alpha(s, t) = (1-t)\psi'(s, 1) + t\psi(s, 1)$$
$$\beta(s, t) = (1-t)\psi(s, 0) + t\psi'(s, 0)$$

Since the values and the derivatives of the two curves $s \to \psi'(s, 1)$ and $s \to \psi(s, 1)$ are equal at $s = 0$, the function

$$A: \quad I \times I \to \mathbb{R}^m$$

$$(s, t) \to \begin{cases} \psi'(s, 1) + \dfrac{t}{s^2}[\psi(s, 1) - \psi'(s, 1)], & s \neq 0 \\ \dfrac{t}{2}\left(\dfrac{d^2\psi(0, 1)}{ds^2} - \dfrac{d^2\psi'(0, 1)}{ds^2}\right), & s = 0 \end{cases}$$

is smooth. α can be factored through A: it is the map $\alpha = A\kappa$,

$$\alpha: \quad (s, t) \xrightarrow{\kappa} (s, s^2 t) \xrightarrow{A} M$$

Consequently, if λ_A is a lift of A, then $\lambda_A \kappa$ is a lift of α, using the reparametrization invariance of the lifting. Since $\partial\kappa(0, 1)/\partial s$ is the same tangent vector as $\partial\kappa(0, 0)/\partial s$, it follows that

$$w_\alpha^1 - w_\alpha^0 = \frac{\partial\alpha(0, 1)}{\partial s} - \frac{\partial\alpha(0, 0)}{\partial s} = 0$$

Similarly, $w_b^1 - w_b^1 = 0$. ■

The next lemma shows that only the difference $v^1 - v^0$ is important.

Lemma 8. Suppose we have two smooth one-parameter families of paths $\tilde{\psi}$ and $\tilde{\psi}'$ in M shrinking to the point $*$, with lifts $\tilde{\lambda}$ and $\tilde{\lambda}'$ shrinking to the point $b \in \pi^{-1}(*)$ just as in Lemma 7, but with $v'^1 - v'^0 = v^1 - v^0$; then it follows that $w'^1 - w'^0 = w^1 - w^0$.

Proof. Again we can suppose that $M = \mathbb{R}^m$, with $* = 0$. Suppose four vectors $v^1, v^0, v'^1, v'^0 \in \mathbb{R}^m$ are given, with $v'^1 - v'^0 = v^1 - v^0$. The two functions ψ and $\psi': I \times I \to M$

$$\psi: \quad (s, t) \to s[(1-t)v^0 + tv^1]$$

$$\psi': \quad (s, t) \to s[(1-t)v'^0 + tv'^1]$$

have boundary vectors $v^1 = \partial\psi(0, 1)/\partial s$, etc. The two-parameter family $\tilde{\Phi}: I \times I \to M^I$ is defined to interpolate between the one-parameter families ψ'

and ψ. The Φ is defined by mapping $I\times I\times I$ into $\mathbb{R}^m$ by the linear map which extends

$$(1,1,0)\to v^0, \qquad (0,1,0)\to v'^0$$
$$(1,1,1)\to v^1, \qquad (0,1,1)\to v'^1$$

ψ and ψ' factor through Φ: $\psi=\Phi\sigma$ and $\psi'=\Phi\sigma'$ with σ and σ': $I\times I\to I\times I\times I$ defined by

$$\sigma(s,t)=(s,s,st)$$
$$\sigma'(s,t)=(0,s,st)$$

The point of the construction is that (Figure 5)

$$\xi=\frac{\partial\sigma(0,1)}{\partial s}-\frac{\partial\sigma(0,0)}{\partial s}=\frac{\partial\sigma'(0,1)}{\partial s}-\frac{\partial\sigma'(0,0)}{\partial s}$$

If λ_Φ lifts Φ, then $\lambda_\Phi\sigma$ and $\lambda_\Phi\sigma'$ lift ψ and ψ', and hence $\lambda_\Phi^\sharp(0,0,0)(\xi)=w^1-w^0=w'^1-w'^0$. The lemma is proved for the particular maps ψ, ψ' constructed here by linear interpolation in t. But Lemma 7 shows that any other families of paths will give the same result. ■

To complete the proof of Lemma 5, we note that Lemmas 6–8 show that the map Γ_b: $T_*M\to T_bE$ is well-defined. The fact that it is linear hinges on the product property of the lifting. Suppose that $\tilde\chi$: $I\to M^I$ is the composition of two families $\tilde\psi$ and $\tilde\psi'$ shrinking to the point $*$, i.e.,

$$\tilde\chi(i)=\tilde\psi(i)\circ\tilde\psi'(i) \qquad \text{and} \qquad \tilde\psi(0)=\tilde\psi'(0)=*$$

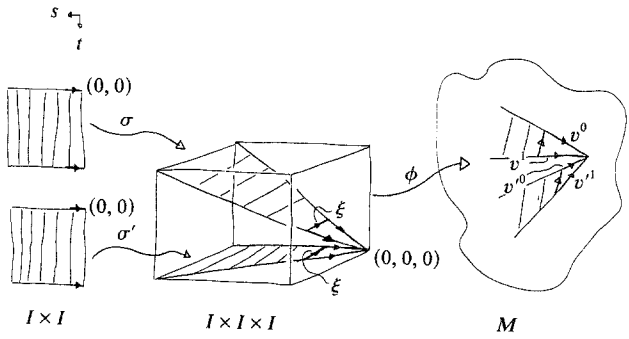


Fig. 5

If v^1, v^0, v'^1, and v'^0 are the boundary vectors of ψ and ψ', then $v^0 = v'^1$, and v^1 and v'^0 are the boundary vectors of χ. The product property of the lifting implies that

$$\Gamma(v^1 - v'^0) = \Gamma(v^1 - v^0) + \Gamma(v'^1 - v'^0)$$

and so Γ is linear.

The conditions of the lemma are thus satisfied at the point $c = b$. The same proof can be applied at other points because the holonomy mapping based at other points in E obeys the same axioms H1–H3. ■

The smoothness of the distribution on E given by the image of Γ is a straightforward consequence of the smoothness property of l_*. This completes Lemma 3. ■

2.1.5. Conclusion

The representation theorem provides an alternative model for the Yang–Mills configuration space, and gives some useful insights into its structure. It is the development of much earlier work on topological bundles (Milnor, 1956; Lashof, 1956), which used homotopy classes of lifting functions to classify topologically distinct bundles.

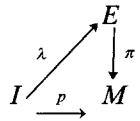
There is a similar type of representation for the gravitational field, which is the subject of the next section. The results here form the basis of the gravitational construction. Apart from this technical role, they also provide an interesting comparison of the construction and structure of the classical Yang–Mills and gravitational fields. This gives a fresh view of the deep similarities and the differences of principle between these two types of physical fields.

Section 2.4 gives some applications of the Yang–Mills results. Section 2.5 relates the curvature two-form to the holonomy of families of loops, and shows how this fits in with the result of Lemma 4.

2.2. Bundles and Liftings

Many physicists are quite happy with the idea that gauge fields are Lie-algebra-valued one-forms on the base manifold, and regard bundles as an extra complication, which one can happily do without. For them, all bundles are trivial, and the connection is a G-invariant one-form on $M \times G$. The extra G's worth looks rather redundant. So why is the language of bundles natural and appropriate here? To answer this question, first the idea of horizontal lifting will be described.

In general terms, if $\pi: E \to M$ is a function, then a lifting assigns a path λ in E to each path p in M in such a way that π projects the path λ onto p:



Usually one can freely specify the starting point $\lambda(0)$ of λ in $\pi^{-1}(p(0))$. The lifting is usually required to vary continuously with respect to both the path in M and the starting point in E, but the details of this will be omitted here.

For a gauge theory (defined by potentials on the base) one can define the parallel transport map $\tau: M^I \to G$. The horizontal lift of a path p into the trivial bundle $M \times G$ starting at $(p(0), g)$ is the path $i \to (p(i), g\tau(K(p, i)))$, where K is the contraction in path space: $K(p, i)$ is the path which traverses the section $[0, i]$ of path p, i.e., $K(p, i)[j] = p(ij)$. The holonomy of a loop, which was defined in Section 2.1, can now be seen to be just the parallel transport operator for the loop.

Any lifting function can be specified by giving, for each path p in the base space, a vector field w in the section of the bundle above the path, so that the lifts λ are the integral curves of the vector field. The vector field must project down onto the tangent vector field of p in the base, and so it is determined up to an arbitrary vertical component. The *horizontal* lifting is a very special type of lifting function. For each point $c \in E$, there is a linear mapping $\Gamma_c: T_{\pi(c)}M \to T_cE$ such that for any path p, $w = d\lambda/di = \Gamma(dp/di)$. The image of Γ is called the horizontal distribution of TE, and at a point c it defines the horizontal subspace of T_cE. To put this another way, the important point is that for *any* two paths passing through the same point, $p(i) = \pi(c)$, and having the same tangent vectors at that point, the tangent vectors of the lifts agree. The vector w depends only on the tangent vector of the path at that point, and on no other feature, local or nonlocal, of the path. That is what distinguishes the horizontal lifting of a connection from any other lifting.

After this slight digression on lifting functions, we come back to the original question: why are bundles essential? The short answer is that in the holonomy representation nontrivial bundles automatically appear alongside the trivial ones (we are supposing that the base space actually has some nontrivial bundles, so that it is not a contractible space, as, for example, $\mathbb{R}^n$ is). The reason for this is that the concept of parallel transport "unifies" the infinitesimal aspect of the connection one-form (on the base space) with the topologically nontrivial global aspect of the finite transformations of the transition functions on overlapping charts. In bundle language it "unifies"

the connection and the topology of the bundle. To put it crudely, the only difference between an electromagnetic field on a circle and a Moebius band, as far as parallel transport and holonomy are concerned, is that in one case the structure group ($U1$ or $\mathbb{R}$) is connected and in the other case (Z_2) it is not. Conventionally, one would ascribe the holonomy in the first case to the presence of a field, and in the second case to the topology of the bundle. When one comes to consider the set of all holonomy mappings, it is rather artificial to separate out those that belong to different bundles. For different connections on the same bundle, the holonomy mappings are in the same homotopy class, so perhaps one could restrict attention to only one homotopy class of maps. This seems a cumbersome idea, and since there is nothing to be lost by considering different bundles, it is not pursued.

There is a second reason for using bundles in Yang–Mills theory. The bundle constructions have a very deep analogy with the reconstruction of the manifold of general relativity, which is described in later sections. The analogy is that the base space of Yang–Mills theory corresponds to the tangent space in gravity, and the total space of the bundle corresponds to the differentiable manifold of general relativity. Thus, to explore the relation between the two theories, it is essential to use the theory of bundles for Yang–Mills fields.

2.3. The *H*-Group Structure of the Loop Space

In homotopy theory the H-group structure of the loop space ΩM is important. The H-group property of the composition law is that the following three homotopy equivalences hold:

$$\omega \sim \omega \circ t$$

$$\omega \circ \omega^{-1} \sim t$$

$$(\alpha \circ \beta) \circ \gamma \sim \alpha \circ (\beta \circ \gamma)$$

where t is the trivial loop, so that if the loop space is factored by homotopy equivalence, the result is an algebraic group.

The important thing to notice is that these homotopies are all of the form

$$I \times I \xrightarrow{h} I \xrightarrow{\Omega} M$$

where Ω is the loop on the left-hand side of the three homotopy equivalences. The map h is not continuous, but the overall map Ωh is. The homotopies are thus actually all thin equivalences, and so $\Omega M/\theta$ (θ is the thin equivalence

relation) is a group. For the analogous constructions in the space of piecewise linear loops on a simplicial complex, see Milnor's work on universal bundles (Milnor, 1956).

2.4. Flat Bundles in Physics

There are many situations in physics where the information about a gauge field is given by specifying its holonomy. One example is a gauge field which has zero curvature, giving homotopic loops the same holonomy element, but with nontrivial holonomy elements for noncontractible loops. The holonomy mapping descends to a group homomorphism

$$H: \quad \pi_1(K) \to G$$

Now any such holonomy mapping obeys conditions H1–H3. H1 is immediate, H2 follows because thin equivalence is a restricted notion of homotopy equivalence, and H3 follows because if $\tilde{\phi}: U \to \Omega M$ is a family of loops, then $H\tilde{\phi}$ is constant on connected components of U, and hence smooth. The reconstruction theorem can be applied, and so it is clear that the gauge field is properly specified by just the information in H.

This type of situation occurs in several different physical models. Some examples are the Aharonov–Bohm effect, vacuum configurations for gauge theories in a Kaluza–Klein context, where the holonomy elements are a symmetry-breaking mechanism, and again in the context of topological field theories.

2.5. The Curvature Two-Form

It is a well-known "fact" that the holonomy of a "small" loop expands as

$$1 + \int_V F$$

where the integral is over the region V which the loop bounds. One can make this expansion precise: If $\tilde{\psi}: I \to \Omega M$ is a smooth family of loops with $\tilde{\psi}(0) = t$, then the second derivative of the holonomy is related to the curvature tensor F on the base manifold by

$$\frac{d^2(H\tilde{\psi})}{ds^2}(0) = F(A)$$

A is a bivector, which contracts with F to give a Lie algebra element. A indicates the asymptotic shape of the area of the loops as they shrink to

zero. Its definition is

$$A^{\mu\nu}=\frac{\partial^2}{\partial s^2}\int_{\partial V} x^{[\mu}dx^{\nu]}=\frac{\partial^2}{\partial s^2}\int_{V} dx^{[\mu}\wedge dx^{\nu]}$$

where x^μ is the coordinate function, and the derivatives are taken at $s=0$. These relations can be proved by differentiating the differential equation for the parallel transport operator. The linear term of the expansion of the holonomy of a family of loops shrinking to the trivial loop is zero:

$$\frac{d(H\tilde{\psi})}{ds}(0)=0$$

This is the content of Lemma 4 of Section 2.1.

Note: The quantity $A^{\mu\nu}$ is a tensor at the point $*$. One can take the $\partial^2/\partial s^2$ inside the integral to define it as

$$\int_0^l dt\,\frac{dx^{[\mu}}{ds}\,\frac{d^2x^{\nu]}}{ds\,dt}$$

with the derivatives evaluated at $s=0$, which is an integral over vectors at $*$, since $x^\mu(0, t)=*$. The other two terms of the differentiation vanish.

3. HOLONOMY AND GRAVITY

3.1. Introduction

The last section was concerned with Yang–Mills theory in its own right. This section presents the analogous constructions for general relativity. The Yang–Mills work provides both the general framework for a similar treatment of gravity and some specific results which are of use in this section. The work on gravity is not complete in the same way as the Yang–Mills results, and so a less formalized presentation is used. The difficulties are of a fairly technical nature; mainly questions about differentiability and differentiable structure, and are still open problems. However, the goals are the same: we are looking for the holonomy representation of a gravitational field configuration, an axiomatization of holonomy, a reconstruction theorem for the field configuration, up to a diffeomorphism, and finally a representation theorem for the gravitational configuration space. Perhaps the reader can keep in mind the more formal development of the Yang–Mills theory.

In order to present properly the analogy which motivates the gravitational constructions, the basic idea of the Yang–Mills theory is presented afresh.

3.2. Yang–Mills

A Yang–Mills field is a connection on a principal bundle $B=(E, \pi, M, G)$. The base space M can be any differentiable manifold, but we have in mind mainly Minkowski space. A curved space-time M involves gravity, and since the point of this section is to represent a gravitational field in a different way, it is better to think of the Yang–Mills theory being in the absence of gravity.

The important aspect of the connection here is that it provides a lifting: given a path in the base space and a point in the fiber over the initial point of the path, it gives a path in the bundle starting at the given point. This point is the one whose tangent vectors are horizontal at every point, and project down onto the tangent vectors of the original path in the base.

Using the idea of lifting, we can describe the points of the total space of the bundle in an unusual manner. We fix a basepoint $*$ in the base once and for all, and also one in the fiber over $*$, so that the fiber becomes a copy of G. A point $c \in E$ is to be described by an equivalence class of $PM \times G$. It will be convenient to call this equivalence class c also. A pair (p, g) is in c if p starts at $*$ and ends at $\pi(c)$, and the lift of p starting at g ends precisely at c.

This description of the bundle B describes the connection neatly, too. The best way to see this is to exhibit the lifting, just using the information given in the equivalence classes. This was described in Section 2.

The final part of the story is that there is an equivalence relation which describes the classes in $PM \times G$ of the points in E. R is the relation $(p, g) \sim (p', g')$ if the endpoints of p and p' coincide, and $g' = H(p^{-1} \circ p')g$. H is the *holonomy mapping* of closed loops in M to their holonomy elements in G. Thus, both the bundle and the Yang–Mills field configuration, the connection, are described by the holonomy mapping. The holonomy mapping involves only the base space M and the group G. To calculate the field configuration it describes, the bundle has to be constructed using the equivalence relation R:

$$E = PM \times G/R$$

and then the lifting function follows.

Section 2 presents this in much more detail, together with axioms for the subspace $\mathscr{H} \subset \mathrm{Map}(\Omega M, G)$ which represents all possible holonomy mappings.

So, to recap, Yang–Mills theory can be said to be about the holonomy of loops in M. The lifts of these loops do not close, the endpoints being in the same fiber, and mapped onto each other by the holonomy element in G. In gravity we shall see that the "lifted" loops do not close by a translation in a space-time direction, and by a rotation of frames.

3.3. Gravity

The gravitational field is described by a differentiable manifold X with a metric, and a connection on the bundle $O(X)$ of orthonormal frames. As before, we pick an arbitrary basepoint $*$ in X and a frame f at $*$, so that the fiber over X is identified with the Lorentz group G. To achieve the holonomy description of gravity, we need the idea of *development*. The development of a curve in the base space $p \in PX$ is defined by horizontally lifting the curve into $O(X)$. Then, using the canonical $\mathbb{R}^n$-valued 1-form on $O(X)$ (the "*vierbein*"), e, the integral $C(t) = \int_0^t e$, integrating along the lift of p, gives a curve in $\mathbb{R}^n$, which is identified via f with the tangent space at $*$, viewed as an affine space with a metric, i.e., Minkowski space M. For the rest of the section, M is now definitely Minkowski space. Intuitively, we can think of the development as the curve in PM with the same geometry as the path p, that is, it bends through the same angles at the same proper distances as p. It has the same intrinsic and extrinsic geometry as p. One can also define the curve C as the unique path in M obeying $C(0) = 0$ and

$$\frac{dC}{dt} = \tau_{K(p,t)}\left(\frac{dp}{dt}\right)$$

where $\tau_{K(p,t)}$ is the parallel transport map: $T_{p(t)}X \to T_*X$. Note that a closed loop in PX will not in general develop a closed loop in M.

Now we can see how to describe the manifold X:

$$\text{point of } X = \text{subset of } PM$$

where p is in $x \in X$ if p is the development of a path which ends at x. The gravitational field is described:

$$\text{point of } O(X) = \text{subset of } PM \times G$$

where the G element is defined in exactly the same manner as for Yang–Mills theories: the frame is parallel transported back to the fiber above $*$, using the connection. This second set of subsets is a very powerful object. It describes the set X, its differentiable structure, its metric, the frame bundle $O(X)$, and the connection on it, as will be shown in detail below. In the same way as for Yang–Mills theory, there is an underlying equivalence relation yielding these subsets of $PM \times G$ as equivalence classes. This relation

is specified by the *holonomy set* $\Omega \subset PM \times G$, which is just the subset of $PM \times G$ which corresponds to the frame f. One can decompose Ω into a set $P \subset PM$ and a mapping $h: P \to G$, so that $(p, g) \in \Omega$ iff $p \in P$ and $g = h(p)$. The set P is the subset of PM corresponding to the point $* \in X$, or in other words, the paths in M which are the development of closed loops in X. The function h is the *Lorentz holonomy mapping* of P, and is simply the holonomy mapping, in the sense described for Yang–Mills theory, of the Lorentz connection in the bundle $O(X)$ for the loops in X corresponding to the elements of P.

The (*affine*) *holonomy mapping* H of P takes a point (p, g) to the isometry of M given by a rotation of g about the origin followed by a translation by an amount $p(1)$, the value of the endpoint of the path:

$$H: \quad P \to A$$

$$p \to p(1)h(p)$$

where A is the Poincaré group. H is in fact the holonomy mapping of the affine connection in the bundle of affine frames. Note that H can be derived from the information in Ω, or, equivalently, the information in P and h.

Now we come to the description of the two equivalence relations, on PM and on $PM \times G$, which reconstruct the manifold and the bundle of orthonormal frames, purely in terms of the information in P and h. The relation R on $PM \times G$ is

$$R: \quad (p, g) \sim (p', g') \qquad \text{if there exists } \pi \in P \text{ such that}$$

$$\pi = (H(\pi)p'^{-1}) \circ p$$

$$\text{and} \quad g = h(\pi)g'$$

The first is rather a curious equation, since π "appears on both sides." If we want to consider $(\pi, h(\pi))$ as a transformation acting on the space $PM \times G$, the condition can be rewritten: There exists $\pi \in P$ such that

$$p \sim (H(\pi)p') \circ \pi$$

$$g = h(\pi)g'$$

with the $\sim$ denoting thin equivalence (see Section 2).

With the relation R we can form the set

$$E = PM \times G/R$$

which is identical to the total space of the bundle. Using the relation R' on PM defined in a similar way,

$$R': \quad p \sim p' \qquad \text{if there exists } \pi \in P \text{ such that}$$

$$\pi = (H(\pi)p'^{-1}) \circ p$$

it is clear that PM/R' is the set X, and that the projection $\pi: E \to X$ defined by $\{p, g\}_R \to \{p\}_{R'}$ is intrinsic to the construction. It is also extremely plausible that the differentiable structure, metric, and connection are implicit in the constructions. As far as the metric goes, we can see that X is "made of" sections of paths in M, and so distances can be measured by measuring the corresponding distances in M. This can be formalized by defining the inverse development map Δ,

$$\Delta: \quad PM \to PX$$
$$\Delta(p)[i] = \{K(p, i)\}$$

where K is the contraction in path space introduced in Section 2: $K(p, i)[j] = p(ij)$. The map Δ gives paths $\Delta(p)$ in the set X with a known metric geometry, i.e., that of $p \in PM$, and hence can be expected to give a metric to X. The differential structure should follow because Δ allows the construction of smooth families of paths in X from smooth families of paths in M. In other words, the differentiable structure is that of $M = \mathbb{R}^4$, patched together in a way determined by the equivalence relation. From these considerations, it is seen that Δ plays the role for the manifold X that l_* plays for the bundle in Yang–Mills theory.

Finally, the connection in the bundle $E \to X$ should be reconstructable in much the same way as for Yang–Mills theory. In fact the strategy is to construct the space-time manifold, and then the holonomy mapping of the (Lorentz) connection, satisfying the axioms H1–H3 of the Yang–Mills theory. Then the Yang–Mills results can be used to construct the connection on the frame bundle.

3.4. Axiomatization and Reconstruction

The previous section showed how to develop the holonomy representation of gravity by analogy with the Yang–Mills construction. It exhibited the holonomy information (P, h), and showed that it is plausible that a complete reconstruction of the gravitational field configuration should follow from it. This section seeks to axiomatize the set $P \subset PM$ and the mapping $h: P \to G$, and provide a reconstruction from the axioms. Actually, the basic ingredients of the reconstruction were spelt out in the previous section. The axioms are not yet complete, and so parts of the reconstruction cannot be done rigorously. The difficulties are pointed out.

Since the Lorentz group is used throughout, it is to be expected that the reconstruction will yield a bundle with a metric compatible connection. However, the notion of torsion has no basic construction in the holonomy scheme of things, and so the connection will be independent of the metric, and the torsion may take any value. Later, when the field equations are

discussed, the equation torsion = 0 will be seen to have a similar status to the Einstein field equation.

While discussing exactly what the gravitational configuration space will be taken to be, there are two further properties of a gravitational field to be mentioned, namely completeness and the Hausdorff property. The question of what happens to the holonomy of a non-Hausdorff manifold is rather interesting. It seems likely, however, that the reconstruction process from the holonomy information will only yield Hausdorff manifolds. This is because in a non-Hausdorff manifold there exist paths $[0, 1)\to X$, open at one end, with two possible endpoints to complete the curve to a path $[0, 1]\to X$. These paths have the same development, which is of finite (Euclidean) length. [It is to be supposed that an arbitrary positive-definite metric is used in the tangent space M, as is usual in discussions of completeness (Hawking and Ellis, 1973).] Such paths can be constructed by the following method. Because the manifold is non-Hausdorff, there exist two points x and x' which cannot be separated by open sets. Now in two charts on open sets U and U' containing the points x and x', respectively, we can consider two sequences of open balls B_n and B'_n, centered on x and x'. The intersection of B_n with B'_n must be nonempty, and we choose a sequence of points p_n, with p_n lying in the intersection of B_n and B'_n. If the radii of the balls are chosen to converge to zero as $n\to\infty$, the sequence of points p_n converges to both x and x'. The path $[0, 1)\to X$ is constructed by joining together the points in a suitable fashion, for example, by linear interpolation in a chart. If the radii of the balls are chosen to converge sufficiently fast, then the development of the path will have finite length. Now in the holonomy description the points of space-time are labeled by the developments of the paths of particles which arrive at that point; the points have no *a priori* existence themselves. The construction of the manifold consists in just grouping together the paths of all the particles which are in coincidence, defined by the equivalence relation R'. So it would be impossible if in the resulting manifold there were two points which were the endpoint of just one particle path. So it would seem that non-Hausdorff manifolds are unphysical, if we believe that the holonomy construction actually represents the physics of the definition of the manifold, as an abstraction from the behavior of particles.

We can still contemplate a non-Hausdorff manifold, however, and ask what the essential difference is in the behavior of the holonomy, in distinction to a Hausdorff manifold. It seems to me that, because two distinct paths in the manifold may have the same development, probably one of two things can happen. Either the holonomy does not obey the axioms to be presented below, or it does, and the reconstruction process identifies regions which have an identical geometry. The standard example of a non-Hausdorff manifold is made by taking two copies of the real line $\mathbb{R}$, and identifying them

on the open interval $(0, \infty)$. The two points 0 in each copy are the two points which cannot be separated by open sets. The tangent space is $\mathbb{R}$. If the metric on the manifold is taken to be just the standard metric on $\mathbb{R}$, the development of a path on the manifold is the path in $\mathbb{R}$ one gets by identifying the two "bottom legs" $(-\infty, 0]$ of the manifold. Now the holonomy set will be identical to that of $\mathbb{R}$, as the base manifold, and so the reconstruction process will give back as the manifold only $\mathbb{R}$. In other words, the two geometrically indistinguishable "legs" $(-\infty, 0]$ get identified. If, on the other hand, the two legs are given different geometries [this is not possible in one dimension!—so imagine a two-dimensional example $\mathbb{R}\times\mathbb{R}$ identified on $\mathbb{R}\times(0, \infty)$], then the holonomy description may be inconsistent. For example, two loops with identical development but which are partly in different "legs" may have different Lorentz group holonomy elements. Hence the mapping h: $P\to G$ would not be definable.

Completeness of a manifold can be defined in terms of the development mapping δ: $PX\to PM$ which takes paths to their developments. A manifold is said to be complete if $\mathrm{Im}(\delta)=PM$. In physical terms, every conceivable particle motion, as defined by its geometry, can take place, and ends at some point in the manifold. This notion of completeness coincides with the notion of b-completeness (Hawking and Ellis, 1973). Completeness is defined in this way because for a pseudo-Riemannian metric the manifold does not have the metric space structure that a positive metric would give. In the case of a positive metric, the metric space completeness coincides with development completeness (Kobayashi and Nomizu, 1963).

For the reasons given above, then, a gravitational field configuration will be defined as a connected, Hausdorff manifold which is complete in the sense that the development map is complete, with a metric of Lorentz signature, and a connection on the bundle of orthonormal frames, so that the connection is metric compatible, but may have nonzero torsion.

Now we turn to the details of the axiomatization. We start with P and h. From it H: $P\to A$ is defined in the same way as before: $H(p)=p(1)h(p)$. Then a *product* operation $*$ on P is defined:

$$p_1 * p_2=(H(p_2)p_1)\circ p_2$$

and an *inverse* operation $p\to\bar{p}$:

$$\bar{p}=H(p)^{-1}p^{-1}$$

The first and second axioms are as follows.

G1 (Homomorphism). If $p_1, p_2\in P$, then $p_1 * p_2\in P$, and $h(p_1 * p_2)=h(p_2)h(p_1)$.

G2 (Inverse). If $p\in P$, then $\bar{p}\in P$.

Now at some stage it has to be shown that if P and h are derived from a gravitational field, i.e., manifold X, metric, etc., then the axioms hold. In order to motivate the axioms, this will be done as the axioms are stated.

Proof of G1 and G2 for a gravitational field configuration. P is the development of the loop space of the manifold, and the product and inverse operations on P are just the product and inverse operations of the loop space. The homomorphism condition on h is just the corresponding condition for the holonomy mapping of the Lorentz connection. ■

Since a path which is thinly equivalent to a loop is also a loop, and the holonomy mapping of a connection agrees on thinly equivalent loops, it is natural to propose the following.

G3 (Thin equivalence). P contains complete thin equivalence classes of paths. The map h agrees on these equivalence classes.

Attempted Proof of G3 for a Gravitational Field. The proof of this is not complete, and is one of the technical problems referred to at the beginning of the section. The proof rests on the following conjecture: For $p, p' \in PX$, p is thinly equivalent to p' iff $\delta(p)$ is thinly equivalent to $\delta(p')$. The problem in proving this is that in general the homotopy which establishes one of the thin equivalences does not develop (or inverse-develop) to a homotopy which establishes the other one. This is easily seen to be the case if the image of one of the thin loops is not a simply connected set. However, one can establish the proof for a special class of thin loops, ones which can be transformed to the trivial loop by a finite number of operations of either (1) reparametrization of the loop, or (2) replacing subsections of the path of the form $p^{-1} \circ p$ with the constant path at $p(0)$. It is not actually clear whether this is a more restricted class of thin loops than the original definition. There may be a pathological example which shows that it is a restricted class.

There are three possible ways out of the problem: establish the conjecture, modify the definition of thin loops to a more restricted set (all the proofs would work with the restricted notion of thin loop given above), or prove that this restricted set is actually all the thin loops. In any event, the lack of a proof does not seem too serious. The proof will, however, be assumed in the following. ■

Returning to the axiomatic development, a few facts can be established, from the axioms.

Proposition. H agrees on thin equivalence classes.

Proposition. If P is not empty, $t \in P$ and $h(t) = \mathrm{id}$, t being the trivial loop.

Proof. There exists $p \in P$, hence $\bar{p} * p = p^{-1} \circ p \in P$, which is thinly equivalent to t, so $t \in P$. Then $h(t) = h(\bar{p} * p) = h(p)h(\bar{p})$. But $h(t) = h(t * t) = h(t)^2$, so $h(t) = \mathrm{id}$. ■

The proof also shows the following result.

Proposition. $h(\bar{p}) = h(p)^{-1}$.

Now we consider the relation R' introduced in the last section. The three axioms introduced are sufficient to prove that R' is an equivalence relation. The proof is rather tedious, and not particularly illuminating, and so is omitted. Then the set X is defined by $X = PM/R'$, with basepoint $* = \{t\}$, and the inverse development map $\Delta: PM \to PX$ is defined as in the last section: $\Delta(p)[i] = \{K(p, i)\}$. Actually it is a bit premature using the space PX, as the set X does not even yet have a topology, let alone a differentiable structure. The map Δ still exists, however, as a map into the space of functions $I \to X$, and so for the moment PX will be defined as just the image of the map Δ, as a subspace of the space of functions $I \to X$. Likewise, ΩX is defined as the subset $\{p: p(1) = *\}$.

Proposition. If $p_1, p_2 \in PM$ and $p_1 \sim p_2$ by thin equivalence, then $p_1 \sim p_2$ by the relation R'.

Proof. Define $\theta = p_1^{-1} \circ p_2$. Then $\theta \in P$ and $H(\theta) = \mathrm{id}$, and so

$$(H(\theta)p_1^{-1}) \circ p_2 = \theta$$

that is, $p_1 \sim p_2$ by R'. ■

Proposition. Δ maps P onto ΩX, and $p_1 * p_2$ is mapped to the composition of loops $\Delta(p_1) \circ \Delta(p_2)$.

Proof. For a general $p \in P$, $\Delta(p)(1) = \{p\}$, so to show that $\Delta(p)$ is a loop, it is necessary to show that $\{p\} = \{t\}$. If q is the path $t \circ p$, which is just a reparametrization of p, then q is thinly equivalent to p, and $q \in P$. Moreover, $q = (H(q)t^{-1}) \circ p$, which means that, by definition, $\{t\} = \{p\}$. Clearly, the reverse argument holds; if $\{t\} = \{p\}$, then $p \in P$, and so $\Delta(p)$ is a loop.

For the product property, consider two general paths $p_1, p_2 \in P$. Then $\Delta(p_1 * p_2) = \Delta(H(p_2)p_1 \circ p_2) = q_1 \circ \Delta(p_2)$, where q_1 is the path $q_1(i) = \{(H(p_2)K(p_1, i)) \circ p_2\}$. But it is easy to show that

$$(H(p_2)q) \circ p_2 \overset{R'}{\sim} q$$

for any path q, and so $q_1(i) = \{K(p, i)\}$, i.e., $q_1 = \Delta(p_1)$. ■

Before going any further, this is the right point to introduce the smoothness axioms. Clearly, so far there is no reason to suppose that X is a manifold; in fact, with the axioms presented so far it may be very far removed from a smooth finite-dimensional manifold. For example, take P to be a proper subgroup of $\Omega\mathbb{R}^n$, for example, say by removing the loops passing through some of the points of $\mathbb{R}^n$: $P=\Omega U$, with U a subset of $\mathbb{R}^n$, and h to be trivial: $h(\omega)=\mathrm{id}$ for all $\omega\in P$. Then P and h will satisfy the axioms G1–G3, but the reconstruction will yield something rather bizarre, certainly not a manifold. The paths which end at points outside U will not be related by R' to any other paths. So if p is such a path, it will form an independent point of X. The set X will contain "too many points" in the sense that if one considers deforming p locally, then X is locally the same "dimension" as the path space PM. Similarly, one can also imagine that too many points may be identified, leading too small a dimension for X. It may also happen that the paths are identified by R' in a chaotic, discontinuous manner, not admitting any smooth structure for the set X.

With these points in mind, the axiom G4 should be tentatively (and imprecisely) stated as:

P is a smooth submanifold of PM, of codimension four.

Before discussing what this might mean, a simple example serves to motivate the axiom. Consider $P\mathbb{R}^4$, with $*=0$. This is a vector space with pointwise addition of paths. The loop space $\Omega\mathbb{R}^4$ is a linear subspace of $P\mathbb{R}^4$, and so by any decent definition of the manifold structure of $P\mathbb{R}^4$, this would be a submanifold. The quotient space $Q=P\mathbb{R}^4/\Omega\mathbb{R}^4$ is a vector space of dimension four, and coincides with set $X=P\mathbb{R}^4/R'$. In fact, Q is isomorphic to $\mathbb{R}^4$ by the endpoint map of the path space.

To arrive at a more precise notion of what a smooth submanifold is, we have to examine the smooth structure, which is relevant here, of the space PM. There is a well-defined notion of a smooth map into PM from a finite-dimensional manifold: the notion of a smooth family of paths. There is also a well-defined notion of a smooth map from PM into a manifold Z: the smooth families ψ: $U\to PM$ give a smooth map $U\to PM\to Z$. So it is natural to suppose that the canonical projection map α: $PM\to PM/R'=X$ should be smooth in this sense. Now $\alpha=e\Delta$, where e is the endpoint map $PX\to X$. This means that the differentiable structure of X should be defined in such a way that for any smooth family of paths $\tilde{\psi}$: $U\to PM$, the map

$$U \xrightarrow{\tilde{\psi}} PM \xrightarrow{\Delta} PX \xrightarrow{e} X$$

is smooth. At this point, to check that we are on the right track, we can

compare this expression with the construction used in the Yang–Mills theory to construct charts on the total space E of the bundle

$$C_\psi\colon U\times G \xrightarrow{(\tilde\psi,\mathrm{id})} PM\times G \xrightarrow{l_*} PE \xrightarrow{e'} E$$

which is very similar. It has already been remarked that the inverse-development function Δ plays the same role for the manifold construction as l_* did for the bundle construction of the Yang–Mills theory.

Without going too deeply into the technical problems involved [which start with the fact that $\tilde\psi(U)$ is not open in PM], we shall assume, as axiom G4, what should be the principal conclusions of the above tentative proposal.

G4. There is a unique four-dimensional smooth structure on X such that for any smooth family $\tilde\psi\colon U\to PM$ the map $e\Delta\tilde\psi\colon U\to X$ is smooth. Paths are mapped nondegenerately into X: if $p\in PM$ and $(dp/di)(i)\neq 0$, then $[d\Delta(p)/di](i)\neq 0$.

The smoothness for the map h is straightforward:

G5. For any smooth family $\tilde\psi\colon U\to P$, the map $h\tilde\psi\colon U\to G$ is smooth.

The proof strategy only for the remainder of the reconstruction will be sketched, as the work is not yet complete, and a more thorough exposition would be premature. The main points to show are:

1. Δ can be used to construct "Riemann normal coordinates" around any point $x\in X$. For example, around the point $*$, the family $\rho\colon M\to PM$, $\rho(m)[i]=im$ (i.e., radial straight lines) should, in some neighborhood of the origin, map invertibly to X. The derivative at the origin is nonzero on account of the nondegeneracy condition in G4.

2. Δ is a 1–1 mapping of PM to PX.

3. For paths $p, q\in PM$, p is thinly equivalent to q if and only if $\Delta(p)$ is thinly equivalent to $\Delta(q)$.

4. The mapping h can, by virtue of the 1–1 correspondence of P and ΩX, be regarded as a holonomy mapping $\Omega X\to G$. The axioms H1–H3 are established by the points above.

5. The reconstruction theorem is applied to h. The construction is the same as that given by the relation R directly on $PM\times G$. The resulting G-principal bundle needs to be interpreted as the frame bundle of X. At a point $c\in E$, the frame is defined as the mapping

$$\theta_c\colon \qquad M\to T_{\pi(c)}X$$

$$g\frac{dp}{di}(1)\to\frac{d\Delta(p)}{di}(1)$$

for any point $(p, g)\in PM\times G$ which is in the equivalence class of c. One has

to show that this mapping agrees on all the different possible paths p, and is linear. This is the same as the problem of defining a connection in the Yang–Mills theory, and one should be able to use the Yang–Mills result by defining the affine holonomy mapping $\Omega X \to A$, and constructing its bundle (which should be the isometric affine frame bundle) and connection. Its connection splits into two parts, the Lorentz connection, and the inverse of the map θ.

The nondegenerateness of θ, which is essential to the notion of a frame, follows directly from the nondegenerate property in G4.

6. Finally, the metric on X follows once the frames are established, essentially from the fact that the reconstructed bundle contains only a subset of frames, the orthonormal ones. The map θ_c is used to map the metric on the Minkowski space M to the tangent space at x on X. These metrics will agree for all the points c in the fiber above x, because the frames are all related by Lorentz transformations in M.

4. THE FIELD EQUATIONS OF GRAVITY

4.1. Introduction

In previous sections, the holonomy description gave rise to a gravitational field with a connection that was naturally metric-compatible, but the fields were otherwise arbitrary. In other words, the Einstein field equation was not imposed, and the torsion was also arbitrary. The aim of this section is to demonstrate a form of the field equations which is naturally suited to the holonomy scheme. There are really two field equations, the Einstein equation and the torsion equation, and, as we shall see, they are naturally paired as the linear momentum field equation and the angular momentum field equation.

So far, the holonomy description has proceeded with the minimal use of tensors; there are displacement vectors and Lorentz group elements, but nothing more complicated than this. The plethora of different tensor types which usually accompanies general relativity is rather a foreign element in this approach. So, to continue in this spirit, maximal use will be made of differential forms in expressing the equations of motion. To give an example, the energy-momentum tensor p^{ab} is best expressed as a vector-valued three-form $p^a = p^{ab}\varepsilon_{bcde}e^c \wedge e^d \wedge e^e$, where e is the unit vector-value one-form. This has a more geometric meaning than the former; when integrated with a "small" three-surface element (over which the curvature can be ignored) it gives the energy-momentum passing through that surface. It also has the technical advantage that the covariant exterior derivative can be applied,

giving different treatment to vectors and differential forms, when there is torsion in general.

Going further than this, the spirit is to introduce the relevant geometric objects over which the differential forms are to be integrated. This has been pursued to its logical conclusion already in previous sections as far as the connection one-form is concerned: one-forms are integrated along paths, and the result is a theory formulated in terms of functions on path space. The strategy here is then to introduce "small" three-surface elements and their bounding two-spheres over which the momenta are integrated. Precise results can be expressed in terms of the limit of a family of spheres which shrink to a point in a smooth way.

4.2. Field Equations

The two equations of motion are (Kibble, 1961; Sciama, 1962)

$$\tfrac{1}{2} R^{ab} \wedge e^c \varepsilon_{abcd} = p_d \qquad \text{(energy-momentum density)}$$

$$\tfrac{1}{2} \tau^a \wedge e^b \varepsilon_{abcd} = S_{cd} \qquad \text{(spin density)}$$

where R is the curvature two-form and τ the torsion two-form. The two quantities on the left-hand sides are the Einstein tensor and the modified torsion tensor. When the right-hand sides are set to zero the equations become the vacuum Einstein equation and the equation $\tau = 0$, expressing the connection form in terms of the metric.

Let us examine the Einstein equation first. What is needed is a precise expression of the idea that over a sufficiently "small" three-surface V, so that the curvature over its extent can be neglected, the integral of the Einstein tensor over V is equal to the matter energy-momentum passing through the surface V,

$$\tfrac{1}{2} \int_V R^{ab} \wedge e^c \varepsilon_{abcd} \sim \text{energy-momentum through } V$$

To perform this integral properly, what is needed is a notion of parallel transport for the vector index of the integrand. The idea is that V is "sufficiently small" for any reasonable parallel transport of the vectors to one point in V to produce a result differing only by corrections of higher order in the size of V from the result of the integration itself. Suppose that V is topologically a three-disk, so that the boundary ∂V is a topological two-sphere. Then V can be filled with a family of curves (a "spray") which each start at the same point v in the interior of V and end at the different points of ∂V. then the vector in the integrand at a point $u \in V$ is parallel transported

along the unique path which links it with v. An alternative way of looking at this is that a special type of gauge has been picked, that in which the connection form ω is zero in the directions along the curves, and the integration is performed in this gauge. At the point v, curves radiate in all directions, and so $\omega = 0$.

The quantity $I(V) = \frac{1}{2}\int_V R^{ab} \wedge e^c \varepsilon_{abcd}$, defined by the aid of a particular spray of curves from v, is the same up to *third order* in a small parameter s, for any spray of curves. What this means precisely is that if one has a smooth one-parameter family of three-disks $V(s)$ shrinking to the point v at $s=0$ (i.e., a smooth map to M of a regular cone whose base is a standard three-disk), and one attaches two different smooth families of sprays to these, then the two different integrals $I(V(s))$ agree up to third order in s at $s=0$. Clearly, since the integration is over a three-dimensional region, this just follows from the equality of the integrands at the point v. Note that for this result it is important that $V(s)$ is parametrized smoothly, as defined above.

This formula is not particularly exciting as it stands, but can be rewritten in an interesting way. The same spray of curves can be developed into Minkowski space M. Suppose that the manifold has a basepoint $*$, and that c is an arbitrary curve which connects $*$ to v. The paths that were chosen in V are connected to c and then developed into M, the tangent space of the point $*$. Then the region V can be mapped into M by mapping a point $u \in V$ to the endpoint of the development of the path leading to u.

Let us formalize this briefly. We started with a family of paths $\tilde{\psi}$: $V \to P_v X$ ($P_v X$ is the path space of the manifold X, based at v). These ended at the point in question: $\psi(u, 1) = u$. The map x: $V \to M$ was defined by $x(u) = \delta(\tilde{\psi}(u) \circ c)(1)$, where δ is the development map $P_* X \to PM$.

Again, one can regard this as a special type of coordinate gauge fixing, x providing a particular type of coordinate chart, such that $dx = e$ for vectors along the family of curves. The point is that ω and x obey

$$\omega(v) = 0, \qquad dx(v) = e$$

so that the "coordinates" are very particular ones: they are adapted to the geometry of the manifold at v.

The local conservation of energy-momentum $dp^a(v) = 0$ (which actually only holds if the torsion or curvature vanishes at v) suggests that the integral over V can be rewritten as a boundary term. In fact, the relevant expression is

$$\tfrac{1}{2}\varepsilon_{abcd} \int_{\partial V} R^{ab} x^c \simeq \int_V p_d$$

which again holds up to third order. This formula is true because, using Stokes' formula, the integrand at v is equal to $\frac{1}{2}\varepsilon_{abcd}$ times

$$d(R^{ab}x^c)(v) = (dR^{ab}x^c + R^{ab} \wedge e^c)(v)$$

At the point v, exterior differentiation d is equivalent to exterior covariant differentiation (D), because $\omega(v) = 0$. Hence the Bianchi identity gives $dR = 0$. The formula is known as Cartan's moment of rotation (Cartan, 1922; Misner *et al.*, 1972). It relates the energy-momentum threading through the two-sphere ∂V to the integral of the "moment of rotation" $R^{[ab}x^{c]}$. So this is our first field equation, that the integral formula holds to third order for any smooth family of two-spheres which shrinks smoothly to a point $v \in X$. It is perhaps a highly inefficient way of stating the Einstein equation, but physically it gives a very appealing picture. Note that, in contrast to the conservation law $Dp^a = 0$, which only holds when $\tau = 0$, the integral expression gives the correct formula with torsion.

The formula for Cartan's moment of rotation found an application to Regge's theory of discrete general relativity (Regge, 1961), which for my part, was a result of considering the formulas here (Barrett, 1985, 1986, 1987, 1988; Miller, 1986). Regge's equations of motion can be understood as a discrete version of the two-sphere expression which rather surprisingly turns out to be an exact formula, rather than approximate to third order as it is here.

In Cartan's time the relativists did not think of the torsion equation as a second field equation, although Cartan and other mathematicians worked out the mathematical theory of torsion. The relativists just set $\tau = 0$, and probably did not ask for the same sort of physical picture that Cartan's moment of rotation gives to the Einstein equation. As far as I am aware, the presentation of the torsion equation as a "small two-sphere" integral expression is a new idea.

The modified torsion tensor, which forms the left-hand side of the torsion equation, is not covariantly conserved, since it is equal only to the matter spin density and not the total angular momentum. However, by using the idea that the total angular momentum should be covariantly conserved, one can arrive at the following formula: If we set

$$M_{ab} = S_{ab} + x_{[a}p_{b]}$$

where the quantities S, p, and $x(v)$ are evaluated at v, then

$$M_{ab} = d(\varepsilon_{abcd}(\tfrac{1}{8}R^{cd}x^2 + \tfrac{1}{2}C^c x^d))(v)$$

where C^a is the translational curvature $C^a(u) = R^{ab}x_b - \tau^a$. The C^a is the

translational part of the affine holonomy for a "small" loop at the end of the curve c.

The corresponding integral expression is

$$\int_V M_{ab} \simeq \varepsilon_{abcd} \int_{\partial V} \tfrac{1}{8} R^{cd} x^2 + \tfrac{1}{2} C^c x^d$$

with the equality again holding to third order. The two terms in the integrand might be called the first moment of translational curvature and a second moment of (rotational) curvature. The equation states that the integral of these two moments of curvature over a small two-sphere ∂V is equal to the matter angular momentum passing through the two-sphere.

This equation can be regarded as the second field equation because if the first field equation holds, one can subtract the "orbital" angular momentum $x(v_{[a})p_{b]}$ from the total angular momentum M_{ab}, and the result is the equation relating the modified torsion tensor to the spin density.

The origin dependence of the total angular momentum M_{ab} is what one would expect. If the angular momentum is measured from a different origin, or indeed the same origin but connected to it by a different path c', then the coordinate of v changes by $x' - x$, and the angular momentum changes by $M' - M = (x' - x_{[a})p_{b]}$. In addition, if there is a change of frame, all the vector indices are rotated by the Lorentz transformation. The linear momentum was independent of change of origin, but behaved as a vector under a change of frame.

Due to the origin dependence of M_{ab}, it is possible to sort out the intrinsic spin of the matter from the orbital angular momentum, thus resolving the ambiguity noted by Kibble (1961). Roughly speaking, the spin part of the angular momentum is the part that cannot be transformed away by the change of origin mentioned above. The details of how this works in ordinary flat space is contained in Penrose and MacCullum (1973).

The two small-sphere expressions for linear and angular momenta presented here have appeared previously in different contexts, due to the fact that the small-sphere limit is the same as the weak-field limit (Penrose, 1982). The expressions, for zero torsion, therefore agree with the twistor expressions for quasilocal momenta (Penrose, 1982; Kelly *et al.*, 1986).

A suitable representation of the equations of motion has been achieved with the aid of the notion of gravitational holonomy, particularly the use of development. The linear and angular momenta of matter passing through small two-spheres are equated with integrals of moments of the two curvatures, rotational and translational. All these quantities have fairly immediate significance in the holonomy description. The curvatures are essentially the holonomy elements for small loops, as explained in Section 2.5. The other

quantity is the displacement vector x defined using the notion of development. Thus, the equations are constraints on the quantities of holonomy and development of paths in a fairly direct way.

ACKNOWLEDGMENTS

Thanks are due to the following for discussing various areas: Jeeva Anandan, Rob Baston, Bernard Kay, Michael Singer (Yang–Mills axioms), Chris Clarke (Hausdorff property), David Elworthy (path space topology), and Luke Hodgkin and Graeme Segal (literature). I am indebted to Chris Isham, Tom Kibble, Robin Tucker, and Nick Webber for general discussion and criticism.

REFERENCES

Aharonov, Y., and Bohm, D. (1959). *Physical Review*, **115**, 485–491.

Anandan, J. (1983). Holonomy groups in gravity and gauge fields, in *Proceedings Conference Differential Geometric Methods in Physics, Trieste 1981*, G. Denardo and H. D. Doebner, eds., World Scientific, Singapore.

Atiyah, M. F. (1980). Geometrical aspects of gauge theories, in *Proceedings International Congress Mathematics*, O. Lehto, ed., Helsinki, pp. 881–885.

Babelon, O., and Viallet, C. M. (1981). *Communications in Mathematical Physics*, **81**, 515–525.

Barrett, J. W. (1985). The holonomy description of classical Yang–Mills theory and general relativity, Ph.D. thesis, University of London.

Barrett, J. W. (1986). *Classical and Quantum Gravity*, **3**, 203–206.

Barrett, J. W. (1987). *Classical and Quantum Gravity*, **4**, 1565–1576.

Barrett, J. W. (1988). *Classical and Quantum Gravity*, **5**, 1187–1192.

Barrett, J. W. (1989). *General Relativity and Gravitation*, **21**, 457–466.

Bialynicki-Birula, I. (1963). *Bulletin de l'Académie Polonaise des Sciences*, **11**, 135.

Cartan, E. (1922). *Comptes Rendus*, **174**, 437–439.

Chan, H.-M., and Tsou, S. T. (1986). *Acta Physica Polonica B*, **17**, 259–276.

Chan, H.-M., Scharbach, P., and Tsou, S. T. (1986). *Annals of Physics*, **166**, 396–421.

Dirac, P. A. M. (1931). *Proceedings of the Royal Society of London A*, **133**, x–xxi.

Dugundji, J. (1966). *Topology*, Allyn and Bacon, Boston.

Durhuus, B. (1980). *Letters in Mathematical Physics*, **4**, 515–522.

Einstein, A. (1922). *The Meaning of Relativity*, 6th ed., Chapman and Hall, London.

Fischer, A. E. (1986). *General Relativity and Gravitation*, **18**, 597–608.

Giles, R. (1981). *Physical Review D*, **24**, 2160–2168.

Hawking, S. W., and Ellis, G. F. R. (1973). *The Large Scale Structure of Space-Time*, Cambridge University Press, Cambridge.

Isham, C. J. (1981). Quantum gravity—An overview, in *Quantum Gravity 2, A Second Oxford Symposium*, C. J. Isham, R. Penrose, and D. W. Sciama, eds., Clarendon Press, Oxford.

Isham, C. J. (1984). Topological and global aspects of quantum theory, in *1983 Les Houches Summer School Lectures "Relativity Groups and Topology"*, North-Holland, Amsterdam.

Kelly, R. M., Tod, K. P., and Woodhouse, N. M. J. (1986). *Classical and Quantum Gravitation*, **3**, 1151–1167.

Kibble, T. W. B. (1961). *Journal of Mathematical Physics*, **2**, 212–221.

Kobayashi, S. (1954). *Comptes Rendus*, **238**, 443–444.
Kobayashi, S., and Nomizu, K. (1963). *Foundations of Differential Geometry*, Volume 1, Interscience, New York.
Lashof, R. (1956). *Annals of Mathematics*, **64**, 436–446.
Mandelstam, S. (1962*a*). *Annals of Physics*, **19**, 1–24.
Mandelstam, S. (1962*b*). *Annals of Physics*, **19**, 25–66.
Mandelstam, S. (1968*a*). *Physical Review*, **175**, 1580–1603.
Mandelstam, S. (1968*b*). *Physical Review*, **175**, 1604–1623.
Miller, W. A. (1986). *Foundations of Physics*, **16**, 143–169.
Milnor, J. (1956). *Annals of Mathematics*, **63**, 272–284.
Misner, C. W., Thorne, K. S., and Wheeler, J. A. (1972). *Gravitation*, Freeman, San Francisco.
Narasimhan, M. S., and Ramadas, T. R. (1979). *Communications in Mathematical Physics*, **67**, 121–136.
Penrose, R. (1982). *Proceedings of the Royal Society of London A*, **381**, 53–63.
Penrose, R., and MacCullum, M. A. H. (1973). *Physics Reports*, **6C**, 241–316.
Polyakov, A. M. (1979). *Nuclear Physics B*, **164**, 171–188.
Regge, T. (1961). *Nuovo Cimento*, **19**, 558–571.
Sciama, D. W. (1962). On the analogy between charge and spin in general relativity, in *Recent Developments in General Relativity*, Pergamon, Oxford.
Singer, I. M. (1981). *Physica Scripta*, **24**, 817–820.
Spanier, E. H. (1966). *Algebraic Topology*, McGraw-Hill, New York.
Teleman, C. (1960). *Annales Scientifiques de l'École Normale Supérieure 3*, **77**, 195–234.
Teleman, C. (1963). *Annali di Matematica, Pura ed Applicata*, **LXII**, 379–412.
Teleman, C. (1969*a*). *Indagationes Mathematicae*, **31**, 89–103.
Teleman, C. (1969*b*). *Indagationes Mathematicae*, **31**, 104–112.
Wilson, K. G. (1974). *Physical Review D*, **10**, 2445–2459.