The large scale structure of space–time

S.W. HAWKING & G.F.R. ELLIS

CAMBRIDGE MONOGRAPHS ON MATHEMATICAL PHYSICS

CAMBRIDGE

more information – www.cambridge.org/9780521200165

This page intentionally left blank

CAMBRIDGE MONOGRAPHS
ON MATHEMATICAL PHYSICS
General Editors: D.W. Sciama, S. Weinberg, P.V. Landshoff

THE LARGE SCALE STRUCTURE OF SPACE-TIME

THE LARGE SCALE STRUCTURE OF SPACE–TIME

S. W. HAWKING, F.R.S.
Lucasian Professor of Mathematics in the University of Cambridge and
Fellow of Gonville and Caius College

AND

G. F. R. ELLIS
Professor of Applied Mathematics, University of Cape Town

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
40 West 20th Street, New York, NY 10011-4211, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521200165

First published 1973
First paperback edition 1974
20th printing 2006

Printed in the United States of America

A catalog record for this publication is available from the British Library.

ISBN-13 978-0-521-20016-5 hardback
ISBN-10 0-521-20016-4 hardback

ISBN-13 978-0-521-09906-6 paperback
ISBN-10 0-521-09906-4 paperback

To
D. W. SCIAMA

Contents

Preface

The subject of this book is the structure of space–time on length-scales from 10^{-13} cm, the radius of an elementary particle, up to 10^{28} cm, the radius of the universe. For reasons explained in chapters 1 and 3, we base our treatment on Einstein's General Theory of Relativity. This theory leads to two remarkable predictions about the universe: first, that the final fate of massive stars is to collapse behind an event horizon to form a 'black hole' which will contain a singularity; and secondly, that there is a singularity in our past which constitutes, in some sense, a beginning to the universe. Our discussion is principally aimed at developing these two results. They depend primarily on two areas of study: first, the theory of the behaviour of families of timelike and null curves in space–time, and secondly, the study of the nature of the various causal relations in any space–time. We consider these subjects in detail. In addition we develop the theory of the time-development of solutions of Einstein's equations from given initial data. The discussion is supplemented by an examination of global properties of a variety of exact solutions of Einstein's field equations, many of which show some rather unexpected behaviour.

This book is based in part on an Adams Prize Essay by one of us (S. W. H.). Many of the ideas presented here are due to R. Penrose and R. P. Geroch, and we thank them for their help. We would refer our readers to their review articles in the *Battelle Rencontres* (Penrose (1968)), Midwest Relativity Conference Report (Geroch (1970*c*)), Varenna Summer School Proceedings (Geroch (1971)), and Pittsburgh Conference Report (Penrose (1972*b*)). We have benefited from discussions and suggestions from many of our colleagues, particularly B. Carter and D. W. Sciama. Our thanks are due to them also.

Cambridge
January 1973

S. W. Hawking
G. F. R. Ellis

1
The role of gravity

The view of physics that is most generally accepted at the moment is that one can divide the discussion of the universe into two parts. First, there is the question of the local laws satisfied by the various physical fields. These are usually expressed in the form of differential equations. Secondly, there is the problem of the boundary conditions for these equations, and the global nature of their solutions. This involves thinking about the edge of space–time in some sense. These two parts may not be independent. Indeed it has been held that the local laws are determined by the large scale structure of the universe. This view is generally connected with the name of Mach, and has more recently been developed by Dirac (1938), Sciama (1953), Dicke (1964), Hoyle and Narlikar (1964), and others. We shall adopt a less ambitious approach: we shall take the local physical laws that have been experimentally determined, and shall see what these laws imply about the large scale structure of the universe.

There is of course a large extrapolation in the assumption that the physical laws one determines in the laboratory should apply at other points of space–time where conditions may be very different. If they failed to hold we should take the view that there was some other physical field which entered into the local physical laws but whose existence had not yet been detected in our experiments, because it varies very little over a region such as the solar system. In fact most of our results will be independent of the detailed nature of the physical laws, but will merely involve certain general properties such as the description of space–time by a pseudo-Riemannian geometry and the positive definiteness of energy density.

The fundamental interactions at present known to physics can be divided into four classes: the strong and weak nuclear interactions, electromagnetism, and gravity. Of these, gravity is by far the weakest (the ratio Gm^2/e^2 of the gravitational to electric force between two electrons is about 10^{-40}). Nevertheless it plays the dominant role in shaping the large scale structure of the universe. This is because the strong and weak interactions have a very short range ($\sim 10^{-13}$ cm or less), and although electromagnetism is a long range interaction, the repulsion of like charges is very nearly balanced, for bodies of macroscopic dimensions, by the attraction of opposite charges. Gravity on the other hand appears to be always attractive. Thus the gravitational fields of all the particles in a body add up to produce a field which, for sufficiently large bodies, dominates over all other forces.

Not only is gravity the dominant force on a large scale, but it is a force which affects every particle in the same way. This universality was first recognized by Galileo, who found that any two bodies fell with the same velocity. This has been verified to very high precision in more recent experiments by Eotvos, and by Dicke and his collaborators (Dicke (1964)). It has also been observed that light is deflected by gravitational fields. Since it is thought that no signals can travel faster than light, this means that gravity determines the causal structure of the universe, i.e. it determines which events of space–time can be causally related to each other.

These properties of gravity lead to severe problems, for if a sufficiently large amount of matter were concentrated in some region, it could deflect light going out from the region so much that it was in fact dragged back inwards. This was recognized in 1798 by Laplace, who pointed out that a body of about the same density as the sun but 250 times its radius would exert such a strong gravitational field that no light could escape from its surface. That this should have been predicted so early is so striking that we give a translation of Laplace's essay in an appendix.

One can express the dragging back of light by a massive body more precisely using Penrose's idea of a closed trapped surface. Consider a sphere $\mathscr{T}$ surrounding the body. At some instant let $\mathscr{T}$ emit a flash of light. At some later time t, the ingoing and outgoing wave fronts from $\mathscr{T}$ will form spheres $\mathscr{T}_1$ and $\mathscr{T}_2$ respectively. In a normal situation, the area of $\mathscr{T}_1$ will be less than that of $\mathscr{T}$ (because it represents ingoing light) and the area of $\mathscr{T}_2$ will be greater than that of $\mathscr{T}$ (because it represents outgoing light; see figure 1). However if a sufficiently large amount of matter is enclosed within $\mathscr{T}$, the areas of $\mathscr{T}_1$ and $\mathscr{T}_2$ will *both* be less than that of $\mathscr{T}$. The surface $\mathscr{T}$ is then said to be a closed trapped surface. As t increases, the area of $\mathscr{T}_2$ will get smaller and smaller provided that gravity remains attractive, i.e. provided that the energy density of the matter does not become negative. Since the matter inside $\mathscr{T}$ cannot travel faster than light, it will be trapped within a region whose boundary decreases to zero within a finite time. This suggests that something goes badly wrong. We shall in fact show that in such a situation a space–time singularity must occur, if certain reasonable conditions hold.

One can think of a singularity as a place where our present laws of physics break down. Alternatively, one can think of it as representing part of the edge of space–time, but a part which is at a finite distance instead of at infinity. On this view, singularities are not so bad, but one still has the problem of the boundary conditions. In other words, one does not know what will come out of the singularity.

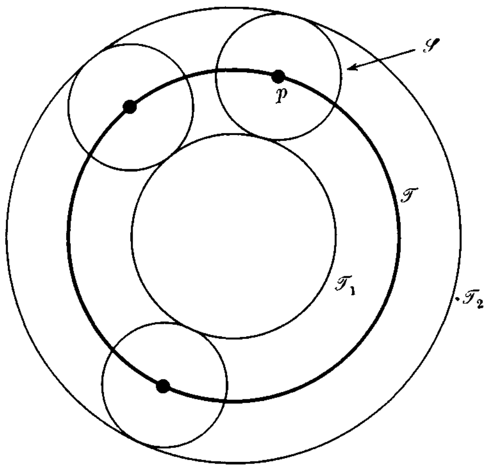

FIGURE 1. At some instant, the sphere $\mathscr{T}$ emits a flash of light. At a later time, the light from a point p forms a sphere $\mathscr{S}$ around p, and the envelopes $\mathscr{T}_1$ and $\mathscr{T}_2$ form the ingoing and outgoing wavefronts respectively. If the areas of *both* $\mathscr{T}_1$ and $\mathscr{T}_2$ are less than the area of $\mathscr{T}$, then $\mathscr{T}$ is a closed trapped surface.

There are two situations in which we expect there to be a sufficient concentration of matter to cause a closed trapped surface. The first is in the gravitational collapse of stars of more than twice the mass of the sun, which is predicted to occur when they have exhausted their nuclear fuel. In this situation, we expect the star to collapse to a singularity which is not visible to outside observers. The second situation is that of the whole universe itself. Recent observations of the microwave background indicate that the universe contains enough matter to cause a time-reversed closed trapped surface. This implies the existence of a singularity in the past, at the beginning of the present epoch of expansion of the universe. This singularity is in principle visible to us. It might be interpreted as the beginning of the universe.

In this book we shall study the large scale structure of space–time on the basis of Einstein's General Theory of Relativity. The predictions of this theory are in agreement with all the experiments so far performed. However our treatment will be sufficiently general to cover modifications of Einstein's theory such as the Brans–Dicke theory.

While we expect that most of our readers will have some acquaintance with General Relativity, we have endeavoured to write this book so that it is self-contained apart from requiring a knowledge of simple calculus, algebra and point set topology. We have therefore devoted chapter 2 to differential geometry. Our treatment is reasonably modern in that we have formulated our definitions in a manifestly coordinate independent manner. However for computational convenience we do use indices at times, and we have for the most part avoided the use of fibre bundles. The reader with some knowledge of differential geometry may wish to skip this chapter.

In chapter 3 a formulation of the General Theory of Relativity is given in terms of three postulates about a mathematical model for space–time. This model is a manifold $\mathscr{M}$ with a metric $\mathbf{g}$ of Lorentz signature. The physical significance of the metric is given by the first two postulates: those of local causality and of local conservation of energy–momentum. These postulates are common to both the General and the Special Theories of Relativity, and so are supported by the experimental evidence for the latter theory. The third postulate, the field equations for the metric $\mathbf{g}$, is less well experimentally established. However most of our results will depend only on the property of the field equations that gravity is attractive for positive matter densities. This property is common to General Relativity and some modifications such as the Brans–Dicke theory.

In chapter 4, we discuss the significance of curvature by considering its effects on families of timelike and null geodesics. These represent the paths of small particles and of light rays respectively. The curvature can be interpreted as a differential or tidal force which induces relative accelerations between neighbouring geodesics. If the energy–momentum tensor satisfies certain positive definite conditions, this differential force always has a net converging effect on non-rotating families of geodesics. One can show by use of Raychaudhuri's equation (4.26) that this then leads to focal or conjugate points where neighbouring geodesics intersect.

To see the significance of these focal points, consider a one-dimensional surface $\mathscr{S}$ in two-dimensional Euclidean space (figure 2). Let p

be a point not on $\mathscr{S}$. Then there will be some curve from $\mathscr{S}$ to p which is shorter than, or as short as, any other curve from $\mathscr{S}$ to p. Clearly this curve will be a geodesic, i.e. a straight line, and will intersect $\mathscr{S}$ orthogonally. In the situation shown in figure 2, there are in fact three geodesics orthogonal to $\mathscr{S}$ which pass through p. The geodesic through the point r is clearly not the shortest curve from $\mathscr{S}$ to p. One way of recognizing this (Milnor (1963)) is to notice that the neighbouring

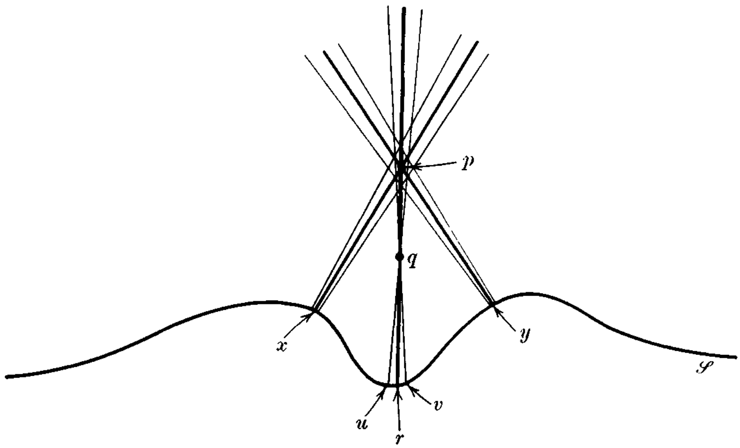

FIGURE 2. The line pr cannot be the shortest line from p to $\mathscr{S}$, because there is a focal point q between p and r. In fact either px or py will be the shortest line from p to $\mathscr{S}$.

geodesics orthogonal to $\mathscr{S}$ through u and v intersect the geodesic through r at a focal point q between $\mathscr{S}$ and p. Then joining the segment uq to the segment qp, one could obtain a curve from $\mathscr{S}$ to p which had the same length as a straight line rp. However as uqp is not a straight line, one could round off the corner at q to obtain a curve from $\mathscr{S}$ to p which was shorter than rp. This shows that rp is not the shortest curve from $\mathscr{S}$ to p. In fact the shortest curve will be either xp or yp.

One can carry these ideas over to the four-dimensional space–time manifold $\mathscr{M}$ with the Lorentz metric $\mathbf{g}$. Instead of straight lines, one considers geodesics, and instead of considering the shortest curve one considers the longest timelike curve between a point p and a spacelike surface $\mathscr{S}$ (because of the Lorentz signature of the metric, there will be no shortest timelike curve but there may be a longest such curve). This longest curve must be a geodesic which intersects $\mathscr{S}$ orthogonally, and there can be no focal point of geodesics orthogonal to $\mathscr{S}$ between

$\mathscr{S}$ and p. Similar results can be proved for null geodesics. These results are used in chapter 8 to establish the existence of singularities under certain conditions.

In chapter 5 we describe a number of exact solutions of Einstein's equations. These solutions are not realistic in that they all possess exact symmetries. However they provide useful examples for the succeeding chapters and illustrate various possible behaviours. In particular, the highly symmetrical cosmological models nearly all possess space–time singularities. For a long time it was thought that these singularities might be simply a result of the high degree of symmetry, and would not be present in more realistic models. It will be one of our main objects to show that this is not the case.

In chapter 6 we study the causal structure of space–time. In Special Relativity, the events that a given event can be causally affected by, or can causally affect, are the interiors of the past and future light cones respectively (see figure 3). However in General Relativity the metric $\mathbf{g}$ which determines the light cones will in general vary from point to point, and the topology of the space–time manifold $\mathscr{M}$ need not be that of Euclidean space R^4. This allows many more possibilities. For instance one can identify corresponding points on the surfaces $\mathscr{S}_1$ and $\mathscr{S}_2$ in figure 3, to produce a space–time with topology $R^3 \times S^1$. This would contain closed timelike curves. The existence of such a curve would lead to causality breakdowns in that one could travel into one's past. We shall mostly consider only space–times which do not permit such causality violations. In such a space–time, given any spacelike surface $\mathscr{S}$, there is a maximal region of space–time (called the Cauchy development of $\mathscr{S}$) which can be predicted from knowledge of data on $\mathscr{S}$. A Cauchy development has a property ('Global hyperbolicity') which implies that if two points in it can be joined by a timelike curve, then there exists a longest such curve between the points. This curve will be a geodesic.

The causal structure of space–time can be used to define a boundary or edge to space–time. This boundary represents both infinity and the part of the edge of space–time which is at a finite distance, i.e. the singular points.

In chapter 7 we discuss the Cauchy problem for General Relativity. We show that initial data on a spacelike surface determines a unique solution on the Cauchy development of the surface, and that in a certain sense this solution depends continuously on the initial data. This chapter is included for completeness and because it uses a number

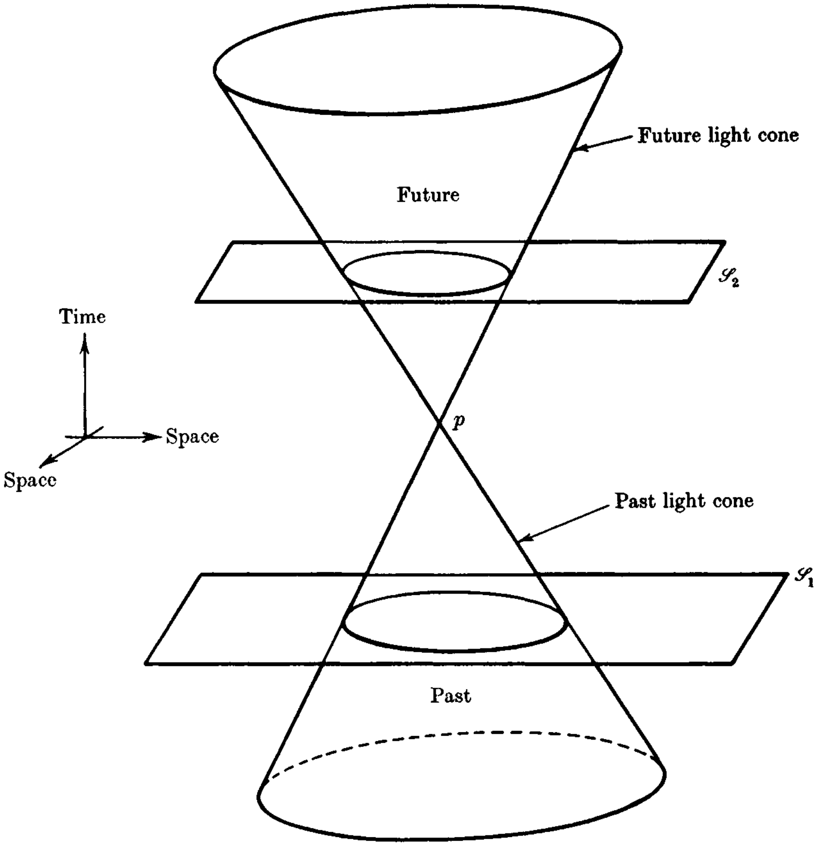

FIGURE 3. In Special Relativity, the light cone of an event p is the set of all light rays through p. The past of p is the interior of the past light cone, and the future of p is the interior of the future light cone.

of results of the previous chapter. However it is not necessary to read it in order to understand the following chapters.

In chapter 8 we discuss the definition of space–time singularities. This presents certain difficulties because one cannot regard the singular points as being part of the space–time manifold $\mathscr{M}$.

We then prove four theorems which establish the occurrence of space–time singularities under certain conditions. These conditions fall into three categories. First, there is the requirement that gravity shall be attractive. This can be expressed as an inequality on the energy–momentum tensor. Secondly, there is the requirement that there is enough matter present in some region to prevent anything escaping from that region. This will occur if there is a closed trapped surface, or if the whole universe is itself spatially closed. The third requirement is that there should be no causality violations. However this requirement is not necessary in one of the theorems. The basic idea of the proofs is to use the results of chapter 6 to prove there must be longest timelike curves between certain pairs of points. One then shows that if there were no singularities, there would be focal points which would imply that there were no longest curves between the pairs of points.

We next describe a procedure suggested by Schmidt for constructing a boundary to space–time which represents the singular points of space–time. This boundary may be different from that part of the causal boundary (defined in chapter 6) which represents singularities.

In chapter 9, we show that the second condition of theorem 2 of chapter 8 should be satisfied near stars of more than $1\frac{1}{2}$ times the solar mass in the final stages of their evolution. The singularities which occur are probably hidden behind an event horizon, and so are not visible from outside. To an external observer, there appears to be a 'black hole' where the star once was. We discuss the properties of such black holes, and show that they probably settle down finally to one of the Kerr family of solutions. Assuming this to be the case, one can place certain upper bounds on the amount of energy which can be extracted from black holes. In chapter 10 we show that the second conditions of theorems 2 and 3 of chapter 8 should be satisfied, in a time-reversed sense, in the whole universe. In this case, the singularities are in our past and constitute a beginning for all or part of the observed universe.

The essential part of the introductory material is that in § 3.1, § 3.2 and § 3.4. A reader wishing to understand the theorems predicting the existence of singularities in the universe need read further only chapter 4, § 6.2–§ 6.7, and § 8.1 and § 8.2. The application of these theorems to collapsing stars follows in § 9.1 (which uses the results of appendix B); the application to the universe as a whole is given in § 10.1, and relies on an understanding of the Robertson–Walker universe models (§ 5.3). Our discussion of the nature of the singularities is contained in § 8.1, § 8.3–§ 8.5, and § 10.2; the example of Taub–NUT space (§ 5.8) plays an important part in this discussion, and the Bianchi I universe model (§ 5.4) is also of some interest.

A reader wishing to follow our discussion of black holes need read only chapter 4, § 6.2–§ 6.6, § 6.9, and § 9.1, § 9.2 and § 9.3. This discussion relies on an understanding of the Schwarzschild solution (§ 5.5) and of the Kerr solution (§ 5.6).

Finally a reader whose main interest is in the time evolution properties of Einstein's equations need read only §6.2–§6.6 and chapter 7. He will find interesting examples given in §5.1, §5.2 and §5.5.

We have endeavoured to make the index a useful guide to all the definitions introduced, and the relations between them.

2
Differential geometry

The space–time structure discussed in the next chapter, and assumed through the rest of this book, is that of a manifold with a Lorentz metric and associated affine connection.

In this chapter, we introduce in § 2.1 the concept of a manifold and in § 2.2 vectors and tensors, which are the natural geometric objects defined on the manifold. A discussion of maps of manifolds in § 2.3 leads to the definitions of the induced maps of tensors, and of submanifolds. The derivative of the induced maps defined by a vector field gives the Lie derivative defined in § 2.4; another differential operation which depends only on the manifold structure is exterior differentiation, also defined in that section. This operation occurs in the generalized form of Stokes' theorem.

An extra structure, the connection, is introduced in § 2.5; this defines the covariant derivative and the curvature tensor. The connection is related to the metric on the manifold in § 2.6; the curvature tensor is decomposed into the Weyl tensor and Ricci tensor, which are related to each other by the Bianchi identities.

In the rest of the chapter, a number of other topics in differential geometry are discussed. The induced metric and connection on a hypersurface are discussed in § 2.7, and the Gauss–Codacci relations are derived. The volume element defined by the metric is introduced in § 2.8, and used to prove Gauss' theorem. Finally, we give a brief discussion in § 2.9 of fibre bundles, with particular emphasis on the tangent bundle and the bundles of linear and orthonormal frames. These enable many of the concepts introduced earlier to be reformulated in an elegant geometrical way. § 2.7 and § 2.9 are used only at one or two points later, and are not essential to the main body of the book.

2.1 Manifolds

A manifold is essentially a space which is locally similar to Euclidean space in that it can be covered by coordinate patches. This structure permits differentiation to be defined, but does not distinguish intrinsically between different coordinate systems. Thus the only concepts defined by the manifold structure are those which are independent of the choice of a coordinate system. We will give a precise formulation of the concept of a manifold, after some preliminary definitions.

Let R^n denote the *Euclidean space of n dimensions*, that is, the set of all n-tuples $(x^1, x^2, \ldots, x^n)$ $(-\infty < x^i < \infty)$ with the usual topology (open and closed sets are defined in the usual way), and let $\frac{1}{2}R^n$ denote the 'lower half' of R^n, i.e. the region of R^n for which $x^1 \leqslant 0$. A map ϕ of an open set $\mathscr{O} \subset R^n$ (respectively $\frac{1}{2}R^n$) to an open set $\mathscr{O}' \subset R^m$ (respectively $\frac{1}{2}R^m$) is said to be of class C^r if the coordinates $(x'^1, x'^2, \ldots, x'^m)$ of the image point $\phi(p)$ in $\mathscr{O}'$ are r-times continuously differentiable functions (the rth derivatives exist and are continuous) of the coordinates $(x^1, x^2, \ldots, x^n)$ of p in $\mathscr{O}$. If a map is C^r for all $r \geqslant 0$, then it is said to be C^∞. By a C^0 map, we mean a continuous map.

A function f on an open set $\mathscr{O}$ of R^n is said to be locally Lipschitz if for each open set $\mathscr{U} \subset \mathscr{O}$ with compact closure, there is some constant K such that for each pair of points $p, q \in \mathscr{U}$, $|f(p) - f(q)| \leqslant K|p-q|$, where by $|p|$ we mean

$$\{(x^1(p))^2 + (x^2(p))^2 + \ldots + (x^n(p))^2\}^{\frac{1}{2}}.$$

A map ϕ will be said to be locally Lipschitz, denoted by C^{1-}, if the coordinates of $\phi(p)$ are locally Lipschitz functions of the coordinates of p. Similarly, we shall say that a map ϕ is C^{r-} if it is C^{r-1} and if the $(r-1)$th derivatives of the coordinates of $\phi(p)$ are locally Lipschitz functions of the coordinates of p. In the following we shall usually only mention C^r, but similar definitions and results hold for C^{r-}.

If $\mathscr{P}$ is an arbitrary set in R^n (respectively $\frac{1}{2}R^n$), a map ϕ from $\mathscr{P}$ to a set $\mathscr{P}' \subset R^m$ (respectively $\frac{1}{2}R^m$) is said to be a C^r map if ϕ is the restriction to $\mathscr{P}$ and $\mathscr{P}'$ of a C^r map from an open set $\mathscr{O}$ containing $\mathscr{P}$ to an open set $\mathscr{O}'$ containing $\mathscr{P}'$.

A C^r *n-dimensional manifold* $\mathscr{M}$ is a set $\mathscr{M}$ together with a C^r *atlas* $\{\mathscr{U}_\alpha, \phi_\alpha\}$, that is to say a collection of charts $(\mathscr{U}_\alpha, \phi_\alpha)$ where the $\mathscr{U}_\alpha$ are subsets of $\mathscr{M}$ and the ϕ_α are one–one maps of the corresponding $\mathscr{U}_\alpha$ to open sets in R^n such that

(1) the $\mathscr{U}_\alpha$ cover $\mathscr{M}$, i.e. $\mathscr{M} = \bigcup_\alpha \mathscr{U}_\alpha$,

(2) if $\mathscr{U}_\alpha \cap \mathscr{U}_\beta$ is non-empty, then the map

$$\phi_\alpha \circ \phi_\beta^{-1}\colon \phi_\beta(\mathscr{U}_\alpha \cap \mathscr{U}_\beta) \to \phi_\alpha(\mathscr{U}_\alpha \cap \mathscr{U}_\beta)$$

is a C^r map of an open subset of R^n to an open subset of R^n (see figure 4).

Each $\mathscr{U}_\alpha$ is a *local coordinate neighbourhood* with the local coordinates x^a ($a = 1$ to n) defined by the map ϕ_α (i.e. if $p \in \mathscr{U}_\alpha$, then the coordinates of p are the coordinates of $\phi_\alpha(p)$ in R^n). Condition (2) is the requirement that in the overlap of two local coordinate neighbourhoods, the coordinates in one neighbourhood are C^r functions of the coordinates in the other neighbourhood, and vice versa.

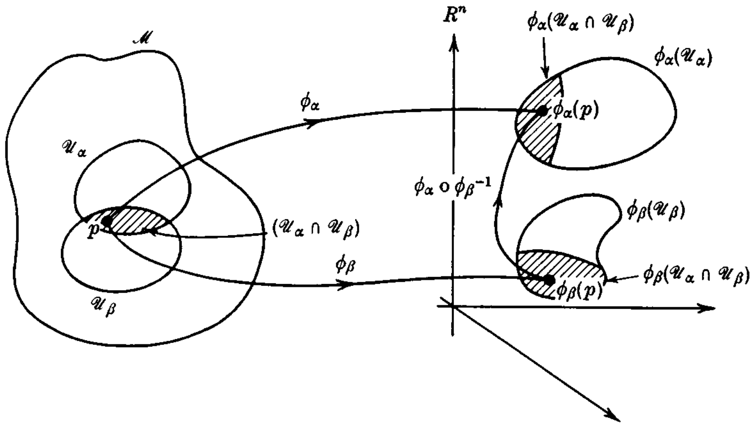

FIGURE 4. In the overlap of coordinate neighbourhoods $\mathscr{U}_\alpha$ and $\mathscr{U}_\beta$, coordinates are related by a C^r map $\phi_\alpha \circ \phi_\beta^{-1}$.

Another atlas is said to be *compatible* with a given C^r atlas if their union is a C^r atlas for all $\mathscr{M}$. The atlas consisting of all atlases compatible with the given atlas is called the *complete atlas* of the manifold; the complete atlas is therefore the set of all possible coordinate systems covering $\mathscr{M}$.

The topology of $\mathscr{M}$ is defined by stating that the open sets of $\mathscr{M}$ consist of unions of sets of the form $\mathscr{U}_\alpha$ belonging to the complete atlas. This topology makes each map ϕ_α into a homeomorphism.

A C^r differentiable manifold with boundary is defined as above, on replacing 'R^n' by '$\frac{1}{2}R^n$'. Then the *boundary of* $\mathscr{M}$, denoted by $\partial\mathscr{M}$, is defined to be the set of all points of $\mathscr{M}$ whose image under a map ϕ_α lies on the boundary of $\frac{1}{2}R^n$ in R^n. $\partial\mathscr{M}$ is an $(n-1)$-dimensional C^r manifold without boundary.

These definitions may seem more complicated than necessary. However simple examples show that one will in general need more than one coordinate neighbourhood to describe a space. The *two-dimensional Euclidean plane* R^2 is clearly a manifold. Rectangular coordinates $(x,\ y;\ -\infty < x < \infty,\ -\infty < y < \infty)$ cover the whole plane in one coordinate neighbourhood, where ϕ is the identity. Polar coordinates (r, θ) cover the coordinate neighbourhood $(r > 0,\ 0 < \theta < 2\pi)$; one needs at least two such coordinate neighbourhoods to cover R^2. The *two-dimensional cylinder* C^2 is the manifold obtained from R^2 by identifying the points (x, y) and $(x+2\pi, y)$. Then (x, y) are coordinates in a neighbourhood $(0 < x < 2\pi,\ -\infty < y < \infty)$ and one needs two such coordinate neighbourhoods to cover C^2. The *Möbius strip* is the manifold obtained in a similar way on identifying the points (x, y) and $(x+2\pi,\ -y)$. The *unit two-sphere* S^2 can be characterized as the surface in R^3 defined by the equation $(x^1)^2+(x^2)^2+(x^3)^2 = 1$. Then

$$(x^2, x^3;\ -1 < x^2 < 1,\ -1 < x^3 < 1)$$

are coordinates in each of the regions $x^1 > 0$, $x^1 < 0$, and one needs six such coordinate neighbourhoods to cover the surface. In fact, it is not possible to cover S^2 by a single coordinate neighbourhood. The *n-sphere* S^n can be similarly defined as the set of points

$$(x^1)^2+(x^2)^2+\ldots+(x^{n+1})^2 = 1$$

in R^{n+1}.

A manifold is said to be *orientable* if there is an atlas $\{\mathscr{U}_\alpha, \phi_\alpha\}$ in the complete atlas such that in every non-empty intersection $\mathscr{U}_\alpha \cap \mathscr{U}_\beta$, the Jacobian $|\partial x^i/\partial x'^j|$ is positive, where $(x^1, \ldots, x^n)$ and $(x'^1, \ldots, x'^n)$ are coordinates in $\mathscr{U}_\alpha$ and $\mathscr{U}_\beta$ respectively. The Möbius strip is an example of a non-orientable manifold.

The definition of a manifold given so far is very general. For most purposes one will impose two further conditions, that $\mathscr{M}$ is Hausdorff and that $\mathscr{M}$ is paracompact, which will ensure reasonable local behaviour.

A topological space $\mathscr{M}$ is said to be a *Hausdorff space* if it satisfies the Hausdorff separation axiom: whenever p, q are two distinct points in $\mathscr{M}$, there exist disjoint open sets $\mathscr{U}, \mathscr{V}$ in $\mathscr{M}$ such that $p \in \mathscr{U}$, $q \in \mathscr{V}$. One might think that a manifold is necessarily Hausdorff, but this is not so. Consider, for example, the situation in figure 5. We identify the points b, b' on the two lines if and only if $x_b = y_{b'} < 0$. Then each point is contained in a (coordinate) neighbourhood homeomorphic to an open subset of R^1. However there are no disjoint open neighbourhoods

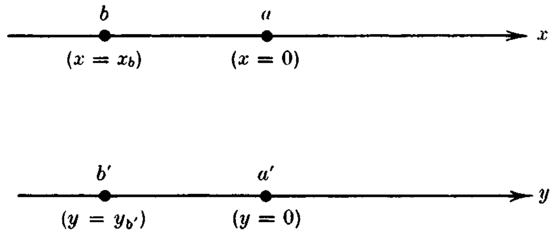

FIGURE 5. An example of a non-Hausdorff manifold. The two lines above are identical for $x = y < 0$. However the two points a $(x = 0)$ and a' $(y = 0)$ are not identified.

$\mathscr{U}, \mathscr{V}$ satisfying the conditions $a \in \mathscr{U}$, $a' \in \mathscr{V}$, where a is the point $x = 0$ and a' is the point $y = 0$.

An atlas $\{\mathscr{U}_\alpha, \phi_\alpha\}$ is said to be *locally finite* if every point $p \in \mathscr{M}$ has an open neighbourhood which intersects only a finite number of the sets $\mathscr{U}_\alpha$. $\mathscr{M}$ is said to be *paracompact* if for every atlas $\{\mathscr{U}_\alpha, \phi_\alpha\}$ there exists a locally finite atlas $\{\mathscr{V}_\beta, \psi_\beta\}$ with each $\mathscr{V}_\beta$ contained in some $\mathscr{U}_\alpha$. A connected Hausdorff manifold is paracompact if and only if it has a countable basis, i.e. there is a countable collection of open sets such that any open set can be expressed as the union of members of this collection (Kobayashi and Nomizu (1963), p. 271).

Unless otherwise stated, *all manifolds considered will be paracompact, connected C^∞ Hausdorff manifolds without boundary*. It will turn out later that when we have imposed some additional structure on $\mathscr{M}$ (the existence of an affine connection, see § 2.4) the requirement of paracompactness will be automatically satisfied because of the other restrictions.

A *function* f on a C^k manifold $\mathscr{M}$ is a map from $\mathscr{M}$ to R^1. It is said to be of class C^r $(r \leqslant k)$ at a point p of $\mathscr{M}$, if the expression $f \circ \phi_\alpha^{-1}$ of f on any local coordinate neighbourhood $\mathscr{U}_\alpha$ is a C^r function of the local coordinates at p; and f is said to be a C^r *function* on a set $\mathscr{V}$ of $\mathscr{M}$ if f is a C^r function at each point $p \in \mathscr{V}$.

A property of paracompact manifolds we will use later, is the following: given any locally finite atlas $\{\mathscr{U}_\alpha, \phi_\alpha\}$ on a paracompact C^k manifold, one can always (see e.g. Kobayashi and Nomizu (1963), p. 272) find a set of C^k functions g_α such that

(1) $0 \leqslant g_\alpha \leqslant 1$ on $\mathscr{M}$, for each α;

(2) the support of g_α, i.e. the closure of the set $\{p \in \mathscr{M}: g_\alpha(p) \neq 0\}$, is contained in the corresponding $\mathscr{U}_\alpha$;

(3) $\sum_\alpha g_\alpha(p) = 1$, for all $p \in \mathscr{M}$.

Such a set of functions will be called a *partition of unity*. The result is in particular true for C^∞ functions, but is clearly not true for analytic functions (an analytic function can be expressed as a convergent power series in some neighbourhood of each point $p \in \mathscr{M}$, and so is zero everywhere if it is zero on any open neighbourhood).

Finally, the *Cartesian product* $\mathscr{A} \times \mathscr{B}$ of manifolds $\mathscr{A}$, $\mathscr{B}$ is a manifold with a natural structure defined by the manifold structures of $\mathscr{A}$, $\mathscr{B}$: for arbitrary points $p \in \mathscr{A}$, $q \in \mathscr{B}$, there exist coordinate neighbourhoods $\mathscr{U}, \mathscr{V}$ containing p, q respectively, so the point $(p,q) \in \mathscr{A} \times \mathscr{B}$ is contained in the coordinate neighbourhood $\mathscr{U} \times \mathscr{V}$ in $\mathscr{A} \times \mathscr{B}$ which assigns to it the coordinates (x^i, y^j), where x^i are the coordinates of p in $\mathscr{U}$ and y^j are the coordinates of q in $\mathscr{V}$.

2.2 Vectors and tensors

Tensor fields are the set of geometric objects on a manifold defined in a natural way by the manifold structure. A tensor field is equivalent to a tensor defined at each point of the manifold, so we first define tensors at a point of the manifold, starting from the basic concept of a vector at a point.

A C^k *curve* $\lambda(t)$ in $\mathscr{M}$ is a C^k map of an interval of the real line R^1 into $\mathscr{M}$. The *vector* (contravariant vector) $(\partial/\partial t)_\lambda|_{t_0}$ tangent to the C^1 curve $\lambda(t)$ at the point $\lambda(t_0)$ is the operator which maps each C^1 function f at $\lambda(t_0)$ into the number $(\partial f/\partial t)_\lambda|_{t_0}$; that is, $(\partial f/\partial t)_\lambda$ is the derivative of f in the direction of $\lambda(t)$ with respect to the parameter t. Explicitly,

$$\left(\frac{\partial f}{\partial t}\right)_\lambda\bigg|_t = \lim_{s\to 0} \frac{1}{s}\{f(\lambda(t+s)) - f(\lambda(t))\}. \tag{2.1}$$

The curve parameter t clearly obeys the relation $(\partial/\partial t)_\lambda t = 1$.

If $(x^1, \ldots, x^n)$ are local coordinates in a neighbourhood of p,

$$\left(\frac{\partial f}{\partial t}\right)_\lambda\bigg|_{t_0} = \sum_{j=1}^{n} \frac{\mathrm{d}x^j(\lambda(t))}{\mathrm{d}t}\bigg|_{t=t_0} \cdot \frac{\partial f}{\partial x^j}\bigg|_{\lambda(t_0)} = \frac{\mathrm{d}x^j}{\mathrm{d}t}\frac{\partial f}{\partial x^j}\bigg|_{\lambda(t_0)}.$$

(Here and throughout this book, we adopt the *summation convention* whereby a repeated index implies summation over all values of that index.) Thus every tangent vector at a point p can be expressed as a linear combination of the coordinate derivatives

$$(\partial/\partial x^1)|_p, \ldots, (\partial/\partial x^n)|_p.$$

Conversely, given a linear combination $V^j(\partial/\partial x^j)|_p$ of these operators, where the V^j are any numbers, consider the curve $\lambda(t)$ defined by

$x^j(\lambda(t)) = x^j(p) + tV^j$, for t in some interval $[-\epsilon, \epsilon]$; the tangent vector to this curve at p is $V^j(\partial/\partial x^j)|_p$. Thus the tangent vectors at p form a vector space over R^1 spanned by the coordinate derivatives $(\partial/\partial x^j)|_p$, where the vector space structure is defined by the relation

$$(\alpha X + \beta Y)f = \alpha(Xf) + \beta(Yf)$$

which is to hold for all vectors X, Y, numbers α, β and functions f. The vectors $(\partial/\partial x^j)_p$ are independent (for if they were not, there would exist numbers V^j such that $V^j(\partial/\partial x^j)|_p = 0$ with at least one V^j non-zero; applying this relation to each coordinate x^k shows

$$V^j \, \partial x^k/\partial x^j = V^k = 0,$$

a contradiction), so the space of all tangent vectors to $\mathscr{M}$ at p, denoted by $T_p(\mathscr{M})$ or simply T_p, is an n-dimensional vector space. This space, representing the set of all directions at p, is called the *tangent vector space* to $\mathscr{M}$ at p. One may think of a vector $\mathbf{V} \in T_p$ as an arrow at p, pointing in the direction of a curve $\lambda(t)$ with tangent vector $\mathbf{V}$ at p, the 'length' of $\mathbf{V}$ being determined by the curve parameter t through the relation $V(t) = 1$. (As $\mathbf{V}$ is an operator, we print it in bold type; its components V^j, and the number $V(f)$ obtained by $\mathbf{V}$ acting on a function f, are numbers, and so are printed in italics.)

If $\{\mathbf{E}_a\}$ ($a = 1$ to n) are any set of n vectors at p which are linearly independent, then any vector $\mathbf{V} \in T_p$ can be written $\mathbf{V} = V^a \mathbf{E}_a$ where the numbers $\{V^a\}$ are the components of $\mathbf{V}$ with respect to the basis $\{\mathbf{E}_a\}$ of vectors at p. In particular one can choose the $\mathbf{E}_a$ as the coordinate basis $(\partial/\partial x^i)|_p$; then the components $V^i = V(x^i) = (\mathrm{d}x^i/\mathrm{d}t)|_p$ are the derivatives of the coordinate functions x^i in the direction $\mathbf{V}$.

A *one-form* (covariant vector) $\boldsymbol{\omega}$ at p is a real valued linear function on the space T_p of vectors at p. If $\mathbf{X}$ is a vector at p, the number into which $\boldsymbol{\omega}$ maps $\mathbf{X}$ will be written $\langle \boldsymbol{\omega}, \mathbf{X} \rangle$; then the linearity implies that

$$\langle \boldsymbol{\omega}, \alpha\mathbf{X} + \beta\mathbf{Y} \rangle = \alpha \langle \boldsymbol{\omega}, \mathbf{X} \rangle + \beta \langle \boldsymbol{\omega}, \mathbf{Y} \rangle$$

holds for all $\alpha, \beta \in R^1$ and $\mathbf{X}, \mathbf{Y} \in T_p$. The subspace of T_p defined by $\langle \boldsymbol{\omega}, \mathbf{X} \rangle =$ (constant) for a given one-form $\boldsymbol{\omega}$, is linear. One may therefore think of a one-form at p as a pair of planes in T_p such that if $\langle \boldsymbol{\omega}, \mathbf{X} \rangle = 0$ the arrow $\mathbf{X}$ lies in the first plane, and if $\langle \boldsymbol{\omega}, \mathbf{X} \rangle = 1$ it touches the second plane.

Given a basis $\{\mathbf{E}_a\}$ of vectors at p, one can define a unique set of n one-forms $\{\mathbf{E}^a\}$ by the condition: $\mathbf{E}^i$ maps any vector $\mathbf{X}$ to the number X^i (the ith component of $\mathbf{X}$ with respect to the basis $\{\mathbf{E}_a\}$).

Then in particular, $\langle \mathbf{E}^a, \mathbf{E}_b \rangle = \delta^a{}_b$. Defining linear combinations of one-forms by the rules

$$\langle \alpha\boldsymbol{\omega} + \beta\boldsymbol{\eta}, \mathbf{X} \rangle = \alpha\langle \boldsymbol{\omega}, \mathbf{X} \rangle + \beta\langle \boldsymbol{\eta}, \mathbf{X} \rangle$$

for any one-forms $\boldsymbol{\omega}$, $\boldsymbol{\eta}$ and any α, $\beta \in R^1$, $\mathbf{X} \in T_p$, one can regard $\{\mathbf{E}^a\}$ as a basis of one-forms since any one-form $\boldsymbol{\omega}$ at p can be expressed as $\boldsymbol{\omega} = \omega_i \mathbf{E}^i$ where the numbers ω_i are defined by $\omega_i = \langle \boldsymbol{\omega}, \mathbf{E}_i \rangle$. Thus the set of all one forms at p forms an n-dimensional vector space at p, the *dual space* $T^*{}_p$ of the tangent space T_p. The basis $\{\mathbf{E}^a\}$ of one-forms is the *dual basis* to the basis $\{\mathbf{E}_a\}$ of vectors. For any $\boldsymbol{\omega} \in T^*{}_p$, $\mathbf{X} \in T_p$ one can express the number $\langle \boldsymbol{\omega}, \mathbf{X} \rangle$ in terms of the components $\boldsymbol{\omega}_i$, $\mathbf{X}^i$ of $\boldsymbol{\omega}$, $\mathbf{X}$ with respect to dual bases $\{\mathbf{E}^a\}$, $\{\mathbf{E}_a\}$ by the relations

$$\langle \boldsymbol{\omega}, \mathbf{X} \rangle = \langle \omega_i \mathbf{E}^i, X^j \mathbf{E}_j \rangle = \omega_i X^i.$$

Each function f on $\mathscr{M}$ defines a one-form $\mathrm{d}f$ at p by the rule: for each vector $\mathbf{X}$,

$$\langle \mathrm{d}f, \mathbf{X} \rangle = Xf.$$

$\mathrm{d}f$ is called the *differential* of f. If $(x^1, \ldots, x^n)$ are local coordinates, the set of differentials $(\mathrm{d}x^1, \mathrm{d}x^2, \ldots, \mathrm{d}x^n)$ at p form the basis of one-forms dual to the basis $(\partial/\partial x^1, \partial/\partial x^2, \ldots, \partial/\partial x^n)$ of vectors at p, since

$$\langle \mathrm{d}x^i, \partial/\partial x^j \rangle = \partial x^i/\partial x^j = \delta^i{}_j.$$

In terms of this basis, the differential $\mathrm{d}f$ of an arbitrary function f is given by

$$\mathrm{d}f = (\partial f/\partial x^i)\,\mathrm{d}x^i.$$

If $\mathrm{d}f$ is non-zero, the surfaces $\{f = \text{constant}\}$ are $(n-1)$-dimensional manifolds. The subspace of T_p consisting of all vectors $\mathbf{X}$ such that $\langle \mathrm{d}f, \mathbf{X} \rangle = 0$ consists of all vectors tangent to curves lying in the surface $\{f = \text{constant}\}$ through p. Thus one may think of $\mathrm{d}f$ as a normal to the surface $\{f = \text{constant}\}$ at p. If $\alpha \neq 0$, $\alpha\,\mathrm{d}f$ will also be a normal to this surface.

From the space T_p of vectors at p and the space $T^*{}_p$ of one-forms at p, we can form the Cartesian product

$$\Pi^s_r = \underbrace{T^*{}_p \times T^*{}_p \times \ldots \times T^*{}_p}_{r \text{ factors}} \times \underbrace{T_p \times T_p \times \ldots \times T_p}_{s \text{ factors}},$$

i.e. the ordered set of vectors and one-forms $(\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s)$ where the $\mathbf{Y}$s and $\boldsymbol{\eta}$s are arbitrary vectors and one-forms respectively.

A *tensor of type* (r, s) *at* p is a function on Π^s_r which is linear in each argument. If $\mathbf{T}$ is a tensor of type (r, s) at p, we write the number into which $\mathbf{T}$ maps the element $(\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s)$ of Π^s_r as

$$T(\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s).$$

Then the linearity implies that, for example,

$$T(\eta^1, \ldots, \eta^r, \alpha\mathbf{X} + \beta\mathbf{Y}, \mathbf{Y}_2, \ldots, \mathbf{Y}_s) = \alpha \,.\, T(\eta^1, \ldots, \eta^r, \mathbf{X}, \mathbf{Y}_2, \ldots, \mathbf{Y}_s) + \beta \,.\, T(\eta^1, \ldots, \eta^r, \mathbf{Y}, \mathbf{Y}_2, \ldots, \mathbf{Y}_s)$$

holds for all $\alpha, \beta \in R^1$ and $\mathbf{X}, \mathbf{Y} \in T_p$.

The space of all such tensors is called the *tensor product*

$$T^r_s(p) = \underbrace{T_p \otimes \ldots \otimes T_p}_{r \text{ factors}} \otimes \underbrace{T^*{}_p \otimes \ldots \otimes T^*{}_p}_{s \text{ factors}}.$$

In particular, $T^1_0(p) = T_p$ and $T^0_1(p) = T^*{}_p$.

Addition of tensors of type (r, s) is defined by the rule: $(\mathbf{T} + \mathbf{T}')$ is the tensor of type (r, s) at p such that for all $\mathbf{Y}_i \in T_p$, $\eta^j \in T^*{}_p$,

$$(T + T')(\eta^1, \ldots, \eta^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s) = T(\eta^1, \ldots, \eta^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s) + T'(\eta^1, \ldots, \eta^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s).$$

Similarly, *multiplication of a tensor by a scalar* $\alpha \in R^1$ is defined by the rule: $(\alpha\mathbf{T})$ is the tensor such that for all $\mathbf{Y}_i \in T_p$, $\eta^j \in T^*{}_p$,

$$(\alpha T)(\eta^1, \ldots, \eta^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s) = \alpha \,.\, T(\eta^1, \ldots, \eta^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s).$$

With these rules of addition and scalar multiplication, the tensor product $T^r_s(p)$ is a vector space of dimension n^{r+s} over R^1.

Let $\mathbf{X}_i \in T_p$ ($i = 1$ to r) and $\omega^j \in T^*{}_p$ ($j = 1$ to s). Then we shall denote by $\mathbf{X}_1 \otimes \ldots \otimes \mathbf{X}_r \otimes \omega^1 \otimes \ldots \otimes \omega^s$ that element of $T^r_s(p)$ which maps the element $(\eta^1, \ldots, \eta^r, \mathbf{Y}_1, \ldots, \mathbf{Y}_s)$ of Π^s_r into

$$\langle \eta^1, \mathbf{X}_1 \rangle \langle \eta^2, \mathbf{X}_2 \rangle \ldots \langle \eta^r, \mathbf{X}_r \rangle \langle \omega^1, \mathbf{Y}_1 \rangle \ldots \langle \omega^s, \mathbf{Y}_s \rangle.$$

Similarly, if $\mathbf{R} \in T^r_s(p)$ and $\mathbf{S} \in T^p_q(p)$, we shall denote by $\mathbf{R} \otimes \mathbf{S}$ that element of $T^{r+p}_{s+q}(p)$ which maps the element $(\eta^1, \ldots, \eta^{r+p}, \mathbf{Y}_1, \ldots, \mathbf{Y}_{s+q})$ of Π^{s+q}_{r+p} into the number

$$R(\eta^1, \ldots, \eta^s, \mathbf{Y}_1, \ldots, \mathbf{Y}_r)\, S(\eta^{s+1}, \ldots, \eta^{s+q}, \mathbf{Y}_{r+1}, \ldots, \mathbf{Y}_{r+p}).$$

With the product $\otimes$, the tensor spaces at p form an algebra over R.

If $\{\mathbf{E}_a\}$, $\{\mathbf{E}^a\}$ are dual bases of T_p, $T^*{}_p$ respectively, then

$$\{\mathbf{E}_{a_1} \otimes \ldots \otimes \mathbf{E}_{a_r} \otimes \mathbf{E}^{b_1} \otimes \ldots \otimes \mathbf{E}^{b_s}\}, \quad (a_i, b_j \text{ run from 1 to } n),$$

will be a basis for $T^r_s(p)$. An arbitrary tensor $\mathbf{T} \in T^r_s(p)$ can be expressed in terms of this basis as

$$\mathbf{T} = T^{a_1 \ldots a_r}{}_{b_1 \ldots b_s} \mathbf{E}_{a_1} \otimes \ldots \otimes \mathbf{E}_{a_r} \otimes \mathbf{E}^{b_1} \otimes \ldots \otimes \mathbf{E}^{b_s}$$

where $\{T^{a_1\ldots a_r}{}_{b_1\ldots b_s}\}$ are the *components* of $\mathbf{T}$ with respect to the dual bases $\{\mathbf{E}_a\}$, $\{\mathbf{E}^a\}$ and are given by

$$T^{a_1\ldots a_r}{}_{b_1\ldots b_s} = T(\mathbf{E}^{a_1}, \ldots, \mathbf{E}^{a_r}, \mathbf{E}_{b_1}, \ldots, \mathbf{E}_{b_s}).$$

Relations in the tensor algebra at p can be expressed in terms of the components of tensors. Thus

$$(T+T')^{a_1\ldots a_r}{}_{b_1\ldots b_r} = T^{a_1\ldots a_r}{}_{b_1\ldots b_s} + T'^{a_1\ldots a_r}{}_{b_1\ldots b_s},$$

$$(\alpha T)^{a_1\ldots a_r}{}_{b_1\ldots b_s} = \alpha \,.\, T^{a_1\ldots a_r}{}_{b_1\ldots b_s},$$

$$(T\otimes T')^{a_1\ldots a_{r+p}}{}_{b_1\ldots b_{s+q}} = T^{a_1\ldots a_r}{}_{b_1\ldots b_s} T'^{a_{r+1}\ldots a_{r+p}}{}_{b_{s+1}\ldots b_{s+q}}.$$

Because of its convenience, we shall usually represent tensor relations in this way.

If $\{\mathbf{E}_{a'}\}$ and $\{\mathbf{E}^{a'}\}$ are another pair of dual bases for T_p and $T^*{}_p$, they can be represented in terms of $\{\mathbf{E}_a\}$ and $\{\mathbf{E}^a\}$ by

$$\mathbf{E}_{a'} = \Phi_{a'}{}^a \mathbf{E}_a \tag{2.2}$$

where $\Phi_{a'}{}^a$ is an $n\times n$ non-singular matrix. Similarly

$$\mathbf{E}^{a'} = \Phi^{a'}{}_a \mathbf{E}^a \tag{2.3}$$

where $\Phi^{a'}{}_a$ is another $n\times n$ non-singular matrix. Since $\{\mathbf{E}_{a'}\}$, $\{\mathbf{E}^{a'}\}$ are dual bases,

$$\delta^{b'}{}_{a'} = \langle \mathbf{E}^{b'}, \mathbf{E}_{a'}\rangle = \langle \Phi^{b'}{}_b \mathbf{E}^b, \Phi_{a'}{}^a \mathbf{E}_a\rangle = \Phi_{a'}{}^a \Phi^{b'}{}_b \delta_a{}^b = \Phi_{a'}{}^a \Phi^{b'}{}_a,$$

i.e. $\Phi_{a'}{}^a$, $\Phi^{a'}{}_a$ are inverse matrices, and $\delta^a{}_b = \Phi^a{}_{b'} \Phi^{b'}{}_b$.

The components $T^{a'_1\ldots a'_r}{}_{b'_1\ldots b'_s}$ of a tensor $\mathbf{T}$ with respect to the dual bases $\{\mathbf{E}_{a'}\}$, $\{\mathbf{E}^{a'}\}$ are given by

$$T^{a'_1\ldots a'_r}{}_{b'_1\ldots b'_s} = T(\mathbf{E}^{a_1'}, \ldots, \mathbf{E}^{a'_r}, \mathbf{E}_{b'_1}, \ldots, \mathbf{E}_{b'_s}).$$

They are related to the components $T^{a_1\ldots a_r}{}_{b_1\ldots b_s}$ of $\mathbf{T}$ with respect to the bases $\{\mathbf{E}_a\}$, $\{\mathbf{E}^a\}$ by

$$T^{a'_1\ldots a'_r}{}_{b'_1\ldots b'_s} = T^{a_1\ldots a_r}{}_{b_1\ldots b_s} \Phi^{a'_1}{}_{a_1} \ldots \Phi^{a'_r}{}_{a_r} \Phi_{b'_1}{}^{b_1} \ldots \Phi_{b'_s}{}^{b_s}. \tag{2.4}$$

The *contraction* of a tensor $\mathbf{T}$ of type (r,s), with components $T^{ab\ldots d}{}_{ef\ldots g}$ with respect to bases $\{\mathbf{E}_a\}$, $\{\mathbf{E}^a\}$, on the first contravariant and first covariant indices is defined to be the tensor $C^1_1(\mathbf{T})$ of type $(r-1, s-1)$ whose components with respect to the same basis are $T^{ab\ldots d}{}_{af\ldots g}$, i.e.

$$C^1_1(\mathbf{T}) = T^{ab\ldots d}{}_{af\ldots g} \mathbf{E}_b \otimes \ldots \otimes \mathbf{E}_d \otimes \mathbf{E}^f \otimes \ldots \otimes \mathbf{E}^g.$$

If $\{\mathbf{E}_{a'}\}$, $\{\mathbf{E}^{a'}\}$ are another pair of dual bases, the contraction $C_1^1(\mathbf{T})$ defined by them is

$$\begin{aligned} C'^1_1(\mathbf{T}) &= T^{a'b'\dots d'}{}_{a'f'\dots g'}\mathbf{E}_{b'}\otimes\dots\otimes\mathbf{E}_{d'}\otimes\mathbf{E}^{f'}\otimes\dots\otimes\mathbf{E}^{g'} \\ &= \Phi^{a'}{}_a\Phi^a{}_{h'}\,T^{h'b'\dots d'}{}_{a'f'\dots g'}\Phi_{b'}{}^b\dots\Phi_{d'}{}^d\Phi^{f'}{}_f\dots\Phi^{g'}{}_g \\ &\qquad\qquad .\mathbf{E}_b\otimes\dots\otimes\mathbf{E}_d\otimes\mathbf{E}^f\dots\otimes\mathbf{E}^g \\ &= T^{ab\dots d}{}_{af\dots g}\mathbf{E}_b\otimes\dots\otimes\mathbf{E}_d\otimes\mathbf{E}^f\otimes\dots\otimes\mathbf{E}^g = C_1^1(\mathbf{T}), \end{aligned}$$

so the contraction C_1^1 of a tensor is independent of the basis used in its definition. Similarly, one could contract $\mathbf{T}$ over any pair of contravariant and covariant indices. (If we were to contract over two contravariant or covariant indices, the resultant tensor would depend on the basis used.)

The symmetric part of a tensor $\mathbf{T}$ of type $(2,0)$ is the tensor $S(\mathbf{T})$ defined by

$$S(\mathbf{T})(\eta_1,\eta_2) = \frac{1}{2!}\{T(\eta_1,\eta_2)+T(\eta_2,\eta_1)\}$$

for all $\eta_1,\eta_2\in T^*{}_p$. We shall denote the components $S(\mathbf{T})^{ab}$ of $S(\mathbf{T})$ by $T^{(ab)}$; then

$$T^{(ab)} = \frac{1}{2!}\{T^{ab}+T^{ba}\}.$$

Similarly, the components of the skew-symmetric part of $\mathbf{T}$ will be denoted by

$$T^{[ab]} = \frac{1}{2!}\{T^{ab}-T^{ba}\}.$$

In general, the components of the symmetric or antisymmetric part of a tensor on a given set of covariant or contravariant indices will be denoted by placing round or square brackets around the indices. Thus

$$T_{(a_1\dots a_r)}{}^{b\dots f}$$
$$= \frac{1}{r!}\{\text{sum over all permutations of the indices } a_1 \text{ to } a_r (T_{a_1\dots a_r}{}^{b\dots f})\}$$

and

$$T_{[a_1\dots a_r]}{}^{b\dots f}$$
$$= \frac{1}{r!}\{\text{alternating sum over all permutations of the indices } a_1 \text{ to } a_r\ (T_{a_1\dots a_r}{}^{b\dots f})\}.$$

For example,

$$K^a{}_{[bcd]} = \tfrac{1}{6}\{K^a{}_{bcd}+K^a{}_{dbc}+K^a{}_{cdb}-K^a{}_{bdc}-K^a{}_{cbd}-K^a{}_{dcb}\}.$$

A tensor is *symmetric* in a given set of contravariant or covariant indices if it is equal to its symmetrized part on these indices, and is *antisymmetric* if it is equal to its antisymmetrized part. Thus, for example, a tensor $\mathbf{T}$ of type $(0,2)$ is symmetric if $T_{ab} = \frac{1}{2}(T_{ab}+T_{ba})$, (which we can also express in the form: $T_{[ab]} = 0$).

A particularly important subset of tensors is the set of tensors of type $(0,q)$ which are antisymmetric on all q positions (so $q \leqslant n$); such a tensor is called a *q-form*. If $\mathbf{A}$ and $\mathbf{B}$ are p- and q-forms respectively, one can define a $(p+q)$-form $\mathbf{A} \wedge \mathbf{B}$ from them, where $\wedge$ is the skew-symmetrized tensor product $\otimes$; that is, $\mathbf{A} \wedge \mathbf{B}$ is the tensor of type $(0, p+q)$ with components determined by

$$(A \wedge B)_{a\ldots bc\ldots f} = A_{[a\ldots b}B_{c\ldots f]}.$$

This rule implies $(\mathbf{A} \wedge \mathbf{B}) = (-)^{pq}(\mathbf{B} \wedge \mathbf{A})$. With this product, the space of forms (i.e. the space of all p-forms for all p, including one-forms and defining scalars as zero-forms) constitutes the Grassmann algebra of forms. If $\{\mathbf{E}^a\}$ is a basis of one-forms, then the forms $\mathbf{E}^{a_1} \wedge \ldots \wedge \mathbf{E}^{a_p}$ (a_i run from 1 to n) are a basis of p-forms, as any p-form $\mathbf{A}$ can be written $\mathbf{A} = A_{a\ldots b}\mathbf{E}^a \wedge \ldots \wedge \mathbf{E}^b$, where $A_{a\ldots b} = A_{[a\ldots b]}$.

So far, we have considered the set of tensors defined at a point on the manifold. A set of local coordinates $\{x^i\}$ on an open set $\mathscr{U}$ in $\mathscr{M}$ defines a basis $\{(\partial/\partial x^i)|_p\}$ of vectors and a basis $\{(\mathrm{d}x^i)|_p\}$ of one-forms at each point p of $\mathscr{U}$, and so defines a basis of tensors of type (r,s) at each point of $\mathscr{U}$. Such a basis of tensors will be called a coordinate basis. A C^k *tensor field* $\mathbf{T}$ *of type* (r,s) on a set $\mathscr{V} \subset \mathscr{M}$ is an assignment of an element of $T^r_s(p)$ to each point $p \in \mathscr{V}$ such that the components of $\mathbf{T}$ with respect to any coordinate basis defined on an open subset of $\mathscr{V}$ are C^k functions.

In general one need not use a coordinate basis of tensors, i.e. given any basis of vectors $\{\mathbf{E}_a\}$ and dual basis of forms $\{\mathbf{E}^a\}$ on $\mathscr{V}$, there will not necessarily exist any open set in $\mathscr{V}$ on which there are local coordinates $\{x^a\}$ such that $\mathbf{E}_a = \partial/\partial x^a$ and $\mathbf{E}^a = \mathrm{d}x^a$. However if one does use a coordinate basis, certain specializations will result; in particular for any function f, the relations $E_a(E_b f) = E_b(E_a f)$ are satisfied, being equivalent to the relations $\partial^2 f/\partial x^a \partial x^b = \partial^2 f/\partial x^b \partial x^a$. If one changes from a coordinate basis $\mathbf{E}_a = \partial/\partial x^a$ to a coordinate basis $\mathbf{E}_{a'} = \partial/\partial x^{a'}$, applying (2.2), (2.3) to x^a, $x^{a'}$ shows that

$$\Phi_{a'}{}^{a} = \frac{\partial x^a}{\partial x^{a'}}, \quad \Phi^{a'}{}_{a} = \frac{\partial x^{a'}}{\partial x^a}.$$

Clearly a general basis $\{\mathbf{E}_a\}$ can be obtained from a coordinate basis

$\{\partial/\partial x^i\}$ by giving the functions $E_a{}^i$ which are the components of the $\mathbf{E}_a$ with respect to the basis $\{\partial/\partial x^i\}$; then (2.2) takes the form $\mathbf{E}_a = E_a{}^i \partial/\partial x^i$ and (2.3) takes the form $\mathbf{E}^a = E^a{}_i \, \mathrm{d}x^i$, where the matrix $E^a{}_i$ is dual to the matrix $E_a{}^i$.

2.3 Maps of manifolds

In this section we define, via the general concept of a C^k manifold map, the concepts of 'imbedding', 'immersion', and of associated tensor maps, the first two being useful later in the study of submanifolds, and the last playing an important role in studying the behaviour of families of curves as well as in studying symmetry properties of manifolds.

A map ϕ from a C^k n-dimensional manifold $\mathscr{M}$ to a $C^{k'}$ n'-dimensional manifold $\mathscr{M}'$ is said to be a C^r *map* ($r \leqslant k$, $r \leqslant k'$) if, for any local coordinate systems in $\mathscr{M}$ and $\mathscr{M}'$, the coordinates of the image point $\phi(p)$ in $\mathscr{M}'$ are C^r functions of the coordinates of p in $\mathscr{M}$. As the map will in general be many–one rather than one–one (e.g. it cannot be one–one if $n > n'$), it will in general not have an inverse; and if a C^r map does have an inverse, this inverse will in general not be C^r (e.g. if ϕ is the map $R^1 \to R^1$ given by $x \to x^3$, then ϕ^{-1} is not differentiable at the point $x = 0$).

If f is a function on $\mathscr{M}'$, the mapping ϕ defines the function $\phi^* f$ on $\mathscr{M}$ as the function whose value at the point p of $\mathscr{M}$ is the value of f at $\phi(p)$, i.e.

$$\phi^* f(p) = f(\phi(p)). \tag{2.5}$$

Thus when ϕ maps points from $\mathscr{M}$ to $\mathscr{M}'$, ϕ^* maps functions linearly from $\mathscr{M}'$ to $\mathscr{M}$.

If $\lambda(t)$ is a curve through the point $p \in \mathscr{M}$, then the image curve $\phi(\lambda(t))$ in $\mathscr{M}'$ passes through the point $\phi(p)$. If $r \geqslant 1$, the tangent vector to this curve at $\phi(p)$ will be denoted by $\phi_*(\partial/\partial t)_\lambda|_{\phi(p)}$; one can regard it as the image, under the map ϕ, of the vector $(\partial/\partial t)_\lambda|_p$. Clearly ϕ_* is a linear map of $T_p(\mathscr{M})$ into $T_{\phi(p)}(\mathscr{M}')$. From (2.5) and the definition (2.1) of a vector as a directional derivative, the vector map ϕ_* can be characterized by the relation: for each C^r ($r \geqslant 1$) function f at $\phi(p)$ and vector $\mathbf{X}$ at p,

$$X(\phi^* f)|_p = \phi_* X(f)|_{\phi(p)}. \tag{2.6}$$

Using the vector mapping ϕ_* from $\mathscr{M}$ to $\mathscr{M}'$, we can if $r \geqslant 1$ define a linear one-form mapping ϕ^* from $T^*{}_{\phi(p)}(\mathscr{M}')$ to $T^*{}_p(\mathscr{M})$ by the condition: vector–one-form contractions are to be preserved under the maps. Then the one-form $\mathbf{A} \in T^*_{\phi(p)}$ is mapped into the one-form $\phi^*\mathbf{A} \in T^*_p$ where, for arbitrary vectors $\mathbf{X} \in T_p$,

$$\langle \phi^*\mathbf{A}, \mathbf{X}\rangle|_p = \langle \mathbf{A}, \phi_* \mathbf{X}\rangle|_{\phi(p)}.$$

A consequence of this is that

$$\phi^*(\mathrm{d}f) = \mathrm{d}(\phi^*f). \tag{2.7}$$

The maps ϕ_* and ϕ^* can be extended to maps of contravariant tensors from $\mathscr{M}$ to $\mathscr{M}'$ and covariant tensors from $\mathscr{M}'$ to $\mathscr{M}$ respectively, by the rules $\phi_*\colon \mathbf{T} \in T^r_0(p) \to \phi_* \mathbf{T} \in T^r_0(\phi(p))$ where for any $\boldsymbol{\eta}^i \in T^*_{\phi(p)}$,

$$T(\phi^*\boldsymbol{\eta}^1, \ldots, \phi^*\boldsymbol{\eta}^r)|_p = \phi_* T(\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^r)|_{\phi(p)}$$

and

$$\phi^*\colon \mathbf{T} \in T^0_s(\phi(p)) \to \phi^*\mathbf{T} \in T^0_s(p),$$

where for any $\mathbf{X}_i \in T_p$,

$$\phi^*T(\mathbf{X}_1, \ldots, \mathbf{X}_s)|_p = T(\phi_*\mathbf{X}_1, \ldots, \phi_*\mathbf{X}_s)|_{\phi(p)}.$$

When $r \geqslant 1$, the C^r map ϕ from $\mathscr{M}$ to $\mathscr{M}'$ is said to be of *rank* s at p if the dimension of $\phi_*(T_p(\mathscr{M}))$ is s. It is said to be *injective* at p if $s = n$ (and so $n \leqslant n'$) at p; then no vector in T_p is mapped to zero by ϕ_*. It is said to be *surjective* if $s = n'$ (so $n \geqslant n'$).

A C^r map ϕ ($r \geqslant 0$) is said to be an *immersion* if it and its inverse are C^r maps, i.e. if for each point $p \in \mathscr{M}$ there is a neighbourhood $\mathscr{U}$ of p in $\mathscr{M}$ such that the inverse ϕ^{-1} restricted to $\phi(\mathscr{U})$ is also a C^r map. This implies $n \leqslant n'$. By the implicit function theorem (Spivak (1965), p. 41), when $r \geqslant 1$, ϕ will be an immersion if and only if it is injective at every point $p \in \mathscr{M}$; then ϕ_* is an isomorphism of T_p into the image $\phi_*(T_p) \subset T_{\phi(p)}$. The image $\phi(\mathscr{M})$ is then said to be an n-dimensional *immersed submanifold* in $\mathscr{M}'$. This submanifold may intersect itself, i.e. ϕ may not be a one–one map from $\mathscr{M}$ to $\phi(\mathscr{M})$ although it is one–one when restricted to a sufficiently small neighbourhood of $\mathscr{M}$. An immersion is said to be an *imbedding* if it is a homeomorphism onto its image in the induced topology. Thus an imbedding is a one–one immersion; however not all one–one immersions are imbeddings, cf. figure 6. A map ϕ is said to be a proper map if the inverse image $\phi^{-1}(\mathscr{K})$ of any compact set $\mathscr{K} \subset \mathscr{M}'$ is compact. It can be shown that a proper one–one immersion is an imbedding. The image $\phi(\mathscr{M})$ of $\mathscr{M}$ under an imbedding ϕ is said to be an n-dimensional *imbedded submanifold* of $\mathscr{M}'$.

The map ϕ from $\mathscr{M}$ to $\mathscr{M}'$ is said to be a C^r *diffeomorphism* if it is a one–one C^r map and the inverse ϕ^{-1} is a C^r map from $\mathscr{M}'$ to $\mathscr{M}$. In this case, $n = n'$, and ϕ is both injective and surjective if $r \geqslant 1$; conversely, the implicit function theorem shows that if ϕ_* is both injective and surjective at p, then there is an open neighbourhood $\mathscr{U}$ of p such that $\phi: \mathscr{U} \to \phi(\mathscr{U})$ is a diffeomorphism. Thus ϕ is a local diffeomorphism near p if ϕ_* is an isomorphism from T_p to $T_{\phi(p)}$.

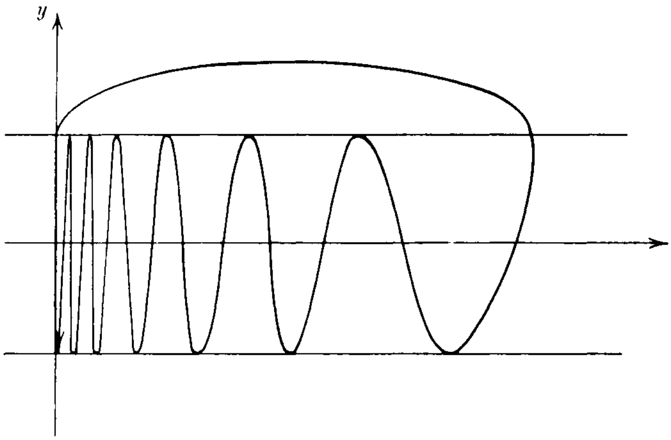

FIGURE 6. A one–one immersion of R^1 in R^2 which is not an imbedding, obtained by joining smoothly part of the curve $y = \sin(1/x)$ to the curve

$$\{(y,0);\ -\infty < y < 1\}.$$

When the map ϕ is a C^r ($r \geqslant 1$) diffeomorphism, ϕ_* maps $T_p(\mathscr{M})$ to $T_{\phi(p)}(\mathscr{M}')$ and $(\phi^{-1})^*$ maps $T^*{}_p(\mathscr{M})$ to $T^*{}_{\phi(p)}(\mathscr{M}')$. Thus we can define a map ϕ_* of $T^r_s(p)$ to $T^r_s(\phi(p))$ for any r, s, by

$$T(\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^s, \mathbf{X}_1, \ldots, \mathbf{X}_r)|_p$$
$$= \phi_* T((\phi^{-1})^*\boldsymbol{\eta}^1, \ldots, (\phi^{-1})^*\boldsymbol{\eta}^s, \phi_*\mathbf{X}_1, \ldots, \phi_*\mathbf{X}_r)|_{\phi(p)}$$

for any $\mathbf{X}_i \in T_p$, $\boldsymbol{\eta}^i \in T^*{}_p$. This map of tensors of type (r, s) on $\mathscr{M}$ to tensors of type (r, s) on $\mathscr{M}'$ preserves symmetries and relations in the tensor algebra; e.g. the contraction of $\phi_*\mathbf{T}$ is equal to ϕ_* (the contraction of $\mathbf{T}$).

2.4 Exterior differentiation and the Lie derivative

We shall study three differential operators on manifolds, the first two being defined purely by the manifold structure while the third is defined (see § 2.5) by placing extra structure on the manifold.

The *exterior differentiation* operator d maps r-form fields linearly to $(r+1)$-form fields. Acting on a zero-form field (i.e. a function) f, it gives the one-form field $\mathrm{d}f$ defined by (cf. §2.2)

$$\langle \mathrm{d}f, \mathbf{X}\rangle = Xf \text{ for all vector fields } \mathbf{X} \tag{2.8}$$

and acting on the r-form field

$$\mathbf{A} = A_{ab\ldots d}\,\mathrm{d}x^a \wedge \mathrm{d}x^b \wedge \ldots \wedge \mathrm{d}x^d$$

it gives the $(r+1)$-form field $\mathrm{d}\mathbf{A}$ defined by

$$\mathrm{d}\mathbf{A} = \mathrm{d}A_{ab\ldots d} \wedge \mathrm{d}x^a \wedge \mathrm{d}x^b \wedge \ldots \wedge \mathrm{d}x^d. \tag{2.9}$$

To show that this $(r+1)$-form field is independent of the coordinates $\{x^a\}$ used in its definition, consider another set of coordinates $\{x^{a'}\}$. Then

$$\mathbf{A} = A_{a'b'\ldots d'}\,\mathrm{d}x^{a'} \wedge \mathrm{d}x^{b'} \wedge \ldots \wedge \mathrm{d}x^{d'},$$

where the components $A_{a'b'\ldots d'}$ are given by

$$A_{a'b'\ldots d'} = \frac{\partial x^a}{\partial x^{a'}}\frac{\partial x^b}{\partial x^{b'}}\cdots\frac{\partial x^d}{\partial x^{d'}}A_{ab\ldots d}.$$

Thus the $(r+1)$-form $\mathrm{d}\mathbf{A}$ defined by these coordinates is

$$\begin{aligned}
\mathrm{d}\mathbf{A} &= \mathrm{d}A_{a'b'\ldots d'}\,dx^{a'} \wedge \mathrm{d}x^{b'} \wedge \ldots \wedge \mathrm{d}x^{d'}\\
&= \mathrm{d}\left(\frac{\partial x^a}{\partial x^{a'}}\frac{\partial x^b}{\partial x^{b'}}\cdots\frac{\partial x^d}{\partial x^{d'}}A_{ab\ldots d}\right) \wedge \mathrm{d}x^{a'} \wedge \mathrm{d}x^{b'} \wedge \ldots \wedge \mathrm{d}x^{d'}\\
&= \frac{\partial x^a}{\partial x^{a'}}\frac{\partial x^b}{\partial x^{b'}}\cdots\frac{\partial x^d}{\partial x^{d'}}\mathrm{d}A_{ab\ldots d} \wedge \mathrm{d}x^{a'} \wedge \mathrm{d}x^{b'} \wedge \ldots \wedge \mathrm{d}x^{d'}\\
&\quad + \frac{\partial^2 x^a}{\partial x^{a'}\,\partial x^{e'}}\frac{\partial x^b}{\partial x^{b'}}\cdots\frac{\partial x^d}{\partial x^{d'}}A_{ab\ldots d}\,\mathrm{d}x^{e'} \wedge \mathrm{d}x^{a'} \wedge \mathrm{d}x^{b'} \wedge \ldots \wedge \mathrm{d}x^{d'} + \ldots + \ldots\\
&= \mathrm{d}A_{ab\ldots d} \wedge \mathrm{d}x^a \wedge \mathrm{d}x^b \wedge \ldots \wedge \mathrm{d}x^d
\end{aligned}$$

as $\partial^2 x^a/\partial x^{a'}\,\partial x^{e'}$ is symmetric in a' and e', but $\mathrm{d}x^{e'} \wedge \mathrm{d}x^{a'}$ is skew. Note that this definition only works for *forms*; it would not be independent of the coordinates used if the $\wedge$ product were replaced by a tensor product. Using the relation $\mathrm{d}(fg) = g\,\mathrm{d}f + f\,\mathrm{d}g$, which holds for arbitrary functions f, g, it follows that for any r-form $\mathbf{A}$ and form $\mathbf{B}$, $\mathrm{d}(\mathbf{A} \wedge \mathbf{B}) = \mathrm{d}\mathbf{A} \wedge \mathbf{B} + (-)^r \mathbf{A} \wedge \mathrm{d}\mathbf{B}$. Since (2.8) implies that the local coordinate expression for $\mathrm{d}f$ is $\mathrm{d}f = (\partial f/\partial x^i)\,\mathrm{d}x^i$, it follows that $\mathrm{d}(\mathrm{d}f) = (\partial^2 f/\partial x^i\,\partial x^j)\,\mathrm{d}x^i \wedge \mathrm{d}x^j = 0$, as the first term is symmetric and the second skew-symmetric. Similarly it follows from (2.9) that

$$\mathrm{d}(\mathrm{d}\mathbf{A}) = 0$$

holds for any r-form field $\mathbf{A}$.

The operator d commutes with manifold maps, in the sense: if $\phi: \mathscr{M} \to \mathscr{M}'$ is a C^r ($r \geqslant 2$) map and A is a C^k ($k \geqslant 2$) form field on $\mathscr{M}'$, then (by (2.7))

$$\mathrm{d}(\phi^* \mathrm{A}) = \phi^*(\mathrm{dA})$$

(which is equivalent to the chain rule for partial derivatives).

The operator d occurs naturally in the general form of Stokes' theorem on a manifold. We first define integration of n-forms: let $\mathscr{M}$ be a compact, orientable n-dimensional manifold with boundary $\partial\mathscr{M}$ and let $\{f_\alpha\}$ be a partition of unity for a finite oriented atlas $\{\mathscr{U}_\alpha, \phi_\alpha\}$. Then if A is an n-form field on $\mathscr{M}$, the integral of A over $\mathscr{M}$ is defined as

$$\int_{\mathscr{M}} \mathrm{A} = (n!)^{-1} \sum_\alpha \int_{\phi_\alpha(\mathscr{U}_\alpha)} f_\alpha A_{12\dots n} \mathrm{d}x^1 \mathrm{d}x^2 \dots \mathrm{d}x^n, \tag{2.10}$$

where $A_{12\dots n}$ are the components of A with respect to the local coordinates in the coordinate neighbourhood $\mathscr{U}_\alpha$, and the integrals on the right-hand side are ordinary multiple integrals over open sets $\phi_\alpha(\mathscr{U}_\alpha)$ of R^n. Thus integration of forms on $\mathscr{M}$ is defined by mapping the form, by local coordinates, into R^n and performing standard multiple integrals there, the existence of the partition of unity ensuring the global validity of this operation.

The integral (2.10) is well-defined, since if one chose another atlas $\{\mathscr{V}_\beta, \psi_\beta\}$ and partition of unity $\{g_\beta\}$ for this atlas, one would obtain the integral

$$(n!)^{-1} \sum_\beta \int_{\psi_\beta(\mathscr{V}_\beta)} g_\beta A_{1'2'\dots n'} \,\mathrm{d}x^{1'}\, \mathrm{d}x^{2'} \dots \mathrm{d}x^{n'},$$

where $x^{i'}$ are the corresponding local coordinates. Comparing these two quantities in the overlap $(\mathscr{U}_\alpha \cap \mathscr{V}_\beta)$ of coordinate neighbourhoods belonging to two atlases, the first expression can be written

$$(n!)^{-1} \sum_\alpha \sum_\beta \int_{\phi_\alpha(\mathscr{U}_\alpha \cap \mathscr{V}_\beta)} f_\alpha g_\beta A_{12\dots n} \,\mathrm{d}x^1\, \mathrm{d}x^2 \dots \mathrm{d}x^n,$$

and the second can be written

$$(n!)^{-1} \sum_\alpha \sum_\beta \int_{\psi_\beta(\mathscr{U}_\alpha \cap \mathscr{V}_\beta)} f_\alpha g_\beta A_{1'2'\dots n'} \,\mathrm{d}x^{1'}\, \mathrm{d}x^{2'} \dots \mathrm{d}x^{n'}.$$

Comparing the transformation laws for the form A and the multiple integrals in R^n, these expressions are equal at each point, so $\int_{\mathscr{M}} \mathrm{A}$ is independent of the atlas and partition of unity chosen.

Similarly, one can show that this integral is invariant under diffeomorphisms:

$$\int_{\mathscr{M}'} \phi_* \mathbf{A} = \int_{\mathscr{M}} \mathbf{A}$$

if ϕ is a C^r diffeomorphism ($r \geqslant 1$) from $\mathscr{M}$ to $\mathscr{M}'$.

Using the operator d, the *generalized Stokes' theorem* can now be written in the form: if $\mathbf{B}$ is an $(n-1)$-form field on $\mathscr{M}$, then

$$\int_{\partial\mathscr{M}} \mathbf{B} = \int_{\mathscr{M}} \mathrm{d}\mathbf{B},$$

which can be verified (see e.g. Spivak (1965)) from the definitions above; it is essentially a general form of the fundamental theorem of calculus. To perform the integral on the left, one has to define an orientation on the boundary $\partial\mathscr{M}$ of $\mathscr{M}$. This is done as follows: if $\mathscr{U}_\alpha$ is a coordinate neighbourhood from the oriented atlas of $\mathscr{M}$ such that $\mathscr{U}_\alpha$ intersects $\partial\mathscr{M}$, then from the definition of $\partial\mathscr{M}$, $\phi_\alpha(\mathscr{U}_\alpha \cap \partial\mathscr{M})$ lies in the plane $x^1 = 0$ in R^n and $\phi_\alpha(\mathscr{U}_\alpha \cap \mathscr{M})$ lies in the lower half $x^1 \leqslant 0$. The coordinates $(x^2, x^3, \ldots, x^n)$ are then oriented coordinates in the neighbourhood $\mathscr{U}_\alpha \cap \partial\mathscr{M}$ of $\partial\mathscr{M}$. It may be verified that this gives an oriented atlas on $\partial\mathscr{M}$.

The other type of differentiation defined naturally by the manifold structure is *Lie differentiation*. Consider any C^r ($r \geqslant 1$) vector field $\mathbf{X}$ on $\mathscr{M}$. By the fundamental theorem for systems of ordinary differential equations (Burkill (1956)) there is a unique maximal curve $\lambda(t)$ through each point p of $\mathscr{M}$ such that $\lambda(0) = p$ and whose tangent vector at the point $\lambda(t)$ is the vector $\mathbf{X}|_{\lambda(t)}$. If $\{x^i\}$ are local coordinates, so that the curve $\lambda(t)$ has coordinates $x^i(t)$ and the vector $\mathbf{X}$ has components X^i, then this curve is locally a solution of the set of differential equations

$$\mathrm{d}x^i/\mathrm{d}t = X^i(x^1(t), \ldots, x^n(t)).$$

This curve is called the *integral curve* of $\mathbf{X}$ with initial point p. For each point q of $\mathscr{M}$, there is an open neighbourhood $\mathscr{U}$ of q and an $\epsilon > 0$ such that $\mathbf{X}$ defines a family of diffeomorphisms $\phi_t\colon \mathscr{U} \to \mathscr{M}$ whenever $|t| < \epsilon$, obtained by taking each point p in $\mathscr{U}$ a parameter distance t along the integral curves of $\mathbf{X}$ (in fact, the ϕ_t form a one-parameter local group of diffeomorphisms, as $\phi_{t+s} = \phi_t \circ \phi_s = \phi_s \circ \phi_t$ for $|t|$, $|s|$, $|t+s| < \epsilon$, so $\phi_{-t} = (\phi_t)^{-1}$ and ϕ_0 is the identity). This diffeomorphism maps each tensor field $\mathbf{T}$ at p of type (r,s) into $\phi_{t*}\mathbf{T}|_{\phi_t(p)}$.

The *Lie derivative* $L_{\mathbf{X}}\mathbf{T}$ of a tensor field $\mathbf{T}$ with respect to $\mathbf{X}$ is