

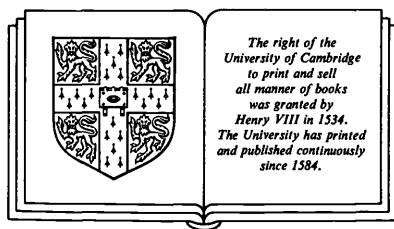
# Applied differential geometry

WILLIAM L. BURKE

# *Applied differential geometry*

WILLIAM L. BURKE

*University of California, Santa Cruz*



CAMBRIDGE UNIVERSITY PRESS

*Cambridge*

*New York New Rochelle*

*Melbourne Sydney*

Published by the Press Syndicate of the University of Cambridge  
The Pitt Building, Trumpington Street, Cambridge CB2 1RP  
32 East 57th Street, New York, NY 10022, USA  
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

© Cambridge University Press 1985

First published 1985

Reprinted 1987 (twice)

Printed in the United States of America

*Library of Congress Cataloging in Publication Data*

Burke, William L.

Applied differential geometry.

Bibliography: p.

1. Geometry, Differential. 2. Mathematical physics.

I. Title.

QC20.7.D52B87 1985 516.3'6 84-14952

ISBN 0 521 26317 4 hard covers

ISBN 0 521 26929 6 paperback

To all those who, like me, have wondered how in hell  
you can change  $\dot{q}$  without changing  $q$ .



# Contents

<i>Preface</i>	<i>page</i> xi
<i>Glossary of notation</i>	xv
<i>Introduction</i>	1
<b>I Tensors in linear spaces</b>	11
1 Linear and affine spaces	12
2 Differential calculus	21
3 Tensor algebra	27
4 Alternating products	31
5 Special relativity	37
6 The uses of covariance	44
<b>II Manifolds</b>	51
7 Manifolds	52
8 Tangent vectors and 1-forms	59
9 Lie bracket	68
10 Tensors on manifolds	72
11 Mappings	77
12 Cotangent bundle	84
13 Tangent bundle	90
14 Vector fields and dynamical systems	94
15 Contact bundles	99
16 The geometry of thermodynamics	108
<b>III Transformations</b>	115
17 Lie groups	115
18 Lie derivative	121
19 Holonomy	132

20	Contact transformations	136
21	Symmetries	141
<b>IV</b>	<b>The calculus of differential forms</b>	<b>147</b>
22	Differential forms	147
23	Exterior calculus	153
24	The $*$ operator	159
25	Metric symmetries	169
26	Normal forms	173
27	Index notation	176
28	Twisted differential forms	183
29	Integration	194
30	Cohomology	202
<b>V</b>	<b>Applications of the exterior calculus</b>	<b>207</b>
31	Diffusion equations	207
32	First-order partial differential equations	213
33	Conservation laws	219
34	Calculus of variations	225
35	Constrained variations	233
36	Variations of multiple integrals	239
37	Holonomy and thermodynamics	245
38	Exterior differential systems	248?
39	Symmetries and similarity solutions	258
40	Variational principles and conservation laws	264
41	When not to use forms	268
<b>VI</b>	<b>Classical electrodynamics</b>	<b>271</b>
42	Electrodynamics and differential forms	272
43	Electrodynamics in spacetime	282
44	Laws of conservation and balance	285
45	Macroscopic electrodynamics	293
46	Electrodynamics of moving bodies	298
<b>VII</b>	<b>Dynamics of particles and fields</b>	<b>305</b>
47	Lagrangian mechanics of conservative systems	306
48	Lagrange's equations for general systems	311
49	Lagrangian field theory	314
50	Hamiltonian systems	320
51	Symplectic geometry	325
52	Hamiltonian optics	333
53	Dynamics of wave packets	338

<b>VIII</b>	<b>Calculus on fiber bundles</b>	347
54	Connections	349
55	Parallel transport	354
56	Curvature and torsion	358
57	Covariant differentiation	365
58	Metric connections	367
<b>IX</b>	<b>Gravitation</b>	371
59	General relativity	372
60	Geodesics	374
61	Geodesic deviation	377
62	Symmetries and conserved quantities	382
63	Schwarzschild orbit problem	387
64	Light deflection	393
65	Gravitational lenses	395
66	Moving frames	402
	<i>Bibliography</i>	409
	<i>Index</i>	411



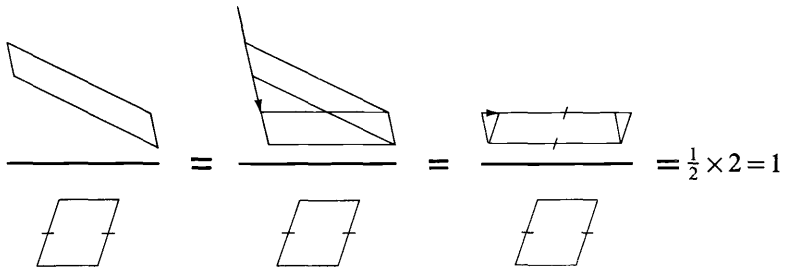


## *Preface*

This is a self-contained introduction to differential geometry and the calculus of differential forms. It is written primarily for physicists. The material complements the usual mathematical methods, which emphasize analysis rather than geometry. The reader is expected to have the standard physics background in mechanics, electrodynamics, and mathematical methods. The mathematically knowledgeable can skip directly to the heart of the book, the calculus of differential forms, in Chapter IV.

This book falls between the usual mathematics and physics texts. On the one hand, proofs are given only when they are especially instructive. On the other hand, definitions, especially of mathematical structures, are given far more carefully than is the usual practice in physics. It is very dangerous to be sloppy in your definitions. I have taken considerable care to give many physical applications and to respect the physical subtleties of these applications. Indeed, my operational rule was to include no mathematics for which I could not produce a useful example. These examples form nearly half the book, and a large part of your learning will take place while reading and thinking about them. I feel that we learn far more from carefully chosen examples than from formal and unnatural deductive reasoning. Most of these examples were originally problems. I wish that I had been left with still more problems for the reader.

I call this a geometric treatment. What do I mean by geometry? One connotation is that of diagrams and pictorial representations. Indeed, I called an early set of notes “The Descriptive Geometry of Tensors.” You will find many diagrams here, and I have gone to some effort to make them honest diagrams: Often they can be substituted for the verbal hints that sometimes constitute a proof.



An illustration of the comparison of two areas. No metric is needed, only the construction of parallel lines and the comparison of lengths along a line.

By geometry, however, I mean more than pictures. After all, pictorial methods grow clumsy in spaces of dimension greater than four, or five. To be geometric connotes for me an emphasis on the structures themselves, rather than on the formal manipulation of representations, especially algebraic ones.

**Examples:** To appreciate the distinction between the simplicity of a geometric concept and the complexity of its numerical representation, consider an ellipse in general position. A simple idea, but represented by an appalling mess of numbers.

A determinant can be defined in terms of explicit rules of computation. But geometrically we can better define the determinant as the factor by which a linear transformation changes volumes. Now, to compare the volumes of two parallelepipeds does not require a metric structure. A linear structure is sufficient. For example, a comparison of the areas of two parallelograms is shown in the accompanying diagram. In this geometric view, the determinant of the composition of the two linear transformations (matrix product) is obviously the product of the determinants. To prove this from the algebraic view requires an involved calculation given in Section 27.

The emphasis on the structures themselves rather than on their representations leads us naturally to use the coordinate-free language of modern mathematics.

This modern language makes the foundations of physical models clear and precise. Surprisingly, it also makes the computations clearer. Further, the coordinate-free language turns out to be very easy to illustrate. As this material developed, there was a useful symbiosis between formalism and concrete calculation. The solutions to concrete problems often led to improvements in the formalism. For example, I taught a Jackson-level electrodynamics class using differential forms, and this led to an improved definition of the Hodge star operator. It also forced me to learn and use twisted tensors. Thermodynamics taught me the importance of contact manifolds and affine structures. Books that remain on the formal level treat these important geometric objects briefly if at all.

This emphasis on concrete applications and proper geometric structures helps us avoid the formal symbol manipulations that so often lead to nonsense or fallacious proofs of correct results. [Look at Figure 3.1 in Soper (1976) or the horrible calculus of variations manipulations and mistakes in Goldstein (1959).] Here we will be able to turn most of the infinitesimals commonly seen in physics into the appropriate geometric objects, usually into either rates (tangent vectors) or gradients (differential forms). The distinction between these is lost in the metric-blinded symbol pushing of tensor calculus. Nor will the funny deltas of the calculus of variations with their ad hoc rules of manipulation be found here.

The material of this book grew out of the first quarter of a fairly ordinary general-relativity course. After teaching the course several times, I realized that general relativity as it is usually taught was bad for the student. The introduction of a metric right at the start obscures the geometric structures and would force us into abstract number shuffling. As I put the metric later and later into the course, I discovered examples from an ever-wider range of applications: classical mechanics, dispersive waves, thermodynamics, and so on. After a while I grew embarrassed at pretending that it was still a general-relativity course and, presto, a course on applied differential geometry appeared.

Although most of this material is fairly standard, and by intent the notation is conventional, two departures from common practice should be mentioned. One is the use of twisted tensors. The importance of twisted tensors in physics has been neglected by nearly everyone. There is not even agreement on the name for these objects. Some of these twisted tensors are related to the axial vectors of physics. The second novelty is the use of contact manifolds. These, rather than symplectic manifolds, are the proper setting for most physical theories. Symplectic geometry is

important for special situations, primarily time-independent Hamiltonian mechanics.

You might ask, why not learn these fancy methods after the usual tensor calculus? Ordinarily I would agree with a progressive, top-down approach. For the material here it doesn't work. The key idea is that here we are removing structure from our geometric objects, not adding it. To think of a space without a metric or a linear structure is much harder than to think about adding structure to a space, for example, adding a multiplication rule to turn a vector space into a Lie algebra. To work hard to erase what you have just worked hard to learn is frustrating and inefficient. Thus I recommend here the bottom-up approach. You should not view this as the overthrow of all that you have learned, however. Rather, view it as a natural development of vector calculus.

It is impossible for me to recall all the sources for this material. Particularly helpful were Frank Estabrook, Jim Swift, Richard Cushman, Ralph Baierlein, Kim Griest, Hume Feldman, and David Fried, but hold them blameless for my mistakes and idiosyncrasies.

W. L. B.

# Glossary of notation

Entries are generally arranged in order of first occurrence in the text.

$\rightarrow$	relates source and target sets of a map, as in $f: A \rightarrow B$
$\mapsto$	relates a typical source element and its corresponding target element, as in $g: x \mapsto x^3$
$\Rightarrow$	logical implication
$\times$	Cartesian product of two sets; rarely, the 3-space vector cross product
$\circ$	composition of maps
$\{ \}$	encloses a list of the elements of a set
$,$	partial differentiation, as in $f_{,x}$
$\sim$	equivalence relation
$[b]$	the equivalence class containing the element $b$
$[a:b:c]$	homogeneous coordinates
$\in$	relates an element to its set, as in $x \in \mathbb{R}$
$\subset$	set inclusion, as in $A \subset B$
$V, V^*$	a vector space, and its dual
$e_i$	basis vectors
$f^i$	dual basis vectors
$f'$	differentiation of a function with respect to its argument
$Df(u)$	the differential of the function $f$ at the point $u$
$\cdot$	action of a linear operator (replaces the generally used parentheses); also a missing argument of a function
$\  \cdot \ $	norm on a vector space
$  \cdot  $	absolute value
$A^t$	transpose of the matrix $A$
$D^2f(u)$	second differential of $f$ at $u$

$\mathbb{R}$	the set of real numbers, also $\mathbb{R}^2$ , and so on
$\otimes$	tensor product
$\binom{p}{q}$	the type of tensor, that has $p$ upper indices and $q$ lower indices; a tangent vector is of type $\binom{1}{0}$
$\wedge$	alternating product, sometimes called exterior product
$\bullet$	metric inner product
$\mathcal{G}$	a metric tensor
$\mathcal{E}$	a Euclidean metric tensor
$z$	the redshift; $(1+z)$ is the ratio of new wavelength to old
$\phi^i$	the $i$ th chart map
$C^\infty$	the set of infinitely differentiable functions
$S^n$	the surface of the sphere in $(n+1)$ -dimensional Euclidean space; $S^1$ is the circle
$T_p(M)$	the set of tangent vectors at the point $p$ of a manifold $M$
$T_p^*(M)$	the set of 1-forms at the point $p$ of a manifold $M$ , dual to $T_p(M)$
$\dot{\gamma}$	the name for an object that is like the rate of change of $\gamma$
$\partial/\partial x$	a basis tangent vector, tangent to the $x$ axis
$dx$	a basis 1-form, the gradient of the $x$ coordinate function
$X(s)$	when the independent variable $x$ is represented by a function, it is usually written $X$
$[u, v]$	the Lie bracket of vector fields $u$ and $v$ ; rarely, the commutator of two operators
$\psi_*$	the pushforward map derived from $\psi$
$\psi^*$	the pullback map derived from $\psi$
$\int_\gamma \omega$	the integration of the differential form $\omega$ along the curve $\gamma$
$\pi$	a projection map
$\pi: E \rightarrow B$	a fiber bundle with projection map $\pi$ , total space $E$ , and base space $B$
$\pi^{-1}(b)$	the set of elements mapped onto $b$ by the map $\pi$
$T^*M$	the cotangent bundle of $M$
$TM$	the tangent bundle of $M$ , dual to $T^*M$
$Tf$	the tangent map derived from $f$
$CM$	the line-element contact bundle of $M$
$C^*M$	the hypersurface-element contact bundle of $M$ , not dual to $CM$
$C(M, n)$	the bundle of $n$ -contact elements of $M$
$F_x$	the name for a variable which is like a partial derivative of $F$
$L_b$	left translation by the Lie-group element $b$
$\mathcal{L}_w \Omega$	the Lie derivative of $\Omega$ by the vector field $w$

$B$	a 3-vector, rarely used
$\nabla$	the 3-vector differential operator $\text{del}$
$\mathcal{I}$	an ideal of differential forms
$\Lambda^r$	the set of $r$ -forms
$\lrcorner$	the contraction operator, as in $v \lrcorner \omega$
$\epsilon^{xyz}$	the permutation symbol; can also be written with lower indices
$d$	exterior derivative
$\theta$	volume element
$\theta$	basis $(n-1)$ -forms derived from the volume element $\theta$
$*$	Hodge star operator
$\star$	the special Hodge star in spacetime, used when the 3-space star is also present
$\#$	the sharp operator mapping 1-forms to tangent vectors using a metric
$  $	used around a set of indices to indicate an ordered summation
$\delta_{\sigma\tau}^{\mu\nu}$	Kronecker delta – may have any number of indices on top, with the same number on the bottom
$\omega^A$	an uppercase index denotes a block index: a string of ordinary indices of unspecified length
$(\beta, \Omega)$	the representation of a twisted tensor using the ordinary tensor $\beta$ and an orientation $\Omega$
$\{B\}$	a differential form representing the orientation of the object $B$
$\partial M$	the boundary of the set $M$
$\delta$	the differential operator adjoint to $d$
$\Delta$	the de Rham operator, $d\delta + \delta d$
$\text{div}$	the divergence operator, as in $\text{div } v$
$\Omega^n$	when $\Omega$ is a differential form, this indicates the $n$ -fold exterior product
$\Gamma_{j\nu}^\mu$	connection components
$\nabla_A B$	the covariant derivative of the vector field $B$ in the $A$ direction
$T(\cdot, \cdot)$	torsion tensor
$R(\cdot, \cdot)$	curvature tensor
$;$	covariant derivative, for example, the $\omega_{\mu;\nu}$ are the components of the covariant derivative of the 1-form $\omega_\mu dx^\mu$
$\stackrel{\cong}{=}$	equality that holds only in special coordinates
$\omega^i$	a frame of orthonormal (or pseudoorthonormal) 1-forms
$\omega^i_j$	connection 1-forms
$\Omega_{ij}$	curvature 2-forms

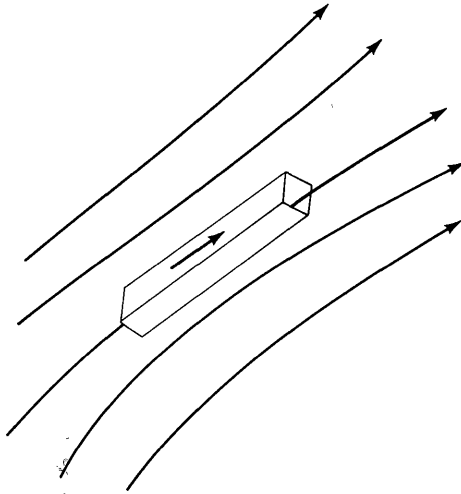




# *Introduction*

No one would try to teach electrodynamics without using vector calculus. It would be crazy to give up so powerful a tool, and it is hard for us to appreciate the resistance to vector calculus shown at the turn of the century. The mathematics of this book can be thought of as the proper generalization of vector calculus, div, grad, curl, and all that, to spaces of higher dimension. The generalization is not obvious. Ordinary vector calculus is misleading because the vector cross product has special properties in three dimensions. This happens because, for  $n=3$ ,  $n$  and  $\frac{1}{2}n(n-1)$  are equal. It is also important to divorce the formalism from its reliance on a Euclidean metric, or any metric for that matter. Other structures are important, and we must make room for them. Also, a metric allows some accidental identifications that obscure the natural properties of the geometric structures. Similarly, the linear and affine structure of Cartesian spaces should be included only if it is appropriate. The mathematics satisfying these conditions is not classical tensor calculus, but what is called calculus on manifolds and, in particular, the calculus of differential forms.

What physical problems does this calculus address? The basic idea of any calculus is to represent the local behavior of physical objects. Suppose you have a smooth distribution of, say, electric charge. The local behavior is called charge density. A graphical representation of this density is to draw a box, a parallelepiped, that encloses a unit amount of charge (on the average) in the limit where the boxes become smaller than the variations in the density. These boxes have volumes but no particular shapes. What they have is the idea of relative volume: Given two such boxes, we can find the ratio of their volumes. (See the example in the Preface.) Charges come positive and negative, and associated with each



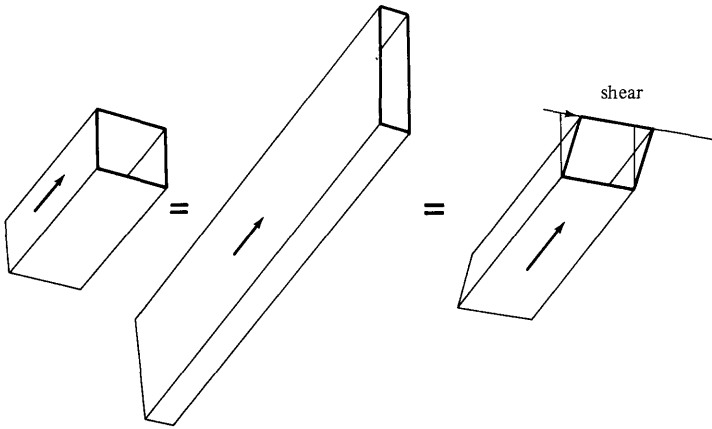
**Figure 1.** A geometric representation of the current density at a point. The box lies along the current lines and encloses one unit of current in the limit where the box shrinks to nothing. The arrow points in the direction of current flow.

box is a sign. These geometric objects form a one-dimensional vector space at each point.

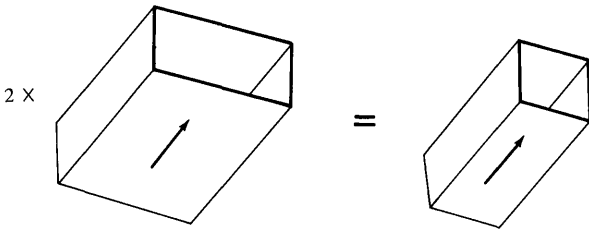
A more complicated situation would be to have a smooth distribution of electric current. To represent this, draw another box, this one of indefinite length in one dimension. The box should be aligned so that no current flows through the sides, and the cross section is such that it encloses a unit amount of current in the same limit as before. In addition to the shape and alignment, the box now needs an arrow pointing in the direction of current flow (Figure 1). In Chapter IV such geometric objects will be discussed; they are called twisted 2-forms.

**Example:** A larger current is represented by a box of smaller cross section. The cumulative particle flux in an accelerator is such a current density, and, reasonably enough, it is commonly measured these days in inverse nanobarns.

Such current densities form a three-dimensional vector space. The rules for equivalence, scaling, and addition are easy to discover, and are illustrated in Figures 2, 3, and 4. These geometric objects also describe electric



**Figure 2.** The equivalence of different representations of current density. Each box encloses the same current in the limit.

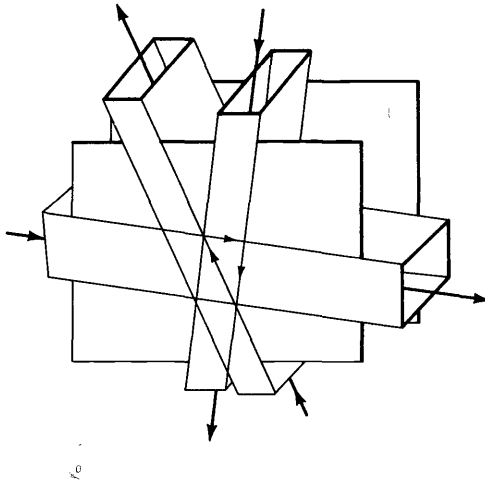


**Figure 3.** The scaling of current densities. Each encloses the same current.

flux. The field lines appear on positive charges and disappear on negative charges.

Another, related geometric object has a representation given by two parallel planes. Now two of the three dimensions are of indefinite extent. With a convention marking one of these planes as the higher, this geometric object can represent a potential gradient. In these last two geometric objects, we see two different aspects of the electric field: first as the quantity of flux, second as the intensity of the field. This pairing of variables is reminiscent of thermodynamics.

All the geometric objects described so far are types of differential forms: Charge density is of the type called a 3-form, current density is a type of 2-form, and the potential gradient is a 1-form. Each, as you can



**Figure 4.** The addition of current densities, shown here in symmetric form. Three current densities that add up to zero are shown. The negative of any one is the sum of the other two.

see, is a natural geometric object in its own right. Although there is some computational convenience to defining, say, 2-forms as objects made up from two 1-forms, this obscures the real meaning of a 2-form. Whenever possible I will try to penetrate the accidental properties inherent in merely convenient representations and constructions, and display the real properties of our geometric objects. For this, pictorial representations and concrete examples will be invaluable.

These geometric objects represent the local behavior of things. Their representations will be in terms of coordinates, but these coordinates will not usually have any intrinsic meaning. Thus their representations must endure under arbitrary smooth changes of the coordinates. Locally these coordinate changes are linear transformations. Note how the constructions given use only objects, such as lines, planes, and volume ratios, that are invariant under linear transformations.

Physical laws relate the behavior of these geometric objects at different points. In electrodynamics, flux lines and current are conserved except where charge density is present or changing. In electrostatics, the electric field can be derived from a potential and has no curl. We will find a natural differential operator, the exterior derivative, with which to express these laws.

Spaces may have additional structure, and often this can be represented in terms of differential forms. Thermodynamics takes place on a space with a special 1-form on it; this carries the content of the first law. Hamiltonian mechanics has a special 2-form, needed, for example, to represent the density in Liouville's theorem. Electrodynamics uses an operator to relate electric flux and electric field intensity; this operator changes 1-forms into 2-forms and numbers into 3-forms. Of course, not everything fits so simply into this scheme, but we will see that an astonishing amount does.

### Preliminaries

This is a section on tools. It is a brief review of the definitions, concepts, and notational conventions to be used in the book. I have tried to follow the advice of Schouten (1954, p. 61) and use only notation in common use. The exception is the extended vector notation, which I will explain.

### Maps

We will refer to many different sets, such as the points of a plane, the lines through a point, or the rotations in three dimensions. One unfamiliar notational usage is that we will refer to points in a  $n$ -dimensional space by a single letter, such as  $x$ . Of central importance is the generalization of the idea of a *function* to such sets. A *map* is a relation between two specified sets that associates a unique element of the second to each element of the first. In different circumstances different notations for the map and its result are useful. To allow complete freedom in choosing this notation, maps will be introduced with a mathematical sentence of the form

$$f: C \rightarrow B; c \mapsto f(c).$$

Here  $f$  is the name of the function. The first phrase specifies the sets involved. This  $f$  takes each element of the set  $C$  and associates with it an element  $b$  in the set  $B$ . The second phrase uses a stopped arrow, and gives either the notation for the specific result or the rules for computing the map. Here  $f(c)$  is the element of  $B$  resulting from  $f$  applied to the element  $c$  of  $C$ . When the sets are obvious, the first phrase will often be omitted.

**Example:** The Euclidean metric is a rule for computing the “dot product” of two vectors. If we call the set of Euclidean vectors  $V$ , then the “dot” is a map:

$$\cdot : V \times V \rightarrow \mathbb{R}; (a, b) \mapsto a \cdot b.$$

The cross denotes the *Cartesian product*. Elements of the set  $A \times B$  are ordered pairs  $(a, b)$ , one from the set  $A$  and one from the set  $B$ . We will make no use of the vector cross product; so no confusion can arise.

Many functions will be written in the usual form

$$f: x \mapsto f(x).$$

When our maps are linear operators, we will usually write the evaluation with a centered dot as  $L \cdot x$  rather than  $L(x)$ . We will use  $f^{-1}(x)$  to denote the set of all those elements that  $f$  maps into  $x$ ; thus  $f^{-1}$  (read this “ $f$  inverse”) is not a function or a map. If the ranges and values are compatible, we can operate with two different maps in succession; this *composition* is written  $g \circ f$ , and means that  $f$  acts first, then  $g$  acts.

I will frequently abuse precise notation by leaving out the arguments of functions in situations where they are obvious. Be careful in such situations not to confuse a function and its values. A precise language would say  $f$  for the function and  $f(x)$  for its value.

One way to construct new maps from old ones is to only partially evaluate a map that requires several arguments. The usual notation for this is the centered dot. Thus if we have a map

$$\phi: E \times F \rightarrow \mathbb{R}; (e, f) \mapsto \phi(e, f),$$

then  $\phi(\cdot, f)$  is a map  $E \rightarrow \mathbb{R}$ , and so on.

### *Ideals*

When a set has an algebraic structure, that is, a rule for the multiplication of two of its elements, then the notion of an ideal is important. An *ideal* is a subset such that whenever any element in the ideal is multiplied by any other element whatsoever, the result is in the ideal. We often can specify an ideal by giving several elements that generate it.

**Example:** The ideal generated in the set of all polynomials by a given polynomial consists of polynomials that have the value zero whenever the generating polynomial has the value zero.

### *Index notation*

An important notational tool is the use of *indices* and the *summation convention*. A set of elements can be labeled by an index. The set over

which the index ranges (often but not necessarily the integers) should be specified if it is not obvious. Indices will be written as subscripts or superscripts. For tensors this is a convenient labeling of their type.

**Example:** The components of ordinary 3-vectors can be written

$$v^a = \{v^x, v^y, v^z\}.$$

Here the index  $a$  ranges over the index set  $\{x, y, z\}$ . Sets are often specified by listing their elements inside braces.

Great simplification comes from using the summation convention on these indices: In an expression such as  $v^a w_{ab}$  where an index appears exactly *twice*, once as a superscript and once as a subscript, a summation is implied over the appropriate index set. Partial derivatives will be denoted by a comma followed by the appropriate subscript

$$\psi_{,x} = \frac{\partial \psi}{\partial x}.$$

An uncommon notation I use in this book is boldface type to denote an “obviously” missing index. This is an extension of the usual vector notation, but without matrices or special ordering. This is a useful heuristic tool in constructing an argument or a calculation. Often the one-dimensional calculation needs only boldface to apply in more dimensions. Be careful: The summation convention will be used on appropriate pairs of missing indices. In detailed calculations we will usually abandon the index notation and, for simplicity, give single-letter names to tensor components. Also, be careful about sums such as  $qq$ . Such a sum usually makes no sense. The missing indices are not one up and one down, and this is a warning that something is wrong.

### *Equivalence relations*

An important concept is that of an equivalence relation. This is a useful tool that allows us to subtract information. An *equivalence relation* on a set is a relation between pairs of elements, denoted usually by “ $\sim$ ,” which is reflexive,

$$x \sim x,$$

symmetric,

$$x \sim y \Rightarrow y \sim x,$$

and transitive,



$$(x \sim y) \text{ and } (y \sim z) \Rightarrow (x \sim z).$$

**Example:** An equivalence relation in the plane is defined by

$$(x_1, y_1) \sim (x_2, y_2) \quad \text{iff } x_1 = x_2.$$

This is called projection on the first coordinate.

The smaller set formed from the set  $S$  by using the equivalence relation  $\sim$  is written  $S/\sim$ . Physicists use the idea of an equivalence relation informally, but rarely appreciate its full power and utility.

We can often deal with an equivalence class by taking a typical member as a representative. The equivalence class containing  $x$  will be written using square brackets out of hierarchy,  $[x]$ . One use of equivalence relations is to reduce a redundant notation. Such notation often simplifies a problem.

**Example:** Look at the set of lines through the origin in three dimensions. Clearly two numbers are sufficient to denote any such line. Still, using three numbers leads to a much simpler picture. Denote the line passing through the point  $(x, y, z)$  by the triple of numbers  $[x:y:z]$ . These are called the homogeneous coordinates of the line. It is a redundant notation, and

$$[x:y:z] \sim [kx:ky:kz]$$

for all nonzero factors  $k$ . This redundant notation cures the problems of dealing with the poles. Problems in projective geometry, such as those needed in computer graphics, are much simpler when expressed in homogeneous coordinates. See Newman and Sproul (1973). The two-dimensional set of lines described here is called the projective plane  $P^2$ .

Equivalence relations appear throughout this book. Some important ones are as follows: An affine space is an equivalence class of vector spaces; a tangent vector is an equivalence class of tangent curves; a 1-form is an equivalence class of tangent functions; and an orientation is an equivalence class of ordered bases.

### *Implicit definitions*

Another important tool that should not be ignored is the implicit definition. Often it is simpler to define something by its properties rather than

by giving a representation or a constructive algorithm. We will often use an implicit definition, one that does not refer to a representation, as an easy way to show that the object is independent of the arbitrary choices involved in its representation. Of course, an implicit definition needs a proof of consistency, which might itself be difficult. Often we will take the best of both worlds. If we give both definitions and show that they are equivalent, then the implicit one shows representation independence, whereas the explicit one shows consistency.

### *Figures*

This book uses pictures as a tool to illustrate many of its points. A picture is necessarily concrete, and this balances the abstraction of the equations. Even when carefully drawn, pictures require careful interpretation. Often several dimensions are suppressed. Also, conventional orthography introduces a spurious orthogonality, which should be ignored. All in all, there is an art to reading these pictures. I have intentionally avoided the use of fancy computer-generated figures. You should learn to draw figures like these for yourself, and use them as an aid to your thinking.

### *Problems*

Although I have written this book with the idea that it will be used for self-instruction much of the time, I have not included many of the usual exercises for the reader. Instead, they appear here as worked examples. Furthermore, students able to use this book should be able to invent routine exercises for themselves, and it would be a disservice not to force them to do so. On the other hand, I have put in suggestions for further work. These are not as cut and dried as the usual exercises, but instead describe interesting and productive directions that the student may wish to pursue. In some cases I have gone a good way in the suggested direction myself; in others, I have just passed on what seems like a good idea to me.



# I

## *Tensors in linear spaces*

Linear spaces and the closely related affine spaces loom large in the literature of mathematics and physics. It is not just that only linear problems are easy to solve. The real reason lies in the fundamental idea of calculus: the local behavior of a smooth function can be represented by a linear function. Calculus is just the technology of these local linear approximations. The spaces in which these local approximations live are important linear spaces.

The most important linear space for us is the space of tangent vectors at a point. Elements of this space are local approximations to smooth curves passing through the point. Because of this, the tangent space, as it is called, is a local picture of the space itself. It is therefore easy to picture. Given any linear space there is an entire algebra of related linear spaces that comprise the various linear and multilinear operators. These operators are called tensors.

Closely related are affine spaces. These spaces have all the structure of a linear space, except that all points are equivalent. They lack a special point for the origin. These spaces are important because they form the arena for much of physical theory. Newtonian mechanics and special relativity are both set in an affine space. The straight lines and uniform parametrization model the uniform motion of free particles. In general relativity, this affine structure results because special relativity is a local approximation to general relativity.

The study of the representations of linear geometric objects is of fundamental importance, as is the covariance of these representations under linear transformations. A smooth coordinate transformation looks locally like a linear transformation. Geometric objects must behave

in an orderly manner under coordinate changes, and their local representations must be covariant under linear transformations.

This covariance under linear transformations shows up in the diagrams. Constructions involving parallelism, tangency, and incidence are allowed; ones involving perpendicularity are not. When teaching, I ask my students to imagine that they control a machine that can make arbitrary linear transformations of the blackboard whenever I turn my back. If my figures are properly covariant, I can always continue my lecture despite these expansions, rotations, and shears.

It is possible to ascribe primary significance to the transformations rather than to the geometric structures. We then study all those geometric structures compatible with a given set of transformations. This is a useful point of view, particularly in special relativity, but not one that I will follow.

### 1. Linear and affine spaces

The qualifier “linear” is used loosely by physicists to refer to several different mathematical structures, all closely related and all useful. There are linear structures associated with vectors, with free vectors, with covectors, and with elements of affine spaces. We will also study affine spaces, since they are an excellent example of the process of subtracting structure from a set.

#### *Linear spaces*

A linear space is a set with two operations, addition and scaling, defined on it. Its elements are often called vectors, and the space a vector space. It must have a special element called the zero vector, and obey simple axioms that force these operations to behave in an orderly manner. Addition commutes and associates. For  $a$ ,  $b$ , and  $c$  vectors,

$$a + b = b + a$$

$$a + (b + c) = (a + b) + c.$$

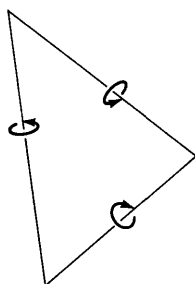
The zero vector must exist, with the property

$$a + 0 = a,$$

as must an addition inverse,

$$a + (-a) = 0.$$

Scaling must associate. For  $k$  and  $m$  numbers,



**Figure 1.1.** The addition of three twisted vectors in three dimensions. The three shown add up to zero.

$$k(ma) = (km)a.$$

Scaling must also distribute:

$$(k + m)a = ka + ma,$$

$$k(a + b) = ka + kb.$$

Finally, scaling by unity must be the identity operation:

$$1a = a.$$

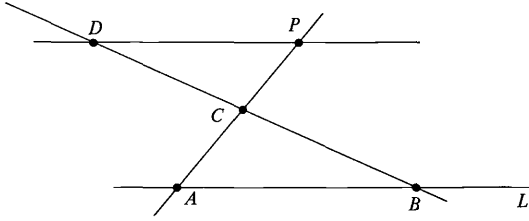
**Example:** Look at the objects in three dimensions that are represented by line segments of definite length. Conventional vectors have a direction; let us instead give these a screw sense (right-handedness or left-handedness). The addition of these objects resembles the addition of vectors, with the signs being specified by the rule shown in Figure 1.1, which shows three such objects that add up to zero.

This is a deliberately unusual example. These objects are called *twisted vectors*. Twisted tensors are used in electrodynamics, and will be discussed in Chapters IV and V.

A vector space is a nice example of a set that is given structure by addition. Next we will look at a structure defined by subtraction.

### *Affine spaces*

An affine space has less structure than a vector space. If we are given any two points of an affine space, the affine structure lets us draw a



**Figure 1.2.** Use of an affine structure to draw a line parallel to  $L$  through the point  $P$ .  $C$  is the midpoint of  $DB$  and  $AP$ .

parametrized straight line through the two points. It defines a notion of uniformity along a family of parallel lines. There is no origin in an affine space, and no idea of scaling. (The reader already familiar with general relativity should not confuse these affine spaces with spaces bearing an affine connection. There is little relation.)

**Examples:** An affine structure lets us draw parallel lines. To draw a line parallel to a given line  $L$  through a point  $P$ , pick any two points  $A$  and  $B$  on  $L$ , and follow the construction shown in Figure 1.2.

Look at a three-dimensional vector space, and add to it the choice of a preferred direction (Figure 1.3). You can think of ordinary 3-space and the vertical direction, but without its Euclidean metric. Now look at the set of all planes that pass through the origin but do not lie in the preferred direction. This set has an affine structure but no linear structure. For example, it has no special plane representing the zero vector.

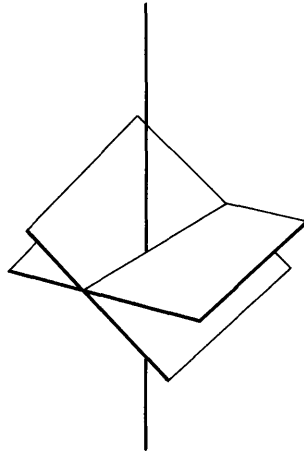
The parametrization along the line passing through two points in an affine space is not unique. Transformations of the parameter  $u$

$$u \mapsto ku + b$$

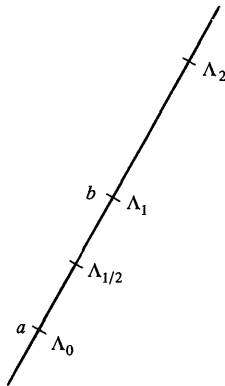
change one uniform parametrization into another. If we single out the parametrization that runs from zero to one between the points, then the structure of an affine space  $A$  is given by a map

$$\Lambda : A \times A \times \mathbb{R} \rightarrow A; (a, b, k) \mapsto \Lambda_k(a, b),$$

with conditions



**Figure 1.3.** In 3-space with a preferred direction (the vertical line here in the figure) singled out, we have a natural affine structure on the planes passing through the origin, but no linear structure.



**Figure 1.4.** The affine map  $\Lambda$ .

$$\Lambda_0(a, b) = a, \quad \Lambda_1(a, b) = b.$$

See Figure 1.4.

**Example:** The structure of the affine space of planes just given can be represented as follows. Let  $e$  be a vector pointing in the preferred



direction, and let  $a, b$  be two other vectors such that  $(e, a, b)$  is a basis for the vector space. Any plane in the set can be represented by the pair of numbers  $(u, v)$  such that the vectors  $a + ue$  and  $b + ve$  lie in the plane. The affine map  $\Lambda_k$  is given by

$$\Lambda_k((u_1, v_1), (u_2, v_2)) = (u_1 + k(u_2 - u_1), v_1 + k(v_2 - v_1)).$$

Note that the representation given has a covariance. The vectors  $a$  and  $b$  can be replaced by any other nondegenerate pair. Although the plane represented by  $(0, 0)$  might appear to be distinguished in one representation, it is not covariant under the above changes. Lacking a zero element, this set does not have a linear structure.

You might now expect a set of axioms for the operation  $\Lambda$ . Curiously, a suitable set of axioms is neither obvious, simple, nor useful. A better way to describe the structure of an affine space is subtractive. An affine space is a linear space minus its origin. Given a linear space, we can easily see that the affine map

$$\Lambda_k(a, b) = a + k(b - a)$$

is invariant under changes of the origin.

The consistency conditions on the operation  $\Lambda$  are best described by saying that the choice of any element  $e$  as the zero element turns an affine space into a linear space if we define scaling by

$$ka = \Lambda_k(e, a),$$

and addition by

$$a + b = \Lambda_2(e, \Lambda_{1/2}(a, b)). \quad (1.1)$$

Were we to make extensive use of this affine structure, we would simplify the notation by defining a special midpoint operation. In the preceding, the choice of the element  $e$  is not essential. If it is true for any  $e$ , it is true for all.

You should think a bit about this subtractive style of defining structure. It is less intuitive than the direct imposition of structure, but often very efficient and natural. If you doubt this, try to give a suitable set of axioms for  $\Lambda_k(a, b)$ , and then try to prove that the preceding construction gives a linear space.

### *Free vectors*

There is a possible confusion here between the vector spaces that a mathematician has in mind – abstract sets whose elements can be added and

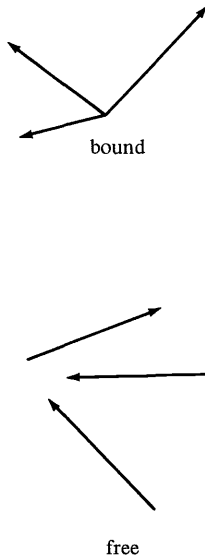


Figure 1.5. Bound and free vectors.

scaled – and the idea coming from 3-vectors that the physicist has in mind: a line with an arrowhead on one end of it. The physicist’s vector is free to roam all over the vector space. To be precise, we call these *free vectors*. Vectors in the sense of a linear vector space, with their tails all at the origin, are called *bound vectors* (Figure 1.5). In an affine space only the free vector is a well-defined object. The two different ideas are useful in different situations, and we will need to use both of them. The idea of putting all the tails at the origin lets us give a less-cluttered view of infinite sets of vectors, such as curves of vectors.

Free vectors can be defined in affine spaces as the set  $(A \times A) / \sim$ , where the equivalence relation  $\sim$  is defined by using the map  $\Lambda_k$ :

$$(a, b) \sim (c, d) \quad \text{iff } d = \Lambda_2(a, \Lambda_{1/2}(b, c)).$$

This is sketched in Figure 1.6. Scaling by  $k$  is the operation

$$[(a, b)] \mapsto [(a, \Lambda_k(a, b))].$$

Here the square brackets stand for an equivalence class;  $[(a, b)]$  is the equivalence class containing  $(a, b)$ . Addition is the operation

$$[(a, b)] + [(c, d)] = [(a, \Lambda_2(a, \Lambda_{1/2}(b, \Lambda_2(c, \Lambda_{1/2}(a, d)))))].$$

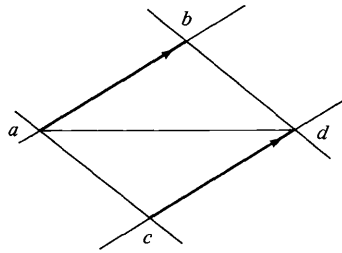


Figure 1.6. Equivalence of free vectors defined by using an affine structure.

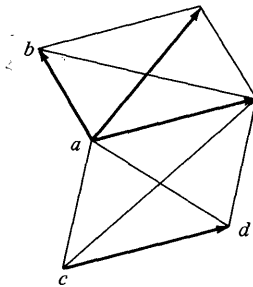


Figure 1.7. The addition of free vectors defined by using an affine structure.

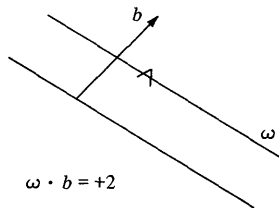


Figure 1.8. The action of the covector  $\omega$  on the vector  $b$ .

See Figure 1.7. This result is not given to be useful, but only to check your understanding of the idea of defining operations on equivalence classes.

### Covectors

For any vector space, the linear operators that map vectors to real numbers are important. They form a vector space themselves; it is of the same

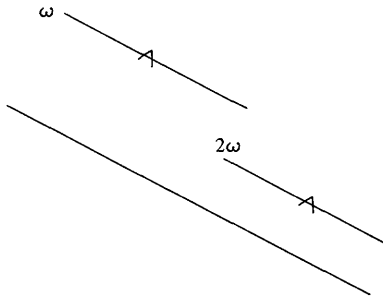


Figure 1.9. The scaling of covectors.

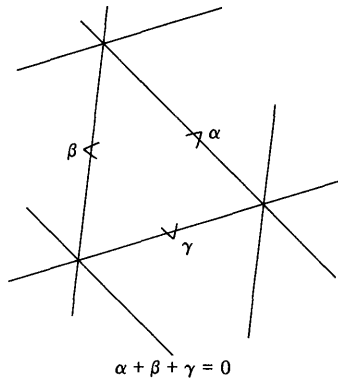


Figure 1.10. The addition of covectors. The three shown add up to zero.

dimension as the vector space, and is called the *dual space*. These linear operators are called *covectors*. A covector can be represented by parallel hypersurfaces. For a given covector  $\omega: V \rightarrow \mathbb{R}$ , the set,  $\hat{\omega}$ , of vectors such that  $\omega \cdot v = 1$ ,

$$\hat{\omega} = \{v \in V \mid \omega \cdot v = 1\},$$

form a hyperplane in the vector space, and provide a faithful representation for  $\omega$  (Figure 1.8). Scaling and addition are easily performed in this representation (Figures 1.9 and 1.10). The space dual to  $V$  is written  $V^*$ .

The unit-contour  $\hat{\omega}$  alone is a useful representation for a *bound covector*. A *free covector*, like a free vector, can roam all over the vector space. It requires a pair of contour lines for its representation, and the equivalent of an arrowhead to flag the “uphill” contour.

**Example:** In Figure 1.10 we drew some free covectors.

A blindness caused by the unnecessary use of metrics afflicts many physicists; so they do not distinguish between vectors and covectors. It does take practice to develop an eye that sees relations in a metric-free fashion. An intuition that is covariant under general linear transformations is an extremely valuable tool, and I encourage you to develop one.

**Example:** One way to appreciate this duality structure is to examine a case where it breaks down. Let  $V$  be the space of continuous functions on the closed interval  $[0, 1]$ . Some elements of  $V^*$  can be written as integrals. For any  $f \in V$  we have the linear operator

$$\Omega_g : f \mapsto \int_0^1 g(x)f(x) dx.$$

There is a natural map  $V \rightarrow V^*$  given by

$$g \mapsto \int_0^1 g(x)(\cdot) dx, \quad (1.2)$$

where the centered dot in parentheses indicates a missing argument. Not all elements of  $V^*$  can be found this way. For example, the map  $\delta_{x_0}$

$$\delta_{x_0} : f \mapsto f(x_0),$$

is a linear operator, and so it must be in  $V^*$ . It cannot be formed as in equation (1.1), however.

We can try to add further functions to  $V$  to restore the symmetry between  $V$  and  $V^*$ , such as the functions called  $\delta$ -functions. Unfortunately, the product of  $\delta(x)$  with  $\delta(x)$  cannot be defined, and this fails.

The correct way to introduce  $\delta$ -functions is to envision them as the results of a limit process that produces a class of sharply peaked functions. This entire class is called (unfortunately) *a*  $\delta$ -function, or *a* generalized function. A generalized function is thus an equivalence class of functions so sharply peaked that it is irrelevant how peaked it is. The product of a single generalized function and an ordinary function is well-defined, and is independent of which element of the class of sharply peaked functions is chosen. Not so for products of generalized functions. This is why  $\delta^2(x)$  really cannot be defined.

The linear structure of a vector space makes calculus easy. A parametrized curve in a vector space  $V$  is a map  $\gamma: R \rightarrow V; s \mapsto \gamma(s)$ . Suppose the vectors are represented by their components in some given basis:

$$v = v^a e_a = \mathbf{v}e.$$

Here the  $v^a$  are numbers, but  $v$  and the  $e_a$  are vectors. Often the  $v^a$  can be thought of loosely as “the vector.” The third expression is the compact “sum over the obviously missing indices” notation. A map  $\gamma$  will be represented by  $n$  functions of one variable if we write

$$\gamma(s) = \gamma^a(s) e_a = \gamma(s) \mathbf{e}.$$

The tangent to  $\gamma$  at  $s$  is the vector  $\dot{\gamma}(s)$  given by the usual expression

$$\dot{\gamma}(s) = \lim_{\epsilon \rightarrow 0} [\gamma(s + \epsilon) - \gamma(s)] / \epsilon.$$

This definition is equivalent to

$$\dot{\gamma}(s) = \frac{d\gamma}{ds}(s) e_a = \frac{d\gamma}{ds}(s) \mathbf{e}.$$

The preceding argument adds vectors that are defined at different points. This cannot be done on all smooth sets (try making sense out of adding a vector at the North Pole to one on the equator). This is what forces us to give the involved definitions of a tangent vector that we will give in Chapter II.

I will often give the basis vectors in a linear space mnemonically convenient names such as  $e_x$ ,  $e_y$ , and so on. The basis covectors will have dual names, such as  $f^x$ ,  $f^y$ , and similarly. The contractions are

$$f^x \cdot e_x = 1, \quad f^y \cdot e_x = 0,$$

and so on.

### Problem

- 1.1 Show how the two-dimensional picture of covectors and covector addition is a section through the general picture in higher dimensions.

## 2. Differential calculus

Our aim in studying manifolds will be to extend the ideas of calculus to sets more general than the real numbers. To prepare for that, let us review ordinary multivariable calculus with an eye to streamlining the

notation and highlighting the concepts. The main aim of this section is to give you practice with the style and notation to be used in this book.

### *Differential*

The two important ideas of calculus are (1) the local linear approximation to a function, and (2) the notion of a constant function. These are both seen in the single-variable Taylor's Series expression

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots.$$

Suppose that  $E$  and  $F$  are two vector spaces, not necessarily of the same dimension, and  $f: E \rightarrow F; u \mapsto f(u)$ . The correct generalization of the Taylor's Series is

$$f(u+h) = f(u) + Df(u) \cdot (h) + \cdots.$$

Here  $h$  is a vector in  $E$ , and  $Df(u)$  is a linear operator depending on the point  $u$ ,  $Df(u): E \rightarrow F$ . Recall our convention of using the centered dot for the evaluation of linear operators. For the right-hand side to be a local linear approximation to  $f$ ,  $Df(u)$  must be the unique linear operator satisfying

$$\lim_{h \rightarrow 0} \frac{\|f(u+h) - f(u) - Df(u) \cdot (h)\|}{\|h\|} = 0.$$

Here  $\|\cdot\|$  denotes any convenient norm on the vector spaces  $E$  and  $F$ . The linear operator  $Df(u)$  is called the *differential* of  $f$ . In spaces of finite dimension, this differential is independent of the norm.

**Examples:** What is the differential of

$$f: \mathbb{R} \rightarrow \mathbb{R}; x \mapsto x^3$$

at  $x=1$ ? Near  $x=1$  we have

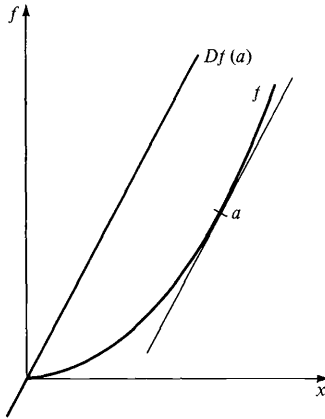
$$f(x) = 1 + 3(x-1) + \cdots.$$

Here the three dots denote second-order terms that vanish in the limit. The differential at  $x=1$  is the linear map

$$Df(1): x \mapsto 3x.$$

In one dimension any such linear map can be represented by a number, here 3, which is called the derivative (Figure 2.1). The correctness of the preceding differential is shown by the fact that

$$\lim_{x \rightarrow 1} \frac{|x^3 - 1 - 3x + 3|}{|x - 1|} = 0.$$



**Figure 2.1.** The differential of the function  $f$  is a linear map, which can be represented by a number, the derivative.

Let  $f$  be a map from  $\mathbb{R}^m \rightarrow \mathbb{R}^n$ ; use Roman indices for coordinates in  $\mathbb{R}^m$ , Greek for those in  $\mathbb{R}^n$ . Thus  $f$  can be represented by the  $m$  functions  $f^\mu(x^a)$ . We are following an index placement that will ultimately make good sense. Near  $x^a = 0$  we have

$$f^\mu(x^a) = f^\mu(0) + \frac{\partial f^\mu}{\partial x^a}(0)x^a,$$

and so the differential is the linear map

$$(x^1, x^2, \dots, x^m) \mapsto \left( \frac{\partial f^1}{\partial x^1}x^1 + \frac{\partial f^1}{\partial x^2}x^2 + \dots, \frac{\partial f^2}{\partial x^1}x^1 + \dots, \dots \right),$$

which I abbreviate

$$x^a \mapsto \frac{\partial f^\mu}{\partial x^a}x^a = \frac{\partial f^\mu}{\partial x}x.$$

The differential is the linear operator whose matrix representation is the matrix of partial derivatives.

Let  $M^n$  be the linear space consisting of all  $n \times n$  matrices. Look at the function  $F$ :

$$F: M^n \rightarrow M^n; A \mapsto AA^t - I.$$

Here  $A^t$  is the transpose of the matrix  $A$ , and  $I$  is the identity matrix. If  $F(A) = 0$ , then  $A$  is an orthogonal matrix. What is the differential of  $F$ ? For any matrix  $A$  we have



$$\begin{aligned}
 F(A+h) &= (A+h)(A+h)^t - I, \\
 &= AA^t - I + Ah^t + hA^t + \dots \\
 &= Ah^t + hA^t.
 \end{aligned}$$

Thus the differential is given by the linear map

$$DF(A): h \mapsto Ah^t + hA^t.$$

Compare this simple explicit statement with the array of second derivatives. We will use this particular differential in Section 11.

### Chain rule

The chain rule is a useful theorem for the differential of the composition of two maps. For vector spaces  $E$ ,  $F$ , and  $G$ , and maps  $f: E \rightarrow F$ , and  $g: F \rightarrow G$ , then the composite function  $g \circ f$  is a map

$$g \circ f: E \rightarrow G; e \mapsto g[f(e)],$$

whose differential is given by

$$D[g \circ f](e) \cdot h = Dg[f(e)] \cdot [Df(e) \cdot h].$$

The bones of the equation, leaving out the obvious arguments, are simply

$$D(g \circ f) = Dg \cdot Df.$$

### Second differential

Higher differentials can be defined similarly. Second differentials  $D^2f$  make the Taylor's Series

$$f(u+h) = f(u) + Df(u) \cdot h + D^2f(u) \cdot (h, h) + \dots$$

valid. Here  $D^2f(u)$  is a map  $E \times E \rightarrow \mathbb{R}$  that is linear in each argument. Such maps are called *bilinear*. Note in the preceding that both  $Df$  and  $D^2f$  are really "single" symbols. We are not thinking of these as the result of an operator "D" applied once or twice to  $f$ .

**Example:** For  $f(x) = x^2$ , we have, at  $x=0$ ,

$$f(x) = f(0) + Df(0) \cdot x + D^2f(0) \cdot (x, x) = x^2,$$

and so the first differential at zero is zero, and the second differential is the bilinear map

$$(x, y) \mapsto xy.$$

Any such bilinear map from  $\mathbb{R}^2 \rightarrow \mathbb{R}$  can again be represented by a single number, and in one-variable calculus this number is called the second derivative.

### *Extrema*

Many physically interesting problems can be formulated as extremum problems. For example, the static equilibrium configurations of a mechanical structure are extrema of its potential energy, and the action integral is an extremum in wave problems. Finding such extrema is a nice exercise in differential calculus. Here we look only at finite dimensional spaces.

The most straightforward problem is to have a map  $f: E \rightarrow \mathbb{R}; x \mapsto f(x)$ , and to seek its *critical points*,  $x_c$ , which satisfy

$$Df(x_c) = 0.$$

This equation provides us with  $n$  equations for the  $n$  coordinates of  $x_c$ . These are necessary conditions for an extremum. In most physical problems it is sufficient to find the critical points rather than true extrema.

A more interesting procedure is to look for *critical points* of  $f$  restricted to a subspace of  $E$ . Look at the problem in which a condition

$$g(x_c) = 0,$$

where  $g: E \rightarrow \mathbb{R}^k$ , is specified as well. We demand that  $Dg$  be a non-singular linear map at the zero value; otherwise the subspace can be pathological.

**Examples:** The case  $E = \mathbb{R}^2$  is simple enough to solve by inspection. Look at Figure 2.2. We are constrained to move only along the curve  $g(x, y) = 0$ . A point  $(x_c, y_c)$  is a critical point of  $f$  if the change in  $f(x, y)$  away from  $(x_c, y_c)$  along the  $g = 0$  curve has no first-order terms.

The geometric interpretation of this is that the contour lines of  $f$  must be parallel to the curve  $g = 0$  at the critical point. Translated into the language of differentials, this states that there must be a constant  $\lambda$ , called a *Lagrange multiplier*, such that

$$Df(x_c, y_c) = \lambda Dg(x_c, y_c). \tag{2.1}$$

Together with the constraint equation

$$g(x_c, y_c) = 0,$$

this gives three equations for the three unknowns  $x_c$ ,  $y_c$ , and  $\lambda$ .

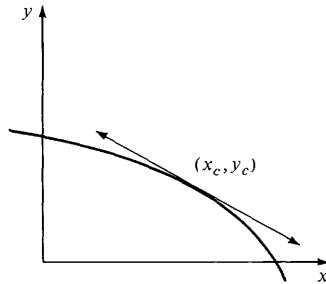


Figure 2.2. A problem of constrained extrema.

Find the extrema of the function

$$f(x, y) = y,$$

on the circle

$$g(x, y) = x^2 + y^2 - 1 = 0.$$

We compute the differentials:

$$Df(x, y) \cdot (u, v) = \frac{\partial f}{\partial x} u + \frac{\partial f}{\partial y} v = v,$$

$$Dg(x, y) \cdot (u, v) = 2xu + 2yv.$$

Condition 2.1 yields the two equations

$$2x_c = 0, \quad 2y_c \lambda = 1,$$

and together with the constraint equation these give us the two critical points

$$(x_c, y_c, \lambda) = (0, 1, 2), (0, -1, -2).$$

See Figure 2.3.

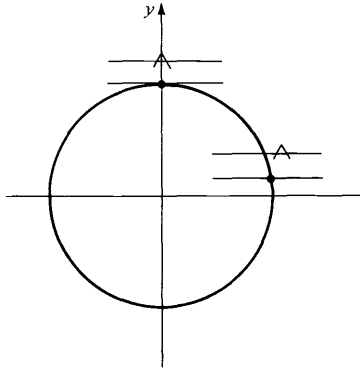
The general case is covered by the following theorem. If  $u_c$  is a critical point of the map  $f: E \rightarrow \mathbb{R}$ , restricted by the  $k$  conditions

$$g(u_c) = 0,$$

where

$$g: E \rightarrow \mathbb{R}^k,$$

then there exists a linear map  $\lambda: \mathbb{R}^k \rightarrow \mathbb{R}$ , such that



**Figure 2.3.** One critical point and one ordinary point for the example.

$$D[f + \lambda \cdot g] \cdot (u_c) = 0,$$

provided that  $Dg(u_c)$  is nonsingular. The linear map  $\lambda$  gives us  $k$  Lagrange multipliers, and the map  $g$  gives us  $k$  additional constraint equations. See Abraham and Marsden (1978) for a proof. Similar Lagrange multiplier results will come up when we study the calculus of variations.

### 3. Tensor algebra

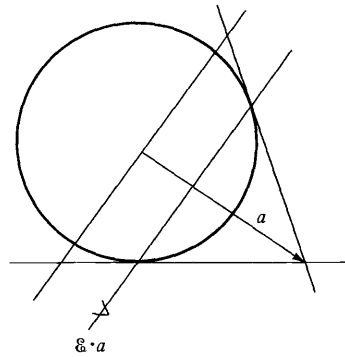
Starting from any given linear vector space, we can construct an algebra of multilinear operators called the *tensor algebra*. The multiplication in this algebra can be specified by a few simple rules. Rather than pull these rules out of the air, let me deduce them first for bilinear operators.

#### *Bilinear operators*

The linear operators on a vector space that map vectors into numbers are called *covectors*. When in particular the vector space is the space of tangent vectors, these dual vectors are called *1-forms*. I will usually use early-alphabet, lower-case Roman letters for vectors, and late-alphabet Greek letters for covectors. A reasonable question to ask is, given the covectors, can we construct bilinear operators from them? A bilinear operator, say,  $\Omega$ , is one that acts on pairs of vectors according to

$$\Omega \cdot (a + b, c) = \Omega \cdot (a, c) + \Omega \cdot (b, c),$$

$$\Omega \cdot (a, b + c) = \Omega \cdot (a, b) + \Omega \cdot (a, c),$$



**Figure 3.1.** The action of the Euclidean metric tensor on vectors lying outside the unit circle. Use scaling if the vector originally lies inside, or reverse the construction, which is just inversion.

and

$$\Omega \cdot (ka, b) = \Omega \cdot (a, kb) = k\Omega \cdot (a, b),$$

where  $k$  is a real number.

A natural way to form a bilinear operator from two given covectors is to let each one act on one of the vectors. Denote the combined operator by  $\omega \otimes \nu$ , and define it by

$$\omega \otimes \nu \cdot (a, b) = (\omega \cdot a)(\nu \cdot b).$$

The multiplication on the right is the ordinary multiplication of numbers.

Objects such as  $\omega \otimes \nu$  are called tensors. The operator  $\otimes : (\omega, \nu) \mapsto \omega \otimes \nu$  is called the tensor product, and the algebra generated by this product is called the tensor algebra. From the preceding equation we see that it has the properties

$$(k\omega) \otimes \nu = \omega \otimes (k\nu) = k(\omega \otimes \nu),$$

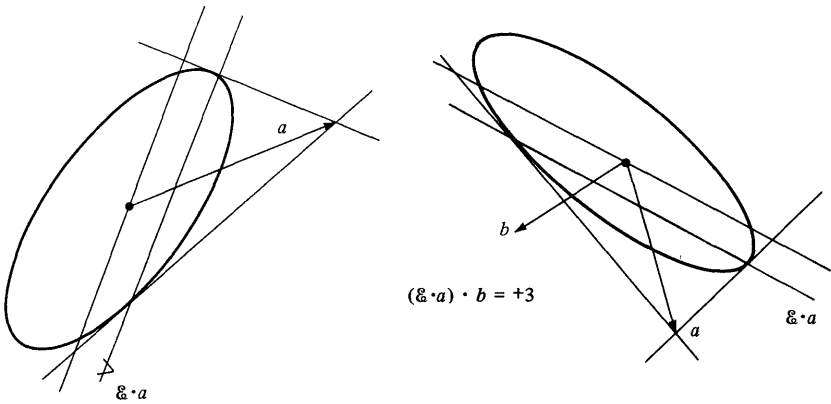
$$(\omega + \sigma) \otimes \nu = \omega \otimes \nu + \sigma \otimes \nu,$$

$$\omega \otimes (\sigma + \nu) = \omega \otimes \sigma + \omega \otimes \nu.$$

Contrast this with the Cartesian product on  $\mathbb{R}^2$ , which is usually given an algebraic structure such that

$$k(x, y) = (kx, ky).$$

Every bilinear operator can be written as a sum of terms such as  $\omega \otimes \nu$ . The single terms are called *monomials*.



**Figure 3.2** (left). The action of the Euclidean metric tensor in general linear coordinates rather than the Cartesian coordinates of Figure 3.1.

**Figure 3.3** (right). The action of the Euclidean metric on pairs of vectors.

**Example:** Let  $e_x, e_y,$  and  $e_z$  be a basis for the Euclidean vectors in 3-space, and let  $f^x, f^y,$  and  $f^z$  be the dual basis. The Euclidean metric is generated by the bilinear operator

$$\mathcal{E} = f^x \otimes f^x + f^y \otimes f^y + f^z \otimes f^z.$$

The action of this tensor as a linear map: vectors  $\rightarrow$  covectors is shown in Figure 3.1. In general coordinates the action is represented by an ellipse (Figure 3.2). The related action: vector  $\times$  vector  $\rightarrow$  number is shown in Figure 3.3.

To see that every bilinear operator can be written as a sum of monomials, just verify that the expansion

$$\Omega = \Omega_{ab} f^a \otimes f^b$$

has the desired action, where the numerical coefficients  $\Omega_{ab}$  are given by

$$\Omega_{ab} = \Omega \cdot (e_a, e_b).$$

The space of these bilinear operators forms a vector space written  $V^* \otimes V^*$ . The elements are called tensors of type  $\binom{0}{2}$ . Tensors of types  $\binom{2}{0}$  and  $\binom{1}{1}$  are also bilinear operators. The former act on pairs of covectors, and the latter on a mixed pair of one vector and one covector.

The mixed tensor product is familiar to physicists as the outer product of row and column vectors; in bra-ket notation it is written

$$\otimes : (|a\rangle, \langle\omega|) \mapsto |a\rangle\langle\omega|.$$

### Tensor product

A tensor multiplication,  $\otimes$ , can be defined on all the tensor spaces by using the rules so far deduced. The multiplication will be associative,

$$a \otimes (b \otimes c) = (a \otimes b) \otimes c = a \otimes b \otimes c,$$

will commute with real numbers,

$$ka \otimes b = a \otimes kb = k(a \otimes b),$$

and will distribute over addition,

$$a \otimes (b + c) = a \otimes b + a \otimes c,$$

where  $a$ ,  $b$ , and  $c$  are tensors, and  $k$  is a real number. This tensor product can easily be extended from sets of basis vectors and covectors to tensors in general. We can multiply tensors of type  $\binom{p}{q}$  with tensors of type  $\binom{r}{s}$ . The result is a tensor of type  $\binom{p+r}{q+s}$ .

From a given set of basis vectors, we can form a basis for the tensor spaces by using all possible tensor products of the correct type. The coefficients in an expansion in terms of these basis tensors are called the *components* of the tensor. These components are often indexed by using superscripts for vector terms and subscripts for covector terms:

$$F = F^{\mu\nu}{}_{\alpha\beta} e_\mu \otimes e_\nu \otimes f^\alpha \otimes f^\beta,$$

where  $f^\alpha$  is the covector dual to  $e_\alpha$ . Notice how the index placement tells us what type of tensor is being represented. Notice also that we are being much *less* abstract than is the usual practice. Usually a tensor is defined to be a quantity with more indices than a vector, transforming in some special manner. To us each tensor is a specific multilinear operator. This multilinearity gives tensors symmetries that differ from the symmetries of vectors. Real numbers act on vectors and tensors by using the scaling operation of a vector space, and they are tensors of type  $\binom{0}{0}$ , often called *scalars*.

### Contraction

If we are given a tensor space of type  $\binom{p}{q}$ , there are natural maps into the tensors of type  $\binom{p-1}{q-1}$ , tensors having one less vector factor and one less covector factor. These natural maps are called *contractions*. They are called natural because every tensor space admits them. To contract, we use term-by-term the natural operation of a covector on a vector.

**Example:** In  $\mathbb{R}^3$  the tensor

$$T = e_x \otimes e_y \otimes f^x + e_y \otimes e_z \otimes f^x + e_x \otimes e_x \otimes f^y$$

is of type  $\binom{2}{1}$ . There are two different ways to contract this tensor. Each results in a vector. If we contract on the first and third factors, we have

$$(f^x \cdot e_x)e_y + (f^x \cdot e_y)e_z + (f^y \cdot e_x)e_x = e_y.$$

If we contract on the second and third, we find zero. It cannot be contracted on the first and second, because these spaces are not dual.

Contractions of complicated tensors require some notation to indicate which terms are being contracted. The most convenient notation for this is the index notation. [See the remarks in Burke (1980), Penrose (1968), and Schouten (1954) on the schizoid view that physicists have of the index notation.]

**Example:** For our tensor  $T$  given in the preceding example, the first contraction would be written  $T_a^{ab}$  and the second  $T_a^{ba}$ .

The one-up, one-down rule in the summation convention was imposed to force us to do the contraction correctly. For a contraction to be well-defined, it must commute with the associative and distributive laws of the tensor product, which is easy to show.

#### 4. Alternating products

Bilinear operators have *symmetric* and *antisymmetric* parts. The parts behave quite differently, and many of the tensors of physical interest are purely of one type or the other. Metric tensors are symmetric. The densities discussed in the Introduction are totally antisymmetric tensors, also called *skew* or *alternating*. We discuss here the geometric representations of some alternating tensors. These geometric representations are quite useful, and will be used throughout this book.

The alternating product of two vectors, denoted by  $\wedge$ , is defined to be

$$a \wedge b = a \otimes b - b \otimes a.$$

We can easily extend  $\wedge$  to triple and higher products. For example,

$$a \wedge b \wedge c$$

$$= (a \otimes b \otimes c - a \otimes c \otimes b - b \otimes a \otimes c + b \otimes c \otimes a + c \otimes a \otimes b - c \otimes b \otimes a).$$





**Figure 4.1** (left). The action of a 2-form  $\Omega$  on the vector  $b$  yields a 1-form  $\Omega \cdot b$ . The data in this figure completely specifies  $\Omega$  in two dimensions.

**Figure 4.2** (right). A properly scaled vector  $c$ .

If  $a$  and  $b$  are vectors, then  $a \wedge b$  is called a *bivector*. Bivectors have many uses, and can represent, for example, pieces of a 2-surface. We can also take the alternating product of two 1-forms

$$\omega \wedge \nu = \omega \otimes \nu - \nu \otimes \omega,$$

and these are called *2-forms*. They will be used, for example, to represent the electromagnetic field in spacetime. The wedge product is similar to the vector cross product, but is associative.

**Example:** Let us find a geometric representation for a 2-form  $\Omega$  in two dimensions. Such a  $\binom{0}{2}$  tensor is a map  $\Omega: V \times V \rightarrow \mathbb{R}$ ;  $(b, c) \mapsto \Omega \cdot (b, c)$ . By partial evaluation it is also a map  $\Omega: V \rightarrow V^*$ ;  $b \mapsto \Omega \cdot b = \Omega \cdot (b, \cdot)$ , and this latter action is the one we will represent. Note the careful and consistent use of the centered dot here for the action of linear operators. To start, pick any vector  $b$ , and ask  $\Omega$  for  $\Omega \cdot b$ , which is a 1-form. Since  $\Omega$  is alternating,

$$\Omega \cdot (b, b) = 0,$$

and the zero contour of  $\Omega \cdot b$  must lie along  $b$  (Figure 4.1). Now take any other vector  $c$ . Can we find  $\Omega \cdot c$  if we know only what is given in Figure 4.1? Yes. Because of linearity we can scale  $c$  so that

$$(\Omega \cdot b) \cdot c = 1,$$

as shown in Figure 4.2. But since  $\Omega$  is alternating, this says that

$$(\Omega \cdot c) \cdot b = -1$$

and together with

$$(\Omega \cdot c) \cdot c = 0,$$

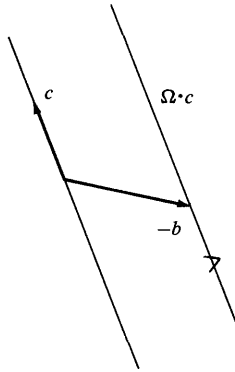


Figure 4.3. Finding  $\Omega$  from the data in Figure 4.1.

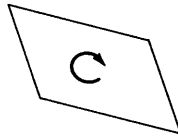


Figure 4.4. The representation of a 2-form in two dimensions.

which is true for any skew tensor, we must have  $\Omega \cdot c$ , as shown in Figure 4.3. Thus any figure like 4.1 or 4.3 is a representation of  $\Omega$ . The common feature is the area shown in Figure 4.4. The arrow indicates the orientation, and tells whether the unit contour of  $\Omega \cdot c$  goes to the right or to the left.

Unlike that of symmetric tensors, the representation of 2-forms depends strongly on dimension, because all 2-forms in two and three dimensions are algebraically special. In two dimensions, every 2-form satisfies

$$\omega \wedge \omega = 0.$$

Normal forms for 2-forms will be covered in Section 26.

**Example:** Let us use the same methods to find a representation of a 2-form in three dimensions. Pick first a vector  $b$ , and ask for  $\Omega \cdot b$ . If it is zero, try again. Then pick another vector  $c$  such that

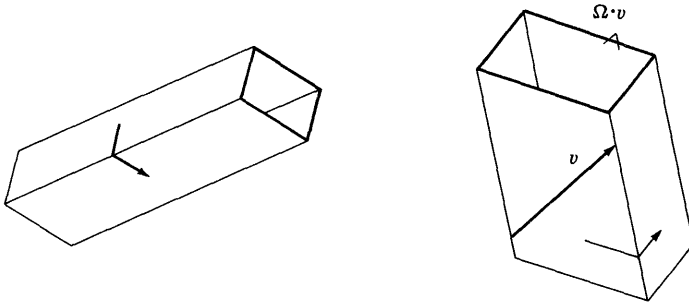


Figure 4.5 (left). A 2-form in three dimensions.

Figure 4.6 (right). The action of a 2-form, represented as in Figure 4.5. Acting on a vector  $d$ , it yields the 1-form  $\Omega \cdot d$ .

$$(\Omega \cdot b) \cdot c = 1.$$

Now pick a vector  $d$  such that

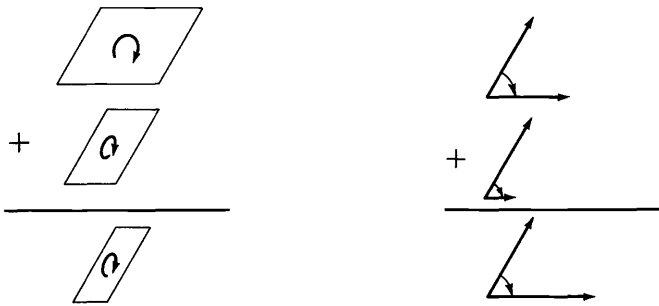
$$(\Omega \cdot b) \cdot d = 0, \quad (\Omega \cdot c) \cdot d = 0.$$

This vector  $d$  points along the “eggcrate” formed by  $\Omega \cdot b$  and  $\Omega \cdot c$ . Now,  $d$  cannot be a linear combination of  $b$  and  $c$ , and so  $b$ ,  $c$ , and  $d$  must form a basis. We have  $\Omega \cdot d = 0$ . Thus we can find  $\Omega \cdot v$  for any vector  $v$ .

A representation based on this uses an oriented prism, as shown in Figure 4.5. To find  $\Omega \cdot v$  for any vector  $v$ , deform the prism, keeping its cross-sectional area constant, until  $v$  lies along one of the sides, as shown in Figure 4.6. Then the opposite side of the prism will be the unit contour of  $\Omega \cdot v$ . This 2-form is similar to the density given in the Introduction, differing only in the type of orientation it has been given. The one in the Introduction is called a twisted 2-form.

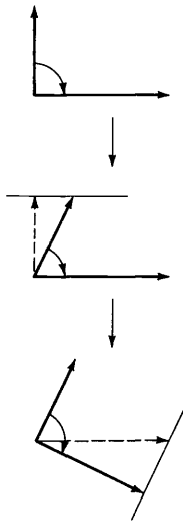
The addition rules in these representations are easy to deduce.

**Example:** The addition of 2-forms and bivectors in two dimensions is sketched in Figures 4.7 and 4.8. A bivector is represented by an ordered pair of vectors, and is not changed by rotations that preserve the ordering of the vectors and the area spanned by the pair, as shown in Figure 4.9.



**Figure 4.7** (left). The addition of 2-forms in two dimensions.

**Figure 4.8** (right). The addition of bivectors in two dimensions. Note the complementary behavior in Figures 4.7 and 4.8.



**Figure 4.9.** Three equivalent representations of the same bivector.

The representations given so far are not typical. Only in four or more dimensions does the full structure of 2-forms and bivectors appear. The general representation of bivectors in  $2n$  dimensions involves  $n$  pairs of vectors, all linearly independent. In four dimensions

$$B = e_x \wedge e_y + e_z \wedge e_t$$

is a typical bivector. A bivector with only one term is called *simple*. In two and three dimensions all bivectors are simple.

Linear transformations  $V \rightarrow V$  are represented by tensor-product terms of the type  $V^* \otimes V$ . Such terms cannot be antisymmetrized; it makes no sense to write

$$e_x \otimes f^y - f^y \otimes e_x.$$

If there is a metric on the space, the metric can be used to convert covectors into vectors, and you can then antisymmetrize the result. An orthogonal rotation is a linear transformation that preserves a metric inner product; whenever we talk about orthogonal rotations there is a metric available, and we can use the preceding trick. Infinitesimal orthogonal rotations can be represented by bivectors. A rotation in the Euclidean plane is represented in Cartesian coordinates by the bivector (unique up to scale)

$$\Omega = e_x \wedge e_y.$$

With a metric the operation  $\Omega \cdot v$  can be defined by

$$\Omega \cdot v = \Omega(\mathfrak{G}(v), \cdot).$$

This gives you the velocity of the tip of the vector  $v$  under a rotation whose rate is  $\Omega$ . This metric inner product, which we will use only rarely, will be represented by a boldfaced centered dot.

**Examples:** The bivector

$$\Omega = e_x \wedge e_t$$

represents a Lorentz transformation. We have

$$\Omega \cdot e_x = (e_x \cdot e_x) e_t = e_t,$$

$$\Omega \cdot e_t = -(e_t \cdot e_t) e_x = e_x.$$

I am using a Lorentz metric of signature  $-+++$ . You can see here how the Lorentz metric straightens out the sign. Using the metric, the electrodynamic 2-form can be turned into a bivector. In this form it represents the rate at which charges are accelerated by the field, an infinitesimal Lorentz transformation.

Consider a situation like well-combed, very fine hair. This can be described in a fluid approximation in which the individual hairs are

ignored and only their mean density and alignment is kept. In  $\mathbb{R}^3$  this density can be represented by a 2-form. Using the graphical representation given in Figures 1 and 2 in the Introduction, align the prism along the lines, and choose the cross-sectional area to include (on the average) one line.

### Problem

- 4.1 Extend the last example of this section to lines in spacetime, the worldlines of particles. Find the 3-form describing their density.

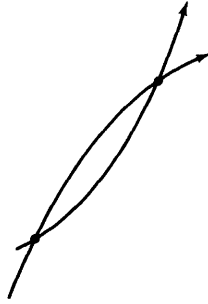
## 5. Special relativity

Special relativity provides a nice example of the use of geometric structure to model physical reality. The geometric structure here was invented by Einstein in one fell swoop, and special relativity is often presented this way. It is more instructive to build up this structure layer by layer, each layer having its own mathematical structure.

Be careful of the naive view that a physical law is a mathematical relation between previously defined quantities. The situation is, rather, that a certain mathematical structure represents a given physical structure. Thus Newtonian mechanics does not assert that  $F = ma$ , with  $F$ ,  $m$ , and  $a$  separately defined. Rather, it asserts that the structure of second-order differential equations applies to the motion of masses. A falsification of Newtonian mechanics would entail showing, for example, that position and velocity were not sufficient to predict the future motion (radiation reaction), or that position alone was sufficient (Aristotelian mechanics). If you miss this point, then our laws here will seem circular.

### *Topological structure*

We start with the broadest structure useful in classical physics: *continuity*. We assert that a continuous map  $\psi: (\text{events}) \rightarrow \mathbb{R}^4$  exists. An *event* is a primitive notion, corresponding to a particular location at some particular time. Events lie in the physical world, and the map  $\psi$  is an operational procedure using physical apparatus. The numbers  $\mathbb{R}^4$  are called *coordinates*. The notion of nearby events must also be taken as a primitive notion. The map  $\psi$  is continuous if nearby events have nearby coordinates. A *particle* is something described by a continuous line of events, called its *worldline*.



**Figure 5.1.** An impossible situation for free particles with a projective structure.

The assertion that such a map  $\psi$  exists is an important physical statement. That the mathematical structure fits the world says that the world is four-dimensional. What appears to be only a definition is often more than that.

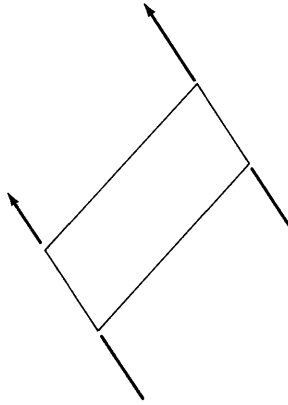
### *Projective structure*

The world has more structure than the topology of events and particles. Among the particles there is a subclass called *free particles*. The structure of these free particles is given by the assertion that in the class of continuous maps  $\psi$  we can find maps such that the worldlines of the free particles are straight lines in  $\mathbb{R}^4$ . This structure is called a *projective structure*. Note that both the straight lines and the free particles are defined by this assertion; nor does there need to be only one projective structure. We assert only that at least one exists.

**Examples:** Not every set of worldlines can be taken to be free particles. No map can straighten out two worldlines if they ever appear as in Figure 5.1.

Suppose that the universe of physics consisted of only two classes of particles, neutral particles and electrons, in a region of space filled with a uniform electric field. Then either the neutrals or the electrons (if far apart from each other) could be taken as free particles.

Coordinates for which the free particles go in straight lines are not unique. Given one such  $\psi$ , we can modify it by any transformation  $\mathbb{R}^4 \rightarrow \mathbb{R}^4$  that preserves the projective structure.



**Figure 5.2.** Affine structure used to compare clock intervals in spacetime.

### *Conformal structure*

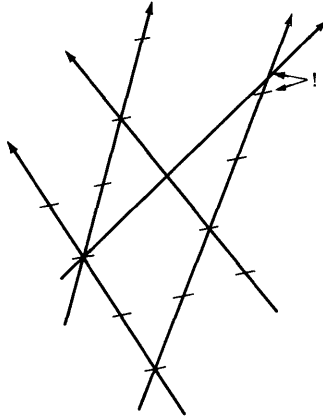
An additional primitive notion is the idea of a special class of worldlines called *light signals*. The structure here is that of the possible directions of light-signal worldlines in spacetime, the light cone. The light sent out in a specified spatial direction travels with a unique speed. The special class of directions gives spacetime what is called a *conformal structure*. I will draw light signals in spacetime diagrams with dashed lines.

### *Affine structure*

There is yet more structure in the world. A clock is another primitive notion. It is an operation that assigns numbers to intervals along a worldline. These intervals are called time intervals or, sometimes, proper-time intervals. These clocks are to have an intrinsic rate. It might be that of an ammonia molecule flipping back and forth, or of an atomic vibration, or one given statistically by the decay of many elementary particles.

The affine structure of these clocks is contained in two assertions. One is *universality*: All clocks are essentially the same. The relative rate between any two clocks carried along the same worldline does not depend on where or when this is done, but only on the types of the clocks. The other is *uniformity*. Among the projective coordinates we can find special coordinates such that the natural affine structure of  $\mathbb{R}^4$  is compatible with the clock readings. Clocks carried along opposite sides of a parallelogram must read the same time interval (see Figure 5.2). Also, the clock





**Figure 5.3.** Ticking objects that cannot be given an affine structure.

ticks must subdivide intervals in the same way that the affine structure does. Notice that we assert that a set of clocks tick uniformly, not by comparing them with some other clocks, but by comparing them among themselves in sets of four. Figure 5.3 shows a hypothetical set of ticking objects that cannot be defined to be uniform clocks, showing that this is a nontrivial assertion.

**Example:** Using only the linear and conformal structures, we can define the relative velocity between two free-particle worldlines. The time intervals shown in Figure 5.4 are to be measured, and the relative velocity  $v$  is then defined to be

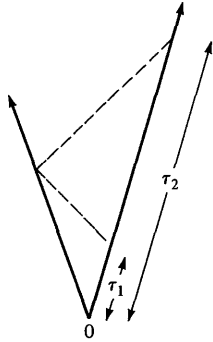
$$v = \frac{\tau_2 - \tau_1}{\tau_2 + \tau_1}.$$

This construction compares time intervals along only a single worldline, and this ratio depends only on the affine structure of the clocks and the affine structure of light signals.

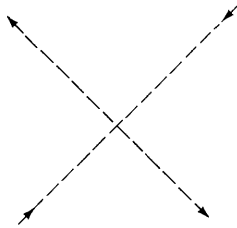
Any representation of free particles and clocks in  $\mathbb{R}^4$  that is compatible with the usual affine structure of  $\mathbb{R}^4$  is called an *inertial reference frame*.

#### *Canonical reference frames*

By combining the affine and conformal structures, an observer can select a map  $\psi$  that is unique up to rotation and overall size. For simplicity we



**Figure 5.4.** The relative velocity between two worldlines defined by using two light signals, shown as dashed lines, and an affine structure that measures the time intervals  $\tau_1$  and  $\tau_2$ . We need only their ratio.

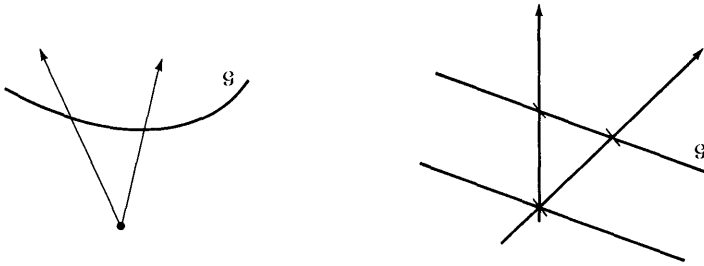


**Figure 5.5.** The action of a Lorentz transformation.

describe this in just two dimensions. Start with maps  $\psi$  compatible with the affine structure. Among these select those in which the light signals move along lines of slope plus and minus one. The remaining freedom consists of expansions and contractions along these  $45^\circ$  axes. Using coordinates along these axes, we find that these transformations are given by

$$(u, v) \mapsto (\alpha u, v/\alpha),$$

as shown in Figure 5.5. These transformations are called *Lorentz transformations*. Note that we are using Lorentz transformations *before* special relativity has even appeared. Even the absolute-time clocks of Newtonian mechanics fit the framework so far set up. Lorentz transformations depend on only the affine and conformal structures. By choosing the appropriate Lorentz transformation, any observer can make his or her worldline vertical. The representation is now unique except for a



**Figure 5.6** (left). The general pattern for describing the behavior of a set of clocks that satisfy our assumptions.

**Figure 5.7** (right). The behavior of clocks compatible with the absolute time of Newton.

possible overall magnification. This representation I call a *canonical reference frame for that observer*. A canonical reference frame provides a useful framework or language in which to describe further physical observations.

### *Special-relativity clocks*

Using a canonical reference frame as defined in the preceding, an observer can now study the detailed behavior of clocks. The observation is to measure unit time intervals along worldlines of different slopes. Such observations can be summarized by using the affine structure to transfer all these unit time intervals so that they start at the origin. The behavior of clocks that satisfy our assumptions is completely contained in the set of points  $\mathcal{G}$  that are a unit time interval away from the origin (Figure 5.6).

Note carefully that a measurement of the set  $\mathcal{G}$  by any one observer is a complete theory of clocks. The transformation from one observer's canonical reference frame to another's is a Lorentz transformation, and is determined solely by the affine and conformal structures of space-time.

When clocks are studied in the real world, the set  $\mathcal{G}$  is adequately described by the hyperbola

$$t^2 - x^2 = 1,$$

or in four dimensions by the hyperboloid

$$t^2 - x^2 - y^2 - z^2 = 1.$$

We are using the second as our unit of time, and the light-second as our unit of length. A marvelous feature of this clock behavior is that it is unchanged by Lorentz transformations. Every observer sees the same clock behavior. This is called Lorentz invariance. Note that the formalism can represent clocks that are not Lorentz-invariant.

**Example:** The absolute-time clocks of Newton are represented by a set  $\mathcal{G}$  such as is shown in Figure 5.7.

An interesting unification is that the asymptotes of this hyperbolic set  $\mathcal{G}$  represent the conformal structure of light signals.

Further physical instances of this type of structure will be given in Section 53, where wave packets of dispersive waves will be discussed. An application to water waves is discussed in Burke (1980).

#### *Metric structure*

The clock structure described by the hyperbola  $t^2 - x^2 = 1$  is special. It can be represented by a symmetric tensor of type  $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ . The set of unit time intervals is given by vectors  $a$  such that

$$\mathcal{G} \cdot (a, a) = -1,$$

where  $\mathcal{G}$  is the tensor

$$\mathcal{G} = f^x \otimes f^x - f^t \otimes f^t.$$

It is easy to show directly that this tensor has the same Lorentz invariance as the clocks that it describes. This metric represents light-signal world-lines by the requirement that time intervals measured along them vanish. The application of this metric structure to the Doppler shift and to aberration will be given in the next section.

### **Problems**

- 5.1 Show by constructing an explicit counterexample that clock uniformity is not a projective invariant.
- 5.2 The Newtonian idea of time as absolute and universal leads to clocks whose unit-time set  $\mathcal{G}$  is given by the line  $t = 1$  rather than by the hyperbola  $t^2 = x^2 + 1$ . can this be described by using a metric tensor? Discuss its invariance, especially under Galilean transformations.
- 5.3 See Figure 5.8.

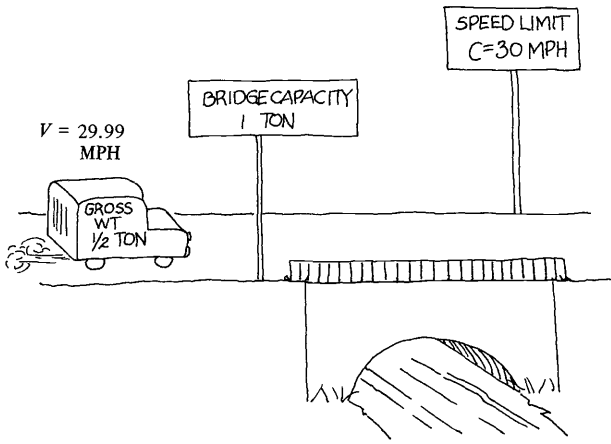


Figure 5.8. What happens?

## 6. The uses of covariance

If we are calculating real physical quantities, then the results must be independent of the representation we use. This is called covariance. Only the final result of calculation need be covariant. Great calculational advantage is often gained, however, by casting the results into a form in which the various pieces are themselves covariant. Often the form of an expression can be deduced in advance, once a covariant notation eliminates spurious coordinate dependencies. This is a matter of style, and I will give here several examples from special relativity of what I consider good style in using covariance in calculations. Manifest covariance is not always the best tactic; the noncovariant coordinates that we will use for contact bundles in Section 15 are a good counterexample.

### *Doppler shift*

Special relativity provides several good examples for us. Suppose an observer moving with 4-velocity  $d$  sends a light signal to another observer moving with 4-velocity  $b$ . Let the direction of the light signal be specified by a vector  $c$ . The Doppler shift observed will be the ratio  $(1+z)$  of received wavelength to emitted wavelength. This ratio can depend on only the three vectors  $d$ ,  $b$ , and  $c$ , and the metric tensor (Figure 6.1). The metric tensor allows us to form dot products, and to normalize the velocity vectors

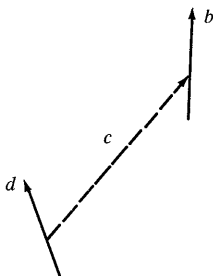


Figure 6.1. The spacetime geometry of the Doppler shift.

$$d \cdot d = b \cdot b = -1.$$

The light signal satisfies

$$c \cdot c = 0.$$

The dot here stands for the inner product using the Lorentz metric. The Doppler shift cannot depend on a generalization of the scalar triple product, since that would introduce a spurious handedness into the problem. Nor can the result depend on the magnitude of  $c$ . The only possibility consistent with these conditions is

$$(1+z) = F\left(\frac{c \cdot d}{c \cdot b}\right).$$

To evaluate the function  $f$ , we could do the calculation in a special reference frame chosen for simplicity, or just notice that, if we think of  $c$  as a 1-form that describes the geometric arrangement of wavecrests in spacetime, then  $(c \cdot d)$  is the rate at which the observer with velocity  $d$  sees wavecrests pass. Thus we must have

$$(1+z) = (c \cdot d)/(c \cdot b).$$

This expression applies to nonplanar motions as well as to plane or linear motions.

### Aberration

A similar problem involves a single observer  $d$  who measures the angle between two light signals  $b$  and  $c$ . Again this angle must be some scalar combination of the various dot products of these vectors. There are now three such dot products that are nontrivial. The angle will not depend on

the magnitudes of either  $b$  or  $c$ . There is only one such combination, and we must have

$$\theta = F\left(\frac{b \cdot c}{(d \cdot b)(d \cdot c)}\right).$$

To evaluate the function  $F$ , we look at two light signals that have an angle  $\theta$  between them:

$$\begin{aligned} d &= e_t, & b &= e_t - e_x, \\ c &= e_t - \cos \theta e_x - \sin \theta e_y; \end{aligned}$$

here we use  $e_x$  to denote the unit vector in the  $x$  direction, and so on. These vectors satisfy

$$d \cdot d = -1, \quad b \cdot b = c \cdot c = 0.$$

The preceding combination reduces to

$$\frac{b \cdot c}{(d \cdot b)(d \cdot c)} = \cos \theta - 1.$$

Both sides of this equation are covariant; so this is the general expression. This same expression could be derived by a more direct approach: projecting the vectors  $b$  and  $c$  perpendicular to  $d$ , normalizing the result, and measuring their angle.

### *The dihedral product*

Look at the geometric situation defined by two planes, one specified by vectors  $d$  and  $b$ , the other by vectors  $b$  and  $c$ . The vector  $b$  is common to the two planes. In Euclidean geometry the intersection of two planes defines a *dihedral angle*. This angle depends on a covariant function of the dot products of the three vectors, with the following additional properties.

1. It is independent of the magnitudes of the three vectors (none of them are assumed to be normalized).
2. It is independent of changes in  $d$  that leave it in the same plane, and similarly for  $c$ .

The second condition can be satisfied by using the alternating tensor products  $d \wedge b$  and  $c \wedge b$ . These alternating products are called bivectors, and have the property that changes in  $d$  such as

$$d \mapsto d + kb$$

leave the bivector unchanged.

These bivectors can be given a metric structure by using the definition

$$(d \wedge b) \cdot (c \wedge d) = (d \cdot c)(b \cdot d) - (b \cdot c)(d \cdot d),$$

which preserves the second condition, no matter whether the metric is Euclidean or Lorentzian. Our problem involves two such bivectors, one for each plane, and thus three possible dot products. The combination that is independent of the magnitudes of  $d$ ,  $b$ , and  $c$  is easily seen in bivector form as

$$\frac{[(d \wedge b) \cdot (c \wedge b)]^2}{[(d \wedge b) \cdot (d \wedge b)][(c \wedge b) \cdot (c \wedge b)]};$$

that is

$$\frac{[(d \cdot c)(b \cdot b) - (b \cdot c)(d \cdot b)]^2}{[(d \cdot d)(b \cdot b) - (d \cdot b)^2][(c \cdot c)(b \cdot b) - (c \cdot b)^2]}.$$

A Euclidean dihedral angle  $\theta$  in special position is given by

$$\begin{aligned} b &= e_z, & d &= e_x, \\ c &= \cos \theta e_x + \sin \theta e_y, \end{aligned}$$

and we have

$$\frac{[ \quad ]^2}{[ \quad ][ \quad ]} = \cos^2 \theta.$$

Since both sides of this equation are covariant separately, we have in general

$$\cos \theta = \frac{(d \cdot c)(b \cdot b) - (b \cdot c)(d \cdot b)}{[(d \cdot d)(b \cdot b) - (d \cdot b)^2]^{1/2} [(c \cdot c)(b \cdot b) - (c \cdot b)^2]^{1/2}}.$$

This combination of three vectors I call the *dihedral product*.

### Spacetime dihedrals

Let us look at three observers who are moving at constant velocity away from an event. They can exchange light signals, and each observer can measure the apparent angle between the other two. We have here a combination of two effects: a geometric diminution caused by their separation, and a magnification caused by aberration. The situation is self-similar in time, and the apparent angles will necessarily be time-independent.

The apparent angle will depend on only the 4-velocity vectors of the three observers. Look at Figure 6.2. For observer  $b$ , only the light signal is important. Changing the source velocity so that it stays in the same



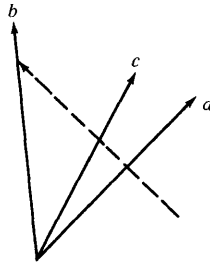


Figure 6.2. The spacetime geometry of aberration.

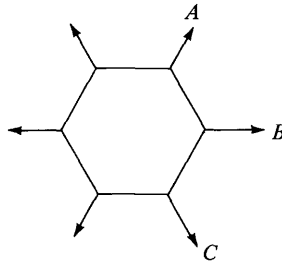


Figure 6.3. A hexagonal array of six observers moving radially outward.

plane will have no effect on the angles. The observer will merely see the source with a different Doppler shift. Again we have a situation with the symmetry of the dihedral product.

A special case for easy computation is

$$b = e_t, \quad d = \cosh \psi e_t + \sinh \psi e_x,$$

$$c = \cosh \psi e_t + \sinh \psi (\cos \theta e_x + \sin \theta e_y),$$

from which we find the general expression

$$\cos \theta = \frac{(b \cdot c)(b \cdot d) - (d \cdot c)(b \cdot b)}{[(d \cdot d)(b \cdot b) - (b \cdot d)^2]^{1/2} [(c \cdot c)(b \cdot b) - (b \cdot c)^2]^{1/2}}.$$

Note the sign changes caused by the Lorentz metric.

**Example:** Suppose we have six equivalent observers arranged in a hexagon in a plane, all moving radially away from the origin (Figure 6.3). What is their radial rapidity if each observer, such as *B*, sees his or her two neighbors, *A* and *C*, as being separated by 90°?

In the frame of symmetry, we have for the normalized tangent vectors

$$b = \cosh \psi e_t + \sinh \psi e_x,$$

$$\left(\frac{a}{c}\right) = \cosh \psi e_t + \sinh \psi \left(\frac{1}{2}e_x \pm \sqrt{\frac{3}{4}}e_y\right).$$

A right angle requires

$$(b \bullet c)(b \bullet a) + (a \bullet c) = 0,$$

which gives us

$$\sinh \psi = \sqrt{2}.$$

### *Pseudospherical geometry*

If we take the radially moving observers of the last example as points in the space of the observers, then we can define a geometry on that space. Take the distance between these points to be the relative velocity between the observers, measured by using rapidity, so that it will be additive. The normalized velocity vectors of these observers form a unit hyperboloid in spacetime. Because of Lorentz invariance, this geometry has the same symmetries as the usual pseudospherical geometry and is therefore equivalent to it. Here we are projecting everything onto the unit hyperboloid by using straight lines through the origin. Light rays between observers project down into geodesics of the unit hyperboloid. The angles between these geodesics correspond to the apparent angular separations.

The geometry of the unit timelike hyperboloid is the non-Euclidean geometry of a surface of uniform negative curvature called the *pseudosphere*. The analogs of the rules for spherical trigonometry are the law of sines (for a triangle with sides  $a$ ,  $b$ , and  $c$ , and opposite angles  $A$ ,  $B$ , and  $C$ , as shown in Figure 6.4),

$$\frac{\sinh a}{\sin A} = \frac{\sinh b}{\sin B} = \frac{\sinh c}{\sin C},$$

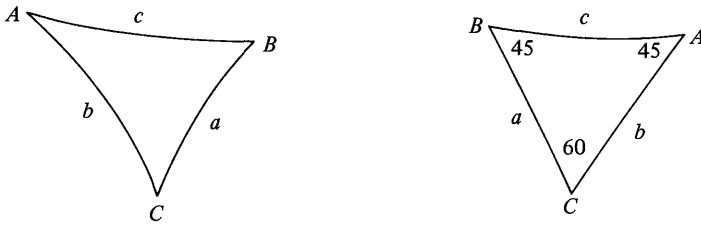
the law of cosines,

$$\cos A = -\cos B \cos C + \sin B \sin C \cosh a$$

and the law of hyperbolic cosines,

$$\cosh a = \cosh b \cosh c - \sinh b \sinh c \cos A.$$

**Example:** The preceding hexagon of observers contains twelve of the pseudospherical triangles  $ABC$  shown in Figure 6.5. We have



**Figure 6.4** (left). A pseudospherical triangle.

**Figure 6.5** (right). The pseudospherical triangle defined by two of our previous six observers.

$$\cosh a = \frac{\cos 45^\circ + \cos 45^\circ \cos 60^\circ}{\sin 45^\circ \sin 60^\circ} = \sqrt{3}.$$

The relative rapidity between observers is

$$\cosh c = \frac{\cos 60^\circ + \cos^2 45^\circ}{\sin^2 45^\circ} = 1 + \sqrt{3}.$$

As an amusing example of non-Euclidean geometry, note that the right-angled hexagons found in the preceding example will tile the pseudo-sphere, with four such hexagons meeting at every vertex.

# II

## *Manifolds*

Most of the sets on which we need to do analysis are not linear spaces. The surface of a sphere is a familiar example of a smooth set that does not have the structure of a linear space. It has no origin (zero vector). Also, we cannot define addition on pairs of points (free vectors) on a sphere in a manner consistent with the axioms for a linear space, nor can we define a constant vector field on such a surface. Nonlinear spaces of cosmological interest are 3-spheres and pseudospheres. In this chapter we develop the basic tools for dealing with such spaces, a subject called *calculus on manifolds*.

The manifolds that come up in applications are of two types. First, we have manifolds such as the configuration space of a rigid body in free fall; the points of this manifold are all possible rotations. The choice of an identity rotation for the body is arbitrary, and the points of this configuration space are all equivalent. Although coordinates will usually be needed to describe concrete situations, the geometric structures will not involve the coordinates directly. Second, we have manifolds like the space of energy, temperature, entropy, pressure, and volume, which describes a thermodynamic system. Here the coordinates all have direct physical interpretation. Surprisingly, the coordinate-free methods developed for the first kind of manifolds are also useful and efficient tools for manifolds with distinguished coordinates.

**Example:** Even though space and time have distinct physical interpretations, the nondispersive wave equation

$$\frac{\partial^2 f}{\partial t^2} = \frac{\partial^2 f}{\partial x^2}$$

is often studied in rotated coordinates

$$u = t - x, \quad v = t + x,$$

where it becomes

$$\frac{\partial^2 f}{\partial u \partial v} = 0.$$

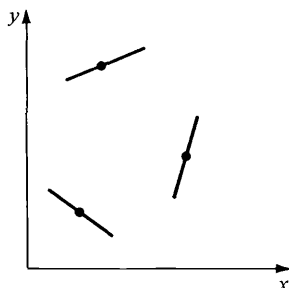
A parallel situation in vector analysis is the usefulness of grad, div, and curl even in situations that do not have rotational symmetry.

Associated with every point in a manifold is a local approximation to the manifold called the tangent space. This is a linear vector space whose elements, called tangent vectors, form the basic building blocks for calculus on manifolds. We will give here two different definitions of a tangent vector. They are equivalent, both are useful, and they have an agreeable duality. A firm grasp of the idea of a tangent space (the space of all tangent vectors) and the tangent bundle (the collection of all tangent spaces) is essential for what follows.

The geometric structures introduced in this chapter go beyond ordinary tensor calculus. Contact structure is a prime example, as is manifold structure itself. Of course, tensor calculus is useful, and some of these additional structures have convenient representations in terms of special tensors.

## 7. Manifolds

Well-behaved sets with enough structure on them so that we can “do calculus” are called differential manifolds, or, for short, *manifolds*. The least structure that a set can have would allow us only to name the points and to discuss the identity of points and their membership in various other sets. Interesting sets usually have more structure. A minimal additional structure is a topology, which is enough structure that continuity of curves and maps can be discussed. The sets called manifolds have even more structure, and the smoothness of curves and maps (that is, their higher derivatives) can also be discussed. Smooth curves in a manifold have local linear approximations called tangent vectors.  $\mathbb{R}^n$  has all this structure and more. The additional structures of  $\mathbb{R}^n$  allow us to define straight lines, global parallelism, and a special point called the origin. They are not structures that we necessarily want to demand of our manifolds. We are going to define manifolds so that locally they look like  $\mathbb{R}^n$ , but do not have this excess structure.



**Figure 7.1.** Three points in the line-element contact bundle of  $\mathbb{R}^2$

**Examples:** Let  $P$  be the set of all straight lines through the origin in Euclidean 3-space. We will soon see that this set is a manifold.

Let  $G$  be the set of all great circles on the sphere. This set too is a manifold.

Let  $Q$  be the set of number triples  $(x:y:z)$  except for  $(0:0:0)$ , modulo the equivalence relation

$$(x:y:z) \sim (kx:ky:kz)$$

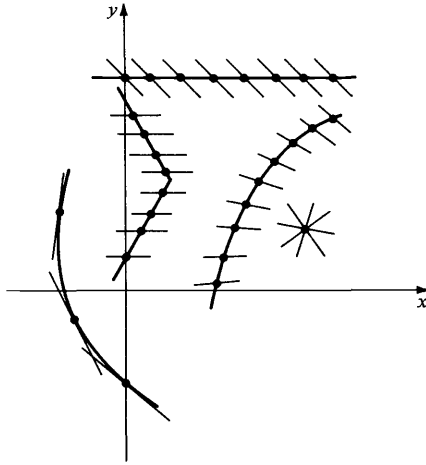
for all real numbers  $k$ .  $Q$  is also a manifold with essentially the same structure as the manifolds  $P$  and  $G$  in the preceding. The common manifold structure is called  $P^2$ , *projective 2-space*. We will use this space as an example many times. (Sometimes it is called  $RP^2$ , real projective 2-space. I will not deal with any complex spaces here, and so I will leave off the  $\mathbb{R}$ .)

Let  $C\mathbb{R}^2$  be the set consisting of pairs  $(q, p)$ , where  $q$  is a point in  $\mathbb{R}^2$  and  $p$  is a straight line in  $\mathbb{R}^2$  through  $q$ . In Figure 7.1 I sketch three points of  $C\mathbb{R}^2$ , which can easily and conveniently be represented in  $\mathbb{R}^2$  itself. In Figure 7.2 I draw five curves in  $C\mathbb{R}^2$ .

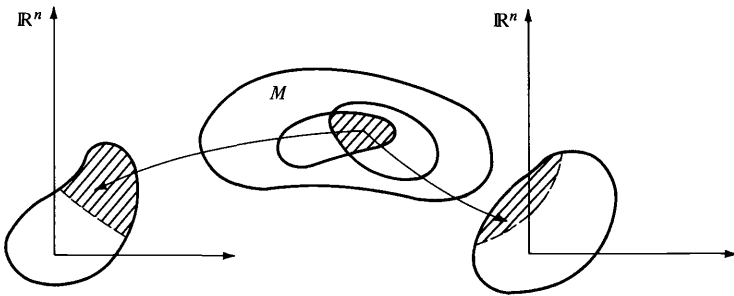
$C\mathbb{R}^2$  is a very useful manifold called the line-element contact bundle of  $\mathbb{R}^2$ . You can see that this is one of a family of such manifolds, and we will meet them repeatedly, both as a useful examples and as building blocks in thermodynamics, mechanics, and partial differential equations.

### Charts

To add the structure of a manifold to a set, we must show how to map an open region around any point in a 1:1 and continuous manner onto an

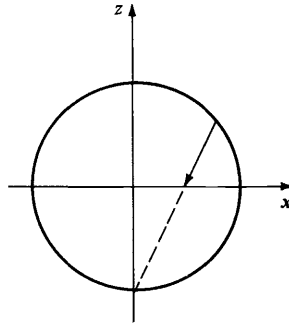


**Figure 7.2.** Five curves in the line-element contact bundle of  $\mathbb{R}^2$ . These could be called partly dotted curves. Only a finite number of points are fully drawn.



**Figure 7.3.** Two charts for a manifold  $M$ . The charts are drawn on separate copies of  $\mathbb{R}^n$ .

open region of  $\mathbb{R}^n$ . The inverse map must also be continuous. Each such map is called a *chart*. The charts must satisfy a compatibility condition. Whenever two charts overlap in the manifold, they define maps of  $\mathbb{R}^n$  into itself, as shown in Figure 7.3. If the set is a smooth manifold, then these maps will be smooth (infinitely differentiable) and have smooth inverses. We subtract the special structure of  $\mathbb{R}^n$  from our manifold by allowing all possible compatible charts on an equal footing. The point in the manifold mapped onto the origin in one chart will not be mapped onto the origin in most other charts. Thus no point is distinguished as the



**Figure 7.4.** The pattern of stereographic projection.

origin of a manifold. The collection of all compatible charts is called an *atlas* for the manifold.

**Examples:** Any linear vector space can be covered with one chart that maps each vector onto the number  $n$ -tuple given by its components in some basis. An atlas consists of all charts derived from this one by smooth transformations, called coordinate transformations.

All of the  $n$ -spheres,  $S^n$ , can be covered by two charts by using stereographic projection. If the  $n$ -sphere is defined as the set of points in an  $(n+1)$ -dimensional Euclidean space that satisfy

$$w^2 + x_1^2 + x_2^2 + \cdots + x_n^2 = 1,$$

then a chart for the region  $w \neq -1$  is the map

$$(w, \mathbf{x}) \mapsto (\mathbf{x}/1+w). \quad (7.1)$$

The geometric picture for this is shown in Figure 7.4.

In chart 7.1 not only are angles preserved, but even finite circles on  $S^n$  appear as Euclidean circles.

The sets  $P$  and  $Q$  defined at the start of this section can be given a manifold structure in the same manner. The Euclidean coordinates of any point on a line in the set  $P$  give a number triple, and the equivalence relation of set  $Q$  identifies number triples belonging to the same line. The set  $CR^2$  in the third example is the Cartesian product of  $\mathbb{R}^2$  and the set  $P$ .



To show that the set  $Q$  is a manifold, look at the point  $[a : b : c]$  in  $Q$ . Suppose that  $c$  is the largest of  $a$ ,  $b$ , and  $c$ . Then a chart around the point  $[a : b : c]$  is given by the map

$$[x : y : z] \mapsto (x/z, y/z)$$

for the open set of points satisfying

$$z \neq 0. \tag{7.2}$$

Both the chart and condition 7.2 are compatible with the equivalence relation, and it maps all of  $Q$  except for a circle onto the entire plane. Two other such charts can be defined, and every point in  $Q$  appears inside at least one of these charts. If we add to these all other compatible charts, we will have an atlas for  $Q$ .

The set of points in Minkowski spacetime satisfying

$$t^2 = x^2 + y^2 + z^2$$

form a manifold called the 3-pseudosphere. It can be covered by a single chart, say, the  $(x, y, z)$  coordinates of its points.

The smoothness of structures on a manifold will be defined in terms of the smoothness of their representations in the charts. The compatibility conditions ensure that this smoothness is chart-independent. The charts must be open so that no point is on an edge. In Section 11 we will find a more efficient way to show that a set can be given a manifold structure, one that does not require us to actually construct the charts. Still, it is important to understand in principle what it means to be a manifold.

**Example:** Let us show that two of the charts for  $P^2$  are compatible:

$$\phi^1: U^1 \rightarrow \mathbb{R}^2; [x : y : z] \mapsto (u, v) = (y/x, z/x),$$

$$U^1 = \{[x : y : z] \mid x \neq 0\},$$

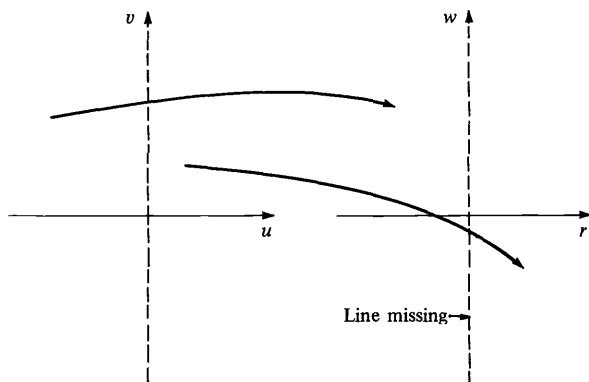
and

$$\phi^2: U^2 \rightarrow \mathbb{R}^2; [x : y : z] \mapsto (r, w) = (x/y, z/y),$$

$$U^2 = \{[x : y : z] \mid y \neq 0\}.$$

The overlap between the charts  $\phi^1$  and  $\phi^2$  is the set

$$U = \{[x : y : z] \mid x \neq 0, y \neq 0\},$$



**Figure 7.5.** The transition map between the two charts for  $P^2$ .

sketched in Figure 7.5. It consists of two disjoint open sets. The transition map is

$$\phi^2 \circ (\phi^1)^{-1}: (u, v) \mapsto (r, w) = (1/u, v/u),$$

which is smooth and is its own inverse. We will continue this example in Section 10.

The utility of these charts is that they enable us to transfer questions from the local neighborhoods in the manifold to open sets in  $\mathbb{R}^n$ . Thus we can graft onto manifolds much of the machinery of multivariable calculus.

### *Defective atlas*

We can often use a defective atlas as a more efficient representation of a manifold. Proper care must be taken with them. If we drop the requirement that the charts be open, then problems can hide on the boundaries. We might not notice discontinuities that coincide with such a boundary. If we drop the 1:1 requirement, then we can end up with functions that are not well-defined, because they take different values at different appearances of the same point.

**Example:** In polar coordinates the function  $\theta$  is either not continuous or not single-valued.

The best of both worlds is to have a single useful defective chart for the manifold, plus some easy local charts for the edges and other bad points



Figure 7.6. A partition of unity for the line.

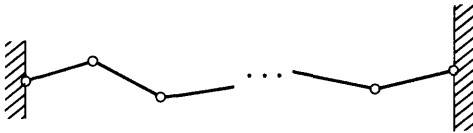


Figure 7.7.

to settle questions of continuity. Often we can translate the message of these edge charts into junction conditions for the single defective chart.

### *Partitions of unity*

The transition functions between different charts are to be smooth, that is, infinitely differentiable ( $C^\infty$ ). It is important that we do not demand that they be analytic. The physicists' emphasis on analytic functions reflects an early training on elliptic equations, a neglect of hyperbolic equations, and a pathological reliance on the crutch of Fourier transforms, which are appropriate only for linear theories.

In the space of  $C^\infty$  functions we can find a set of functions that are each strictly zero outside a bounded region, of which only a finite number are nonzero at any point, and that add up to unity at every point. Such a set is called a *partition of unity*. Given a set of charts that cover the manifold, we can always find a partition of unity such that each function is nonzero on only a single chart. Figure 7.6 sketches the idea.

Such a partition of unity lets us break down an integral on the manifold into a set of pieces that are each contained in a single chart. It also lets us assemble a smooth geometric object on the manifold by piecing together smooth pieces in the charts. We will appeal to this construction to show, for example, that there is a smooth, global contact structure on a contact bundle.

### Problems

- 7.1 The configuration space of the system of  $n$  rigid links shown in Figure 7.7 is a familiar manifold. What is it?
- 7.2 Is the cone a 2-manifold? What about the tip?
- 7.3 Show that there is no natural isomorphism between the manifolds  $P$  and  $G$  defined in the example at the beginning of this section. A natural isomorphism would be independent of the Euclidean metric.
- 7.4 Find a partition of unity for the torus.

### 8. Tangent vectors and 1-forms

One of the basic concepts in the calculus of manifolds is the tangent space, the space of tangent vectors. There are two complementary ideas merged in the idea of a tangent vector. One is that of the local behavior of a curve or a family of curves. The other is that of a differential operator acting on functions. We will base our development of the idea of a tangent vector on the tangent vector as the description of the local behavior of curves. This is the geometric approach and lets us, for example, easily make sense of the common practice of drawing the tangent vectors right on the manifold itself. The differential-calculus idea of local linear approximations does not work directly. In a vector space, curves inherit a linear structure from the global linear structure of the vector space itself. In a manifold the local linear structure comes directly from the ideas of tangency and contact. By constructing our tangent vectors from curves in the manifold rather than as a space outside the manifold, we automatically arrive at a concept that is invariant under mappings.

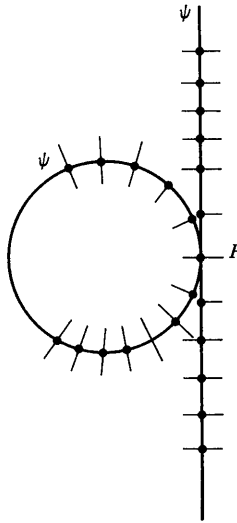
#### *Tangency of maps*

Tangency is the local agreement of maps. Consider two maps,  $\phi$  and  $\psi$ , both  $\mathbb{R}^m \rightarrow \mathbb{R}^n$ . At any point they can be represented by Taylor's Series expansions. If these expansions agree up to terms of order  $p$ , then the maps are said to have  $p$ th-order tangency at that point. Tangency requires agreement not only in direction, but also in parameter rate. We will use tangency here to mean only first-order tangency.

**Example:** The maps  $\mathbb{R} \rightarrow \mathbb{R}^2$ :

$$u \mapsto (u, u^3), \quad u \mapsto (\sin u, 0)$$

are tangent at  $u = 0$ .



**Figure 8.1.** Two curves in  $\mathbb{C}\mathbb{R}^2$  that are tangent.

Tangency for maps between manifolds is defined in terms of charts by using the preceding definition. It is a routine calculation to show that tangency in one chart implies tangency in all compatible charts.

**Example:** In the manifold  $\mathbb{C}\mathbb{R}^2$  introduced in Section 7, let  $\phi$  be the right-hand curve shown in Figure 8.1. The line element at every point is horizontal. The other curve  $\psi$  is the circle, with line elements always pointing to the center. These curves are tangent at the point  $P$ .

Tangency is a structure that is preserved by smooth maps. We will also use the weaker notion of contact, when the maps agree in direction but not necessarily in parameter rate.

### *Tangent vector*

A tangent vector is an abstraction meant to represent the structure common to a class of parametrized curves all tangent at a point. It is the local structure of maps of the form  $\mathbb{R} \rightarrow M$ , in a manifold  $M$ . A tangent vector will be defined to be an equivalence class of tangent curves. We can represent this equivalence class in several ways: by some arbitrarily selected

member, by some canonical member, or by some numerical algorithm. The tangent to the curve  $\gamma: s \mapsto \gamma(s)$  at  $\gamma(s)$  will be written  $\dot{\gamma}(s)$ . My convention is to use the dot for a mnemonic label, and the apostrophe for differentiation with respect to an argument. This  $\dot{\gamma}$  is not a derivative but an equivalence class of tangent curves. The space of tangent vectors at the point  $p$  in the manifold  $M$  will be written  $T_p(M)$ . It is the space of rates.

**Example:** In the space  $\mathcal{Q}$  of number triples modulo scaling (from Section 7) look at three curves through the point  $[1:1:1]$ ,

$$u \mapsto [1+u:1:1], \quad u \mapsto [1+2u:1:1],$$

and

$$u \mapsto [1 + \sin u : \cosh u : \cos u].$$

Only the first and third are tangent. The first and second are obvious choices for canonical representatives of equivalence classes. All three have first-order contact.

### *1-Forms*

We can also discuss tangency of functions, maps of the form  $f: M \rightarrow \mathbb{R}$ . The local behavior of a function can be represented by an equivalence class of such maps, all tangent at a point. The space of such tangent functions is the space of local gradients, written  $T_p^*(m)$ .

**Example:** The functions

$$\mathbb{R}^3 \rightarrow \mathbb{R}; \quad (x, y, z) \mapsto x,$$

and

$$(x, y, z) \mapsto \sin x \cos y$$

are tangent at  $(0, 0, 0)$ .

These local gradients are called 1-forms, a name that will never really make sense. The space of 1-forms is naturally dual to the tangent space. The operation taking a tangent vector and a 1-form to a number is the natural operation of taking the rate of change of a function along a curve.

**Example:** The curve  $\mathbb{R} \rightarrow \mathbb{R}^3$ :

$$\gamma: s \mapsto (\cos s, \sin s, s)$$

has a tangent vector at  $(1, 0, 0)$

$$\dot{\gamma} = [s \mapsto (1, s, s)].$$

Remember, a tangent vector is an equivalence class of curves. Here  $s \mapsto (1, s, s)$  is a curve, and the square brackets denote the equivalence class containing this curve. In this equation I am representing this equivalence class by a particularly simple curve in the class. Similarly, the function

$$f: (x, y, z) \mapsto x^2 + y \cos z$$

has a local gradient at  $(1, 0, 0)$  given by the equivalence class

$$[(x, y, z) \mapsto (1 + 2x + y)].$$

The function  $f$  evaluated along the curve  $\gamma$  is the map

$$s \mapsto \cos^2 s + \sin s \cos s,$$

and the rate of change of  $f$  with respect to the parameter  $s$  at  $s = 0$  is just the ordinary derivative

$$2 \sin s \cos s + \cos^2 s - \sin^2 s,$$

which at  $s = 0$  has the value 1. This local rate can also be deduced from the local approximations

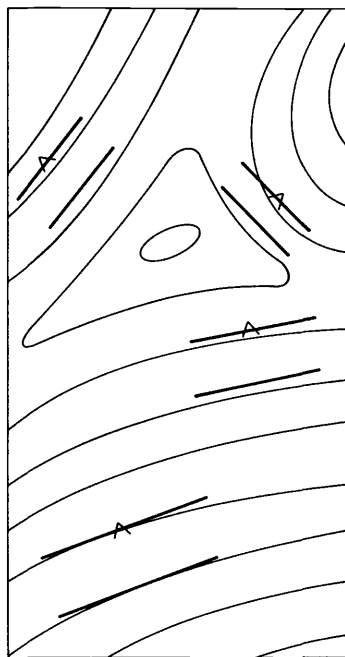
$$s \mapsto 1 + 2x + y = 1 + s.$$

Because of this duality, the space of 1-forms is called the cotangent space. For spaces of finite dimension this duality relation is reciprocal.

### *Graphical representations*

As defined, the tangent space and the cotangent space are abstract spaces. Each has convenient representations in the charts for the manifold. For a tangent vector, take that curve in the equivalence class that appears straight in that chart. This special curve in turn can be represented by the piece of it that extends for one unit of parameter change. With an arrowhead on the end, this is the usual idea of a tangent vector, drawn right on the chart of the manifold. The curve itself is chart-dependent, and there is no intrinsic significance to the point in the manifold that lies under the arrowhead. That point changes as you change charts. What remains invariant is the equivalence class represented by that special curve.

A similar trick allows us to represent 1-forms directly on a chart also. We now pick the function in the equivalence class that appears linear in



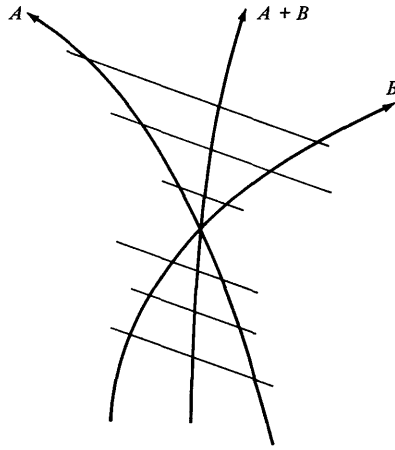
**Figure 8.2.** Gradient 1-forms and contours for a topographic map.

that chart. It can be conveniently represented graphically by drawing two contour lines a unit increment apart, one passing through the given point. Again, the points in the manifold under these contour lines have no intrinsic significance.

**Example:** To see how natural this representation of a 1-form is, consider the problem of making a topographic map for a billiard table. The local gradient at any point can be found by taking a level and a shim of unit thickness and moving the shim out until the level is indeed level. Repeat in any other direction. This gives you two points on the unit contour of the gradient 1-form at the point – provided, of course, that the level and shim are small compared to the hills and valleys of interest, but large compared to the irregularities in the felt. From a gridwork of gradients, it is easy to sketch in the contours of a topographic map of the surface. See Figure 8.2.

1-forms are important geometric objects. They can represent, among other things, force, momentum, and phase velocity. We will see that the





**Figure 8.3.** The addition of two curves in  $\mathbb{R}^2$ .

connection between classical mechanics and wave mechanics through Hamilton-Jacobi theory is remarkably simplified by a geometric treatment that uses 1-forms. The representation of classical thermodynamics to be developed in Section 16 will base its structure on a special 1-form.

### *Linear structure*

Both the tangent space and the cotangent space have the structure of a linear vector space, and these linear structures are compatible with the taking of local linear approximations.

One way to add two tangent vectors is to go to any chart and there add their linear representatives. Or you could add any two representations as shown in Figure 8.3. Nor would the interpolating curves have to be straight lines: any smooth parametrized curves would do. Of course, the addition of curves is chart-dependent, except for the part that is the tangent vector.

**Examples:** The sum of the tangent vectors

$$[s \mapsto (1, s, s)]$$

and

$$[s \mapsto (1 + 2s, 0, -s)]$$

at the point  $(1, 0, 0)$  is

$$[s \mapsto (1 + 2s, s, 0)].$$

In the space  $Q$  given earlier the three curves

$$\gamma^1: u \mapsto [1+u:1:1],$$

$$\gamma^2: u \mapsto [1:1+u:1],$$

$$\gamma^3: u \mapsto [1:1:1+u]$$

might seem to be a natural choice for a basis at the point  $[1:1:1]$ . Clearly every curve through that point can be written as a linear combination of these curves; or, rather, any curve is tangent to some curve that is a linear combination of these curves. But these curves are not independent. In fact, we have

$$\dot{\gamma}^1 + \dot{\gamma}^2 + \dot{\gamma}^3 = 0,$$

where the zero curve is the equivalence class of curves with no linear term, such as

$$u \mapsto [\cosh u : \cos u : 1].$$

The tangent space of  $Q$  is, in fact, only two-dimensional. Any two of these tangent vectors form a basis for the tangent space at most points.

It is even easier to add 1-forms. For this, merely add any two functions in the equivalence class. Of course, adding the special linear representations is usually the easiest. Because the linear structure of the cotangent space is more apparent than that of the tangent space, it is sometimes easier to find the latter from the cotangent space and duality.

You can see that working on manifolds is very nearly the same as working in a vector space; if you use the right concepts, all it takes is the occasional addition of equivalence-class brackets. But if you use the wrong concepts, such as adding vectors defined at different points, you get nonsense.

### *Basis vectors*

In any given chart we can find natural basis vectors for the tangent and cotangent spaces. These basis vectors will not be invariant objects, because they depend on the selection of a particular coordinate chart. This chart-dependent basis uses the tangents to the curves

$$\frac{\partial}{\partial x} = [s \mapsto (x_0 + s, y_0, \dots)],$$

$$\frac{\partial}{\partial y} = [s \mapsto (x_0, y_0 + s, \dots)],$$

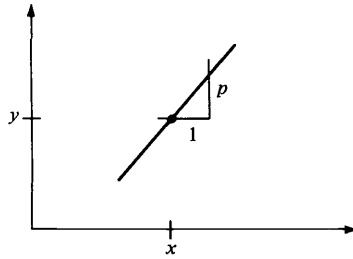


Figure 8.4. A chart for  $\mathbb{C}\mathbb{R}^2$ .

and so on, where the square brackets stand for equivalence class. I will explain soon why these curves have been given differential operators as names. All we are doing here is formalizing the graphical construction given earlier.

In a parallel manner we can find a chart-dependent basis for the cotangent space by using the local gradients of the coordinate functions

$$dx = [x], \quad dy = [y],$$

and so on.

The cotangent space is nearly always simpler than the tangent space. For example, the tangent vector  $\partial/\partial x$  depends on the choice of all the coordinate functions. The 1-form  $dx$  depends only on the coordinate function  $x$ .

Usually several charts are needed to cover a manifold, and the basis vectors and 1-forms of a single chart need not cover the entire manifold. Smooth vector fields are constructed by using a partition of unity to piece together several chart-dependent parts.

**Example:** A chart for most of  $\mathbb{C}\mathbb{R}^2$  is given by  $(x, y, p)$ , where  $p$  is the slope of the line element as defined in Figure 8.4. In Figure 8.5 the three basis vectors defined by these coordinates are drawn. The basis vectors are equivalence classes of curves. These are represented here by simple members of the equivalence class. Note the construction used to get the correct parametrization of  $\partial/\partial p$ , using equal intervals on the vertical line.

### *Numerical representation*

If we are given a chart, the coefficients in the Taylor's Series expansion can be used as a numerical representation for the tangent vector. For a

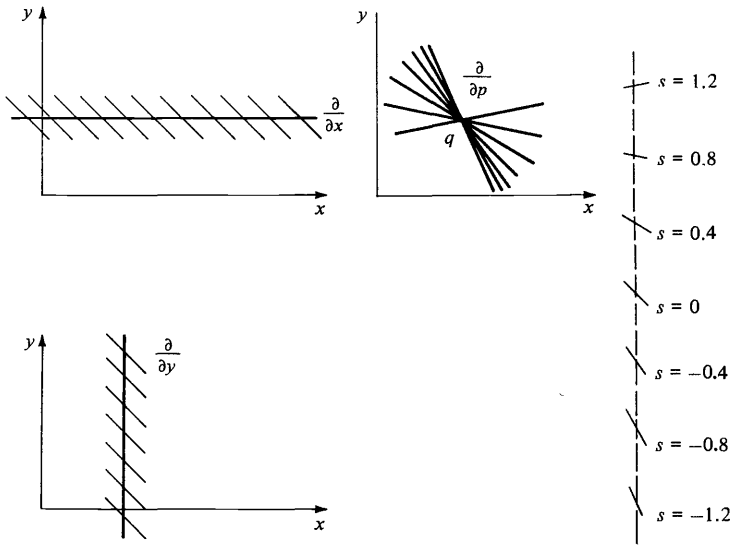


Figure 8.5. Basis vectors defined by the coordinates of the chart of Figure 8.4.

curve  $\gamma: \mathbb{R} \rightarrow M$  and a chart  $\phi: M \rightarrow \mathbb{R}^m$ , we have  $m$  functions of one variable,  $X^1(s)$  and so on:

$$(\phi \circ \gamma): \mathbb{R} \rightarrow \mathbb{R}^m; s \mapsto (X^1(s), X^2, \dots).$$

The  $m$ -tuple

$$\left( \frac{dX^1}{ds}(u), \frac{dX^2}{ds}(u), \dots \right)$$

represents the tangent to  $\gamma$  at  $u$ . The tangent vector corresponding to the  $m$ -tuple  $(a, b, \dots)$  is the equivalence class

$$[s \mapsto (x_0 + as, y_0 + bs, \dots)].$$

In different charts different  $m$ -tuples will represent this tangent vector, of course.

### Derivations

A differential operator is an operator that measures some of the terms in the Taylor's Series expansion of a map. A first-order differential operator, called a *derivation*, is sensitive to the linear terms. From the binomial theorem, we have

$$(1 + \epsilon x)(1 + \epsilon x') = 1 + \epsilon x + \epsilon x',$$

and to capture this algebraic structure, a derivation  $X$  must operate on products of functions according to

$$X(fg) = fX(g) + gX(f).$$

This is the Leibniz rule for derivatives. It guarantees that  $X$  is insensitive to quadratic and higher-order terms.

We showed earlier how a curve operates on a function as a derivation. The derivation aspect is emphasized in the notation  $\partial/\partial x$  for coordinate-chart basis vectors. We could have defined tangent vectors as derivations. Many authors do. I prefer the more geometric approach, using curves. By reading the Taylor's Series backward, we can see that in spaces of finite dimension every linear derivation is also a tangent vector.

### Problems

- 8.1** We can graphically represent points of  $\mathbb{R}^4$  as pairs: a point in  $\mathbb{R}^2$  and a vector at that point. Invent a representation of tangent vectors compatible with this, and show how to add them.
- 8.2** Represent  $P^2$  using homogeneous coordinates  $[x:y:z]$ . Is  $dx$  a smooth 1-form field?

## 9. Lie bracket

The Lie bracket is important in the study of symmetries and transformation groups. This discussion should also help clarify for you the idea of a tangent vector on a manifold, and prepare you for the discussion of tensors on manifolds to follow. An unfortunate fallout from the "pile of numbers transforming properly" approach to tensors is the impression that only tensors are well-defined geometric objects, and, conversely, that any well-defined, coordinate invariant object (to speak old tensor) must be a tensor.

Suppose that we have two vector fields,  $A$  and  $B$ . A vector field is a rule for picking a tangent vector at every point in the manifold. If we are given in addition a function  $f: M \rightarrow R$ , then we can operate on  $f$  with  $A$  to find at every point the derivative of  $f$  in the  $A$  direction,  $A(f)$ .

**Example:** The derivative of  $f = x^2 + y^2$  in the  $(\partial/\partial x + \partial/\partial y)$  direction is

$$\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)(x^2 + y^2) = 2(x + y).$$

Thus  $A(f)$  is another function; so we can consider the vector field  $B$  operating on it.

What about this double operation? It maps functions to functions. Is it a vector? That is, is it a derivation? To be a vector is to be a differential operator sensitive to only the linear terms in a Taylor's Series expansion. Clearly  $AB$  is not a vector.

**Example:** A vector at the point  $(0, 0)$  operating on  $f(x, y) = x^2 + y^2$  must give zero. But we have

$$\frac{\partial}{\partial x} \frac{\partial}{\partial x} (x^2 + y^2) = 2.$$

So far this is not surprising. What *is* remarkable is that there *is* a vector concealed in the second derivatives. A person familiar with quantum mechanics will think “spin 1 times spin 1 surely has some spin 1 in it” and be on the right track. Look in more detail at what goes wrong with  $AB$ . In a coordinate chart  $(x^1, x^2, \dots)$  we have

$$A = a^\mu \frac{\partial}{\partial x^\mu}, \quad B = b^\mu \frac{\partial}{\partial x^\mu},$$

and operating on a product of two functions  $f$  and  $g$  we have

$$(BA)(fg) = b^\mu \frac{\partial}{\partial x^\mu} \left[ a^\nu \frac{\partial}{\partial x^\nu} (fg) \right].$$

Note that the index on the  $A$  summation must be different from the index on the  $B$  summation:

$$(BA)(fg) = b^\mu \frac{\partial}{\partial x^\mu} [a^\nu (f_{,\nu} g + fg_{,\nu})].$$

Subscript commas stand for partial derivatives.

$$(BA)(fg) = b^\mu a^\nu [f_{,\nu} g + f_{,\nu} g_{,\mu} + f_{,\mu} g_{,\nu} + fg_{,\nu\mu}] \\ + b^\mu a_{,\mu}^\nu (f_{,\nu} g + gf_{,\nu}),$$

and now we try to put this back together again. We have

$$BA(fg) = g(BA)f + f(BA)g + b^\mu a^\nu (f_{,\mu} g_{,\nu} + f_{,\nu} g_{,\mu}).$$

The third term contains the garbage that ruins the Leibniz rule. Look at it. It is symmetric. Thus if we define the *Lie bracket* operator by

$$[B, A] = (BA - AB),$$

then it will satisfy

$$[B, A](fg) = f[B, A]g + g[B, A]f.$$

And so it will be a derivation, sensitive only to the linear terms in the functions that it operates on.

The preceding calculation is better done without using explicit coordinates. We have

$$\begin{aligned} BA(fg) &= B[fA(g) + gA(f)] \\ &= f(BA)g + g(BA)f + B(f)A(g) + B(g)A(f), \end{aligned}$$

and so on.

The Lie bracket is easily computed if the vectors are given explicitly, using

$$\left[ \frac{\partial}{\partial x^\mu}, \frac{\partial}{\partial x^\nu} \right] = 0$$

and

$$[fB, gA] = fg[B, A] + fB(g)A - gA(f)B. \quad (9.1)$$

The general expression for the components can be read off from the above calculation

$$[B, A] = (b^\mu a^\nu{}_{,\mu} - a^\mu b^\nu{}_{,\mu}) \frac{\partial}{\partial x^\nu}.$$

In 3-vector notation this would be written

$$[B, A] = (B \cdot \nabla)A - (A \cdot \nabla)B.$$

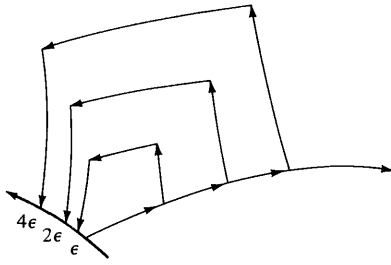
The Lie bracket satisfies the Jacobi identity

$$[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0.$$

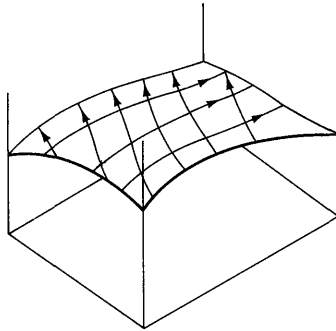
### *Geometric construction*

The Lie bracket is a tangent vector, that is, an equivalence class of curves. We can construct it directly as a curve. For any vector field we can find parametrized curves whose tangents everywhere match the vector field. These are called integral curves (discussed further in Section 14). For two given vector fields  $A$  and  $B$ , suppose that we know their integral curves. The curve representing the Lie bracket  $[A, B]$  at a point  $p$  is the map  $\epsilon \mapsto x^\mu(\epsilon)$ , constructed as follows.

1. Go out along an integral curve of  $A$  from  $p$  a parameter distance  $+\sqrt{\epsilon}$ .
2. Go out from there along an integral curve of  $B$  a distance  $+\sqrt{\epsilon}$ .



**Figure 9.1.** The geometric construction of the Lie bracket.



**Figure 9.2.** Two vector fields with zero Lie bracket. They are the tangents to curves that form a 2-surface.

- 3: Come back from there along an integral curve of  $A$  a distance  $-\sqrt{\epsilon}$ .
4. Finally, come back from there along an integral curve of  $B$  a distance  $-\sqrt{\epsilon}$ . This point is  $x^\mu(\epsilon)$  for  $\epsilon > 0$ . For  $\epsilon < 0$ , just change the signs of all the square roots above.

The limit  $\epsilon \rightarrow 0$  is a curious one here. The squares shrink to zero, but less rapidly than the tangent vectors. This is sketched in Figure 9.1.

### *Frobenius Theorem*

Another geometric interpretation of the Lie bracket is shown in Figure 9.2. If at every point the Lie bracket of the tangent vectors to two families of curves is a linear combination of the two vectors, then these curves fit together to form 2-surfaces. This result is called the Frobenius Theorem and holds for any number of vectors, with submanifolds of the



appropriate dimension. The proof is not illuminating, and can be found in most standard books. This will be important when we discuss exterior differential systems.

### 10. Tensors on manifolds

Once we have constructed a vector space, such as the tangent space, then an entire algebra of tensors becomes available. Here we will highlight the ideas of tensor fields on manifolds. A vector field, for example, is a choice of a tangent vector at every point. Each of these tangent vectors lives in its own tangent space. The fundamental notion is linearity. For tensors on manifolds, we demand linearity over functions, not just over numbers. This forces our tensors to be *local* linear operators.

**Example:** The Lie bracket operator  $V \times V \rightarrow V$ ;

$$(a, b) \mapsto [a, b]$$

is *not* a tensor of type  $\binom{1}{2}$  since

$$[fa, b] \neq f[a, b]$$

for functions  $f$ .

When the manifold requires several charts to cover it, we must be careful to connect the tensor fields smoothly from chart to chart.

**Example:** Continue with the manifold  $P^2$  from Section 7. Use first the coordinate chart

$$[x : y : z] \mapsto (u, v) = (y/x, z/x).$$

Are the 1-forms  $du$  and  $dv$ , a basis in this chart, continuous over all of  $P^2$ ?

No, they are not. In the chart

$$[x : y : z] \mapsto (r, w) = (x/y, z/y)$$

we have the function  $v$ , given by

$$v = w/r,$$

and its gradient is

$$dv = dw/r - w dr/r^2.$$

This is sketched in Figure 10.1. It is infinite on the dotted line and not continuous along the  $x$  axis.

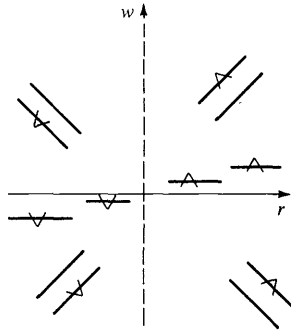


Figure 10.1. The discontinuity in the chart-dependent basis 1-form  $dv$ .

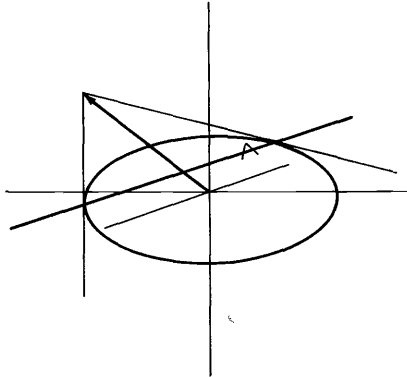


Figure 10.2. A symmetric tensor of type  $\binom{0}{2}$ .

### Symmetric tensors

Symmetric tensors of type  $\binom{0}{2}$ , tensors in the space  $T_p^*(M) \otimes T_p^*(M)$ , are very important because they define metric structures. They also describe crystal optics, electrical conductivity, and some wave-propagation problems. These tensors also have a representation in the tangent space. To represent the tensor, use the second-degree curve

$$T(u, u) = 1,$$

where  $u$  is a tangent vector.

**Example:** In Figure 10.2 I sketch the tensor

$$T: \left( a \frac{\partial}{\partial x} + b \frac{\partial}{\partial y}, a \frac{\partial}{\partial x} + b \frac{\partial}{\partial y} \right) \mapsto a^2 + 4b^2.$$

The action of this tensor as a map

$$T: T_p(M) \rightarrow T_p^*(M); u \mapsto [v \mapsto T(u, v)]$$

is also shown;  $v \mapsto T(u, v)$  is a 1-form, taking in the vector  $v$  and producing a number.

Since 1-forms appeared as the gradients of functions, you might expect that some appropriate combination of the second partial derivatives will produce these second-rank tensors. In general they do not, and it should help you to grasp the idea of tensors to see how they fail.

The index notation might tempt us to use the second derivative as a map  $V \times V \rightarrow \mathbb{R}$ ;

$$(u, v) \mapsto \frac{\partial^2 f}{\partial x^\mu \partial x^\nu} u^\mu v^\nu.$$

To see that this cannot work, look at the following counterexample.

**Example:** Let  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ ;

$$(x, y) \mapsto x + \frac{1}{2}x^2 + \frac{1}{2}y^2.$$

At the point  $(0, 0)$  the second derivative is an operator that in these coordinates is identical with the tensor

$$B = dx^2 + dy^2.$$

In another coordinate system this tensor may not coincide with the second derivative operator. Go, for example, to coordinates  $(u, v)$ ,

$$u = x + y^2, \quad v = y.$$

Then at  $(0, 0)$  the tensor  $B$  is given by

$$du^2 + dv^2,$$

since at  $(0, 0)$

$$du = dx + 2y dy = dx, \quad dv = dy.$$

But in these new coordinates

$$f = u + \frac{1}{2}u^2 - \frac{1}{2}v^2 - uv^3 + \frac{1}{2}v^4,$$

and the second-derivative operator now corresponds to the tensor  $(du^2 - dv^2)$ . This is an important example. Please think through it carefully.

On the other hand, the index-free notation suggests the construction of an operator

$$H_f: (u, v) \mapsto v(uf).$$

This  $H_f$  will be a tensor if it is function-linear, that is, if

$$v(gu)f \stackrel{?}{=} gv(uf),$$

for any function  $g$ . We have

$$v((gu)f) = v(g(uf)),$$

and, because  $v$  is a derivation,

$$v(g(uf)) = gv(uf) + (uf)(vg).$$

The second term here prevents  $H_f$  from being a tensor. However, at a critical point of  $f$ , a point where we have

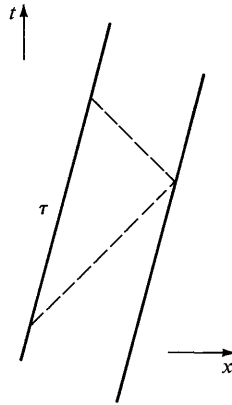
$$df = 0,$$

$H_f$  does form a tensor, called the Hessian. If you think about the contours around a critical point, you will see that they are ellipses or some second-order curve, just like the curves used to represent symmetric tensors.

### *Metric tensors*

The metric structure of a set can be represented (except for some exotic cases) by a symmetric second-rank tensor. People used to think, several centuries ago, that they were studying *the* geometry of space. We now take a more pragmatic viewpoint. For such-and-such purposes, we can put a metric on set so-and-so. A nice example of this is the metric on velocity space in special relativity that makes velocity space isometric with the 3-pseudosphere defined in Section 7.

**Example:** The geometry of a rotating disc has generated an enormous literature, much of it an excellent catalog of mistaken interpretations of special relativity. In particular, many have tried to answer the question of whether that space is curved or not. A crucial prior question is “what space are we talking about?” There are no gravitational fields in the problem, and no need for general relativity, and no *spacetime* curvature. The rotating disc can be adequately represented in any of the inertial reference frames (which are linear vector spaces) of special relativity. We can also define a



**Figure 10.3.** The physical operation represented by a metric tensor on the space of observers on a rotating disc. The dashed lines represent light signals. The time interval  $\tau$  is being measured.

new space for the problem, the space of observers on the disc. Each “point” of this space is a spacetime helix. Thus the three-parameter family of helices,

$$\gamma(\rho, \phi, \zeta) : \mathbb{R} \rightarrow \mathbb{R}^4; u \mapsto (\rho \cos(\phi + \omega u), \rho \sin(\phi + \omega u), \zeta, u),$$

for any value of the parameter  $\omega$  describes a family of uniformly rotating observers. Clearly the 3-tuple  $(\rho, \phi, \zeta)$  is a coordinate chart for this manifold.

To measure the distance between any two nearby observers, we adopt the geometric construction that corresponds to a radar-ranging experiment, shown in Figure 10.3. Geometrically, we find any vector  $\eta$  connecting the two worldlines, and project this onto the subspace perpendicular to the vector  $\lambda$ , which is the tangent to the observers’ worldlines. This is the operation

$$\eta \mapsto (\lambda \cdot \lambda)\eta - (\eta \cdot \lambda)\lambda,$$

where the dot represents the Lorentz metric of special relativity. Here  $u$  is not proper time, and we have not bothered to normalize this vector. We define the “distance” between the observers by an operator  $\mathcal{R}$ ,

$$\begin{aligned} (\lambda \cdot \lambda)^2 \mathcal{R} &= [(\lambda \cdot \lambda)\eta - (\eta \cdot \lambda)\lambda] \cdot [(\lambda \cdot \lambda)\eta - (\eta \cdot \lambda)\lambda] \\ &= (\lambda \cdot \lambda)^2 (\eta \cdot \eta) - (\lambda \cdot \lambda)(\eta \cdot \lambda)^2. \end{aligned}$$

The dot here denotes the metric inner product. For our observers we have, at a point corresponding to the observer  $(\rho, \phi, \zeta)$ ,

$$\lambda = -\rho\omega \sin(\phi + \omega u) \frac{\partial}{\partial x} + \rho\omega \cos(\phi + \omega u) \frac{\partial}{\partial y} + \frac{\partial}{\partial t},$$

and for the connecting vector for a nearby observer  $(\rho + \Delta\rho, \phi + \Delta\phi, \zeta + \Delta\zeta)$  we have a vector in spacetime

$$\begin{aligned} \eta = & [\Delta\rho \cos(\phi + \omega u) - \rho \Delta\phi \sin(\phi + \omega u)] \frac{\partial}{\partial x} \\ & + [\Delta\rho \sin(\phi + \omega u) + \rho \Delta\phi \cos(\phi + \omega u)] \frac{\partial}{\partial y} + \Delta\zeta \frac{\partial}{\partial z}. \end{aligned}$$

Here is one of the few good uses of the traditional infinitesimals of physics. It would be clumsy to avoid using them here. To avoid confusing them with 1-forms, I use “ $\Delta$ ” for them. Now we have

$$\begin{aligned} \lambda \cdot \lambda &= \rho^2 \omega^2 - 1, \\ \eta \cdot \eta &= (\Delta\rho)^2 + \rho^2 (\Delta\phi)^2 + (\Delta\zeta)^2, \\ \eta \cdot \lambda &= \rho^2 \omega \Delta\phi. \end{aligned}$$

Grinding through the algebra, we find the distance between these observers to be

$$(\Delta\rho)^2 + \frac{\rho^2}{1 - \rho^2 \omega^2} (\Delta\phi)^2 + (\Delta\zeta)^2,$$

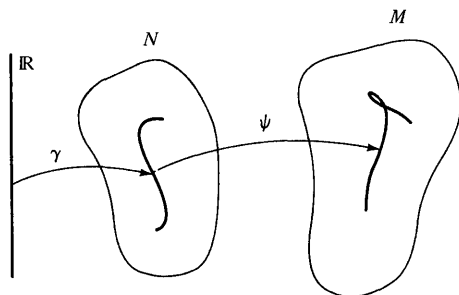
and on the space of observers this is represented by a metric tensor

$$d\rho \otimes d\rho + \frac{\rho^2}{1 - \rho^2 \omega^2} d\phi \otimes d\phi + d\zeta \otimes d\zeta.$$

This metric *is* curved. It is not a spacetime metric, however, and its curvature has no more to do with the Einstein field equations than does the curvature of the velocity-space metric.

## 11. Mappings

Suppose that  $M$  and  $N$  are manifolds, and that we have a map  $\psi: N \rightarrow M$ . What about the calculus of such maps? How do we describe their local behavior? This is an important problem. If  $N$  is smaller than  $M$ , then the map  $\psi$  defines a subspace of  $M$ . It may be a submanifold, depending on



**Figure 11.1.** The pattern of the pushforward map, pushing a curve  $\gamma$  in the direction of the map  $\psi$ .

the good behavior of the map  $\psi$ . If  $M$  is smaller than  $N$ , then  $\psi$  can be thought of as a function on  $N$ , perhaps a vector-valued function. If  $N$  and  $M$  are the same manifold, then  $\psi$  specifies an automorphism, a transformation of the manifold into itself. If  $M$  and  $N$  are different charts for a manifold, then  $\psi$  represents a coordinate transformation. The behavior of our geometric structures under mappings is very simple.

### *Pushforward*

Suppose that in addition to the map  $\psi$  we have a parametrized curve in  $N$ ,  $\gamma: \mathbb{R} \rightarrow N$ . Then this leads automatically to a curve in  $M$ ,

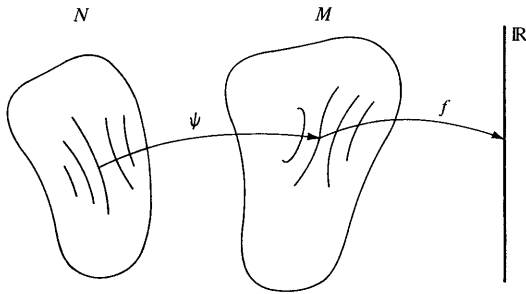
$$(\psi \circ \gamma): \mathbb{R} \rightarrow M,$$

called the *pushforward* of the curve  $\gamma$  (Figure 11.1). As a local linear approximation, a tangent vector in  $N$  can be pushed forward to  $M$  as long as the map  $\psi$  is smooth. The smoothness of  $\psi$  is established by examining the smoothness of its representation in terms of charts for  $M$  and  $N$ . Pushforward is a linear operator, written  $\psi_*$ , and the pushforward of the tangent vector  $v$  by the map  $\psi$  is the linear transformation written  $\psi_* \cdot v$ . Let  $x$  be  $n$  coordinates for  $N$  and  $y$  be  $m$  coordinates for  $M$ .  $\psi$  is represented by  $m$  functions  $Y^1(x^1, x^2, \dots, x^n), \dots$ . The curve  $\gamma$  is represented by functions  $\gamma(u)$ , and its tangent vector is

$$\dot{\gamma} = \frac{d\gamma}{du} \frac{\partial}{\partial x}.$$

Using the chain rule we have

$$\psi_* \cdot \dot{\gamma} = \left( \frac{\partial Y^a}{\partial x^\mu} \right) \left( \frac{\partial \gamma^\mu}{du} \right) \frac{\partial}{\partial y^a},$$



**Figure 11.2.** The pattern of the pullback map, pulling a function  $f$  in the opposite direction to the map  $\psi$ .

and the general rule for pushing forward a basis vector is thus

$$\psi_*: \frac{\partial}{\partial x} \mapsto \frac{\partial Y^a}{\partial x} \frac{\partial}{\partial y^a}.$$

Note the mnemonic simplicity. It is easy to see that the pushforward map is compatible with the linear structure of the tangent space. For maps  $\psi: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , the pushforward map  $\psi_*$  is just the differential  $D\psi$ .

### *Pullback*

The dual situation is to have a function on  $M$ ,  $f: M \rightarrow \mathbb{R}$ . This leads immediately to a function on  $N$ ,  $(f \circ \psi): N \rightarrow \mathbb{R}$ , called the *pullback* of  $f$  by  $\psi$  (Figure 11.2). The gradients of functions are equivalence classes of functions, and for smooth maps the pullback is compatible with the linear structure and the equivalence relation. The pullback operator is written  $\psi^*$ . The pullback of the basis 1-forms can easily be found to be

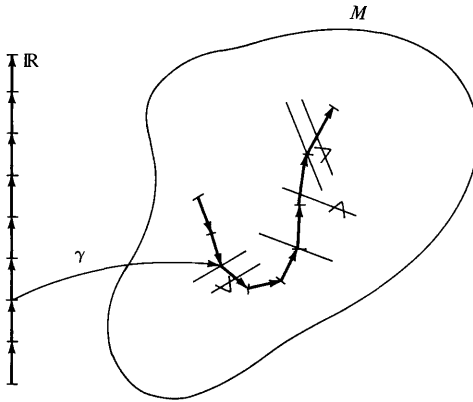
$$\psi^*: dy \mapsto \frac{\partial Y}{\partial x^\alpha} dx^\alpha.$$

These two expressions are easily remembered; they are just analogs of the chain rule. The linear-operator structure of these spaces is preserved by mapping, and we have

$$\omega \cdot (\psi_* \cdot v) = (\psi^* \cdot \omega) \cdot v.$$

**Example:** We will use pullback to define integrals on manifolds. Let  $\gamma$  be a curve in  $M$ ,  $\gamma: I \rightarrow M$ ;  $s \mapsto \gamma(s)$ , for some interval  $I \subset \mathbb{R}$ , and let  $\omega$  be a 1-form field on  $M$ . We want to define the integral





**Figure 11.3.** The geometric idea of integration on a manifold, here a line integral.

$$\int_{\gamma} \omega$$

to be the net number of contour lines of  $\omega$  crossed by the curve  $\gamma$ . If we push forward the vector  $\partial/\partial s$ , we have the tangent to the curve,  $\dot{\gamma}$ , and

$$\omega \cdot \dot{\gamma} = \omega \cdot \left( \gamma_* \cdot \frac{\partial}{\partial s} \right)$$

gives us the net rate at which contour lines are being crossed relative to the parameter  $s$  (Figure 11.3). We also have

$$\omega \cdot \dot{\gamma} = (\gamma^* \cdot \omega) \cdot \frac{\partial}{\partial s}.$$

Integrating the rate over the interval gives us the net change

$$\int_{\gamma} \omega = \int_{s_1}^{s_2} (\gamma^* \cdot \omega) \cdot \frac{\partial}{\partial s} ds = \int_{s_1}^{s_2} \gamma^* \cdot \omega,$$

which will be taken as the definition of the integral of a 1-form. The third form is a convenient shorthand. In coordinates

$$\int_{\gamma} \omega = \int_{s_1}^{s_2} \omega_{\mu} \frac{dX^{\mu}}{ds} ds,$$

where the functions  $X^{\mu}$  specify the curve. Integration will be studied carefully in Section 29.

We pull back a tensor product by pulling back each separate factor.

### *Coordinate transformations*

If  $M$  and  $N$  have the same dimension, and if the map has a smooth inverse, then 1-forms can be pushed forward and tangent vectors pulled back. A legitimate coordinate transformation does have these properties, and our preceding formulae lead immediately to the usual coordinate-transformation laws of tensor analysis:

$$\bar{v}^\alpha = \frac{\partial y^\alpha}{\partial x^\mu} v^\mu, \quad \bar{\omega}_\alpha = \frac{\partial x^\mu}{\partial y^\alpha} \omega_\mu,$$

where the  $\bar{v}^\alpha$  are the components of  $v$  in the  $y$ -coordinate chart, and so on. In practice it is usually easier to make explicit transformations than to use these expressions.

### *Automorphisms*

An automorphism is a map of a space into itself that preserves whatever geometric structures of the space are of interest. The local behavior of an automorphism is described by the pushforward of its basis vectors.

**Example:** An instructive map, much studied in ergodic theory, is the Henon map,

$$\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2; (x, y) \mapsto (y+1-ax^2, b^2x),$$

which depends on two parameters,  $a$  and  $b$ . We can find out about its local behavior, what it does to volumes, shapes, and orientations, by studying its local approximation.

Tangent vectors are pushed forward according to

$$\begin{aligned} \frac{\partial}{\partial x} &\mapsto -2ax \frac{\partial}{\partial x} + b^2 \frac{\partial}{\partial y}, \\ \frac{\partial}{\partial y} &\mapsto \frac{\partial}{\partial x}. \end{aligned}$$

A unit volume element is described by the bivector  $\partial/\partial x \wedge \partial/\partial y$ , and this is pushed forward into

$$\frac{\partial}{\partial x} \wedge \frac{\partial}{\partial y} \mapsto b^2 \frac{\partial}{\partial y} \wedge \frac{\partial}{\partial x}.$$

Thus volumes are magnified by a factor of  $b^2$  for each iteration of the map, and the orientation is reversed.

Another local property is the rate of spreading of nearby points. For this we need to find the eigenvectors of the linear transformation of the tangent space

$$\begin{bmatrix} -2ax & b^2 \\ 1 & 0 \end{bmatrix}$$

This is just the Jacobian of the transformation. It has a positive eigenvector with eigenvalue

$$(a^2x^2 + b^2)^{1/2} - ax;$$

so some vector is magnified at each iteration by this factor. From this we can estimate the distortion that a small circle would undergo. In fact, the map stretches in one direction and compresses in another, like bread being kneaded or milk being stirred into coffee. Thus even a smooth map can still stir up a space and destroy information about the initial state.

### Submanifolds

A map  $\psi: N \rightarrow M$ , where  $N$  is smaller than  $M$ , can be viewed as inserting  $N$  into  $M$  as a subspace. Such an insertion can have cusps, corners, and other pathologies. A good way to study such a map is to look at the manner in which the tangent spaces of  $N$  are mapped into  $M$ . At a pathological point the tangent vectors of  $N$  will be compressed, and not span a subspace of the tangent space of  $M$  of the correct dimension.

**Example:** Steiner's surface is a 2-surface in  $\mathbb{R}^3$  defined informally by the equation

$$x^2y^2 + x^2z^2 + y^2z^2 - xyz = 0.$$

Actually, the  $x$ ,  $y$ , and  $z$  axes also satisfy this equation, but we do not want to include them. A better specification is given by the map

$$\psi: Q \rightarrow \mathbb{R}^3; [\lambda : \mu : \nu] \mapsto \frac{(\mu\nu, \lambda\nu, \lambda\mu)}{\lambda^2 + \mu^2 + \nu^2}.$$

The manifold  $Q$  is a projective 2-space represented in homogeneous coordinates, defined in Section 7. First, we should check that the map  $\psi$  is properly homogeneous, because  $[\lambda : \mu : \nu]$  is an equivalence class. This is obvious.

Except for the circle of points  $\lambda = 0$ , we can represent  $Q$  by the chart

$$\phi: [\lambda: \mu: \nu] \mapsto (u, v) = \left( \frac{\mu}{\lambda}, \frac{\nu}{\lambda} \right),$$

$$\phi^{-1}: (u, v) \mapsto [1: u: v].$$

The representation of  $\psi$  in this chart is the map

$$(\psi \circ \phi^{-1}): \mathbb{R}^2 \rightarrow \mathbb{R}^3; (u, v) \mapsto (uv, v, u)/(1+u^2+v^2),$$

and if we call the denominator  $\Delta$ , we have the pushforward map

$$\Delta^2 \frac{\partial}{\partial u} \mapsto v(1-u^2+v^2) \frac{\partial}{\partial x} - 2uv \frac{\partial}{\partial y} + (1-u^2+v^2) \frac{\partial}{\partial z},$$

$$\Delta^2 \frac{\partial}{\partial v} \mapsto u(1-v^2+u^2) \frac{\partial}{\partial x} + (1-v^2+u^2) \frac{\partial}{\partial y} - 2uv \frac{\partial}{\partial z}.$$

Now, the map  $\psi$  will be pathological anywhere that either of these vectors vanishes; that will indicate a crease or worse. At the points

$$(u, v) = (1, 0), (-1, 0), (0, 1), (0, -1),$$

we have a crease. We will also have bad points wherever these two vectors are parallel. There is a good discussion of this surface in Hilbert and Cohn-Vossen (1952).

### *Implicit arguments for manifolds*

There is a nice implicit way to show that a set is a manifold. It is much easier than the actual construction of an atlas of charts. As in the example of the Steiner surface, many manifolds of interest are formed from larger manifolds by imposing some condition.

**Example:** The 3-sphere is the subset of  $\mathbb{R}^4$  given by

$$x^2 + y^2 + z^2 + w^2 - 1 = 0.$$

The following theorem handles many such cases. Let  $\psi: N \rightarrow M$  be a smooth map, let  $m$  be a point in  $M$ , and  $P$  the inverse image of  $m$ :

$$P = \psi^{-1}(m).$$

If  $\psi_*: T_n(N) \rightarrow T_m(M)$  is always *surjective* for  $n \in P$ , then  $P$  is a manifold. Surjective means that for every  $b$  in  $T_m(M)$ , the equation

$$\psi_* \cdot a = b$$

has at least one solution  $a$ .

**Examples:** For the 3-sphere we consider the map

$$\psi: \mathbb{R}^4 \rightarrow \mathbb{R}; (x, y, z, w) \mapsto (x^2 + y^2 + z^2 + w^2 - 1),$$

and look at  $\psi^{-1}(0)$ . We have

$$\psi_* \cdot \frac{\partial}{\partial x} = 2x \frac{\partial}{\partial s},$$

and so on.  $\psi_*$  fails to be surjective only if we have  $x = y = z = w = 0$ , and that point is not in the inverse image. Thus  $S^3$  and, in fact, all of the  $S^n$  are manifolds.

In Section 2 we computed the differential of the matrix map

$$F: M^n \rightarrow M^n; A \mapsto AA^t - I$$

to be

$$DF(A): h \mapsto Ah^t + hA^t,$$

where  $M^n$  is the space of  $n \times n$  matrices. The matrices  $(AA^t - I)$  are always symmetric. Both  $M^n$  and also the symmetric matrices are manifolds, even vector spaces. The orthogonal matrices go to zero under the above map. They form a manifold if the pushforward is surjective, that is, if we can always find a matrix  $h$  such that, for every symmetric matrix  $C$ , and for  $A$  satisfying  $AA^t = I$ ,

$$Ah^t + hA^t = C.$$

This is easy. Pick, for example

$$h = \frac{1}{2}CA.$$

Thus the orthogonal matrices form a manifold.

### Problem

11.1 Sketch Steiner's surface near a bad point.

## 12. Cotangent bundle

We are now going to consider a larger space, the space of all cotangent spaces to a manifold  $M$ , called the *cotangent bundle* and written  $T^*M$ . Points in this space are pairs: a point in the original space together with a 1-form at that point. In Figure 12.1, I sketch two points in the cotangent

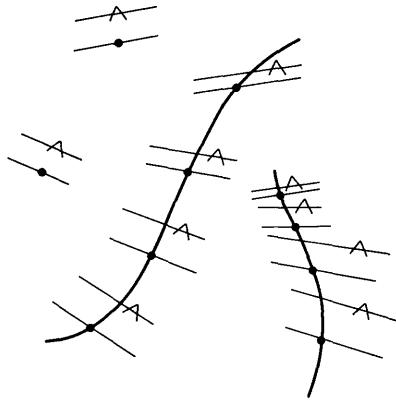


Figure 12.1. A few points and curves in the cotangent bundle.

bundle and two curves. There is a natural projection map on the cotangent bundle,

$$\pi: T^*M \rightarrow M; (q, \omega) \mapsto q,$$

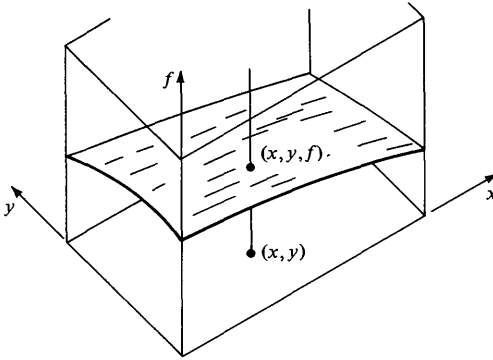
which forgets the 1-form. The points in  $T^*M$  are grouped into subsets that share the same base point  $q$ , called *fibers*. Such a mathematical structure is called a *fiber bundle*. These special manifolds will be used often and are introduced informally in this section.

### *Fiber bundles*

A common mathematical structure in physical theories is the idea of a *field*. At each point of some space some geometric object is specified.

**Example:** Inside a strained elastic body there exists a temperature field, which assigns to each point in the body a real number, the temperature at that point. There also exists a stress field inside the body; it assigns to each point a symmetric tensor.

The field is a simple extension of the idea of a function. It is often useful to study real-valued functions in terms of their graphs. The graph is in a space with one additional dimension for the function value. Any function is now a hypersurface in this larger space (see Figure 12.2). Similarly, we will study fields by going to a larger space where both the original space (this is generally called the *base space*) and the field variables



**Figure 12.2.** The pattern of a bundle illustrated by the graph of a function in  $\mathbb{R}^2$ .

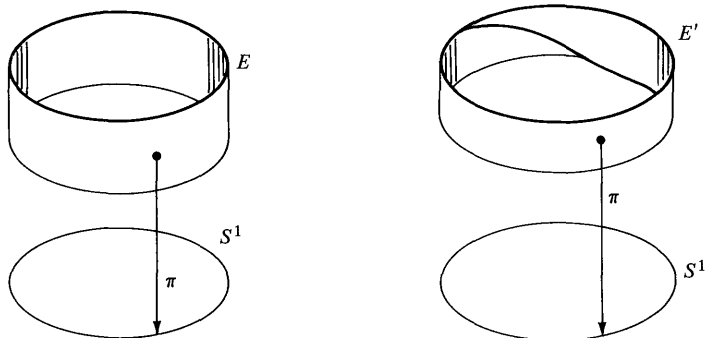
are independent coordinates (degrees of freedom). Such a larger space containing the original space and also the field is an example of a fiber bundle.

An additional generalization is needed to make the idea of a fiber bundle useful. The values of a real-valued function at different points can be directly compared. The local rate of change is the geometric structure, known as the gradient. We need to consider fields for which the field values at different points are not directly comparable. The canonical example of this is a vector field. Unless additional structure (representing new physics) is added to the manifold, vectors at different points cannot be compared. Thus we will demand that the field values all lie in spaces of similar structure, but not that they all lie in the same space. Thus you should look at Figure 12.2 as a two-parameter collection of copies of  $\mathbb{R}$ . The idea that a given field value is associated with a particular point of the base space is crucial to the idea of a bundle.

A fiber bundle is a pair of manifolds, say,  $E$  and  $B$ , and a projection map  $\pi: E \rightarrow B$ . The space  $B$ , the base space, is to be the space on which the field is defined;  $E$  is the space in which we graph the field; and  $\pi$  is the map that assigns to each point on the graph the point in  $B$  where that field variable is defined. For any point  $b \in B$ , the space  $\pi^{-1}(b)$  is called the fiber over  $b$ . It is the space of field variables. Usually these fibers will have further geometric structure. If they are all vector spaces, then we refer to  $\pi: E \rightarrow B$  as a vector bundle, and so on.

A particular field is called a *section* and is given by a map

$$\Gamma: B \rightarrow E$$



**Figure 12.3** (left). An  $\mathbb{R}$ -bundle over the circle.

**Figure 12.4** (right). Another  $\mathbb{R}$ -bundle over the circle. This one has a different topology from the one shown in Figure 12.3.

such that

$$\pi \circ \Gamma(b) = b.$$

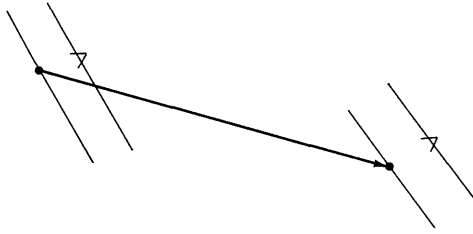
The map  $\Gamma$  assigns to each point a field quantity at that point.

The demand that the space  $E$  be a manifold lets us discuss the smoothness of curves in  $E$  and also the smoothness of sections. In contrast to this local orderliness, a bundle allows more large-scale freedom; this is crucial for the applications of these bundles, say, in gauge theories. In Figures 12.3 and 12.4, I sketch two different bundles of the type  $\pi: E \rightarrow S^1$ . In both the fibers are just  $\mathbb{R}$ . They differ because the second has a global twist. Notice that the second bundle has the property that every continuous section has a zero. The first bundle has the topology of a Cartesian product,  $S^1 \times \mathbb{R}$ . Such a bundle is called *trivial*. A fiber bundle must have the property that around each point in the base space there is a region  $U$  in the base space such that  $\pi^{-1}(U)$  is a trivial bundle. The set  $\pi^{-1}(U)$  is the set of points in the bundle that the map  $\pi$  sends onto  $U$ .

### The cotangent bundle

The simplest bundle of interest is the one that represents a 1-form field on a manifold. The fibers will be the various cotangent spaces. The cotangent bundle of a space  $M$  is written  $T^*M$ . Although each 1-form represents a gradient in the infinitesimal neighborhood of a point, not every 1-form field is the gradient of a function. If  $q$  are coordinates for the points in  $M$ , and the components of the 1-form  $\omega$  are  $p$ ,





**Figure 12.5.** A tangent vector to the cotangent bundle, using the representation where the tangent vectors are drawn right on the manifold.

$$\omega = p dq = p_\alpha dq^\alpha, \tag{12.1}$$

then reasonable coordinates for the point in  $T^*M$  that is the 1-form  $\omega$  at the point  $q$  are  $(q^1, \dots, p_1, \dots)$ . This index placement looks peculiar, but it will keep the summation indices paired up and one down. This use of  $p$  for the 1-form coordinates will fit in with the applications both to contact manifolds and to classical mechanics.

**Example:** Figure 12.2 represents a section of the cotangent bundle  $T^*R^2$ .

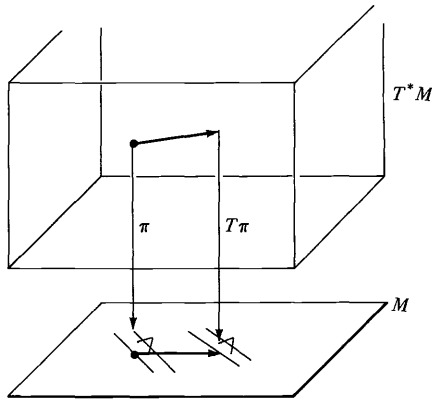
*Natural structures*

$T^*M$  has more structure than just its manifold structure. There is a natural way to lift a function on  $M$  into a section of  $T^*M$ . Just assign to each point of  $M$  the 1-form that is the gradient of the function.

Cotangent bundles have even more structure. In Figure 12.5, I sketch a tangent vector to  $T^*M$ . In Figure 12.6, I sketch the tangent bundle itself. Note that this picture of a tangent vector suggests a natural operation. A tangent vector in  $T^*M$  can be pushed forward by the projection map to give a tangent vector in  $M$ , and from the base point of the tangent vector in  $T^*M$  we find a 1-form on  $M$ . Let them act on each other. From each tangent vector in  $T^*M$  we can thus form a number. This operation is the map

$$\dot{q} \frac{\partial}{\partial q} + \dot{p} \frac{\partial}{\partial p} \mapsto p\dot{q}.$$

It is a linear operator on tangent vectors, obviously, and so must be represented by a 1-form on  $T^*M$ ; this 1-form is an element of the cotangent space to the cotangent bundle. By inspection this 1-form is



**Figure 12.6.** A geometric sketch of the entire cotangent bundle. The fibers are lacking one dimension.

$$\theta = p dq.$$

There is some chance of confusion here. The 1-form  $p dq$  in the preceding equation is a point in the space  $T^*(T^*M)$ . The 1-form  $p dq$  in equation 12.1 is a point in  $T^*M$ . The degenerate notation comes from our using  $q$  in two different ways. As long as we are careful to know what space we are using, we can exploit this degenerate notation without danger.

The 1-form  $\theta$  on  $T^*M$  is called the canonical 1-form on  $T^*M$ . It is independent of the coordinates  $q$  used on  $M$ . To see this, we can translate our verbal description into a coordinate-free prescription for the action of  $\theta$ . At the point  $(a, \omega)$ , the action of  $\theta$  on a vector  $v$  is given by

$$\theta \cdot v = \omega \cdot (\pi_* \cdot v).$$

The projection map  $\pi: (a, \omega) \mapsto a$  is used to push forward the tangent vector  $v$  from  $T^*M$  to  $M$ .

The 1-form  $\theta$  is related to the natural lift we have been discussing. The pullback of the 2-form

$$dp \wedge dq$$

onto a lifted section vanishes.

The cotangent bundle most familiar to you is, no doubt, the phase space of time-independent Hamiltonian mechanics. The canonical 1-form  $p dq$  comes up repeatedly. The full geometric significance of this

1-form involves the exterior derivative, and will be developed in Chapter IV. There are very close relations between cotangent bundles and the contact manifolds that are to be developed in the next chapter and that form the basic structure of mechanics and classical thermodynamics.

### 13. Tangent bundle

The next important space for us is the tangent bundle. This is what the physicist calls velocity space. It is the collection of all the tangent spaces of a manifold. The tangent bundle of  $M$  is written  $TM$ . A point in  $TM$  is a pair, a point in  $M$  and a tangent vector at that point.

**Example:** In Figure 13.1 I sketch three points in  $T\mathbb{R}^2$ ; in Figure 13.2 I draw two curves in  $T\mathbb{R}^2$  in this representation.  $T\mathbb{R}^2$  is  $\mathbb{R}^4$ . It is better for many uses to use a representation on  $\mathbb{R}^2$  itself rather than attempt a four-dimensional drawing.

$TM$  is a manifold. An atlas for it can be generated from the atlas of  $M$  by taking the Cartesian product of a chart for  $M$  and an open set in  $\mathbb{R}^n$ . Globally the tangent bundle need not have the structure of a Cartesian product.

**Example:**  $TS^2$  is not  $S^2 \times \mathbb{R}^2$ ; otherwise we could find a vector field on  $S^2$  that does not vanish anywhere.

Since  $TM$  is a manifold, it has its own tangent space, and so on.

**Example:** Tangent vectors to  $T\mathbb{R}^2$  can be represented as pairs of points, because  $T\mathbb{R}^2$  happens to be a vector space. In Figure 13.3 I draw two tangent vectors in  $T(T\mathbb{R}^2)$  and add them. The base points follow ordinary vector addition. The vectors have to be added relative to the vector at the base point. Use the affine midpoint construction, equation 1.1, for this.

#### *Natural coordinates*

A coordinate chart ( $q$ ) on a manifold  $M$  generates natural coordinates on  $TM$ . Call the tangent vector components  $\dot{q}$ , using the dot as a mnemonicly convenient part of the symbol, not a time derivative. It is to remind us that  $T_q(M)$  is the space of rates. This was Newton's notation for fluxions, which is what these are. Leibnizian partials, on the other

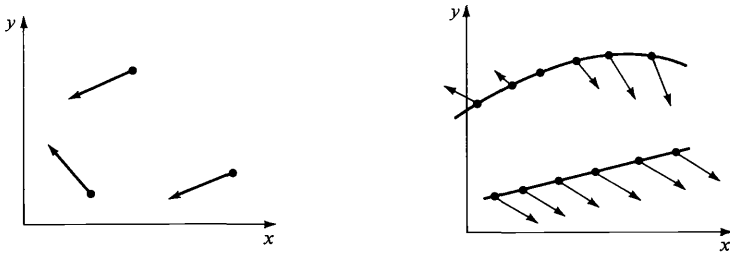


Figure 13.1 (left). Three points in the tangent bundle to  $\mathbb{R}^2$ .

Figure 13.2 (right). Two curves in  $T\mathbb{R}^2$  in the same representation as Figure 13.1.

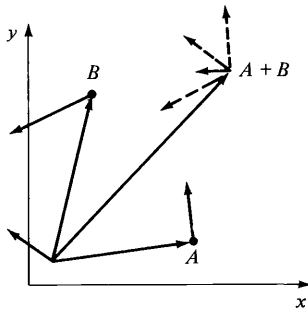


Figure 13.3. Two tangent vectors to  $T\mathbb{R}^2$  and their sum.

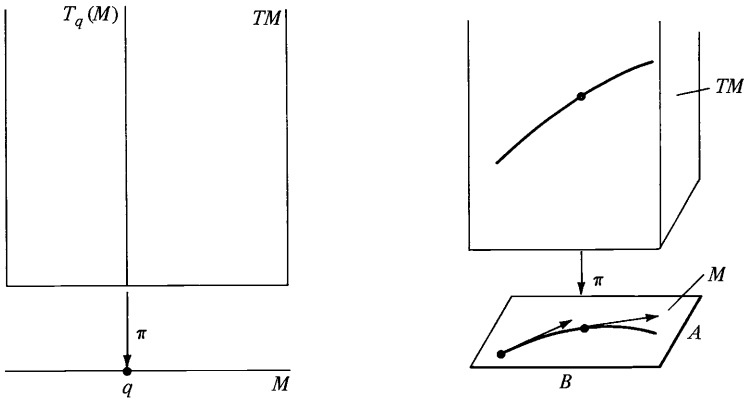
hand, are cotangent space objects. In one dimension these can be, and were, confused. The point in  $TM$  with coordinates  $(q, \dot{q})$  is the vector

$$u = \dot{q} \frac{\partial}{\partial q} = \dot{q}^\alpha \frac{\partial}{\partial q^\alpha},$$

at the point in  $M$  with coordinates  $q$ . This is the same as using coordinates  $p$  for the 1-form  $p dq$  in the cotangent bundle. There is again the convenient abuse of notation here. I am using the  $q$  both for the coordinates on  $M$  and for half the coordinates on  $TM$ . It should cause no trouble.

### Structure of $TM$

These tangent bundles have more structure than just that of a manifold. There is a natural projection,



**Figure 13.4** (left). The bundle structure of  $TM$ .

**Figure 13.5** (right). Lifting a curve from  $M$  to  $TM$ .

$$\pi: TM \rightarrow M; (q, \dot{q}) \mapsto (q).$$

See Figure 13.4. This projection is what lets us call it a *bundle*. The inverse images of this projection are just the tangent spaces

$$\pi^{-1}(q) = T_q(M).$$

These tangent spaces are the fibers of the tangent bundle.

A tangent bundle has yet more structure. If we are given a curve in  $M$ , there is a natural curve in  $TM$  given by

$$s \mapsto \left( Q(s), \frac{dQ}{ds}(s) \right)$$

in our natural coordinates. This is called *lifting* the curve into the tangent bundle (Figure 13.5). It is similar to lifting a function into a section of the cotangent bundle. Not every curve in  $TM$  is the lift of some curve in  $M$ .  $TM$  has structure given by this special class of curves. A curve  $\gamma: \mathbb{R} \rightarrow TM; s \mapsto (Q, \dot{Q})$  is the lift of a curve in  $M$  if we have

$$\dot{Q} = dQ/ds.$$

The coordinate-free statement of this is that for the map

$$\gamma: \mathbb{R} \rightarrow TM; s \mapsto (m, v),$$

where  $v$  is a tangent vector at  $m$  in  $M$ ,

$$\pi_* \cdot \dot{\gamma} = v.$$

The geometric structure of  $TM$  is more subtle than that of  $T^*M$ . Although  $p \, dq$  turned out to be *both* a definition of the coordinates  $p$  and a 1-form on  $T^*M$ , this is not true for

$$v = \dot{q} \frac{\partial}{\partial q},$$

which only defines the coordinates  $q$ . It is not a vector field on  $TM$ ; that is, using different coordinates  $q$  for  $M$  leads to different vector fields  $\dot{q} \partial/\partial q$  on  $TM$ . A counterexample appears in the next section.

Further geometric structure for  $TM$  is the vertical-rotation map that takes vectors in  $TM$  and rotates them up into the fiber. In coordinates this is

$$a \frac{\partial}{\partial q} + b \frac{\partial}{\partial \dot{q}} \mapsto a \frac{\partial}{\partial \dot{q}}.$$

To see that this does not depend on coordinates, note that it is just the pushforward map  $\pi_*$ . The disparity between the tangent and cotangent bundles, which might at first offend your sense of duality, comes from the projection map  $\pi$ . For any bundle,  $\pi_*$  can act only on vectors. This breaks the duality.

#### Vector fields

A vector field  $X$  is defined as a smooth map

$$X: M \rightarrow TM$$

such that

$$\pi \circ X = \text{identity}.$$

That is, it is a section of the tangent bundle.

#### Tangent map

Given a map  $f: M \rightarrow N$ , we can collect all the various pushforward maps into a map  $Tf$  called the *tangent map*:

$$Tf: TM \rightarrow TN.$$

Here the point  $(m, u)$  in  $TM$ , that is, the vector  $u$  at the point  $m$ , is mapped according to

$$(m, u) \mapsto (f(m), f_* \cdot u).$$

Note that  $f_* \cdot u$  is a vector in  $N$  at the point  $f(m)$ . In coordinates, if we have

$$f: q \mapsto F,$$

then

$$Tf: (q, \dot{q}) \mapsto \left( F, \frac{\partial F}{\partial \dot{q}^i} \dot{q}^i \right)$$

Here both  $F$  and  $\partial F/\partial \dot{q}^i$  are evaluated at  $q$ . The map  $Tf$  is really just a bookkeeping trick for automatically keeping track of the base points of the pushforward maps.

#### 14. Vector fields and dynamical systems

One important way to add mathematical structure to a manifold is to give a vector field on it. This is the basic framework for ordinary differential equations, dynamical systems theory, and part of the structure of Hamiltonian mechanics.

The local structure of a vector field is quite uninteresting. We can almost always find coordinates such that the vector field is given by  $\partial/\partial x^1$ . Such coordinates need not cover the entire manifold, or even very much of it. Also, there are special points called critical points, around which this cannot be done because the vector field vanishes at the point. Still, the interesting qualitative questions to ask about a vector field are global and quite difficult. Here I am basically interested in establishing just the equivalence between vector fields, systems of ordinary differential equations, and infinitesimal transformations.

##### *Integral curves*

A curve  $\gamma: s \mapsto Q(s)$  is called an *integral curve* of a vector field  $v: M \rightarrow TM$  if the tangent to the curve at every point is equal to the vector field at that point:

$$\dot{\gamma}(s) = \frac{dQ}{ds}(s) \frac{\partial}{\partial q} = v[Q(s)].$$

**Examples:** In  $\mathbb{R}^2$  the vector field

$$v = y \frac{\partial}{\partial x} - (y+x) \frac{\partial}{\partial y}$$

has integral curves satisfying

$$\frac{dX}{ds} = Y(s), \quad \frac{dY}{ds} = -[Y(s) + X(s)],$$

which are equivalent to

$$\frac{d^2X}{ds^2} + \frac{dX}{ds} + X = 0.$$

The integral curves are spirals representing damped harmonic motion.

To see that  $\dot{q}(\partial/\partial q)$  is not a coordinate-independent vector field on the tangent bundle, look at  $\mathbb{R}^2$  in both rectangular and polar coordinates. Project the vector field down onto  $\mathbb{R}^2$ . In rectangular coordinates the vector field

$$\dot{x} \frac{\partial}{\partial x} + \dot{y} \frac{\partial}{\partial y}$$

has integral curves that are Euclidean straight lines. The vector field

$$\dot{r} \frac{\partial}{\partial r} + \dot{\theta} \frac{\partial}{\partial \theta}$$

has some circular integral curves, those having  $\dot{r} = 0$ . Thus the two vector fields must be different, because they have qualitatively different integral curves.

A vector field and an autonomous system of first-order ordinary differential equations are equivalent. An autonomous system is one in which the dependent variable does not appear explicitly. Any system can be written as an autonomous system by adding one more dimension, taking the dependent variable as a new independent variable whose rate of change is set equal to unity.

### Flows

Well-behaved vector fields have integral curves that can be extended for all values of the parameter  $s$ . For simple systems we can sometimes even find general expressions that describe all the integral curves. If the integral curves exist for all parameter values, backward and forward, then we can define a transformation of  $M$  called the *flow* of the vector field as

$$\Phi: M \times \mathbb{R} \rightarrow M; (q, s) \mapsto \Phi_s(q),$$

such that for a fixed point  $q$  in  $M$  the curve

$$s \mapsto \Phi_s(q)$$

is an integral curve.



Since  $s$  does not appear explicitly, we have

$$\Phi_s \circ \Phi_t = \Phi_{(s+t)},$$

and these transformations form a one-parameter group. A vector field is often called an infinitesimal transformation. Note that an infinitesimal transformation is *not* a transformation; it is the limiting behavior of a family of transformations. Limits usually are new objects. A smooth family of unparametrized curves, one passing through each point of a space, is called a *congruence*.

### *Dynamical systems*

A dynamical system is a physical system that can be described by a point in a state space, and whose future evolution is determined by the location of that point. These systems are thus described by first-order ordinary differential equations, i.e., vector fields. These state spaces are sometimes  $\mathbb{R}^n$ , but more often are manifolds of more general type. When we study dynamical systems, we are more interested in global questions of topology, stability, and qualitative behavior than in the study of ordinary differential equations. For an elementary but sophisticated introduction, see Hirsch and Smale (1974).

**Example:** A dynamical system described by a time-independent Hamiltonian  $H$  is represented by the vector field

$$v = \frac{\partial H}{\partial p} \frac{\partial}{\partial q} - \frac{\partial H}{\partial q} \frac{\partial}{\partial p}$$

on a cotangent bundle  $(q, p)$ . These will be studied in Sections 50, 51, and 52. Time-independent Hamiltonian systems live naturally in the cotangent bundle.

### *Singular systems*

Even if the true state space of a dynamical system is  $\mathbb{R}^n$ , some of the dynamics may operate so fast that they appear in the description of the system only as a constraint. Useful approximations for such systems often live on manifolds.

**Example:** Look at the singular dynamical system on  $\mathbb{R}^2$

$$w = y \frac{\partial}{\partial x} + \lambda(y - y^3 - x) \frac{\partial}{\partial y}$$

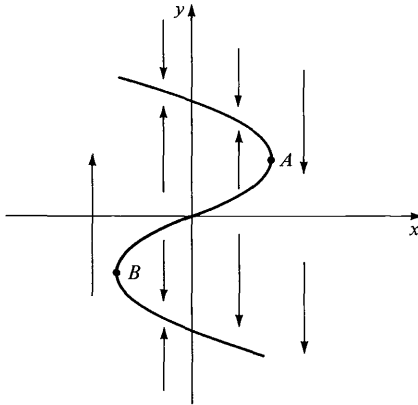


Figure 14.1. The vector field and the slow manifold.

for large values of the parameter  $\lambda$ . The vector field has an enormous component

$$\lambda(y - y^3 - x) \frac{\partial}{\partial y}$$

that acts to move the system rapidly unless the condition

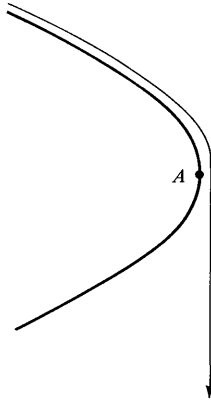
$$x = y - y^3 \tag{14.1}$$

is satisfied. The action of the fast dynamics is to “snap” the system onto the set that satisfies equation 14.1, which is sometimes called the slow manifold. After rapidly snapping onto the slow manifold, the system then crawls slowly along it.

The dynamics on the slow manifold is described by the slow vector

$$u = y \frac{\partial}{\partial x}$$

if we use  $x$  as a local coordinate for pieces of the slow manifold. From the sketch of these vector fields in Figure 14.1, we see that the system faces a crisis at points  $A$  and  $B$ . To see what happens there, we must look more closely at the approach of the system to these points. In Figure 14.2 we show the details of the approach of the system to point  $A$ . The system is riding a small distance  $O(1/\lambda)$  above the slow manifold. It does this to pick up the small downward component needed to stay on the slow manifold. You might



**Figure 14.2.** The crisis point  $A$ . From here the system snaps down onto the lower branch of the slow manifold.

worry about stability and oscillations about this path. Remember that these are first-order systems; the dynamics has no inertia, and is more Aristotelian than Newtonian. Upon reaching the neighborhood of point  $A$ , the system overshoots and snaps down to the lower curve. It continues along this curve until it hits  $B$ , snaps up, and so continues around in what is called a limit cycle. This system is called a relaxation oscillator.

Most physical systems have dynamical processes that occur on a large variety of time scales. The ones that are too fast to be followed in detail are put in as constraints. The hiker does not worry about the electrical forces that act to keep his feet from penetrating the ground. Some processes are too slow to be of any interest. Dynamical systems theory studies those with interesting intermediate time scales. This is the way that strange manifolds can arise in mechanics, although dynamical systems often pick out sets that are much less well-behaved than manifolds.

**Examples:** The motion of a rigid body takes place on a state space that is the tangent bundle of the orthogonal group described earlier. We ignore the fast interactions that hold the rigid body together, thus we eliminate about  $10^{24}$  unnecessary equations.

A single first-order partial differential equation is equivalent to a vector field whose integral curves are called characteristics. The full

theory of this takes more technical machinery than we have yet developed, but let us consider here a single linear partial differential equation. It can be put in the form

$$f \frac{\partial u}{\partial q} + au = 0,$$

where  $a$  and the  $f$ 's can depend on the position coordinates  $q$ .

Consider the vector field

$$v = f \frac{\partial}{\partial q}.$$

The partial differential equation gives the rate of change in this direction. The behavior of  $u$  along an integral curve of this vector field (called a characteristic) is known and does not depend on what happens at points off the curve. The partial differential equation breaks up into ordinary differential equations along the characteristics. We will use this in Section 16 and return to the full theory in Section 32.

## 15. Contact bundles

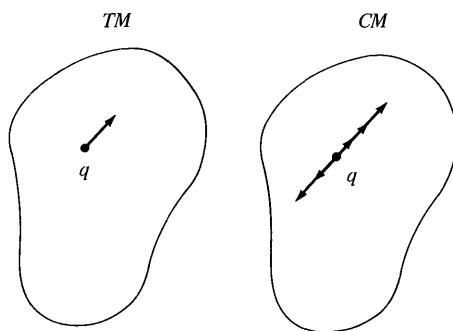
There is a manifold that describes the local behavior of unparametrized curves, just as the tangent bundle describes parametrized curves. This type of manifold is called a *contact bundle*. There is also a contact bundle that represents the direction of the local gradients of functions without regard to strength; it is an analog of the cotangent bundle. In between there are lots of other contact bundles that represent the calculus of maps of various dimensions. These contact bundles are very useful, although they are not usually given sufficient attention by either mathematicians or physicists.

### *Line-element contact bundles*

Locally an unparametrized curve is just a one-dimensional submanifold. The local approximation to this is called a *line element*; it is like a tangent vector of unspecified length. Be careful: books written in "old tensor" unfortunately call a metric tensor a line element. If you wish to be precise, a line element is an equivalence class of curves, and is also, therefore, an equivalence class of tangent vectors, say,  $[v]$ , under

$$v \sim kv,$$

where  $k$  can be positive or negative but not zero.



**Figure 15.1.** A comparison of the tangent bundle  $TM$  and the line-element contact bundle  $CM$ .

The line-element contact bundle that I call  $CM$  consists of pairs, a point in the manifold  $M$ , and a line element at that point (Figure 15.1).

**Example:** The contact bundle  $CR^2$  was discussed in Sections 7 and 8.

There are several natural coordinate systems for the  $CM$ . One of them uses homogeneous coordinates for the line element. For any line element  $[v]$ , represented by a vector

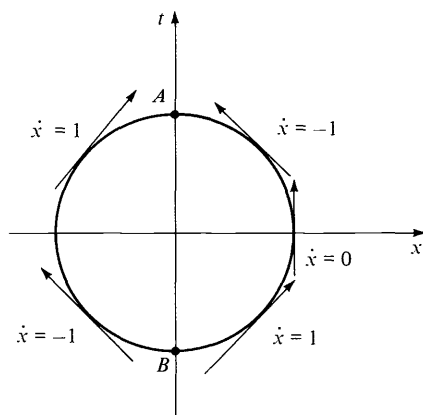
$$v = \dot{q} \frac{\partial}{\partial q}$$

we use the homogeneous coordinates  $[\dot{q}^1 : \dot{q}^2 : \dots]$ . Although homogeneous coordinates preserve the symmetry of the coordinates and do not require numerous coordinate charts, they do not provide a natural basis for the tangent and cotangent spaces of  $CM$ .

We find another natural coordinate system for  $CM$  if we arbitrarily select one homogeneous coordinate and set it equal to 1. We will call the special coordinate  $t$  and the others  $q$ . Our unparametrized curve can be written in terms of  $(m-1)$  functions  $Q(t)$ , and the derivatives

$$\dot{q} = \frac{dQ}{dt}$$

describe the line element. Despite their lack of symmetry, these  $(t, q, \dot{q})$  are the most useful coordinates for  $CM$ . Although these coordinates do not cover all of  $CM$ , they leave out only points that can be easily added



**Figure 15.2.** Line elements tangent to the circle as represented by special tangent vectors.

on the edges. A complete atlas can be formed from several such charts, each one selecting a different special coordinate.

**Example:** The line elements tangent to a circle in  $\mathbb{R}^2$  can all be described in these natural coordinates, except for the points  $A$  and  $B$  shown in Figure 15.2.

### *Hypersurface-element contact bundles*

Another important class of contact bundles are the hypersurface-element contact bundles of manifolds. They are closely related to the cotangent bundles. We will denote the hypersurface-element contact bundle of  $M$  by  $C^*M$ . Be careful: this is not the dual of  $CM$ . Elements of  $C^*M$  are pairs: a point in  $M$  and a hypersurface contact element at the point. Elements of  $T^*M$ , on the other hand, are pairs of points and 1-forms. A hypersurface contact element is a ray of 1-forms, all having the same alignment but different magnitudes (Figure 15.3).

We can find convenient coordinates for  $C^*M$  by again breaking the symmetry of the coordinates. Pick one dimension, call it  $f$ , and describe a hypersurface implicitly by giving  $f$  as a function of the other coordinates  $q$ . The local behavior of the hypersurface is given by the partial derivatives

$$p = \frac{\partial f}{\partial q}$$

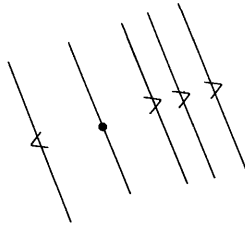


Figure 15.3. A ray of 1-forms.

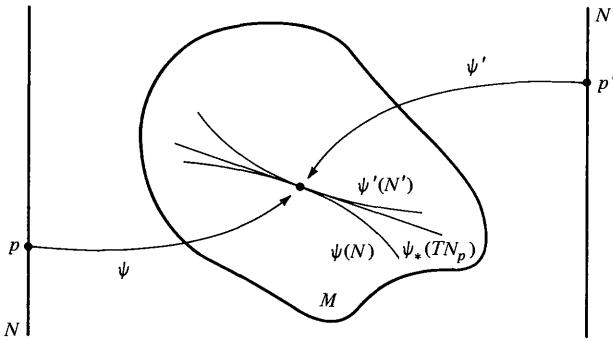


Figure 15.4. Two submanifolds having first-order contact.

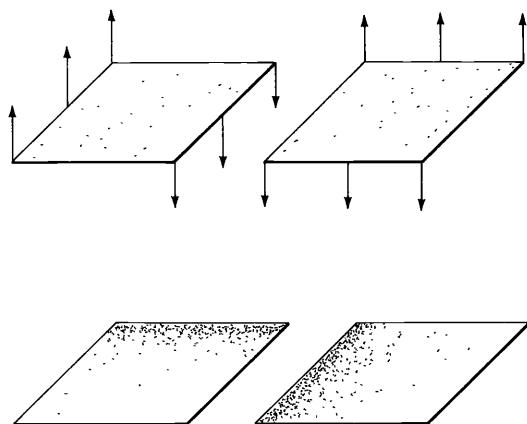
and we maintain notational consistency by putting the index on  $p$  down rather than up. A coordinate chart for most of  $C^*M$  is given by  $(q, f, p)$ .

*Contact bundles*

A submanifold is a subset of a manifold that is itself a manifold. It can be represented by a map  $\psi: N \rightarrow M$  such that both  $\psi$  and  $T\psi$  are  $1:1$ . The pair  $(N, \psi)$  is often called a submanifold. First-order contact (which I will just call contact) between two submanifolds  $(N, \psi)$  and  $(N', \psi')$  at a common point  $\psi(p) = \psi'(p')$  means that

$$T\psi[T_p(N)] = T\psi'[T_{p'}(N')],$$

as shown in Figure 15.4.  $T\psi$  is the tangent map described in Section 13. The arguments of the maps are sets, not points; equated on the two sides are the sets resulting from transforming all the points of each tangent space. The equality is not point by point. This would make no sense. An



**Figure 15.5.** The four degrees of freedom of a 2-surface contact element in  $\mathbb{R}^4$ . The fourth dimension is indicated here by the shading.

equivalence class of submanifolds in contact at a point  $q \in M$  will be called a *contact element* at  $q$ . We will write this equivalence class  $[N, \psi]$ . A contact element is equivalent to a linear subspace of the tangent space.

This notion of contact is a weaker equivalence relation than tangency. Two curves are tangent if their tangent vectors are equal; the curves are in contact if their tangent vectors agree in direction but not necessarily in magnitude or sign.

For any given manifold  $M$ , and a given dimension  $n$  for the submanifold  $N$ , the  $n$ -element contact bundle of  $M$ ,  $C(M, n)$ , is the manifold of pairs, consisting of a point in  $M$  and an  $n$ -contact element at that point.

**Example:** Look at 2-surface elements in  $\mathbb{R}^4$ . The 2-element contact bundle here has eight dimensions. The four degrees of freedom of the contact element are sketched in Figure 15.5. The fourth dimension is indicated by the shading, and these are three-dimensional perspective drawings.

### *Contact structure*

These contact bundles have two natural geometric structures on them. First, they are bundles, using the natural projection that forgets the contact element. Second, they have what is called a contact structure. Look at  $C(M, n)$ , the bundle of  $n$ -dimensional contact elements to



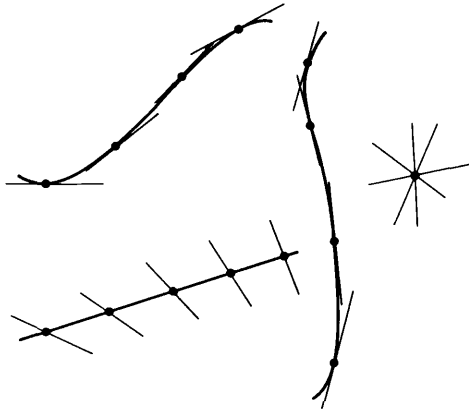


Figure 15.6. Four curves in  $\mathbb{C}\mathbb{R}^2$ . Two of these are lift of curves in  $\mathbb{R}^2$ .

$M$ . For every  $n$ -dimensional submanifold  $(N, \psi)$  in  $M$ , there is a natural lift

$$\sigma: M \rightarrow C(M, n); q \mapsto (q, [N, \psi]).$$

where  $[N, \psi]$  is an  $n$ -dimensional contact element. On the other hand, not every  $n$ -submanifold of  $C(M, n)$  is such a lift. A contact structure picks out these special submanifolds.

We can find a simple chart for most of  $C(M, n)$  by picking  $n$  of the coordinates of  $M$  as special, calling them  $q$ , and looking at the remaining  $(m - n)$  coordinates  $y$  as functions of the  $q$ . The graphs of the functions  $Y(q)$  can represent most of the  $n$ -dimensional submanifolds of  $M$ . Note here the use of upper case for the  $y$  when they are functions rather than coordinates. The contact elements are represented by the matrix of partial derivatives

$$p^a_{\mu} = \frac{\partial Y^a}{\partial q^{\mu}}.$$

Latin indices run over  $n$  labels; Greek indices run over  $(m - n)$  labels. The index placement agrees with that chosen before for  $CM$  and  $C^*M$ , where the  $p$ s were called  $\dot{q}$  and  $p$ ; and the  $y$ s were called  $q$  and  $f$ .

In Figure 15.6, I sketch some curves in  $\mathbb{C}\mathbb{R}^2$ , two that are lifts and two that are not. For any  $m$ -submanifold to be a lift, we must have the  $p$ s equal to the partial derivatives. Look at the 1-forms

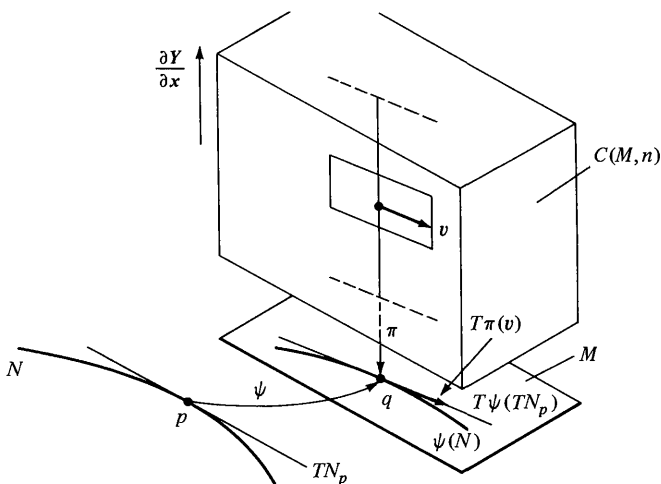


Figure 15.7. The geometric structure of  $CM$ . One fiber dimension has been left out.

$$\theta^a = dy^a - p^a_{\mu} dq^{\mu}.$$

and pull these back onto a submanifold of  $C(M, n)$ :

$$\psi: N \rightarrow C(M, n); q \mapsto (q, Y(q), P(q)).$$

Here we are using special coordinates in  $N$  as well. Pull back the  $\theta$  using

$$\psi^* \cdot dq = dq, \quad \psi^* \cdot dy^a = \frac{\partial Y^a}{\partial q^{\mu}} dq^{\mu},$$

and we have

$$\psi^* \cdot \theta^a = \left( \frac{\partial Y^a}{\partial q^{\mu}} - P^a_{\mu} \right) dq^{\mu}$$

(see Figure 15.7). If this is a lifted submanifold, then these pullbacks all vanish. This is also a sufficient condition. We will see later that these lifted submanifolds are what are called integral submanifolds of the contact ideal.

A contact structure is represented by these 1-forms, which are called contact 1-forms. We must be a little careful here. These  $\theta$  are not coordinate-independent, nor are they continuous over  $M$ . These difficulties are resolved if we form an ideal of 1-forms based on these  $\theta$  and on ones from other charts, and smoothly patch it together by using a partition of

unity. (See Preliminaries for the definition of an ideal.) The ideal is coordinate-independent, and, because we have used a partition of unity to smoothly patch it together, is composed only of smooth 1-forms. This is all usually just a matter of principle. In practice, we generally avoid the problem by picking coordinates such that we do not have to work at a point of discontinuity.

The simplicity of the representation of contact structure by an ideal of differential forms contrasts with the difficulty of recognizing those curves in the tangent bundle that are lifts. This is why we find contact bundles so useful. In addition,  $CM$  is one dimension smaller than  $TM$ .

**Example:** In statistical mechanics, we have a manifold  $M$  with particles going in all possible directions. In the space  $CM$  the worldlines are unscrambled, and we have there a smooth congruence of lifted worldlines. Despite the special significance of  $t$  in the representation of  $CM$ , this is fully compatible with special relativity.

The density of particles per unit volume per unit direction is represented by a  $(2m-2)$ -form,  $\rho$ , on  $CM$ , which has dimension  $(2m-1)$ . Not all congruences in  $CM$  are lifts of particle worldlines. The conditions on  $\rho$  are that

$$\rho \wedge \theta = 0, \quad (15.1)$$

where the  $\theta$  are the contact forms. (This is not supposed to be obvious until you get to Section 23.)

### *Jet bundles*

Given any bundle  $\pi: E \rightarrow B$ , the sections of this bundle define submanifolds of  $E$  whose contact structure is often important. A section, recall, is a map  $\psi: B \rightarrow E$  that commutes with the projection  $\pi$ .

**Example:** Look at functions on the plane that are maps

$$F: \mathbb{R}^2 \rightarrow \mathbb{R}.$$

The graphs of these functions are sections of the bundle

$$\pi: \mathbb{R}^3 \rightarrow \mathbb{R}^2; (s, y, f) \mapsto (s, y).$$

The sections are the maps

$$\psi: \mathbb{R}^2 \rightarrow \mathbb{R}^3; (x, y) \mapsto (x, y, f) = (x, y, F(x, y)),$$

which define 2-submanifolds of  $\mathbb{R}^3$ . Note my convention of using lowercase letters for independent variables, and the corresponding

uppercase letter for that variable as a dependent variable [here,  $f$  and  $F(x, y)$ ].

A contact element at the point  $p = (x_0, y_0, F(x_0, y_0))$  is an equivalence class of 2-submanifolds having first-order contact at  $p$ . This class can be represented by the linear submanifold (in these coordinates)

$$(x, y) \mapsto \left( x, y, F + \frac{\partial F}{\partial x} (x - x_0) + \frac{\partial F}{\partial y} (y - y_0) \right).$$

The partial derivatives  $\partial F/\partial x$  and  $\partial F/\partial y$  are evaluated at  $(x_0, y_0)$ . They form useful coordinates for the contact elements.

The jet bundle is the bundle formed from the original bundle by adding a fiber of contact elements at each point. The structure of the jet bundle is again given by contact forms that specify where the contact elements are in contact with a lifted submanifold.

**Example:** Coordinates for the jet bundle of the bundle  $\mathbb{R}^3 \rightarrow \mathbb{R}^2$  in the preceding example are  $(x, y, f, f_x, f_y)$ . Here  $f_x$  is a single symbol, the name for a particular coordinate, not a partial derivative. The contact 1-form is

$$df - f_x dx - f_y dy.$$

If we pull back this 1-form onto a submanifold

$$(x, y) \mapsto (x, y, F, F_x, F_y),$$

we find

$$df - f_x dx - f_y dy \mapsto \frac{\partial F}{\partial x} dx + \frac{\partial F}{\partial y} dy - F_x dx - F_y dy,$$

which vanishes only if

$$F_x = \frac{\partial F}{\partial x}, \quad F_y = \frac{\partial F}{\partial y}.$$

Since sections can never be tangent to the fibers, these coordinates cover all of the jet bundle.

If the fibers of the original bundle are  $n$ -dimensional, then  $n$  contact 1-forms form the ideal that represents the contact structure. We can look at the coordinates used for contact manifolds as being the result of artificially imposing a bundle structure on the manifold by singling out a

special coordinate. Jet bundles differ from contact bundles in having bundles for their base spaces whose projection map  $\pi$  automatically singles out special directions.

### Problems

- 15.1 Draw a 1-form on  $C\mathbb{R}^2$ . Draw one that is in the contact ideal. Use the representation of Figure 15.1.
- 15.2 Show how to form an atlas for  $CM$  from the coordinates  $(t, q, \dot{q})$ .
- 15.3 Describe the manifold of directed line elements, that is, vectors with a direction (plus or minus) but no particular length. What are the natural coordinates? What is its contact structure?
- 15.4 Show that the contact ideal  $dt - p dq$  on  $C^*M$  does not single out the  $t$  coordinate. Study the question first in  $C\mathbb{R}^2$  and there explicitly rotate coordinates.

### 16. The geometry of thermodynamics

A thermodynamic system is a homogeneous assembly that includes a very large number (say,  $10^{24}$ ) of unexamined and mutually interacting subparts. The energy shared among the unmeasured degrees of freedom of these subparts is called heat. Think, for instance, of a sample of gas. The atoms are subparts with translational degrees of freedom whose kinetic energy is shared among them. In addition, a thermodynamic system has some degrees of freedom (or combinations) that are measured, such as its volume, magnetization, electric charge, and number of particles. These aggregate variables are called *extensive variables*. Along with these extensive variables are forces or potentials that describe the energy transfer into and out of these degrees of freedom. The energy changes associated with volume changes depend on the pressure, changes associated with the magnetization depend on the magnetic field, and so on. We are going to describe thermodynamic systems in a contact bundle whose coordinates are total energy, entropy, temperature, all the extensive variables, and the potentials associated with the extensive variables. These potentials are called *intensive variables*.

Three important geometric structures will be needed to describe a thermodynamic system: a linear structure, a contact structure, and a convexity structure. The linear structure models the physical idea of short-range interactions and the possibility of homogeneous systems with a scaling symmetry. The contact structure incorporates energy conservation and the First Law. The convexity incorporates the Second Law and the increase of entropy when systems are mixed. The proper geometric

setting for thermodynamics is a contact bundle. In this geometric setting we will be able to rationally handle the profusion of different differentials. Some will go to  $dx$ s and some to  $(\partial/\partial x)$ s, a useful distinction lost in the usual notation. None of the funny deltas will remain.

### *Fundamental equation*

The *state space* for the system consists of all the preceding extensive variables, plus one aggregate variable for the unexamined degrees of freedom. This aggregate variable is called *entropy*,  $S$ , and its corresponding potential is called *temperature*,  $T$ . State space is a manifold whose coordinates have physical meaning by themselves.

**Example:** For a system consisting of a fixed amount of an ideal gas, the state space is two-dimensional, with coordinates entropy and volume ( $S, V$ ). A system with a variable amount of gas would need another coordinate,  $N$ , to represent the amount of gas in the system.

The state space describes the current condition of a given system. The system itself is specified by giving the energy stored in the system for all possible states. This relation is called the *fundamental equation*, and gives the internal energy  $U$  as a function of the state variables. It is a complete description of the system.

**Example:** For a system consisting of  $N$  molecules of an ideal gas, the fundamental equation is

$$U(S, V) = N^{5/3} V^{-2/3} \exp(2S/3Nk). \quad (16.1)$$

Here  $k$  is Boltzmann's constant. This equation is probably not familiar to you, and it is certainly not supposed to be obvious. We will discuss the relation of this equation to  $PV = NkT$  in a moment.

### *Linear structure*

Thermodynamics deals with homogeneous systems, systems that can be subdivided into subsystems that behave like the whole. The physical idea is that the system is much larger than the range of its interactions, and that for most atoms (or whatever elementary pieces it has, spins, macromolecules, etc.) these interactions are saturated. The energy stored in them will just be proportional to their number. These systems for which

the shape does not matter are sometimes called limp systems. The internal energy of a subsystem is proportional to its size. This is a scaling symmetry, and it requires that the energy function  $U$  must be homogeneous of degree one in all the extensive variables taken together. Thus we must have, by Euler's theorem, for a fundamental equation  $U(S, V, \dots)$

$$U = S \frac{\partial U}{\partial S} + V \frac{\partial U}{\partial V} + \dots \quad (16.2)$$

The geometric interpretation of this will follow once we have an interpretation of these partial derivatives in a contact manifold. You might wonder how we are to change both the volume and the surface area by the same factor. It is important to realize that the shape of the system is not fixed. If energy were needed to deform the body, then this energy must be included in the fundamental equation, along with additional degrees of freedom describing the shape of the body.

### *Contact structure*

In thermodynamics the potential or force on a given degree of freedom is the rate at which work must be done to change that degree of freedom. Since energy is conserved, which is the First Law of thermodynamics, this work must change the internal energy.

**Example:** Thus for the preceding example we must have for the pressure

$$P = - \frac{\partial U}{\partial V}$$

and for the temperature

$$T = \frac{\partial U}{\partial S}.$$

The force or potential associated with every extensive variable is called an intensive variable. In a system twice the size, these intensive variables are the same, not doubled.

The geometric structure here is clearly that of a surface given by the fundamental equation and its various slopes, the potentials. This is naturally described in a contact bundle. The graph of the fundamental equation is an  $n$ -surface in  $(n+1)$ -space, and the potentials are the components of the contact element to that surface. The contact bundle is the  $(2n+1)$ -space with coordinates of internal energy, which are

the potentials, the degrees of freedom, and a contact ideal generated by the 1-form

$$\alpha = dU + \sum (\text{forces}) d(\text{degrees of freedom}).$$

Again we see that we can geometrize a problem with infinitesimals by finding the correct space in which the infinitesimals become honest 1-forms.

**Example:** The ideal gas system given earlier is described in the five-dimensional contact bundle with coordinates  $(U, T, S, P, V)$  and a contact ideal generated by

$$\alpha = dU - T dS + P dV.$$

The homogeneity condition 16.2 becomes

$$U = TS - PV,$$

and is known as the Gibbs–Duhem relation.

A thermodynamic system in the contact bundle is an  $n$ -dimensional integral submanifold of the contact ideal. The pullback of  $\alpha$  onto the system submanifold must be zero.

### *Convexity*

To complete the geometric interpretation of classical thermodynamics, we must include the constraint of the Second Law. This appears in our geometric structure as the requirement that the fundamental equation for the energy function  $U$  must be a convex function, which implies that, in the fundamental equation,  $U$  must be a convex function of its arguments. There is an excellent discussion of the geometric role of convexity in the introduction by A. S. Wightman to R. B. Israel (1979).

### *Equation of state*

A thermodynamic system is usually specified not by its fundamental equation, but by an equation of state describing the potentials. Such a relation is more convenient, but contains less information.

**Example:** The equation of state for an ideal gas is the familiar relation

$$PV = NkT.$$



To fully describe the system, we also need

$$U = (3/2)NkT.$$

The system in the contact manifold will be described by the map  $\psi: \mathbb{R}^2 \rightarrow \mathbb{R}^5$ ,

$$\psi: (S, V) \mapsto (U, T, S, P, V) = (U(S, V), T(S, V), S, P(S, V), V).$$

The three functions  $U$ ,  $T$ , and  $P$  must satisfy the first two equations, as well as

$$\psi^* \cdot \theta = 0.$$

We can write our map as

$$(S, V) \mapsto (3NkT/2, T, S, NkT/V, V),$$

involving now the single function  $T(S, V)$ . For the pullbacks we have

$$\begin{aligned} \psi^* \cdot dU &= (3/2)Nk \left( \frac{\partial T}{\partial S} dS + \frac{\partial T}{\partial V} dV \right), \\ \psi^* \cdot dS &= dS, \quad \psi^* \cdot dV = dV; \end{aligned}$$

so

$$\begin{aligned} \psi^* \cdot \theta &= (3/2)Nk \left( \frac{\partial T}{\partial S} dS + \frac{\partial T}{\partial V} dV \right) \\ &\quad - T dS + (NkT/V) dV = 0. \end{aligned}$$

$T$  must satisfy

$$\frac{1}{T} \frac{\partial T}{\partial S} = 2/3Nk, \quad \frac{1}{T} \frac{\partial T}{\partial V} = -2/3V.$$

The first gives us

$$\ln T = 2S/3Nk + C(V)$$

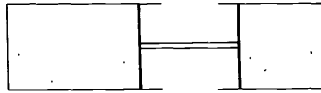
and the second determines the function  $C(V)$ , and leads to

$$T = AV^{-2/3} \exp(2S/3Nk).$$

This leads to the fundamental equation given in equation 16.1. The factor of  $N^{5/3}$  ensures that a system of twice the volume and with twice the number of molecules has twice the energy.

### Interactions

The degrees of freedom of different thermodynamic systems can be coupled. We show this in Figure 16.1 for two ideal gas systems. Equilibrium



**Figure 16.1.** Two thermodynamic systems, here gases, with their volume degrees of freedom coupled.

demands that the respective potentials be equal. Two containers of gas in contact must be at the same pressure. If they can also exchange energy between their thermal degrees of freedom, then in equilibrium they must also be at the same temperature. We often measure the properties of a system by allowing it to interact with special systems such as pressure regulators, volume regulators, heat baths (temperature regulators), and batteries (electric potential regulators).

#### *A magnetic system*

Look at a magnetic system with state variables of entropy  $S$  and magnetization  $M$ . The contact structure here is

$$\alpha = dU - TdS - HdM.$$

Some of these systems can be described by the empirical Curie law

$$M = aH/T,$$

where a magnetic field magnetizes the system, but an increased temperature disorders the system and reduces the magnetization. Suppose further that the thermal properties in zero magnetic field are described by a heat capacity

$$C_H = (\beta T)^3, \quad (16.3)$$

typical of a crystalline substance storing energy in lattice vibrations.

We can find the fundamental equation at zero field,  $U(S)$ , from the given heat capacity. We work in the contact bundle  $(U, T, S)$ , which describes the system at zero magnetic field. The system lies along a curve in  $(U, T, S)$  space, which we parametrize by  $S$  itself:

$$(U, T, S) = (U(S), U_S(S), S).$$

The subscripts here denote partial derivatives. The tangent vector is

$$U_S \frac{\partial}{\partial U} + U_{SS} \frac{\partial}{\partial T} + \frac{\partial}{\partial S}.$$

The heat capacity is the rate at which  $U$  changes relative to the rate at which  $T$  changes along this line. Equation 16.3 gives us

$$U_S/U_{SS} = (\beta U_S)^3.$$

A solution to this is

$$U(S) = (3S)^{4/3}/4\beta. \quad (16.4)$$

For nonzero magnetic field, the Curie law is a partial differential equation for  $U$ :

$$aU_M = U_S M,$$

that is,

$$\left( a \frac{\partial}{\partial M} - M \frac{\partial}{\partial S} \right) U = 0.$$

This states that  $U$  is constant along the parabolas

$$(S - S_0) = -M^2/2a.$$

$S_0$  parametrizes the parabolas in terms of their intersection with the  $S$  axis. This is the method of characteristics mentioned in Section 14. The value of  $U$  at the  $M = 0$  axis is given by equation 16.4, and so we have the fundamental equation for our magnetic system

$$U(S, M) = (1/4\beta)(3S + 3M^2/2a)^{4/3}.$$

The temperature is given by

$$T = (1/\beta)(3S + 3M^2/2a)^{1/3}.$$

Such a system can be used as a refrigerator. We start with the system unmagnetized at some temperature  $T_0$ . While keeping it in contact with a heat bath at this temperature, we increase the magnetic field strength to  $H$ . Heat flows out, and the entropy changes to

$$S = (\beta T_0)^3/3 - aH^2/2T_0^2.$$

Now insulate the system from the heat bath and reduce the field to zero. No heat flow means that  $S$  is constant. The final temperature is lower and satisfies

$$(T/T_0)^3 = 1 - 3aH^2/2\beta^3 T_0^5.$$

Of course, the Curie law is not true down to  $T = 0$ , but this procedure, called adiabatic demagnetization, can in practice cool systems from a few degrees Kelvin down to  $10^{-3}$  or so.

We will continue the discussion of classical thermodynamics in Sections 20 and 37.

# III

## *Transformations*

Transformations are maps of a set into itself. The symmetries of a mathematical structure are given by those transformations that do not change the structure. We can even define a mathematical structure by starting with the transformations, and finding the mathematical structure compatible with these transformations. Thus we can define Euclidean geometry as the study of those mathematical structures invariant under orthogonal transformations. Similarly, the mathematics of special relativity can be defined as the study of all structures compatible with Lorentz transformations. This approach is quite abstract and far removed from the world of experiment. Still, it is an interesting approach.

We can also use symmetry transformations to solve problems. Sometimes a problem can be transformed into a simpler problem. If the structure of the problem is invariant under the transformation, then the solution of the simpler problem gives us a solution of the original problem. At one time it was hoped that all problems in mechanics could be solved by finding transformations that reduced them to trivial problems. As might have been expected, such transformations were even harder to find than solutions of the original problem.

The prime tool in the study of transformations is the infinitesimal transformation, described by a vector field. The operation of this vector field on geometric objects is given by the Lie derivative, one of the most important ideas in this book. Like contact structure, it is sometimes slighted by both mathematicians and physicists.

### **17. Lie groups**

A *Lie group* is a set that has the structure of both a group and a manifold, and for which the rules of group multiplication are given by smooth

functions. The classification of such groups is an involved mathematical field; fortunately, we do not need it here. A scant familiarity with a few of the simpler groups will do. We usually discuss Lie groups in terms of their matrix representations (and now we must be careful, because in group theory “representation” has a technical meaning). Discussions of Lie groups for physicists focus on the computational details of the representations – again, detail that we do not need here.

**Example:** Without a doubt, the most important group to a physicist is the orthogonal group in three dimensions, called  $O(3)$ . This is the set of all rotations and inversions in three dimensions that preserve the lengths according to the Euclidean metric. The argument of Section 11 showed that all the orthogonal groups are manifolds. One coordinate chart for part of  $O(3)$  would be the Euler angles, although these are really of limited use. Group multiplication is just matrix multiplication. Since it involves only polynomials, it is clearly smooth. Thus the  $O(n)$  are Lie groups.

It is not altogether clear to me why Lie groups are so important. For example, they are the fundamental building blocks for gauge theories. One additional structure that groups have is a set of global translations. For a given group  $G$  and element  $a$  in the group, the map

$$L_a: G \rightarrow G; b \mapsto ab$$

is called left translation. It is a transformation of the group, and it leaves no element unmoved.

### Transformation groups

For our purposes the most important use of Lie groups is to represent the symmetries of mathematical structures on manifolds. We want transformations that leave the structure of interest unchanged.

We have a Lie group *action* on a manifold  $M$  if we have a map

$$\mu: G \times M \rightarrow M; (a, q) \mapsto \mu_a(q),$$

where  $G$  is a Lie group, satisfying

$$\mu_e(q) = q$$

where  $e$  is the identity element in  $G$ , and

$$\mu_a \circ \mu_b = \mu_{ab}. \tag{17.1}$$

This last rule could have been given in the other order. Our action is called a *left action*.

**Examples:** The left translation just defined is an action of a Lie group upon itself.

The proper Lorentz group acting on a two-dimensional spacetime can be represented by the one-parameter family of matrices

$$\begin{bmatrix} \cosh \psi & \sinh \psi \\ \sinh \psi & \cosh \psi \end{bmatrix}.$$

The group is one-dimensional, and the group manifold is just  $\mathbb{R}$ . The identity element is given by  $\psi = 0$ . We have an action on  $\mathbb{R}^2$

$$(t, x) \mapsto (\cosh \psi t + \sinh \psi x, \sinh \psi t + \cosh \psi x).$$

The composition of two transformations is just the addition of the rapidities,  $\psi_a + \psi_b$ . Relation 17.1 is easily verified.

### *Lie algebra*

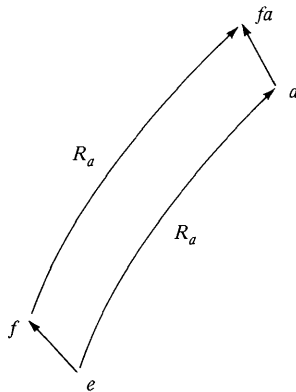
The behavior of a Lie group is determined largely by its behavior in the neighborhood of the identity element  $e$ . The behavior can be represented by an algebraic structure on the tangent space of  $e$ , called the *Lie algebra* of the group. Each element of the tangent space of the identity corresponds to an infinitesimal transformation. On the manifold  $M$  on which the group acts, these infinitesimal transformations are represented by vector fields. The Lie bracket of Section 9 gives these vector fields an algebraic structure. The structure depends only on the Lie group, as you can see by using equation 17.1.

Any Lie group acts on itself by left translation, and the vector fields generated by this action are a convenient representation of the Lie algebra. These vector fields are invariant under right translation:

$$TL_a(e, v(e)) = (a, v(a)),$$

where  $TL_a: TG \rightarrow TG$  is the tangent map of Section 13.

**Examples:** This relation is easily seen heuristically by thinking of the head and tail of a vector as points in the Lie group. Figure 17.1 shows that the infinitesimal left translation generated by  $f$  takes  $a \mapsto fa$ .  $R_a$ , the finite right translation by the element  $a$ , takes  $e$  to  $a$  and  $f$  to  $fa$ .



**Figure 17.1.** The finite right translation of an infinitesimal left translation, which is a vector field.

This heuristic trick can be justified by going to a matrix representation of the Lie group, and using the matrix exponential to generate a one-parameter subgroup satisfying equation 17.1.

It is a famous result called the hairy-ball theorem that every continuous vector field on  $S^2$  has a zero. But an infinitesimal translation on a Lie group leaves no element unmoved. Thus  $S^2$  cannot be a Lie group. Of all the spheres, only  $S^1$  and  $S^3$  are Lie groups.

The Lie algebra of the group is the algebra of infinitesimal transformations, which are vector fields on  $M$ . To find these, look at the path of a single point of  $M$  under a one-parameter family of transformations. The tangent vectors to this path are the vector field. We have a vector at every point, and hence an infinitesimal transformation. The paths are integral curves of the vector field.

**Examples:** The Lorentz-group action defined earlier has an integral curve through the point  $(t_0, x_0)$ ,

$$s \mapsto (\cosh s t_0 + \sinh s x_0, \cosh s x_0 + \sinh s t_0),$$

and the infinitesimal transformation is the vector field

$$v = x \frac{\partial}{\partial t} + t \frac{\partial}{\partial x}.$$

The set of all points in a space  $(t, x, y)$  with Lorentz metric satisfying the condition

$$t^2 = x^2 + y^2, \quad t > 0,$$

form a manifold  $\Sigma$  called the *future null cone*. The Lorentz-group action on this spacetime is generated by the three infinitesimal transformations

$$L_x = x \frac{\partial}{\partial t} + t \frac{\partial}{\partial x},$$

$$L_y = y \frac{\partial}{\partial t} + t \frac{\partial}{\partial y},$$

$$R = -x \frac{\partial}{\partial y} + y \frac{\partial}{\partial x}.$$

The finite transformations are easily written down.

All three of these vector fields are tangent to the future null cone, and points on  $\Sigma$  remain on  $\Sigma$ . Thus we also have an action of the Lorentz group on  $\Sigma$ . If we use the coordinates  $(x, y)$  as coordinates for  $\Sigma$ , then we can find vector fields on  $\Sigma$  representing these infinitesimal transformations. These are

$$L_x = \sqrt{x^2 + y^2} \frac{\partial}{\partial x},$$

$$L_y = \sqrt{x^2 + y^2} \frac{\partial}{\partial y},$$

$$R = y \frac{\partial}{\partial x} - x \frac{\partial}{\partial y}.$$

These are all smooth vector fields.

Look at the future null cone in four-dimensional spacetime, and reduce this to the 2-manifold  $\Omega$  of null directions. This is the 2-manifold that you see when you look around you in various directions. There is a Lorentz group action on it as well.

Let us use a stereographic representation to find a coordinate chart for  $\Omega$ :

$$\phi: (t, x, y, z) \mapsto (u, v) = \left( \frac{x}{t+z}, \frac{y}{t+z} \right)$$



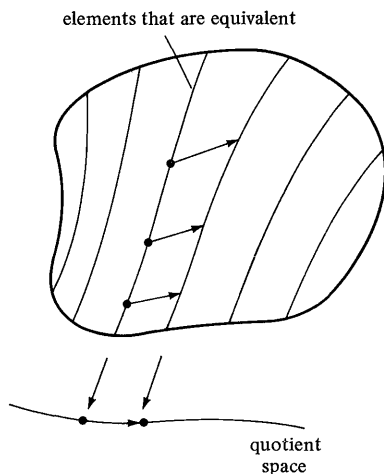


Figure 17.2. A group action on a quotient space.

An infinitesimal Lorentz transformation in the  $z$  direction is represented by the vector field in Lorentz 4-space:

$$w = z \frac{\partial}{\partial t} + t \frac{\partial}{\partial z}.$$

To show that we have a group action on  $\Omega$ , we must check that the action on 4-space passes to the quotient of the equivalence relation defining  $\Omega$ . The general idea is sketched in Figure 17.2. Here this is obvious. Since the map  $\phi$  is homogeneous in the coordinates, all rays map onto the same  $(u, v)$  point. As an infinitesimal action on 4-space,  $w$  describes the equivalence class of curves

$$[s \mapsto (t + sz, x, y, z + st)],$$

and in our chart this representative curve is the curve

$$s \mapsto (u, v) = (x, y)/(t + z)(1 + s).$$

Using the binomial theorem, we find an equivalent curve

$$s \mapsto (u - su, v - sv);$$

so the action of the Lorentz group on the space  $\Omega$  in a stereographic chart is just the dilation given by

$$-u \frac{\partial}{\partial u} - v \frac{\partial}{\partial v}.$$

We have just done a transformation of tensor components by hand.

### Problem

- 17.1** There is a Lorentz-group action on the pseudospheres of Section 6. The equations describing it, however, are nonlinear and complicated. The pseudosphere can be embedded into a Minkowski spacetime, and there the Lorentz-group action is linear and represented by matrices. Show how to use this to simplify drawing accurate figures of the pseudosphere, such as the tiling mentioned in Section 6.

## 18. Lie derivative

A group action that acts on the points of a manifold also acts on sets of points – on curves and surfaces, for example – and also on tangent vectors and 1-forms. These are, remember, just equivalence classes of curves and functions.

Given a transformation, we naturally ask how it modifies our various geometric objects. Most of our geometric objects have representations in the tangent space. Once the action of a transformation on points is given, the action on tangent vectors and other geometric objects is determined. The geometric operation measuring this change is called the *Lie derivative*.

Tensors and most other geometric objects can be compared only if they are defined at the same point in the manifold. To define the change in a 1-form field, say, under a transformation, we use the pullback map to pull back the 1-form at the new point to the old point; then we compare this pulled-back 1-form with the 1-form already there (see Figure 18.1).

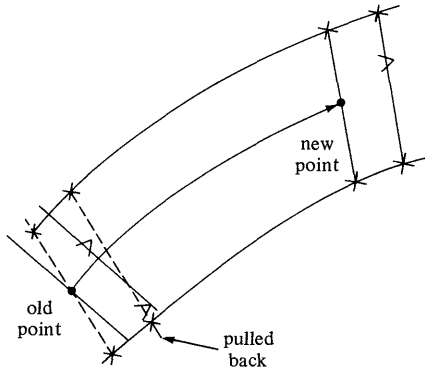
For a 1-parameter transformation

$$\Phi_\epsilon: M \rightarrow M; q \mapsto q_\epsilon,$$

we define the Lie derivative of a 1-form field  $\omega$  to be

$$\text{Lie derivative}(\omega) = \lim_{\epsilon \rightarrow 0} \frac{(\Phi_\epsilon^* \cdot \omega(q_\epsilon) - \omega(q))}{\epsilon}.$$

For tangent vectors we must use the pushforward map; since we are demanding a group action, the transformation is invertible, and the



**Figure 18.1.** Using a flow to compare 1-forms at different points.

inverse transformation can be used to push tangent vectors back from the new point to the old point.

In this way, any geometric objects that are defined in the tensor product spaces of the tangent space can be Lie-differentiated. The sign is given by the rule

$$\text{Lie derivative} = \text{Lim} (\text{value pulled back} - \text{value already there}).$$

For a continuous transformation, the Lie derivative depends only on the vector field that describes the infinitesimal transformation. The Lie derivative of the geometric object  $\Omega$  with respect to the vector field  $w$  is written  $\mathcal{L}_w \Omega$ .

### Functions

If we are given a vector field  $w$  that represents an infinitesimal transformation, the Lie derivative of a function  $f$  is just its derivative in the  $w$  direction:

$$\begin{aligned} \mathcal{L}_w f &= \text{Lim}_{\epsilon \rightarrow 0} [f(x + \epsilon w) - f(x)] / \epsilon \\ &= f_{,\alpha} w^\alpha = df \cdot w = wf. \end{aligned}$$

There are many different notations for this, because there are several different ideas that happen to coincide here in their simplest cases.

**Example:** Look at a transformation that moves you around in the room. If the Lie derivative of the temperature vanishes, then it is

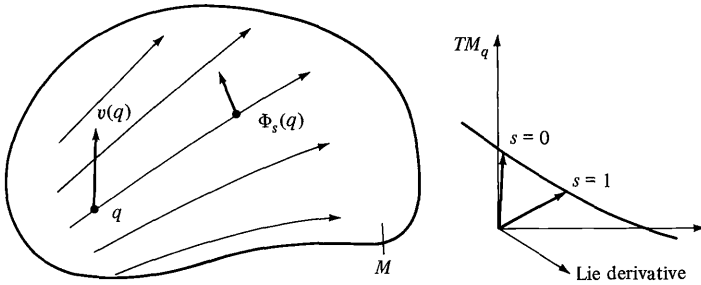


Figure 18.2. The pattern of a Lie derivative.

the same temperature everywhere along the integral curves of the transformation.

### Tangent vectors

Now, carrying along a number takes no thought at all, but carrying along a vector takes some care. Still, it should be possible to do this. After all, once you know how to transform points, you can transform curves and hence tangents, which are equivalence classes of curves. Assume that our transformation is given by the flow  $\Phi_s$ ; then, for fixed  $s$ ,

$$\Phi_s : M \rightarrow M$$

is a transformation of  $M$ . Suppose now that a vector field  $v$  is given on  $M$ . Using the tangent map  $T\Phi_{(-s)}$ , we can bring back vectors from along an integral curve

$$\phi : s \mapsto T\Phi_{(-s)} \cdot v(\Phi_s(q)). \tag{18.1}$$

This says to take the vector  $v(\Phi_s(q))$  at the point  $\Phi_s(q)$  and push it back to  $q$  by using the inverse of the tangent map. Note that our vectors are going against the flow. They can do so only because  $\Phi_s$  is invertible. Now, the map  $\phi$  in equation 18.1 is a curve in the vector space  $TM_q$ ; all the vectors are in the same tangent space. The Lie derivative is just the derivative of this curve at  $s = 0$ :

$$\mathcal{L}_w v = \frac{d}{ds} (\phi).$$

See Figure 18.2.

An immediate and important conclusion that results from defining Lie derivatives in terms of maps is that Lie differentiation commutes with pullback and pushforward.

To show that there is nothing conceptually difficult here, I will first calculate the Lie derivative of a vector field in a totally pedestrian manner. A more elegant calculation will then follow. In a coordinate chart with origin at the point of interest, we have the vector field

$$v = (v^\mu + v^\mu_{,\alpha} x^\alpha + \dots) \frac{\partial}{\partial x^\mu},$$

the infinitesimal transformation

$$w = (w^\mu + w^\mu_{,\alpha} x^\alpha + \dots) \frac{\partial}{\partial x^\mu},$$

and, for small  $s$ ,

$$\Phi_s(0) = sw.$$

$$v(\Phi_s(0)) = v^\mu + sv^\mu_{,\alpha} w^\alpha \frac{\partial}{\partial x^\mu}.$$

The transformation is locally,

$$x^\mu \mapsto x^\mu + w^\mu + sw^\mu_{,\alpha} x^\alpha.$$

The basis vectors are pushed forward according to

$$\frac{\partial}{\partial x^\alpha} \mapsto \frac{\partial}{\partial x^\alpha} + w^\mu_{,\alpha} \frac{\partial}{\partial x^\mu},$$

and in components this is

$$v^\alpha \frac{\partial}{\partial x^\alpha} \mapsto (v^\mu + w^\mu_{,\alpha} v^\alpha) \frac{\partial}{\partial x^\mu}.$$

Note carefully the indices on  $w^\mu_{,\alpha}$ . The use of  $\partial/\partial x^\mu$  for the basis vectors at different points makes this confusing. If you use  $\partial/\partial y^\mu$  for the new basis vectors, things are much clarified in the middle of the computation. Thus the inverse pushforward of  $v$  gives us the curve

$$s \mapsto v^\mu + sv^\mu_{,\alpha} w^\alpha - sw^\mu_{,\alpha} v^\alpha.$$

The Lie derivative is just the linear term, and this is our old friend the Lie bracket of Section 9:

$$\mathcal{L}_w v = [w, v]. \quad (18.2)$$

Now, how could we have found that out with less trouble? Since all the operations are specified in coordinate-free language, we are free to choose any coordinates we wish. It is always possible to pick coordinates  $(q, t)$  such that any nonzero vector field  $w$  appears in the normal form

$$w = \frac{\partial}{\partial t}.$$

This will be discussed in Section 26. The pushforward map on tangent vectors is now just the identity, and so

$$\mathfrak{L}_w \left( Q \frac{\partial}{\partial q} + T \frac{\partial}{\partial t} \right) = \frac{\partial Q}{\partial t} \frac{\partial}{\partial q} + \frac{\partial T}{\partial t} \frac{\partial}{\partial t}.$$

But we also have for the Lie bracket

$$\left[ \frac{\partial}{\partial t}, Q \frac{\partial}{\partial q} + T \frac{\partial}{\partial t} \right] = \frac{\partial Q}{\partial t} \frac{\partial}{\partial q} + \frac{\partial T}{\partial t} \frac{\partial}{\partial t}.$$

Hence in these special coordinates

$$\mathfrak{L}_w v = [w, v],$$

and since both sides are coordinate-independent, this is true in general. Although the normal form  $\partial/\partial t$  cannot be used where  $w$  has a zero, the linearity of the Lie derivative allows us to add a constant field and remove any zero we wish.

A vector field  $v$  such that

$$\mathfrak{L}_w v = 0$$

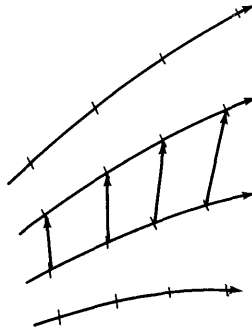
is said to be *Lie-transported* or *dragged* along the vector field  $w$ . The geometric interpretation of dragging a vector is to picture its head and tail being carried along neighboring integral curves, as in Figure 18.3.

**Example:** We have

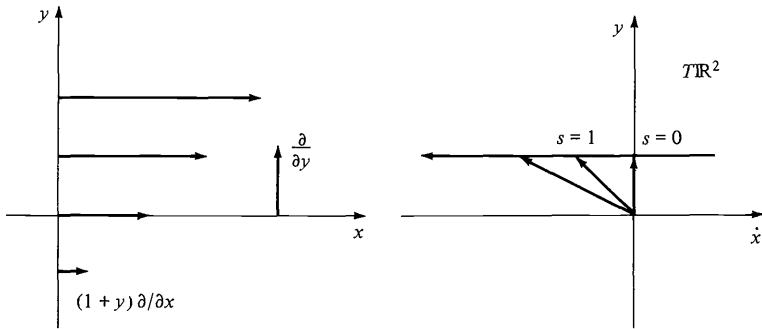
$$\mathfrak{L}_{(1+y)(\partial/\partial x)} \frac{\partial}{\partial y} = \left[ (1+y) \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right] = -\frac{\partial}{\partial x}.$$

In Figure 18.4 I sketch this.

We will see in Section 46 that the magnetic field in a conducting fluid obeys



**Figure 18.3.** A vector field that is Lie-transported by a flow. Tick marks indicate equal parameter increments.



**Figure 18.4.** The Lie derivative of a constant vector field with respect to a shear.

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E},$$

and for high conductivity we have

$$\mathbf{E} = -\mathbf{V} \times \mathbf{B},$$

where  $\mathbf{V}$  is the 3-velocity of the fluid. Thus the evolution of the field for slow fluid motions follows

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{V} \times \mathbf{B}),$$

and for incompressible fluid motion this becomes

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla_{\mathbf{B}} \mathbf{V} - \nabla_{\mathbf{V}} \mathbf{B}.$$

For steady motions we have the Lie bracket

$$-[\mathbf{B}, \mathbf{V}] = \mathcal{L}_{\mathbf{V}} \mathbf{B} = 0.$$

For unsteady motions we define a nonrelativistic 4-velocity

$$v = \frac{\partial}{\partial t} + \mathbf{V} \frac{\partial}{\partial \mathbf{q}},$$

and now take  $\mathbf{B}$  to be a four-vector with no  $t$  component. Then again we have a Lie derivative

$$\mathcal{L}_v \mathbf{B} = \nabla_v \mathbf{B} + \frac{\partial \mathbf{B}}{\partial t} - \nabla_{\mathbf{B}} v = 0.$$

For both steady and unsteady motions the magnetic field vector is dragged along by the fluid. Note how the Lie derivative gives directional changes as well.

### 1-Forms

The preceding rules let us drag along the local tangent spaces; so we can easily discuss the Lie derivative of any geometric object that can be represented in the tangent space. Since the Lie derivative is to be a local approximation, it must satisfy Leibniz's rule (i.e., be a derivation). This lets us break up a product, such as a tensor product, into factors.

$$\mathcal{L}_w a \otimes b = (\mathcal{L}_w a) \otimes b + a \otimes (\mathcal{L}_w b).$$

The Lie derivative of a 1-form  $\alpha$  can be found by taking any vector field  $v$ , and the function

$$f = \alpha \cdot v.$$

We Lie-differentiate  $f$ :

$$\mathcal{L}_w f = (\mathcal{L}_w \alpha) \cdot v + \alpha \cdot \mathcal{L}_w v.$$

From this we find

$$(\mathcal{L}_w \alpha) \cdot v = d(\alpha \cdot v)w - \alpha \cdot [w, v]. \quad (18.3)$$

All the operations on the right-hand side have been defined before; so this is an explicit, coordinate-free definition of  $\mathcal{L}_w$  as a 1-form. This is a typical coordinate-free definition of a 1-form, given in terms of its action



on a vector. The vector must appear explicitly, of course. In terms of a coordinate basis,

$$\mathcal{L}_w \alpha = (\alpha_{\mu, \nu} w^\nu + \alpha_\nu w^\nu{}_{, \mu}) dx^\mu. \quad (18.4)$$

**Example:** Consider in  $\mathbb{R}^2$

$$\mathcal{L}_{x(\partial/\partial x)} dx.$$

First think about the sign. The transformation  $x(\partial/\partial x)$  is an expansion. The pullback is a contraction, and so the pullback of  $dx$  has its contour lines closer together. The Lie derivative is thus positive.

From

$$dx \cdot \frac{\partial}{\partial x} = 1$$

we have, differentiating

$$(\mathcal{L}_{x(\partial/\partial x)} dx) \cdot \frac{\partial}{\partial x} + dx \cdot \left[ x \frac{\partial}{\partial x}, \frac{\partial}{\partial x} \right] = 0,$$

$$(\mathcal{L}_{x(\partial/\partial x)} dx) \cdot \frac{\partial}{\partial x} - dx \cdot \frac{\partial}{\partial x} = 0,$$

hence

$$\mathcal{L}_{x(\partial/\partial x)} dx = dx.$$

Note that

$$\mathcal{L}_{x(\partial/\partial x)} dx \neq x \mathcal{L}_{\partial/\partial x} dx.$$

As can be seen in the preceding example, the Lie derivative does not act like a directional derivative. It is not linear on the functions that appear in the directional argument. For 1-forms we have, from equation 18.3,

$$\mathcal{L}_g \alpha = g \mathcal{L}_w \alpha + (\alpha \cdot w) dg,$$

for  $g$  any smooth function. This leads to the useful computational formula

$$\mathcal{L}_{X(\partial/\partial x)}(f dx + g dy) = X \frac{\partial}{\partial x} (f dx + g dy) + f dX,$$

where  $X$  is any function.

**Example:** What are the infinitesimal symmetries of the 1-form field  $dx$  in three dimensions? Let  $w$  be a vector field that represents an infinitesimal transformation that leaves  $dx$  unchanged:

$$w = X \frac{\partial}{\partial x} + Y \frac{\partial}{\partial y} + Z \frac{\partial}{\partial z}.$$

Then to have

$$\mathfrak{L}_w dx = 0,$$

the only requirement on  $w$  is that the  $X$  be constant.

### Tensors

From repeated application of Leibniz's rule, we can Lie-differentiate any tensor index-by-index:

$$\begin{aligned} & \mathfrak{L}_w \left( T^{\mu\nu\dots}_{\sigma\tau\dots, \alpha} \frac{\partial}{\partial x^\mu} \otimes \frac{\partial}{\partial x^\nu} \otimes \dots \otimes dx^\sigma \otimes dx^\tau \otimes \dots \right) \\ &= [T^{\mu\nu\dots}_{\sigma\tau\dots, \alpha} w^\alpha - T^{\alpha\nu\dots}_{\sigma\tau\dots} w^\mu_{, \alpha} - T^{\mu\alpha\dots}_{\sigma\tau\dots} w^\nu_{, \alpha} - \dots \\ & \quad + T^{\mu\nu\dots}_{\alpha\tau\dots} w^\alpha_{, \sigma} + T^{\mu\nu\dots}_{\sigma\alpha\dots} w^\alpha_{, \tau} + \dots] \frac{\partial}{\partial x^\mu} \otimes \frac{\partial}{\partial x^\nu} \otimes \dots \otimes dx^\sigma \otimes dx^\tau \otimes \dots. \end{aligned}$$

Other methods for calculating Lie derivatives of differential forms use the exterior calculus to be developed in Chapter IV.

### Transformations lifted to $TM$

An infinitesimal transformation of a manifold  $M$  can be lifted to give an infinitesimal transformation of  $TM$  by using the tangent map of Section 13. If  $w(\partial/\partial Q)$  is the infinitesimal transformation on  $M$  with coordinates  $Q$ , then the infinitesimal transformation on  $TM$  with coordinates  $(Q, \dot{Q})$  is given by the vector field

$$\tilde{w} = w^\mu \frac{\partial}{\partial q^\mu} + w^\mu_{, \alpha} \dot{q}^\alpha \frac{\partial}{\partial \dot{q}^\mu}.$$

The integral curves of this vector field generate a surface that is a section of  $TM$ , which is a vector field. The Lie derivative of this vector field vanishes. To see this, write down the ordinary differential equations for the integral curves of  $\tilde{w}$ :

$$\frac{dq^\mu}{ds} = w^\mu, \quad \frac{d\dot{q}^\mu}{ds} = w^\mu_{, \alpha} \dot{q}^\alpha. \quad (18.5)$$

For a vector field  $v$  whose graph is generated by integral curves of  $w$ , the partial derivatives of the components of  $v$  satisfy

$$\frac{dv^\mu}{ds} = v^\mu{}_{,\alpha} w^\alpha.$$

The rate of change of these components is given by equation 18.5; that is,

$$\frac{dv^\mu}{ds} = w^\mu{}_{,\alpha} v^\alpha.$$

Comparing these, we have

$$w^\mu{}_{,\alpha} v^\alpha - v^\mu{}_{,\alpha} w^\alpha = 0,$$

which is the equation of Lie transport. An integral curve of the lifted vector field contains information not only about a single solution, but also about the local behavior of nearby solutions. This is useful in discussing the spreading of solutions, for example. It is also useful in two-point boundary-value problems for suggesting how to improve successive approximations.

Similarly, we can lift a transformation up into the cotangent bundle. The vector field there comes from writing the Lie-derivative condition

$$\alpha_{\mu,\nu} w^\nu + \alpha_\nu w^\nu{}_{,\mu} = 0$$

and, if we use canonical coordinates  $(q, p)$  on  $T^*M$ , this leads to the vector field

$$w^\mu \frac{\partial}{\partial q^\mu} - w^\nu{}_{,\mu} p_\nu \frac{\partial}{\partial p_\mu}.$$

**Example:** What happens in the shear

$$w = y^2 \frac{\partial}{\partial x}?$$

The transformation in  $TM$  is

$$y^2 \frac{\partial}{\partial x} + 2y \dot{y} \frac{\partial}{\partial \dot{x}}.$$

This shows that a vector above the  $x$  axis with a positive  $y$  component is rotated clockwise (see Figure 18.5).

The transformation in  $T^*M$  is

$$y^2 \frac{\partial}{\partial x} + 2yk \frac{\partial}{\partial l},$$

where we use coordinates  $(x, y, k, l)$  for  $T^*M$ . In the cotangent space this transformation is shown in Figure 18.6, in contrast with

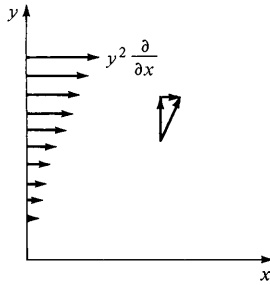


Figure 18.5. The transport of a vector by a shear.

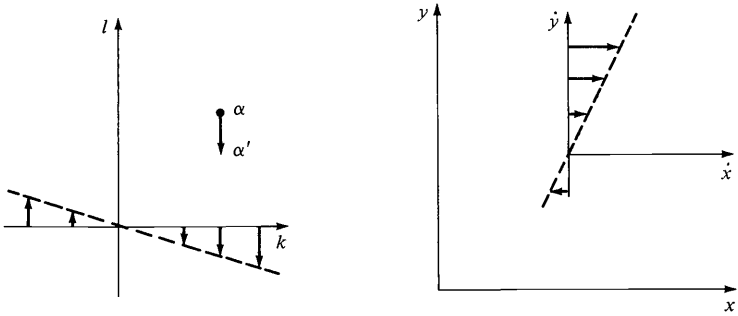


Figure 18.6 (left). The action of the same shear as in Figure 18.5, this time acting on an element  $\alpha$  of the cotangent space.

Figure 18.7 (right): The action of the same shear on the tangent space.

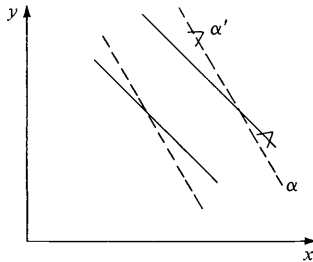


Figure 18.8. The action on a 1-form, sketched on the manifold. Contrast with Figure 18.5.

the tangent space transformation shown in Figure 18.7. The transformation of the tangent-space representation of the 1-form  $\alpha$  is sketched in Figure 18.8.

### 19. Holonomy

A surprising and delightful feature of transformations is that successive transformations often do not commute. When they do commute, they are called *holonomic*. Holonomy is a general idea that describes what happens when we perform a cycle of transformations:  $ABA^{-1}B^{-1}$  (Figure 19.1).

The geometry of parking a car is a nice example of the exploitation of noncommutation. This fun example is taken from Nelson (1967), an otherwise very serious book.

Specify the state of a car by four coordinates: the  $(x, y)$  coordinates of the center of the rear axle, the direction  $\theta$  of the car, and the angle  $\phi$  between the front wheels and the direction of the car (Figure 19.2).

The driver has two different transformations at his disposal. He can turn the steering wheel, or he can drive the car forward or back. The infinitesimal transformations will be the vector fields

$$\text{STEER} = \frac{\partial}{\partial \phi},$$

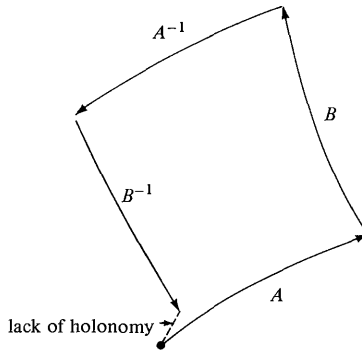
and

$$\text{DRIVE} = \cos \theta \frac{\partial}{\partial x} + \sin \theta \frac{\partial}{\partial y} + \frac{\tan \phi}{L} \frac{\partial}{\partial \theta}.$$

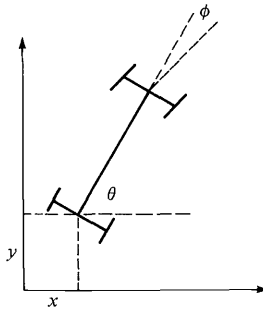
The geometry used to compute  $\text{DRIVE}$  is shown in Figure 19.3. Now,  $\text{STEER}$  and  $\text{DRIVE}$  do not commute; otherwise you could do all your steering at home before driving off on a trip. We have, in fact,

$$[S, D] = \frac{1}{L \cos^2 \phi} \frac{\partial}{\partial \theta} = \text{ROTATE}.$$

The operation  $[S, D]$  is the infinitesimal version of the sequence of transformations: steer, drive, steer back, drive back. Now,  $\text{ROTATE}$  can get you out of some parking spaces, but not tight ones. As Figure 19.4 shows, you may not have enough room to  $\text{ROTATE}$  out. The usual tight parking space restricts the  $\text{DRIVE}$  transformation, but not  $\text{STEER}$ . A truly tight parking space restricts  $\text{STEER}$  as well by putting your front wheels against the curb.



**Figure 19.1.** The pattern for lack of holonomy as the lack of commutativity of transformations.

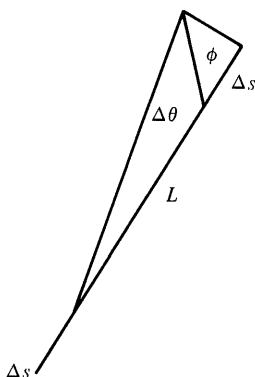


**Figure 19.2.** The state-space variables for a simplified model of an automobile.

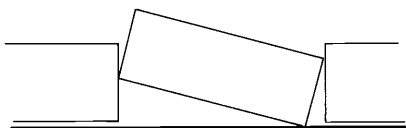
Fortunately, there is still another commutator available:

$$[D, R] = \frac{1}{L \cos^2 \phi} \left( \sin \theta \frac{\partial}{\partial x} - \cos \theta \frac{\partial}{\partial y} \right) = \text{SLIDE.}$$

SLIDE is a displacement at right angles to the car, and can get you out of any parking place. Just remember to steer, drive, steer back, drive some more, steer, drive back, steer back, drive back  $[SDS^{-1}DSD^{-1}S^{-1}D^{-1}]$ . You have to reverse steer in the middle of the parking place. This is not intuitive, and no doubt is part of the problem with parallel parking. Thus from only two controls you can form the vector fields STEER, DRIVE, ROTATE, and SLIDE, which span the full tangent space at every point. You can move anywhere in the four-dimensional configuration space.



**Figure 19.3.** The action of the DRIVE transformation on our car.



**Figure 19.4.** A parking space too tight to ROTATE out of.

Vector fields that do not commute are called *anholonomic*. If two transformations commute, then the system could never leave a 2-surface. This obvious result is called the *Frobenius Theorem*. Suppose we are given a field of contact elements on a manifold  $M$ . These contact elements are linear subspaces of the tangent spaces  $T_p(M)$ , sometimes called *distributions* (nothing to do with  $\delta$ -functions). A submanifold is called a *complete integral submanifold* of the field of contact elements, if at each point the contact element of the submanifold equals the given contact element. A submanifold of lower dimension is called an *integral submanifold* if its contact element is contained within the given contact element.

A given field of contact elements is said to be *completely integrable* if for any two vector fields lying in the contact-element field, their commutator does also. The Frobenius Theorem says that a completely integrable contact-element field has a complete integral submanifold. A proof can be found in Choquet–Bruhat (1977).

If you think of two-dimensional contact elements in three dimensions as little “floor tiles,” then an integrable field is one where these tiles fit together to form a surface.

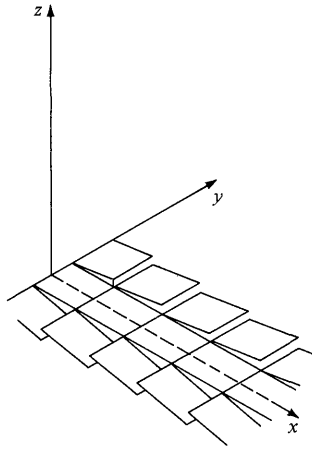


Figure 19.5. A nonintegrable field of contact 2-elements.

**Examples:** A field of two-dimensional contact elements on  $\mathbb{R}^3$  can be specified by giving two basis vectors, and then adding to these all possible linear combinations. For the two vector fields  $\partial/\partial y$  and  $\partial/\partial x + y(\partial/\partial z)$ , sketched in Figure 19.5, we have

$$\left[ \frac{\partial}{\partial y}, \frac{\partial}{\partial x} + y \frac{\partial}{\partial z} \right] = \frac{\partial}{\partial z};$$

so they do not have a two-dimensional integral submanifold.

In classical mechanics we often consider constraints that force the system to move along only a linear subspace of the tangent space. If these subspaces are integrable, then the system must remain on a subspace of smaller dimension. These are called holonomic constraints.

An important application of the Frobenius Theorem is to the computation of integrability conditions for partial differential equations.

**Example:** Suppose we want to solve the partial differential equation

$$\frac{\partial \psi}{\partial \mathbf{x}} = v$$

for a specified vector field  $v$ . The geometric view of this equation is found by taking the space  $(\psi, \mathbf{x})$ , with a contact-element field generated by the set of vector fields



$$\frac{\partial}{\partial x} - v \frac{\partial}{\partial \psi}.$$

A solution to this equation is an integral submanifold of this contact-element field. Each vector points in a direction that keeps all but one coordinate constant; this is the definition of a partial derivative.

The Frobenius Theorem states that an integral submanifold exists if the contact-element field is completely integrable. This requires that for every pair of vector fields in the contact element, such as

$$a = a \left( \frac{\partial}{\partial x} - v \frac{\partial}{\partial \psi} \right) \quad \text{and} \quad b = b \left( \frac{\partial}{\partial x} - v \frac{\partial}{\partial \psi} \right),$$

we have  $[a, b]$  also in the contact element. This is equivalent to the demand that

$$\left[ \frac{\partial}{\partial x^i} - v_i \frac{\partial}{\partial \psi}, \frac{\partial}{\partial x^j} - v_j \frac{\partial}{\partial \psi} \right] = -(v_{j,i} - v_{i,j}) \frac{\partial}{\partial \psi}$$

be in the contact element, using linearity. Since  $\partial/\partial\psi$  is not in the ideal, we must have

$$v_{j,i} - v_{i,j} = 0;$$

that is,

$$\nabla \times v = 0$$

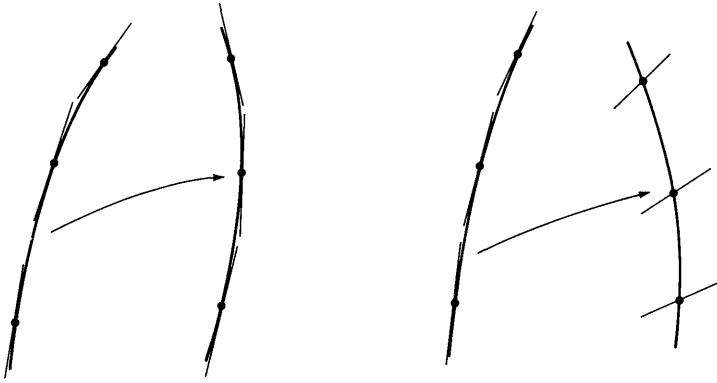
in 3-vector notation. The Frobenius Theorem assures us that this is a sufficient condition as well as a necessary one.

### Problems

- 19.1** Discuss the parking problem for a semitrailer, a full trailer, and an elephant train of  $n$  cars. Can all of these be parked?
- 19.2** Will the number of cycles around the commutator loop be of order  $1/\epsilon$  if the parking space is longer than the car by only  $\epsilon$ ? What if you are tight against the curb as well?

### 20. Contact transformations

Contact transformations are those transformations that preserve the contact structure of a contact bundle. The most common contact transformations are the Legendre transformations.



**Figure 20.1** (left). A contact transformation preserves the contact structure by taking lifted curves to other lifted curves.

**Figure 20.2** (right). A transformation that is not a contact transformation.

### Contact transformations

Contact structure singles out those special submanifolds that are lifted from the base space. A transformation between contact bundles

$$K: C(M, n) \rightarrow C(M', n)$$

preserves the contact structure if it maps these special submanifolds into other such special submanifolds. Figure 20.1 sketches a transformation that does preserve the contact structure, and Figure 20.2 sketches one that does not. A submanifold  $(N, \psi)$ , with  $\psi: N \rightarrow C(M, n)$ , is the lift of a submanifold of  $M$  if we have

$$\psi^* \cdot \theta = 0,$$

where  $\theta$  are the 1-forms that generate the contact ideal in  $C(m, n)$ , and  $\psi^*$  is the pullback operator for the map  $\psi$ . The image of the submanifold must also satisfy this condition

$$(K \circ \psi)^* \cdot \theta' = 0,$$

where  $\theta'$  are 1-forms generating the contact ideal in  $C(M', n)$ . Thus

$$\psi^* \cdot (K^* \cdot \theta') = 0,$$

and the condition for us to have a contact transformation is that

$$K^* \cdot \mathcal{G} = \mathcal{G},$$

where  $\mathcal{G}$  is the ideal generated by all possible linear combinations of

### Legendre transformations

**Examples:** Look at the transformation of  $C\mathbb{R}^2$  into itself given in our standard coordinates by

$$K: (x, y, p) \mapsto (X, Y, P) = (p, px - y, x).$$

Uppercase letters denote the new coordinates of the point. They are functions of the old coordinates. The contact 1-form in the new coordinates is

$$\Theta = dY - P dX;$$

the pullbacks are

$$K^* \cdot dX = dp, \quad K^* \cdot dY = x dp + p dx - dy;$$

and so the pullback of the contact 1-form

$$K^* \cdot \Theta = p dx - dy$$

is just the negative of the original contact 1-form. Thus  $K$  is a contact transformation, called a *Legendre transformation*.

Let us use a contact transformation to transform the nonlinear partial differential equation

$$\frac{\partial f}{\partial x} + f \frac{\partial f}{\partial y} = 0$$

into a linear equation. We go to a large-enough manifold,  $C(\mathbb{R}^3, 2)$ , with coordinates  $(x, y, f, p, q)$  and a contact 1-form

$$\theta = df - p dx - q dy.$$

This 1-form tells us what the coordinates mean. On a lifted submanifold we have  $p = f_{,x}$  and  $q = f_{,y}$ . In this space a solution to our problem is a 2-submanifold  $\psi: \mathbb{R}^2 \rightarrow C(\mathbb{R}^3, 2)$  satisfying

$$g \circ \psi = 0$$

for

$$g: C(\mathbb{R}^3, 2) \rightarrow \mathbb{R}; (x, y, f, p, q) \mapsto p + fq,$$

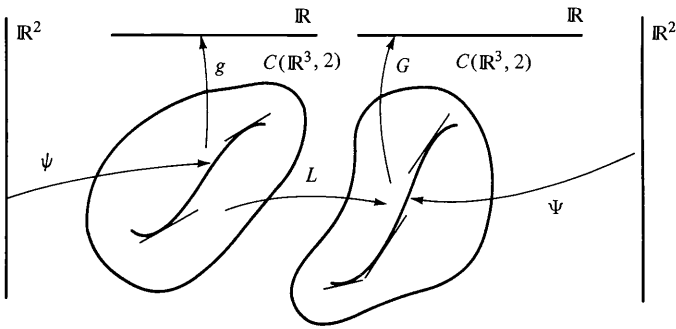


Figure 20.3. A sketch of our use of a contact transformation to simplify a nonlinear partial differential equation.

and  $\psi$  must also be an integral submanifold of  $\theta$ :

$$\psi^* \cdot \theta = 0.$$

The trick now is to move the nonlinearity from the original dependent variable to the independent variables. We use the contact transformation

$$L: (x, y, f, p, q) \mapsto (X, Y, F, P, Q) = (p, q, px + qy - f, x, y),$$

which is a higher-dimensional form of the Legendre transformation. The transformed problem is to find an integral submanifold  $(\mathbb{R}^2, \Psi)$  of the contact ideal satisfying

$$G \circ \Psi = 0$$

(see Figure 20.3). We must have

$$G = g \circ L^{-1}$$

to reproduce solutions of the original problem. We have

$$L^{-1}: (X, Y, F, P, Q) \mapsto (x, y, f, p, q) = (P, Q, PX + QY - F, X, Y),$$

and so

$$G: (X, Y, F, P, Q) \mapsto X + (PX + QY - F)Y.$$

As a partial differential equation, this is

$$X + XY \frac{\partial F}{\partial X} + Y^2 \frac{\partial F}{\partial Y} - YF = 0,$$

which is linear in  $F$ . The nonlinearity has been traded for variable coefficients. Solutions to the new equation can be superposed. This is not true of the original equation.

Look at the line-element contact bundle to two-dimensional space-time,  $\mathbb{C}\mathbb{R}^2$ , and let us find the contact transformation that is the lift of a Lorentz transformation on spacetime.

Using coordinates  $(t, x, \dot{x})$ , we find that the contact 1-form is

$$\theta = \dot{x} dt - dx.$$

The contact transformation will be of the form

$$(t, x, \dot{x}) \mapsto (\cosh \psi t + \sinh \psi x, \cosh \psi x + \sinh \psi t, \dot{X}(t, x, \dot{x})),$$

for some function  $\dot{X}(t, x, \dot{x})$ . We can find  $\dot{X}$  by demanding that the contact 1-form  $\theta$  pull back to a multiple of itself. From

$$\dot{x} dt - dx \mapsto (\cosh \psi dt + \sinh \psi dx) \dot{X} - \cosh \psi dx - \sinh \psi dt,$$

we have

$$\dot{x} dt - dx \mapsto \left( \frac{\dot{X} \cosh \psi - \sinh \psi}{\dot{x}} \right) \dot{x} dt - (\cosh \psi - \dot{X} \sinh \psi) dx,$$

and we see that we must have

$$\dot{X} = \frac{\dot{x} \cosh \psi + \sinh \psi}{\cosh \psi + \dot{x} \sinh \psi}.$$

This is the usual velocity-addition rule of special relativity. To see that, introduce  $\alpha$  and  $\alpha'$  such that

$$\dot{x} = \tanh \alpha, \quad \dot{X} = \tanh \alpha',$$

We have then

$$\alpha' = \alpha + \psi.$$

### Enthalpy

Many thermodynamic systems are studied under conditions of constant pressure. Their thermodynamic description will be simplified if we take entropy and pressure as independent variables, rather than entropy and volume.

Ordinary gas thermodynamics takes place in the 2-surface contact bundle to the bundle  $(U, S, V) \mapsto (S, V)$ . The contact structure is given by the 1-form

$$dU - T dS + P dV, \tag{20.1}$$

which defines the potentials

$$T = \frac{\partial U}{\partial S} \quad \text{and} \quad P = -\frac{\partial U}{\partial V}.$$

Now, make a Legendre transformation to a new contact bundle, reversing the roles of  $P$  and  $V$ . The new bundle is

$$(H, T, S, V, P) \mapsto (H, S, P),$$

and the new variable  $H$  is called *enthalpy*. The contact structure is now given by the 1-form

$$dH - TdS - VdP, \quad (20.2)$$

which defines the new coordinates as partial derivatives of the fundamental equation  $H(S, P)$ :

$$T = \frac{\partial H}{\partial S}, \quad V = \frac{\partial H}{\partial P}.$$

The map between these contact bundles is given by

$$(U, T, S, P, V) \mapsto (H, T, S, V, P) = (U + PV, T, S, V, P).$$

The pullback of the new contact 1-form given in equation 20.2 is exactly the original contact 1-form given in equation 20.1.

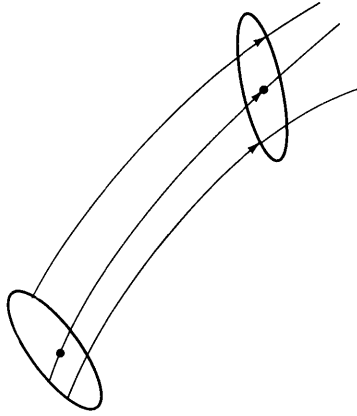
A gas system can be fully described by the enthalpy tabulated as a function of entropy and pressure. When heat is added to a system at constant pressure, the change in enthalpy is equal to the heat added. The  $PdV$  work is automatically accounted for.

## 21. Symmetries

The various structures that we will study have symmetries that are an important part of their description. A *symmetry* is a transformation that leaves the structure essentially unchanged. Some symmetries, such as reflection in a plane, are discrete. Others, such as translations and rotations, form continuous sets. Continuous symmetries are best approached via their infinitesimal generators, and these are represented by vector fields. Structures that have representations in the tangent space are modified by infinitesimal transformations according to the Lie derivative.

### *Metric symmetry*

The most familiar symmetry of a structure is metric symmetry. A transformation is a metric symmetry, called an *isometry*, if the distance between transformed points is the same as the original distance between them.



**Figure 21.1.** The pattern of an isometry. The metric ellipses are carried along by the transformation.

**Example:** Using the graphical representation of a metric tensor from Section 3, we sketch an isometry in Figure 21.1.

A vector field that generates an infinitesimal isometry is called a *Killing vector*. It must satisfy Killing's equation

$$\mathfrak{L}_k \mathcal{G} = 0,$$

where  $\mathcal{G}$  is the metric tensor.

**Example:** Take Euclidean plane geometry:

$$\mathcal{G} = dx \otimes dx + dy \otimes dy.$$

For a vector field  $k$  to be a Killing vector field

$$k = X \frac{\partial}{\partial x} + Y \frac{\partial}{\partial y},$$

we must have

$$\mathfrak{L}_k \mathcal{G} = 0,$$

that is,

$$\mathfrak{L}_k dx \otimes dx + dx \otimes \mathfrak{L}_k dx + \mathfrak{L}_k dy \otimes dy + dy \otimes \mathfrak{L}_k dy = 0.$$

The Lie derivative, like any derivative, obeys a Leibniz rule for products; so we must have

$$dX \otimes dx + dx \otimes dX + dY \otimes dy + dy \otimes dY = 0.$$

Now,

$$dX = \frac{\partial X}{\partial x} dx + \frac{\partial X}{\partial y} dy;$$

so we have, finally,

$$2 \frac{\partial X}{\partial x} dx \otimes dx + \left( \frac{\partial X}{\partial y} + \frac{\partial Y}{\partial x} \right) (dx \otimes dy + dy \otimes dx) + 2 \frac{\partial Y}{\partial y} dy \otimes dy = 0.$$

Any solution will be a linear combination of the translations

$$k = \frac{\partial}{\partial x}, \quad k = \frac{\partial}{\partial y},$$

and the rotation

$$k = y \frac{\partial}{\partial x} - x \frac{\partial}{\partial y}.$$

Killing's equation is best handled in components. Using equation 18.4 we have

$$L_k(g_{\mu\nu} dx^\mu \otimes dx^\nu) = (g_{\mu\nu, \sigma} k^\sigma + g_{\sigma\nu} k^\sigma{}_{, \mu} + g_{\mu\sigma} k^\sigma{}_{, \nu}) dx^\mu \otimes dx^\nu = 0. \quad (21.1)$$

**Examples:** For the Euclidean plane the  $g_{\mu\nu}$  are constant, and we have the equations

$$X_{,x} = 0, \quad Y_{,y} = 0, \quad X_{,y} + Y_{,x} = 0,$$

found earlier.

For spacetime we have

$$\mathfrak{G} = dt^2 - dx^2,$$

and equations for the Killing vector

$$k = T \frac{\partial}{\partial t} + X \frac{\partial}{\partial x}$$

are now

$$T_{,t} = 0, \quad X_{,x} = 0, \quad T_{,x} - X_{,t} = 0.$$

These are satisfied by translations and the Lorentz transformation



$$k = x \frac{\partial}{\partial t} + t \frac{\partial}{\partial x}.$$

Four-dimensional spacetime has ten fundamental symmetries. Four are translations, three are rotations, and three are Lorentz transformations.

A 2-cylinder has only two global symmetries, not three as does the plane. A rotation, which is locally a solution of Killing's equation, cannot be made continuous over the whole manifold.

The Killing equations can also be used in reverse, to find metrics with prescribed symmetries. This is complicated by the fact that we do not know what the coordinates mean until we have the metric. Isometries will be discussed more in Section 25.

### *Contact symmetry*

Transformations that preserve the contact structure of a contact bundle are called contact transformations. A Legendre transformation is a discrete contact transformation. An infinitesimal contact transformation must satisfy

$$\mathcal{L}_v \mathcal{G} = \mathcal{G},$$

where  $\mathcal{G}$  is the contact ideal.

**Example:** Look at a mechanical system with a state space  $(q, p)$  evolving under the action of the Hamiltonian vector field

$$v = \frac{\partial H}{\partial p} \frac{\partial}{\partial q} - \frac{\partial H}{\partial q} \frac{\partial}{\partial p}.$$

This evolution can be represented by an infinitesimal contact transformation in the  $(2n+1)$ -dimensional space  $(q, h, p)$ . The  $h$  dimension is a "Hamiltonian" dimension, which allows the Hamiltonian  $H(q, p)$  to be represented by a hypersurface. First, we show that  $\mathcal{L}_v(p dq)$  is a perfect differential. (This calculation is much easier using the exterior calculus to be developed in Chapter IV. The coordinate calculation is a bit tricky and a good exercise in index shuffling. A lot happens between the lines here, and you should fill it in. It will be redone using the more efficient calculus of differential forms in Section 23.)

We let Latin indices run over the full  $2nq$  and  $p$  variables, and let Greek run over either the  $nq$ s or the  $np$ s. Using equation 18.3, we have

$$\mathfrak{L}_v(\mathbf{p} d\mathbf{q}) = (p_{\mu, a} v^a) dq^\mu + p_\nu d\left(\frac{\partial H}{\partial p_\nu}\right).$$

Now the  $p$  are independent variables, and we have

$$\frac{\partial p_\nu}{\partial q^\mu} = 0, \quad \frac{\partial p_\mu}{\partial p_\nu} = \delta_\nu^\mu,$$

and so we have

$$\mathfrak{L}_v(\mathbf{p} d\mathbf{q}) = -\frac{\partial H}{\partial q^\mu} dq^\mu + p_\nu d\left(\frac{\partial H}{\partial p_\nu}\right) = -dH + d\left(p_\nu \frac{\partial H}{\partial p_\nu}\right).$$

Thus we do have a perfect differential.

Now look at the infinitesimal transformation

$$u = \frac{\partial H}{\partial p} \frac{\partial}{\partial q} - \frac{\partial H}{\partial q} \frac{\partial}{\partial p} + X \frac{\partial}{\partial h}.$$

Can we pick  $X$  so that this is a contact transformation? We compute  $\mathfrak{L}_u(dh - p d\mathbf{q})$ . We have

$$\mathfrak{L}_u dh = dX,$$

and

$$\begin{aligned} \mathfrak{L}_u(\mathbf{p} d\mathbf{q}) &= \mathfrak{L}_v(\mathbf{p} d\mathbf{q}) + \mathfrak{L}_{X(\partial/\partial h)}(\mathbf{p} d\mathbf{q}) \\ &= -dH + d\left(p \frac{\partial H}{\partial p}\right); \end{aligned}$$

so we have

$$\mathfrak{L}_u(dh - p d\mathbf{q}) = d\left(X + H - p \frac{\partial H}{\partial p}\right) = 0$$

if we take

$$X = p \frac{\partial H}{\partial p} - H.$$

Thus every motion of a Hamiltonian system is a contact transformation in  $(q, h, p)$  space. The contact structures of mechanics will be discussed further in Chapter VII.



# IV

## *The calculus of differential forms*

This chapter develops a set of tools for manipulating differential forms. This calculus of differential forms is the promised generalization of ordinary vector calculus. It carries to manifolds such basic notions as gradient, curl, and integral.

In this chapter, I will develop these tools. In the next chapter, I will present some simple applications of them. The true value of these tools will be seen in Chapter VI, where they are applied to electrodynamics, and in the following chapters on mechanics and gravitation. Some specialized tools only make sense when seen in the context of a specific problem. These have been deferred. Thus the concept of an exterior system waits until we have some partial differential equations to play with. These tools of the calculus of differential forms are so efficient that we now go to some lengths to cast problems into a form amenable to these methods.

### **22. Differential forms**

The calculus of differential forms, often called exterior calculus, is an efficient system for manipulating functions, vectors, and differential forms, both ordinary and twisted. It is natural and very efficient for calculations, especially ones that do not involve a metric or a covariant derivative. In a later section we will bring Riemannian geometry into the exterior calculus by using the method of moving frames. The exterior calculus is particularly efficient for computations in which tensors are represented explicitly rather than as indexed arrays. One important use of differential forms is in integration. In fact, one book defines differential forms as those objects that can be integrated. This is a bit extreme. A differential  $r$ -form is a totally antisymmetric tensor of type  $\binom{0}{r}$ ; it

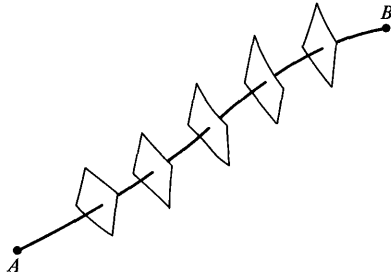


Figure 22.1. The geometric idea of a line integral.

takes in  $r$  vectors and yields a number. The space of  $r$ -forms is denoted by  $\Lambda^r$ .

One use for differential forms is to describe subspaces of the tangent space. The vectors  $v$  in a  $p$ -dimensional subspace can be efficiently described by giving  $(n-p)$  differential 1-forms  $\omega$  such that

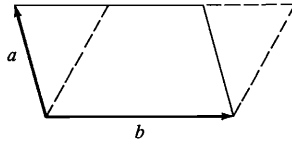
$$\omega \cdot v = 0.$$

We have used this before in describing contact structure.

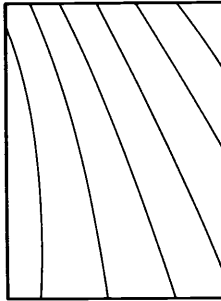
The antisymmetry of forms is important and useful, particularly in integration. Integration is the summation of objects described by a density over some region.

**Example:** A line integral measures the number of hyperplanes crossed by the line, as shown in Figure 22.1. The density of hyperplanes can be represented locally by a 1-form. This 1-form cuts off a standard length on the line. The line integral counts up the number of standard lengths along the line, in terms of the ideas about limits that are usual in calculus. Areas and volumes can be treated similarly.

We see from this example that we need a coordinate-free measure of the ratio of the sizes of two regions. This measure is called its *extension*, and this ratio is independent of metric notions. The measure of extension should have the following properties: (1) linearity in the size of the region, and (2) invariance under shears, since these transformations preserve areas. Thus, to represent an element of 2-extension, we could use two vectors, say,  $a$  and  $b$ , lying on the sides of the area as shown in Figure 22.2. Condition (1) suggests using the tensor product of  $a$  and  $b$  to represent the extension spanned by  $a$  and  $b$ , but condition (2) says that



**Figure 22.2.** The geometric representation of an element of 2-extension. The dashed element is an equivalent representation.



**Figure 22.3.** A situation with a finite number of lines in the plane.

we need instead to take the wedge product  $a \wedge b$ . Then we will have the expected invariance of area under shears:

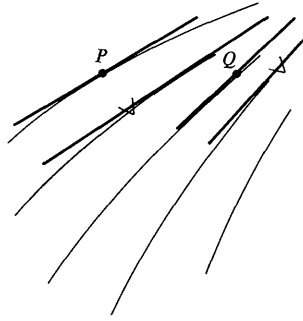
$$(a + kb) \wedge b = a \wedge b,$$

for all constants  $k$ . The extensions of small pieces of surface can thus be represented by bivectors.

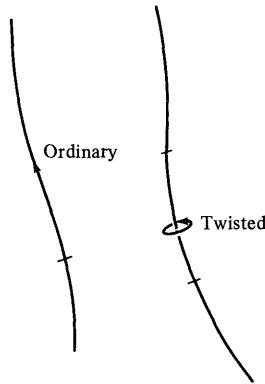
Dual to the idea of extension is that of *density*. The combination of a density and an extension is to be a number: the number of standard sizes contained in the given piece. Thus the densities will be antisymmetric, covariant tensors, that is,  $r$ -forms. Since they are being contracted with totally skew multivectors, any part not skew would be of no importance.

These  $r$ -forms also have a nice geometric interpretation. An  $r$ -form can represent the local density and alignment of a family of  $(n-r)$  surfaces in  $n$ -space.

**Example:** Suppose the plane is filled with a finite number of non-intersecting lines, for example, the grain lines in a piece of wood. Figure 22.3 sketches the situation. All the lines are drawn there;



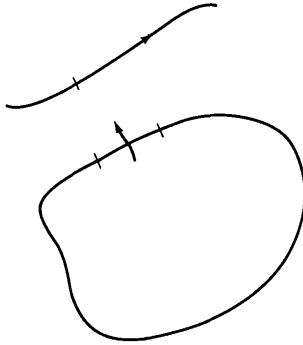
**Figure 22.4.** Using a 1-form to represent the density of lines at a point.



**Figure 22.5.** An ordinary tangent vector and a twisted tangent vector.

there are no lines that are between those drawn. At any point the local density and alignment can be represented by drawing a 1-form lying along the lines and with a spacing equal to the local line spacing (see Figure 22.4). Note that the density at point  $Q$  is higher than that at  $P$ , and is represented by a larger 1-form.

So far we have a representation of the magnitude and alignment of objects. What about the sign? This depends on the objects whose density is being represented, and we need to generalize the idea of a sign to the idea of an orientation.



**Figure 22.6.** A closed curve in two dimensions has an inside and an outside. This gives it a transverse orientation and leads to twisted tangent vectors.

**Examples:** Lines in 3-space can have two types of orientations. They can have a direction specified along the line. A piece of such a line is the usual idea of a vector in physics. Alternatively, the lines can have a circulation specified around the line. This leads to the twisted vectors given as an example in Section 1. See Figure 22.5.

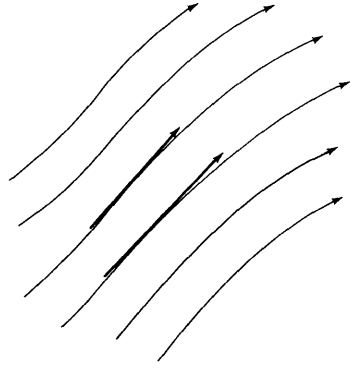
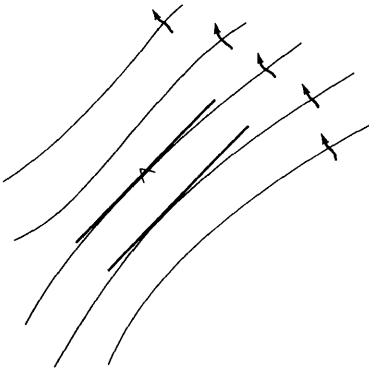
Lines in two dimensions also have two possible types of orientations. Twisted lines have a transverse direction specified. This is the natural orientation for the boundary of a two-dimensional region. The transverse direction specifies the distinction between inside and outside. See Figure 22.6.

For the plane filled with lines, if we had lines with a twisted orientation, then the density would be specified by an ordinary 1-form as shown in Figure 22.7. If the lines have the ordinary orientation, such as would represent fluid-flow streamlines, then to represent the density we will have to use twisted 1-forms, as shown in Figure 22.8.

Most of our densities will be represented by twisted differential forms. In fact, some authors call twisted forms densities. We will discuss twisted forms in detail in Section 28. For now we will mention them only occasionally and informally.

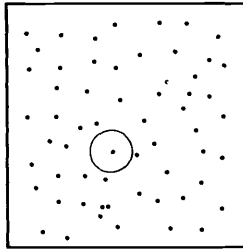
**Examples:** Suppose the plane is filled with points, say, fly specks. The density of the points is represented by a twisted 2-form. The





**Figure 22.7** (left). An ordinary 1-form represents the density of twisted lines in two dimensions.

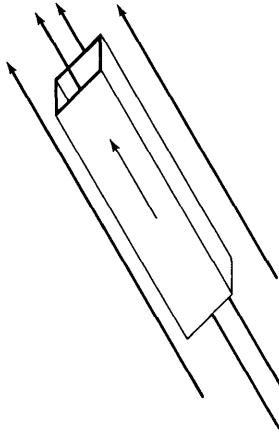
**Figure 22.8** (right). A twisted 1-form represents the density of directed lines in two dimensions.



**Figure 22.9.** Using a 2-form to represent the local density of small, point-like objects.

2-form is an area of the plane that includes on the average one point (see Figure 22.9). A point has no orientation, but we can assign a weight, positive or negative, to the points if we wish. This also works for a continuous distribution of material. The 2-form now encloses a unit amount of material, in the appropriate limit.

A twisted 2-form in 3-space represents the density of a field of lines, as shown in Figure 22.10. We will use these 2-forms in electrodynamics. Note that a 2-form is best thought of as a thing-in-itself, not necessarily something made from two 1-forms.



**Figure 22.10.** A twisted 2-form in three dimensions represents the density of directed lines.

An  $n$ -form in  $n$  dimensions gives us a standard oriented volume. As a tensor this operator takes in  $n$  vectors and yields a number. The standard  $n$ -form

$$\theta = dx^1 \wedge dx^2 \wedge dx^3 \cdots$$

takes the determinant of the components of  $n$  vectors.

### 23. Exterior calculus

The exterior calculus defines several operations on differential forms. We can *contract* an  $r$ -form with a vector. We can multiply forms by using the *exterior product*. We can extend the usual idea of a curl to an *exterior derivative*. We can take the *Lie bracket* of two vector fields, and the *Lie derivative* of differential forms. Metric ideas can be brought in by means of a duality operation, the *Hodge star* (to be discussed in the next section). Here we will give coordinate-free definitions of these operations. In Section 27 we will translate these operations into the index notation, and there provide the proofs that are messy in coordinate-free notation.

#### *Contraction*

The contraction operation  $\lrcorner$  (read this “angle”) is a map

$$\lrcorner: V \times \Lambda^r \rightarrow \Lambda^{r-1}; (a, \omega) \mapsto a \lrcorner \omega.$$

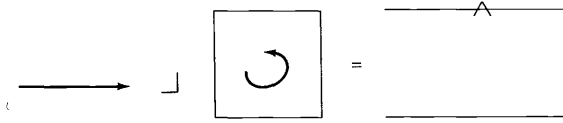


Figure 23.1. The pattern for the contraction of a 2-form with a vector.

Some use the notation  $i_a(\omega)$  for  $a \lrcorner \omega$ . Here  $V$  is the space of tangent vectors, and  $\Lambda^r$  is the space of  $r$ -forms. Now, remember that an  $r$ -form is a tensor of type  $\binom{0}{r}$ , and hence maps  $r$  vectors into a real number. Here  $a \lrcorner \omega$  is the  $(r-1)$ -form such that

$$a \lrcorner \omega \cdot (a_2, a_3, \dots, a_r) = \omega \cdot (a, a_2, \dots, a_r). \tag{23.1}$$

Because  $\omega$  is an  $r$ -form, the right-hand side is well-defined. To show that  $a \lrcorner \omega$  is an  $(r-1)$ -form, we must show that the operation is both multilinear and skew. These follow from the fact that  $\omega$  is an  $r$ -form. We also have

$$a \lrcorner f\omega = fa \lrcorner \omega = f(a \lrcorner \omega),$$

for any function  $f$ . I am trying to follow the convention of using first part of the alphabet to denote vectors, Greek for forms, and  $f$  and  $g$  for functions (which are also 0-forms).

**Example:** The contraction  $(\partial/\partial x) \lrcorner dx \wedge dy$  is sketched in Figure 23.1.

### Exterior product

The most fundamental operation is the extension of the wedge operation from pairs of 1-forms to any pair of forms:

$$\wedge: \Lambda^p \times \Lambda^q \rightarrow \Lambda^{p+q}.$$

We want the resulting algebra to be associative

$$\alpha \wedge (\beta \wedge \gamma) = (\alpha \wedge \beta) \wedge \gamma,$$

and to satisfy (not so obvious)

$$a \lrcorner (\omega \wedge \nu) = (a \lrcorner \omega) \wedge \nu + (-)^p \omega \wedge (a \lrcorner \nu),$$

where  $\omega$  is a  $p$ -form. This last equation lets us compute contractions by reducing them to contractions of single forms, which have already been defined.

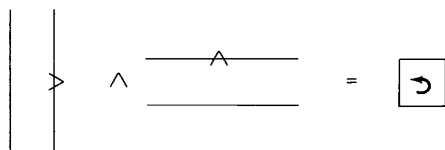


Figure 23.2. The pattern for the wedge product of two 1-forms.

**Examples:** The product  $dx \wedge dy$  is sketched in 2-space in Figure 23.2.

Consider  $(dx \wedge dy) \wedge dz$ . From equation 23.1 we must have

$$\begin{aligned} (dx \wedge dy) \wedge dz \cdot \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) &= \frac{\partial}{\partial x} \lrcorner [(dx \wedge dy) \wedge dz] \cdot \left( \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \\ &= \left[ \frac{\partial}{\partial x} \lrcorner (dx \wedge dy) \right] \wedge dz \cdot \left( \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \\ &= dy \wedge dz \cdot \left( \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) = 1. \end{aligned}$$

We are about to show that  $(dx \wedge dy) \wedge dz$  is a 3-form; so it must be, and in fact is,

$$\begin{aligned} (dx \wedge dy) \wedge dz &= dx \otimes dy \otimes dz - dx \otimes dz \otimes dy - dy \otimes dx \otimes dz \\ &\quad + dy \otimes dz \otimes dx + dz \otimes dx \otimes dy - dz \otimes dy \otimes dx. \end{aligned}$$

To show that  $\omega \wedge \nu$  as calculated by the preceding rule is a  $(p+q)$ -form, we must show that it is multilinear, which is obvious, and skew. Remember, tensors are linear operators. Note how irrelevant the coordinate transformation properties are for this discussion. To show that it is skew on the first two indices, we calculate

$$\begin{aligned} b \lrcorner [a \lrcorner (\omega \wedge \nu)] &= b \lrcorner [(a \lrcorner \omega) \wedge \nu + (-)^p \omega \wedge (a \lrcorner \nu)] \\ &= (b \lrcorner a \lrcorner \omega) \wedge \nu + (-)^{p-1} (a \lrcorner \omega) \wedge (b \lrcorner \nu) \\ &\quad + (-)^p (b \lrcorner \omega) \wedge (a \lrcorner \nu) - \omega \wedge (b \lrcorner a \lrcorner \nu). \end{aligned}$$

This is clearly antisymmetric in  $a$  and  $b$ . We can extend this to all the other indices by using induction, but it is not an interesting proof. Associativity can also be shown by a routine induction argument [see Hermann (1968) for details]. We will prove associativity in Section 27 by using the index notation.

We have

$$\omega \wedge \nu = (-)^{pq} \nu \wedge \omega.$$

This is easy to see and remember if you write this out as a sum of terms of the form  $dx \wedge dy \wedge \dots$ . These are antisymmetric, and there must be  $p$  times  $q$  sign changes in passing  $\omega$  through  $\nu$ . Calling a function a 0-form field is consistent. It passes right through because of linearity.

It is convenient to omit the  $\wedge$  and write  $dx dy$  for  $dx \wedge dy$ , and we will usually do this.

We will often need the volume element in  $n$ -space,

$$\theta = dq^1 dq^2 \dots dq^n, \quad (23.2)$$

as well as its contractions,

$$\theta_i = \frac{\partial}{\partial q^i} \lrcorner \theta. \quad (23.3)$$

These satisfy the identity

$$dq^i \wedge \theta_j = \theta \delta_j^i. \quad (23.4)$$

In three dimensional rectangular coordinates we also have

$$dq^i \wedge dq^j = \epsilon^{ijk} \theta_k$$

where  $\epsilon^{ijk}$  is the totally antisymmetric permutation symbol, with

$$\epsilon^{xyz} = 1.$$

The familiar vector cross product can be seen in the wedge product of two 1-forms in 3-space:

$$A_i dq^i \wedge B_j dq^j = \epsilon^{ijk} A_i B_j \theta_k.$$

The output here is a 2-form. To produce another 1-form requires the Hodge star operator, which will be discussed in the next section.

#### *Exterior derivative*

Just as we extended  $\wedge$  to operate on forms of any rank, so too we can extend the operator  $d$ , defined thus far as a map from 0-forms to 1-forms, to an operator

$$d: \Lambda^r \rightarrow \Lambda^{r+1}; \quad \omega \mapsto d\omega,$$

called the exterior derivative. The operator  $d$  acts on monomials according to the rule

$$d(f dq^i dq^j \dots) = df dq^i dq^j,$$

and linearly on sums of terms. The operator so defined is the unique such operator satisfying the conditions:

(1)  $d(\omega + \nu) = d\omega + d\nu,$

(2)  $d(\omega \wedge \nu) = d\omega \wedge \nu + (-)^p \omega \wedge d\nu,$  where  $\omega$  is a  $p$ -form,

(3)  $d$  is the usual operator on 0-forms,

(4)  $d(df) = 0$  on functions.

This invariant characterization shows that the  $d$  is independent of coordinates. This  $d$  is a curl operator, and condition (4) is the familiar curl-grad  $\phi = 0$  identity. The exterior derivative is as much differentiating as we can do on a manifold without giving it further structures, such as a vector field (Lie derivative) or a metric (covariant derivative). Uniqueness is obvious, since these four rules are a sufficient algorithm for the calculation of  $d$  in any basis.

**Example:** For  $\omega = 3x^2y dy + dx$ , we have

$$\begin{aligned} d\omega &= d[(3x^2y) \wedge dy] + ddx \\ &= d(3x^2y) \wedge dy + 3x^2y \wedge ddy \\ &= (6xy dx + 3x^2 dy) \wedge dy \\ &= 6xy dx dy. \end{aligned}$$

An important result is that

$$d(d\omega) = 0$$

for  $\omega$  a form of any rank. The proof will be an easy calculation in the index notation. Another important result is that  $d$  and pullback commute. We also have the useful explicit formula for a 1-form  $\alpha$

$$d\alpha \cdot (w, v) = w \lrcorner d(v \lrcorner \alpha) - v \lrcorner d(w \lrcorner \alpha) - [w, v] \lrcorner \alpha = v \lrcorner w \lrcorner d\alpha. \quad (23.5)$$

and when  $\alpha$  is a 2-form

$$\begin{aligned} d\alpha \cdot (w, v, u) &= w \lrcorner d(\alpha \cdot (v, u)) - v \lrcorner d(\alpha \cdot (w, u)) + u \lrcorner d(\alpha \cdot (w, v)) \\ &\quad - \alpha \cdot ([w, v], u) + \alpha \cdot ([w, u], v) - \alpha \cdot ([v, u], w). \end{aligned}$$

The general formula can be found in Abraham and Marsden (1978, page 121).

The familiar curl operator appears when we apply the exterior derivative to a 1-form in 3-space:

$$\begin{aligned} d(A dq) &= \frac{\partial A_x}{\partial y} dy dx + \frac{\partial A_x}{\partial z} dz dx + \cdots \\ &= \left( \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) dx dy + \cdots. \end{aligned}$$

The divergence operator appears when we apply  $d$  to a 2-form

$$\begin{aligned} d(A\theta) &= d(A_x dy dz + A_y dz dx + A_z dx dy) \\ &= \left( \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \right) \theta. \end{aligned}$$

Here  $\theta$  is the contracted volume element defined in equation 23.3.

### *Lie derivative*

This operation was defined earlier for geometric objects in general. For  $r$ -forms we have the important Cartan identity

$$\mathfrak{L}_v \omega = v \lrcorner d\omega + d(v \lrcorner \omega). \quad (23.6)$$

This is easily shown in coordinates where  $v$  is in normal form,  $\partial/\partial q^1$ . (Such normal forms will be discussed in Section 26.) Some useful theorems with uninteresting and often messy proofs are:

$$\begin{aligned} \mathfrak{L}_v d\omega &= d\mathfrak{L}_v \omega; \\ \mathfrak{L}_v(a \lrcorner \omega) &= [v, a] \lrcorner \omega + a \lrcorner \mathfrak{L}_v \omega; \\ \mathfrak{L}_{f v} \alpha &= f \mathfrak{L}_v \alpha + df \wedge (v \lrcorner \alpha); \\ u \lrcorner \mathfrak{L}_v d\omega - v \lrcorner \mathfrak{L}_u d\omega &= [u, v] \lrcorner d\omega; \\ \mathfrak{L}_v(\omega \wedge \nu) &= (\mathfrak{L}_v \omega) \wedge \nu + \omega \wedge (\mathfrak{L}_v \nu); \\ \mathfrak{L}_v \mathfrak{L}_w - \mathfrak{L}_w \mathfrak{L}_v &= \mathfrak{L}_{[v, w]}. \end{aligned} \quad (23.7)$$

Using equations 23.6 and 18.2, we can prove 23.5. Another useful relation is

$$\begin{aligned} v \lrcorner d[\omega \cdot (a, v, c, \dots)] &= \mathfrak{L}_v \omega \cdot (a, b, c, \dots) \\ &\quad + \omega \cdot ([v, a], b, c, \dots) + \omega \cdot (a, [v, b], c, \dots) + \cdots, \end{aligned}$$

where parentheses here delimit arguments; both sides of this equation are numbers.

The explicit form of equation 23.6 is useful in calculations:

$$\mathfrak{L}_{X(\partial/\partial x)} F dz = X \frac{\partial F}{\partial x} dz + dX \wedge \left( \frac{\partial}{\partial x} \lrcorner F dz \right), \quad (23.8)$$

where  $dz$  is some basis  $r$ -form, with  $d(dz) = 0$ .

**Example:** If  $F$  in equation 23.8 is independent of some particular coordinate, say,  $x$ , then the vector  $\partial/\partial x$  is a symmetry:

$$\mathfrak{L}_{(\partial/\partial x)} F dz = 0.$$

Let us redo the calculation of Section 21. For the vector field

$$V = \frac{\partial H}{\partial p} \frac{\partial}{\partial q} - \frac{\partial H}{\partial q} \frac{\partial}{\partial p}$$

with  $H$  a function of the  $qs$  and the  $ps$ , using the Cartan identity (equation 23.6), we compute

$$\begin{aligned} \mathfrak{L}_V p dq &= v \lrcorner dp \wedge dq + d\left(p \frac{\partial H}{\partial p}\right) \\ &= -\frac{\partial H}{\partial p} dp - \frac{\partial H}{\partial q} dq + d\left(p \frac{\partial H}{\partial p}\right) \\ &= d\left(-H + p \frac{\partial H}{\partial p}\right). \end{aligned}$$

Recall also that Lie derivatives commute with smooth maps. This is important, and shows that symmetries are invariant under pullback.

### Problem

**23.1** Show that equation 15.1 is correct.

## 24. The \* operator

The exterior calculus is so efficient for calculations that its extension to spaces with a metric is worthwhile. In  $n$  dimensions we do this by introducing an operator called star

$$*: \Lambda^p \rightarrow \Lambda^{n-p}.$$

This star contains both metric and orientation structure. To avoid any confusion with star for pullback or pushforward, this star operator,



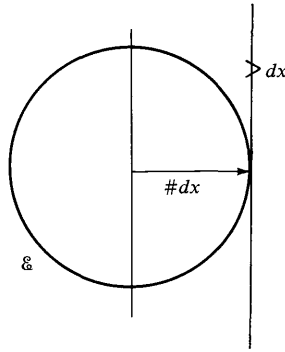


Figure 24.1. The sharp operator for two-dimensional Euclidean space.

called the Hodge star, is centered vertically and precedes the differential form.

Given a metric, we can define the *sharp* operator, which maps any 1-form, say,  $\alpha$ , into a tangent vector, which we will call  $\#\alpha$ .

**Example:** We illustrate  $\#dx$  in Euclidean 2-space in Figure 24.1.

We define the  $*$  operator to be the unique linear operator satisfying, for  $\beta$  any  $r$ -form and  $\alpha$  a 1-form,

$$*(\beta \wedge \alpha) = \#\alpha \lrcorner * \beta.$$

This reduces the problem of starring an  $(r+1)$ -form to that of starring an  $r$ -form. To begin the sequence, we have to define the star of a 0-form. Pick  $*1$  to be an  $n$ -form of unit density and positive orientation.

**Example:** In two-dimensional Euclidean space we have

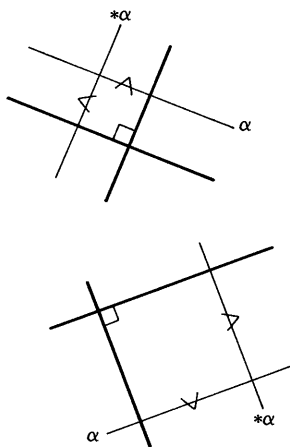
$$\#dx = \frac{\partial}{\partial x}, \quad \#dy = \frac{\partial}{\partial y},$$

and

$$*1 = dx \, dy.$$

This leads to

$$*dx = dy, \quad *dy = -dx.$$



**Figure 24.2.** The star operator for two-dimensional Euclidean space.

The operation of \* is sketched in Figure 24.2. The 1-forms  $\alpha$  and  $*\alpha$  form an oriented square, as shown.

An equivalent definition of the Hodge star operator uses an orthonormal basis of 1-forms,  $\omega^1, \omega^2, \dots, \omega^n$ , and defines the star operator first on monomials

$$*(\omega^1 \omega^2 \dots \omega^k) = (-)^p \omega^{k+1} \omega^{k+2} \dots \omega^n,$$

where  $p$  is the number of 1-forms in the set  $\omega^1, \dots, \omega^k$  of negative squared length, and

$$\Omega = \omega^1 \omega^2 \dots \omega^n$$

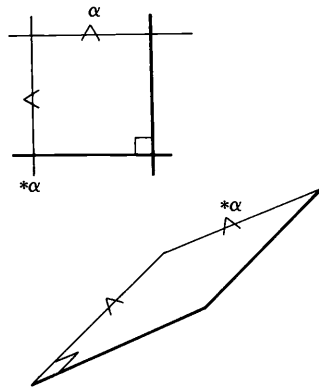
is an  $n$ -form of positive orientation. Using the antisymmetry of  $\Omega$ , we can star any monomial term; and using the linear and distributive rules, we can star any differential form whatsoever.

To prove this equivalence, apply the definition  $k$  times to find

$$*(\omega^1 \omega^2 \dots \omega^k) = \# \omega^k \lrcorner \dots \lrcorner \# \omega^2 \lrcorner \# \omega^1 \lrcorner *1.$$

**Examples:** In Minkowski 2-space we have

$$\#dt = -\frac{\partial}{\partial t}, \quad \#dx = \frac{\partial}{\partial x},$$



**Figure 24.3.** The star operator for two-dimensional spacetime.

and we take

$$*1 = dt \, dx.$$

Then

$$*dt = -dx,$$

and

$$*dx = -dt.$$

The operation of this  $*$  is sketched in Figure 24.3. The 1-forms  $\alpha$  and  $*\alpha$  now form Lorentz-perpendicular squares.

In spherical polar coordinates we have

$$\mathcal{G} = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2$$

and a suitable choice for  $*1$  will be

$$*1 = dr \wedge r d\theta \wedge r \sin \theta d\phi.$$

In three-dimensional Euclidean space, if we take

$$*1 = dx \, dy \, dz,$$

then we have

$$*dx = dy \, dz, \quad *dy = dz \, dx, \quad *dz = dx \, dy,$$

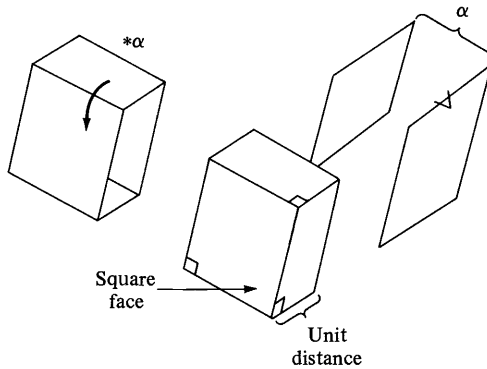


Figure 24.4. The star operator for three-dimensional Euclidean space. The 1-form  $\alpha$  turns into the 2-form  $*\alpha$ .

and

$$*(dx\,dy) = dz$$

and cyclic; and finally

$$*(\ast 1) = 1.$$

If we use the  $dq, \theta, \theta$  notation (equations 23.2 and 23.3), we have succinctly

$$\ast 1 = \theta, \quad \ast dq = \theta, \quad \ast \theta = dq,$$

and

$$\ast \theta = 1.$$

See Figure 24.4 for the action of this star operator.

In spacetime we have another Hodge star operator, which we write as  $\star$ . We can relate it to the 3-space operator  $\ast$  if we take

$$\star 1 = dt \wedge \ast 1.$$

If  $E$  is a spatial 1-form,

$$\begin{aligned} \star E &= \star(1 \wedge E) = \#E \lrcorner \star 1 \\ &= \#E \lrcorner (dt \wedge \ast 1) = -dt \wedge [\#E \lrcorner \ast 1] \end{aligned}$$

and so

$$\star E = -dt \star E.$$

Similarly,

$$\star dt = -\star 1, \quad \star dt E = -\star E,$$

and if  $B$  is a purely spatial 2-form,

$$\star B = dt \star B, \quad \star dt B = -\star B, \quad \star \star 1 = -dt.$$

Note finally that

$$\star(\star 1) = -1.$$

Using the 3-space  $dq, \theta, \theta$  notation, we have

$$\begin{aligned} \star 1 &= dt \theta, & \star dq &= -dt \theta, \\ \star dt &= -\theta, & \star \theta &= dt dq, \\ \star dt dq &= -\theta, & \star \theta &= -dt, \\ \star dt \theta &= -dq, & \star \star 1 &= -1. \end{aligned}$$

Although these may seem peculiar, such hybrid relations will be very useful when we come to discuss electrodynamics.

The vector cross product can be represented by

$$\alpha \times \beta = *(\alpha \wedge \beta).$$

Since  $*$  and  $\wedge$  do not distribute, there is no way to reduce multiple products, and hence  $\times$  is not associative. Much of the efficiency of the exterior calculus arises because the wedge operator is associative.

**Example:** The rarely used differential operator

$$(A \times \nabla) \times B,$$

where  $A, B$  are vector fields, cannot be translated into the exterior calculus.

The operator  $(r \times \nabla)$  does come up when discussing rotations, but is easily translated into a Lie derivative.

The star operator can be defined even for degenerate metrics, which are those for which  $\#$  cannot be inverted.

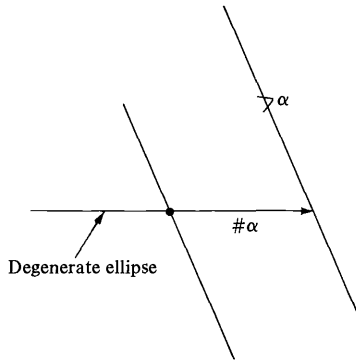


Figure 24.5. The sharp operator # mapping 1-forms into vectors for a degenerate metric.

**Example:** We illustrate # for the degenerate metric  $(\partial/\partial x) \otimes (\partial/\partial x)$  in Figure 24.5.

We show now that for a positive-definite, nondegenerate metric,

$$**\omega = (-)^{(n-r)r}\omega = (-)^{nr+r}\omega \tag{24.1}$$

where  $\omega$  is an  $r$ -form in  $n$  dimensions. Thus for  $n$  odd we have

$$** = 1,$$

and for  $n$  even

$$** = (-)^r.$$

Some authors define a conjugation operation

$$\bar{\alpha} = (-)^r\alpha,$$

which simplifies the formulas a little bit. We will not use it here.

To prove equation 24.1, take an orthonormal basis  $\omega^1, \omega^2, \dots, \omega^n$ , with

$$*1 = \omega^1\omega^2 \cdots \omega^n,$$

and look at a monomial  $r$ -form  $\alpha = \omega^1 \cdots \omega^r$ . We have lost no generality by taking these to be the leading terms. We have from our definition

$$*\alpha = (-)^q \omega^{r+1} \cdots \omega^n,$$

where  $q$  is the number of timelike 1-forms in the first  $r$  basis 1-forms. Now permute  $*1$  into the form

$$*1 = (-)^r (-)^{n-r} \omega^{r+1} \dots \omega^n \omega^1 \dots \omega^r$$

so that we have

$$**\alpha = (-)^p (-)^{r(n-r)} \alpha,$$

where  $p$  is the total number of timelike basis 1-forms, proving the result.

We also need the following theorem: if  $\alpha$  and  $\beta$  are both  $r$ -forms, then for a nondegenerate metric

$$*\alpha \wedge \beta = *\beta \wedge \alpha.$$

Since these are linear operators, it suffices to prove this for monomials. Both sides vanish unless  $\alpha$  and  $\beta$  contain the same basis forms. When they do, the result is obvious. Since  $*$  is a linear operator, it is true for  $r$ -forms in general, not just monomials. This operator extends a metric naturally to  $r$ -forms. We define  $\alpha \bullet \beta$  such that

$$\alpha \wedge *\beta = (\alpha \bullet \beta) *1$$

and from the preceding calculation this is symmetric. Many authors take this as the definition of the star operator. It gives  $*$  only as the solution of a set of simultaneous equations, and is not as direct as our definition.

#### Laplace's equation

The exterior calculus provides a natural language for discussing the familiar partial differential equations of mathematical physics. We will look at two examples here, starting with Laplace's equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = 0.$$

We write this as the first-order system

$$p = \frac{\partial \phi}{\partial x}, \quad q = \frac{\partial \phi}{\partial y}, \quad r = \frac{\partial \phi}{\partial z},$$

$$\frac{\partial p}{\partial x} + \frac{\partial q}{\partial y} + \frac{\partial r}{\partial z} = 0,$$

and introduce the 2-form

$$\alpha = p \, dy \, dz + q \, dz \, dx + r \, dx \, dy;$$

we use the Hodge star operator for the Euclidean metric, so that

$$*\alpha = p \, dx + q \, dy + r \, dz.$$

Then our first-order system is equivalent to the equations

$$\begin{aligned}d\alpha &= 0, \\d*\alpha &= 0, \\ \alpha &= *d\phi.\end{aligned}$$

Note that  $\phi$  looks like a potential for  $\alpha$ . Here  $\alpha$  is a closed 2-form, and so describes the density of conserved lines. These are the familiar flux lines of electrostatics. Laplace's second-order equation can be written

$$d*d\phi = 0.$$

The same pair of equations also describes Laplace's equation in two dimensions. Here a nice pictorial representation of the solution can be given. Recall the construction of  $*\alpha$  from  $\alpha$  using a Euclidean square. Any 1-form  $\omega$  describes the density of lines in 2-space, and one with  $d\omega = 0$  describes lines that are conserved, neither beginning nor ending. Here the 1-form  $\alpha$  describes these lines. The 1-form  $*\alpha$  describes hyper-surfaces, which in two dimensions are also curves, orthogonal to these lines. These are sketched in Figure 24.6.

### Linear wave equation

The equation

$$\frac{\partial^2 \phi}{\partial t^2} - \frac{\partial^2 \phi}{\partial x^2} - \frac{\partial^2 \phi}{\partial y^2} - \frac{\partial^2 \phi}{\partial z^2} = 0 \quad (24.2)$$

describes linear waves in three dimensions. Again, it is easier to spot the correct geometric objects if we write the equation as a first-order system. Use contact variables  $u, v, w, r$  defined by the contact form

$$d\phi - u dx - v dy - w dz - r dt.$$

Define a 3-form  $\alpha$  by

$$\star\alpha = u dx + v dy + w dz + r dt,$$

using the spacetime star operator. We have

$$d\phi = \star\alpha,$$

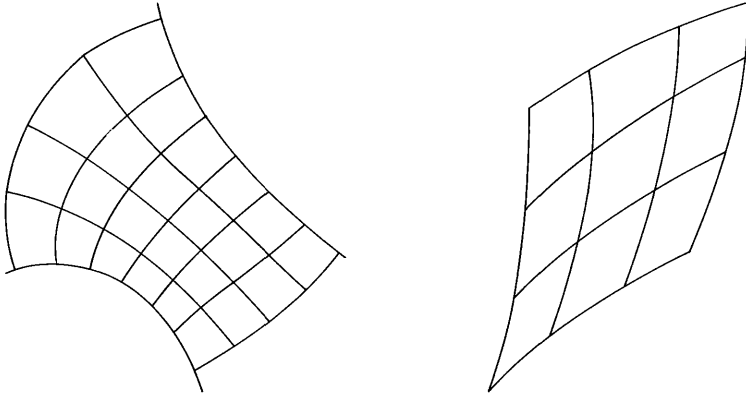
and so

$$d\star\alpha = 0.$$

Equation 24.2 turns into

$$d\alpha = 0.$$





**Figure 24.6** (left). The pattern for Laplace's equation. The two families of lines meet at right angles and form curvilinear squares.

**Figure 24.7** (right). The pattern for the wave equation. The two families of lines now are Lorentz-perpendicular.

Solutions of the wave equation can be described by conserved flux lines, crossed now by Lorentz-orthogonal hypersurfaces. In two dimensions, one space and one time, this is easily sketched, and leads to a construction using curvilinear Lorentzian squares, as shown in Figure 24.7.

### *Laplace-de Rham operator*

Contrary to what you might guess, the operator  $*d*d$  does not produce the Laplacian on differential forms in general, but only on 0-forms. The correct operator is, up to an overall sign,

$$\Delta = *d*d + (-)^n d*d*,$$

where  $n$  is the dimension of the space. There are numerous conventions for the overall sign. This operator, called the Laplacian or the Laplace-de Rham operator, yields the vector Laplacian or the vector wave equation, depending on the metric.

If you need more applications, you can now skip to Chapter V. We have covered everything needed for the first four sections of that chapter.

### **Problem**

**24.1** Use  $*$  to define a cross product on a three-dimensional Lorentz space. Does this lead to an easy derivation of the rules of pseudospherical trigonometry?

## 25. Metric symmetries

Here we examine further the symmetries of a metric. We will use differential forms and the  $*$  operator to study both isometries and conformal symmetries.

### *Isometries*

A transformation that preserves the metric is called an isometry. Such a transformation maps a figure into an image that is the same metric size and shape. Recall the discussion of Figure 21.1. An infinitesimal isometry is represented by a vector field that satisfies

$$\mathfrak{L}_k \mathcal{G} = 0.$$

This is equivalent to the commutator

$$\mathfrak{L}_k \# \omega = \# \mathfrak{L}_k \omega, \quad (25.1)$$

which has an elegant translation into the exterior calculus if we use the star operator to represent the metric,

$$\mathfrak{L}_k * \omega - * \mathfrak{L}_k \omega = [\mathfrak{L}_k, *] \omega = 0, \quad (25.2)$$

for all differential forms  $\omega$ . Although the methods of Section 18 are adequate and efficient for the discussion of isometries, it will be useful for us to reformulate those ideas in terms of the exterior calculus.

To verify equation (25.2), first note that the commutator is linear over the functions. That is

$$[\mathfrak{L}_k, *] f \omega = f [\mathfrak{L}_k, *] \omega.$$

The calculation is similar to that for the Lie bracket. Now, a Killing vector field must preserve the unit density  $*1$ , and so for all 1-forms

$$[\mathfrak{L}_k, *] f = 0.$$

Now we use induction. For  $\beta$  a 1-form,  $\alpha$  an  $r$ -form, we assume

$$[\mathfrak{L}_k, *] \alpha = 0, \quad (25.3)$$

and show that this implies that

$$[\mathfrak{L}_k, *](\alpha \wedge \beta) = 0.$$

We have

$$\mathfrak{L}_k *(\alpha \wedge \beta) - * \mathfrak{L}_k(\alpha \wedge \beta) = \mathfrak{L}_k[\# \beta \lrcorner * \alpha] - * \mathfrak{L}_k(\alpha \wedge \beta). \quad (25.4)$$

Since  $\mathfrak{L}_k$  is a derivation, this is equal to

$$\begin{aligned} & (\mathfrak{L}_k \# \beta) \lrcorner * \alpha + \# \beta \lrcorner \mathfrak{L}_k * \alpha - * [(\mathfrak{L}_k \alpha) \wedge \beta] - * [\alpha \wedge \mathfrak{L}_k \beta] \\ & = \mathfrak{L}_k \# \beta \lrcorner * \alpha + \# \beta \lrcorner \mathfrak{L}_k * \alpha - \# \beta \lrcorner * \mathfrak{L}_k \alpha - \# (\mathfrak{L}_k \beta) \lrcorner * \alpha, \end{aligned} \quad (25.5)$$

and this vanishes by virtue of equations 25.1 and 25.3.

The isometries of a given manifold with a metric form a Lie group, and the infinitesimal isometries form a Lie algebra. The Lie-algebra product is the Lie bracket of Section 9. If vector fields  $k$  and  $k'$  are both Killing vector fields, then so is the bracket  $[k, k']$ . This follows from an easy calculation using equation 23.7.

Equation 25.2 can be used to find Killing vector fields. It is sufficient to satisfy this for a 0-form and a basis of 1-forms. The bracket operation  $[\mathfrak{L}_k, *]$  is function-linear, and an induction similar to that we have just done shows that if the operation is satisfied for 0-forms and 1-forms, then it is satisfied for all higher-rank forms. This is no more efficient than a straightforward use of equation 21.1 in components.

### Conformal symmetry

A weaker requirement for a transformation is that it preserves only the shapes but not the absolute sizes of figures. In terms of the metric, this is the requirement that

$$[\mathfrak{L}_k, \#] = \phi \# \quad (25.6)$$

for some scalar function  $\phi$ . Again, this translates into a requirement on the Hodge star. We start with

$$\mathfrak{L}_k * 1 = -\frac{n\phi}{2} * 1, \quad (25.7)$$

where  $n$  is the dimension of the space. This follows from a consideration of the unit volume element, which requires  $n$  unit vectors. Each of these scales like  $(g_{\mu\nu})^{1/2}$ . This is opposite to the scaling of  $\#$ , which scales like  $g^{\mu\nu}$ . Thus we have here the factors of  $n$ , two, and the minus sign. An induction similar to that in equations 25.4 and 25.5, but using 25.6 instead of 25.1, gives us

$$[\mathfrak{L}_k, *] \omega = \left( r - \frac{n}{2} \right) \phi * \omega \quad (25.8)$$

for any  $r$ -form. This is the same scale factor  $\phi$  that appears in equation 25.6. Equation 25.6 can be solved for conformal Killing vector fields, and again it is sufficient to satisfy the equation for 0-forms and 1-forms.

*Conformal symmetries of Euclidean space*

Let us find the conformal symmetries of Euclidean 3-space. Write the symmetry vector field

$$k = X \frac{\partial}{\partial x} + Y \frac{\partial}{\partial y} + Z \frac{\partial}{\partial z}.$$

Equation 25.7 gives us

$$(X_{,x} + Y_{,y} + Z_{,z}) dx dy dz = -\frac{3\phi}{2} dx dy dz.$$

Equation 25.8 applied to  $dx$  gives us

$$(Y_{,x} + X_{,y}) dx dz + (Z_{,x} + X_{,z}) dy dx + (Y_{,y} + Z_{,z} - X_{,x} + \frac{1}{2}\phi) dy dz = 0.$$

Setting the components of these two equations and the two cyclic permutations of the second to zero gives us the system

$$X_{,x} = Y_{,y} = Z_{,z}, \quad (25.9)$$

$$X_{,y} + Y_{,x} = 0, \quad (25.10)$$

$$X_{,z} + Z_{,x} = 0, \quad (25.11)$$

$$Y_{,z} + Z_{,y} = 0. \quad (25.12)$$

These are now a bit too complicated to solve by inspection. We can derive, after following a loop through the equations, that

$$X_{,yz} = 0,$$

but this is not much of a start; we can replace  $X$ , a function of three variables, with two functions each of two variables.

A better approach is to use the Lie-algebra properties of the solutions. First, note that the solutions can be broken up into pieces that are each homogeneous in the variables  $x$ ,  $y$ , and  $z$ . Examples of such homogeneous solutions are the translations (degree zero) and the rotations (degree one), which are isometries and hence are also conformal isometries.

It is easy to find all conformal Killing vector fields of degree one by an exhaustive search. The most general such vector field has nine arbitrary constants, and we can just slug it out. We can use the discrete symmetry of permuting  $x$ ,  $y$ , and  $z$  to reduce the labor. In any event we turn up one additional independent symmetry, the expansion

$$k = x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} + z \frac{\partial}{\partial z}.$$

What about quadratic (degree two) solutions? Suppose  $k$  is such a solution. Then  $[(\partial/\partial x), k]$  will also be a solution, of degree one; so it must be some linear combination of the four that we know. We can reverse this process to find candidate solutions by integration. This involves arbitrary constants of integration, which must be chosen so that the equations are satisfied.

**Example:** From the expansion vector field we generate the candidate solution

$$k = x^2 \frac{\partial}{\partial x} + 2xy \frac{\partial}{\partial y} + 2xz \frac{\partial}{\partial z}$$

plus terms independent of  $x$ . Equation 25.9 is satisfied. To satisfy equation 25.10, we must add a term  $-y^2$  to  $X$ . To satisfy equation 25.11 we must add a term  $-z^2$  to  $X$  as well. Equation 25.9 remains satisfied, and so does equation 25.12. Thus we have found one quadratic solution

$$k = (x^2 - y^2 - z^2) \frac{\partial}{\partial x} + 2xy \frac{\partial}{\partial y} + 2xz \frac{\partial}{\partial z}$$

and two more by cyclic permutation.

The same trick can be used on the other linear solutions. By symmetry only two are needed, and these just give us the same three solutions. There can thus be no others.

We can proceed to search for higher and higher solutions. Whenever we find none, as happens here for cubic solutions, then there are no further solutions anywhere. Any higher solution could have been lowered by successive differentiations until it passed through our cubic “window,” and would be seen in the above analysis.

#### *Conformal symmetries of spacetime*

These calculations can be done for the Minkowski metric. Beyond the ordinary Killing vector fields, we find dilations

$$k = t \frac{\partial}{\partial t} + x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} + z \frac{\partial}{\partial z}$$

and four transformations called *proper conformal transformations*:

$$k = (t^2 + x^2 + y^2 + z^2) \frac{\partial}{\partial t} + 2t \left( x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} + z \frac{\partial}{\partial z} \right),$$

$$k = (x^2 + t^2 - y^2 - z^2) \frac{\partial}{\partial x} + 2x \left( t \frac{\partial}{\partial t} + y \frac{\partial}{\partial y} + z \frac{\partial}{\partial z} \right),$$

and the two cyclic permutations of the last equation.

### *Symmetries of 2-spaces*

You might have expected us to study 2-spaces first as simpler examples. Unfortunately, 2-spaces are special and cast no insight on the general case. Of course, Euclidean and Minkowski metrics are themselves special. For the general metric there will be no symmetries at all, and only very rarely any conformal symmetries except isometries.

For Euclidean 2-space we have equations for conformal Killing vector fields:

$$X_{,x} = Y_{,y}, \quad X_{,y} + Y_{,x} = 0.$$

No longer do these form an overdetermined system. Any solution of Laplace's equation generates a symmetry, for example,

$$k = \sin x \cosh y \frac{\partial}{\partial x} + \cos x \sinh y \frac{\partial}{\partial y}.$$

The algebra of symmetries is no longer finite-dimensional. This corresponds to the well-known invariance of Laplace's equation under conformal mapping.

## 26. Normal forms

A useful tactic is to pick coordinates in which the structure of the problem is particularly simple. If the structure of Euclidean geometry is part of a problem, then orthonormal coordinates will simplify this part of the problem. Naturally we are not forced to use special coordinates, and sometimes the structure of the rest of the problem may be so messed up by the special coordinates that they become useless. We will discuss here the use of special coordinates that simplify either a smooth, nonzero tangent vector field, or a smooth 1-form field, or a closed  $(n-1)$ -form field. We will find these *normal forms* useful later, and they illustrate very well the differences between tangent vectors and 1-forms.

### *Tangent vectors*

We can always find coordinates locally such that any given smooth nonzero tangent vector field can be written as  $(\partial/\partial x^1)$ , where  $x^1$  is the first coordinate in the system. It is easy to see how to construct such coordinates.

The  $x^2, \dots, x^n$  coordinates should have level surfaces containing the vector field. The  $x^1$  coordinate will have level surfaces that can be found by taking any surface transverse to the vector field and using the flow generated by the vector field to transport it over some neighborhood. This normal form for the vector field holds only locally, of course.

**Example:** Consider the equation

$$\mathcal{L}_v \alpha = \beta \tag{26.1}$$

for  $\beta$  a specified  $r$ -form field. Can this equation always be solved for  $\alpha$ ? Yes, locally. To see this, work in coordinates where

$$v = \frac{\partial}{\partial x^1}.$$

Expand the  $r$ -form  $\beta$  in a sum of terms

$$\beta = F dx^\mu dx^\nu \dots + \dots.$$

Since the Lie-derivative operation is linear, we have, expanding  $\alpha$  similarly,

$$\alpha = G dx^\mu dx^\nu \dots + \dots.$$

$$\mathcal{L}_v \alpha = \frac{\partial G}{\partial x^1} dx^\mu dx^\nu \dots + \dots$$

and the coefficients  $G$ , etc. are given by integration

$$\frac{\partial G}{\partial x^1} = F.$$

Therefore there are no obstructions that prevent us from solving equation 26.1 in some local region.

### *1-Forms*

The situation here is much richer. A nonzero differential form has a normal form in  $n$  dimensions that looks like

$$\omega = x^1 dx^2 + x^3 dx^4 + \dots + x^{2m-1} dx^{2m} + dx^n.$$

The value of  $m$  depends on the algebraic structure of the 2-form  $d\omega$ , and the factor  $dx^n$  may or may not be present.

Taking for granted the result, we can distinguish the cases as follows. We must have

$$(d\omega)^m = d\omega \wedge d\omega \wedge \cdots \wedge d\omega \neq 0 \quad (m \text{ factors}),$$

but

$$(d\omega)^{m+1} = 0.$$

Thus we can find  $m$ . Also, we must have

$$\omega \wedge (d\omega)^m = 0$$

if the factor  $dx^n$  is absent, but not otherwise. See Slebodzinski (1970) or Sternberg (1964) for proofs and fuller discussion. A special case will be proven as an example in Section 38.

**Example:** Let us find the necessary and sufficient condition for a scalar function  $e^\phi$  to be an integrating factor for a 1-form  $\omega$ . That is, we wish to find  $e^\phi$  such that

$$d(e^\phi \omega) = 0. \quad (26.2)$$

The condition is

$$\omega \wedge d\omega = 0 \quad (26.3)$$

called an integrability condition. To see that it is necessary, we expand equation 26.2, find

$$d\omega + d\phi \wedge \omega = 0,$$

and wedge in  $\omega$  to find equation 26.3. To see that it is sufficient, suppose we put  $\omega$  into normal form:

$$\omega = x^1 dx^2 + x^3 dx^4 + \cdots + dx^n.$$

Now clearly it cannot have a  $dx^n$  term and still satisfy the integrability condition; we would have terms in  $\omega \wedge d\omega$  like  $dx^n \wedge dx^1 \wedge dx^2$ . Neither can we have more than a single monomial, or we would find terms such as  $x^1 dx^2 \wedge dx^3 \wedge dx^4$ . Thus any 1-form satisfying equation 26.3 can be written in the form

$$\omega = x^1 dx^2,$$

and this has the integrating factor  $1/x^1$ .

The normal form for 1-forms leads immediately to a normal form for closed 2-forms:

$$d\omega = dx^1 \wedge dx^2 + \cdots + dx^{m-1} \wedge dx^m.$$



This result is called the Darboux theorem, and will be used when we discuss symplectic geometry in Section 51.

### *Clebsch form*

This is a local normal form for closed  $(n-1)$ -forms and general  $(n-2)$ -forms. The  $(n-1)$ -form can be written

$$d\omega = dx^1 dx^2 \cdots dx^{n-1}.$$

This is geometrically obvious, and corresponds to the simple normal form for tangent vectors. Locally, then, we have

$$\omega = d\alpha + x^1 dx^2 \cdots dx^{n-1},$$

where  $\alpha$  is some  $(n-3)$ -form.

**Example:** In  $\mathbb{R}^3$  we can write any closed 2-form as

$$d\omega = dx dy$$

and any 1-form as

$$\omega = d\alpha + x dy.$$

This agrees with our previous normal form for 1-forms.

### **Problem**

**26.1** What conditions must a 1-form  $\omega$  satisfy in order to generate a contact structure in a  $(2n+1)$ -dimensional space? Would

$$\omega = x dy + y dz + w dx + t dw$$

do in  $\mathbb{R}^5$ ? Would

$$\omega = (3x^2 + t) dx + dy - (y+z) dz + t dt + w(3wt - 2y - 2z) dw?$$

### **27. Index notation**

Both index-free and index-intensive notations are useful. Some problems fall easily to one approach, some to the other. Although the coordinate-free methods of the exterior calculus are excellent for dealing with differential forms, they must be strained quite a bit to handle stress-energy tensors, as we will see in Section 41. The index notation, although typographically baroque, is the most versatile notation, and should be used for those few cases where the abridged notations become clumsy or

ambiguous. [There is a good discussion of all this in Schouten (1954). This section is mainly for reference. Don't get ill reading it.]

### *Differential forms*

With a coordinate basis  $dx^\mu$ , an  $r$ -form  $\omega$  has components

$$\omega = \omega_{\mu\nu\dots} dx^\mu \otimes dx^\nu \dots$$

The components are totally antisymmetric. This can also be written

$$\omega = \omega_{|\mu\nu\dots|} dx^\mu \wedge dx^\nu \dots,$$

where the vertical bars indicate an *ordered* collection of indices  $\mu < \nu < \dots$ .

**Example:** A 2-form  $\omega$  in  $\mathbb{R}^3$  has a tensor expansion

$$\omega = \omega_{\mu\nu} dx^\mu \otimes dx^\nu,$$

with components

$$\begin{aligned} \omega_{xx} = \omega_{yy} = \omega_{zz} &= 0, \\ \omega_{xy} = -\omega_{yx}, \quad \omega_{yz} = -\omega_{zy}, \quad \omega_{xz} = -\omega_{zx}. \end{aligned}$$

Only three of these nine components are independent. If we use an ordered summation, we can ignore all components but  $\omega_{xy}$ ,  $\omega_{yz}$ , and  $\omega_{xz}$ , writing

$$\omega = \omega_{|\mu\nu|} dx^\mu \wedge dx^\nu = \omega_{\mu\nu} dx^{|\mu} \wedge dx^{\nu|}.$$

### *Exterior product*

We introduce for this the permutation tensor  $\delta_{\mu\nu\sigma\dots}^{\alpha\beta\gamma\dots}$ . This function of the labels  $\alpha, \beta, \gamma, \dots$  and  $\mu, \nu, \sigma, \dots$  equals  $+1$  if the labels  $\alpha, \beta, \gamma, \dots$  are all different and are an even permutation of the labels  $\mu, \nu, \sigma, \dots$ ; equals  $-1$  for an odd permutation; and vanishes otherwise. They are called *Kronecker deltas*. To avoid the cumbersome summations over an indeterminate number of indices, we introduce block indices: subscript and superscript capital letters will stand for a string of indices of unspecified length. We will use the obvious summation convention on these block indices.

The exterior product  $\gamma = \alpha \wedge \beta$  is written in terms of block indices

$$\gamma_A = \delta_A^{BC} \alpha_{|B|} \beta_{|C|}. \quad (27.1)$$

Not all strings represented by block indices need to have the same length.

**Example:** If  $\alpha$  and  $\beta$  are 2-forms, we have, in ordinary indices,

$$\gamma_{\mu\nu\sigma\tau} = \delta_{\mu\nu\sigma\tau}^{\xi\zeta\eta\theta} \alpha_{|\xi\zeta|} \beta_{|\eta\theta|}.$$

If we take the specific example

$$\alpha = dx dy, \quad \beta = dx dz + dz dw,$$

and an ordering  $(x, y, z, w)$ , then the sum includes only the single term

$$\gamma_{\mu\nu\sigma\tau} = \delta_{\mu\nu\sigma\tau}^{xyzw},$$

and we have

$$\gamma = dx dy dz dw.$$

Equation 27.1 is easily proven. It is a linear operator on  $\alpha$  and  $\beta$ ; so it is sufficient to prove it on the basis  $r$ -forms, and this is easy.

We can now easily prove the associative rule for exterior multiplication. Recall the use of  $\delta_\nu^\mu$  to relabel indices:

$$T^\nu \delta_\nu^\mu = T^\mu,$$

which can be extended to all the deltas,

$$T^A \delta_{|A|}^B = T^B = (1/N!) T^A \delta_A^B, \quad (27.2)$$

where an ordered summation avoids multiple counting;  $N$  is the number of indices in a block. Thus we wish to show that

$$\alpha \wedge (\beta \wedge \gamma) = (\alpha \wedge \beta) \wedge \gamma,$$

which is equivalent to

$$\alpha_{|A|} \delta_C^{AE} \delta_E^{BD} \beta_{|B|} \gamma_{|D|} = \alpha_{|A|} \delta_E^{AB} \beta_{|B|} \delta_C^{ED} \gamma_{|D|}.$$

This follows from

$$\delta_C^{AE} \delta_E^{BD} = \delta_E^{AB} \delta_C^{ED} = \delta_C^{ABD}$$

using equation 27.2. Note that the ordering bars can be shifted from one term in a sum to the other.

Many books use the antisymmetrization operation denoted by square brackets

$$T_{[A]} = (1/N!) T_B \delta_A^B,$$

where  $N$  is the number of indices in the string.

*Exterior derivative*

Take a  $p$ -form  $\omega$

$$\omega = \omega_{|A|} dx^A = (1/p!) \omega_A dx^A,$$

where

$$dx^A = dx^\alpha dx^\beta dx^\gamma \dots$$

and

$$A = \{\alpha, \beta, \gamma, \dots\}.$$

We have the explicit formula

$$d\omega = d\omega_{|A|} \wedge dx^A.$$

Now we can verify that this definition satisfies the four properties we demanded in Section 23. Number two is the only one that is not immediate. If we have

$$\gamma = \omega \wedge \nu$$

for  $\omega \in \Lambda^p$  and  $\nu \in \Lambda^r$  (that is,  $\omega$  is a  $p$ -form and  $\nu$  is an  $r$ -form), then

$$\gamma = \gamma_{|A|} dx^A = \frac{1}{(r+p)! r! p!} \delta_A^{BC} \omega_B \nu_C dx^A$$

and

$$d\gamma = \frac{1}{r! p!} (\nu_C d\omega_B + \omega_B d\nu_C) \wedge \frac{\delta_A^{BC} dx^A}{(r+p)!}$$

But we have

$$\frac{1}{(r+p)!} \delta_A^{BC} dx^A = \delta_{|A|}^{BC} dx^A = dx^{BC} = dx^B \wedge dx^C$$

and so

$$d\gamma = \frac{d\omega_B \wedge dx^B}{p!} \wedge \frac{\nu_C dx^C}{r!} + (-)^p \frac{\omega_B dx^B}{p!} \wedge \frac{d\nu_C \wedge dx^C}{r!}$$

where the  $(-)^p$  comes from pulling the 1-form  $d\nu_C$  through the  $p$ -form  $dx^A$ . Finally

$$d\gamma = d\omega \wedge \nu + (-)^p \omega \wedge d\nu.$$

This shows the consistency of our definition in Section 23.

### Determinants

A geometric discussion of the determinant was given in the Preface. The algebraic discussion of determinants is done by introducing *permutation symbols*  $\epsilon^A$  and  $\epsilon_A$  with exactly as many indices as the space has dimensions. Both  $\epsilon^A$  and  $\epsilon_A$  are totally antisymmetric, and are +1 if  $A$  contains an even permutation of the coordinate labels. We will see soon that if these definitions are used in different coordinate systems, they are the components of different tensors. Clearly we have

$$\epsilon^A \epsilon_B = \delta_B^A$$

and

$$\epsilon^A \epsilon_A = N!$$

The determinant of a matrix can be written in a number of ways:

$$\begin{aligned} M^1_\alpha M^2_\beta M^3_\gamma \cdots \epsilon^{\alpha\beta\gamma\cdots} &= \text{Det}(M), \\ M^\mu_\alpha M^\nu_\beta M^\sigma_\gamma \cdots \epsilon^{\alpha\beta\gamma\cdots} &= \epsilon^{\mu\nu\sigma\cdots} \text{Det}(M), \\ M^\mu_\alpha M^\nu_\beta M^\sigma_\gamma \cdots \epsilon^{\alpha\beta\gamma\cdots} \epsilon_{\mu\nu\sigma\cdots} &= N! \text{Det}(M); \end{aligned}$$

and if we extend the idea of block indices to include multiple copies of the tensor when it does not have enough indices otherwise, we can write these last equations as

$$M^B_A \epsilon^A = \epsilon^B \text{Det}(M), \quad (27.3)$$

$$M^B_A \epsilon^A \epsilon_B = N! \text{Det}(M), \quad (27.4)$$

where the index blocks have  $N$  indices. In fact, if our index blocks are shorter than the dimension of the space, the left-hand side of equation 27.4 gives us all the algebraic invariants of a matrix from the trace on up.

An important theorem is that the determinant of the product of two matrices is the product of the determinants. This is easily shown by means of block indices. The product of two matrices  $M^\alpha_\beta$  and  $N^\gamma_\delta$  is  $M^\alpha_\gamma N^\gamma_\delta$ . Using block indices, we find it is  $M^A_G N^G_D$ . Thus the product of the determinants is

$$\text{Det}(M) \text{Det}(N) = (1/N!)^2 M^A_B \delta^B_A N^C_D \delta^D_C.$$

But

$$\delta^B_A \delta^D_C = \epsilon^B \epsilon_A \epsilon^D \epsilon_C = \delta^B_C \delta^D_A$$

and so

$$\begin{aligned}\text{Det}(M)\text{Det}(N) &= (1/N!)^2 (M^A_B \delta_C^B) N^C_D \delta_A^D \\ &= (1/N!) M^A_C N^C_D \delta_A^D = \text{Det}(MN).\end{aligned}$$

Cramer's rule follows easily from these definitions. We define the *minor* of the matrix  $a_{ij}$  to be the array

$$A^{ij} = \epsilon^{jp\cdots} \epsilon^{iq\cdots} a_{pq\cdots}.$$

The inverse of the matrix  $a_{ij}$  is then

$$a^{ij} = A^{ij}/\text{Det}(a),$$

because we have

$$\begin{aligned}a^{ij} a_{jk} &= \frac{A^{ij} a_{jk}}{\text{Det}(a)} \\ &= \frac{\epsilon^{jp\cdots} \epsilon^{iq\cdots} a_{jk} a_{pq\cdots}}{\text{Det}(a)} \\ &= \epsilon_{kp\cdots} \epsilon^{iq\cdots} = \delta_k^i.\end{aligned}$$

We can also find the partial derivative of a determinant

$$\begin{aligned}\text{Det}(a) &= \epsilon^{12\cdots} \epsilon^{pq\cdots} a_{1p} a_{2q} \cdots, \\ \frac{\partial}{\partial x^k} (\text{Det } a) &= \epsilon^{12\cdots} \epsilon^{pq\cdots} \frac{\partial a_{1p}}{\partial x^k} a_{2q} \cdots + \cdots \\ &= A^{pj} \frac{\partial a_{jp}}{\partial x^k};\end{aligned}$$

so in terms of the inverse we have

$$\frac{1}{\text{Det}(a)} \frac{\partial}{\partial x^k} \text{Det}(a) = a^{pj} a_{jp,k}.$$

### Permutation tensors

To see that these permutation symbols are not the components of tensors, we check to see if they transform like the components of a tensor. If they were tensors, then in a new, primed coordinate system

$$\epsilon_{A'} \stackrel{?}{=} J_{A'}^B \epsilon_B,$$

where  $J$  is the Jacobian matrix

$$J_{\nu'}^{\mu} = \frac{\partial x^{\mu}}{\partial x'^{\nu'}}.$$

But, using equation 27.3, we find that this leads to

$$\epsilon_{A'} \stackrel{?}{=} \text{Det}(J) \epsilon_A \quad (27.5)$$

which contradicts the definition that gives  $\epsilon_A$  the same numerical components in all coordinate systems. Only for coordinates related by transformations preserving volume and orientation will it define the same tensor.

On a space with a metric and an orientation, we can compensate for the determinant in the transformation law (equation 27.5) by using the determinant of the metric tensor. We define

$$g = \text{Det}(g_{\mu\nu}) = (1/N!) g_{AB} \epsilon^A \epsilon^B.$$

Note that  $N$  copies of  $g_{\mu\nu}$  are needed when you write out the summation. This determinant in a new system of coordinates will be given by

$$\begin{aligned} g' &= \frac{1}{N!} g_{A'B'} \epsilon^{A'} \epsilon^{B'} = \frac{1}{N!} g_{CD} J_{A'}^C J_{B'}^D \epsilon^{A'} \epsilon^{B'} \\ &= \frac{1}{N!} g_{CD} \epsilon^C \epsilon^D \text{Det}^2(J) = \text{Det}^2(J) g. \end{aligned} \quad (27.6)$$

Thus we are led to define the quantity

$$\eta_A = \sqrt{|g|} \epsilon_A.$$

As a tensor  $\eta_A$  maps  $r$ -multivectors into  $(n-r)$ -forms. In new coordinates its components are given by

$$\epsilon_{A'} \sqrt{|g'|} \stackrel{?}{=} J_{A'}^A \epsilon_A \sqrt{|g|},$$

and from equation 27.6 we have

$$\epsilon_{A'} |\text{Det}(J)| \sqrt{|g|} \stackrel{?}{=} \text{Det}(J) \epsilon_{A'} \sqrt{|g|}.$$

The determinants of the two Jacobian matrices cancel up to a sign. Similarly we can define

$$\eta^A = \frac{\epsilon^A}{\sqrt{|g|}} \text{sgn}(g)$$

and these are related by

$$g_{AB} \eta^B = \eta_A.$$

What about the lost sign? The sign changes under odd permutations, and this means that  $\eta_A$  represents not an  $n$ -form but a twisted  $n$ -form,

which will be discussed in detail in the next section. We can also define  $\eta_A$  on a space with a volume element but no metric, the transformation of the volume element being also a source of  $\text{Det } J$ .

### Problems

- 27.1 Show that the geometric definition of a determinant given in the Preface is independent of the parallelepiped chosen.
- 27.2 Does  $\eta_A$  exist on a space, such as the phase space of mechanics, that has only a distinguished volume element but no metric?

## 28. Twisted differential forms

It is now time to carefully discuss twisted tensors and twisted differential forms. Although most of the material in this book is quite standard, twisted tensors are not. This is partly the fault of tensor analysis and its myopic emphasis on coordinate transformations. Twisted tensors do transform with a sign change under inversion of coordinates, and that may seem unnatural. But the fact is that under such an inversion a twisted tensor does not itself change, and its coordinate representation changes sign only to follow the changing coordinates. Most of the geometric objects that we will need in our physical theories are actually twisted tensors. However, the coordinate representations of ordinary tensors are simpler; so we will represent our twisted tensors in terms of ordinary tensors. The precise representation will be developed in this section. (You might now like to leaf through Section 42 to see how twisted forms will be used.)

Twisted tensors were introduced by Hermann Weyl (1922). Schouten (1951) developed them, and called them Weyl tensors or  $W$ -tensors. Synge and Schild (1949) refer to them as oriented tensors, and de Rham (1960) called them tensors of odd kind. Nowadays they are usually called twisted tensors. We could make out a good case that the usual differential forms are actually the twisted ones, but the language is forced on us by history. Twisted differential forms are the natural representations for densities, and sometimes actually are called densities, which would be an ideal name were it not already in use in tensor analysis. I agonized over a notation for twisted tensors, say, a different typeface. In the end I decided against it. There is too much notation already floating around here. Most of the time twisted forms will be written just like ordinary forms. When I need to identify a form  $\alpha$  as twisted, it will be written  $(\alpha, \Omega)$  to emphasize that it differs from an ordinary form  $\alpha$  by an orientation  $\Omega$ .



I will discuss twisted tensors in two ways: first, in terms of their intrinsic properties, to emphasize that they are geometric objects and are as natural and fundamental as ordinary tensors; then in terms of a representation involving ordinary tensors, which is the easiest representation to manipulate although it unfortunately makes them seem like subsidiary objects. A careful definition of twisted tensors will appear only in the second discussion.

### *Orientations*

To orient an  $n$ -dimensional vector space is to choose an ordered basis, which is an ordered set of  $n$  vectors that span the space. Even permutations of these vectors specify the same orientation; odd ones define the opposite orientation. Further, any basis related by a transformation with a positive determinant defines the same orientation. A basis deformed in any way that maintains a nonzero volume will still represent the same orientation.

**Example:** In three dimensions we can deform a basis as long as no vector passes through the plane defined by the other two.

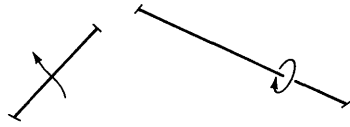
An orientation can also be specified by giving an ordered set of  $n$  1-forms that span the dual space; that is, the  $n$ -form that results from multiplying them together using wedge is nonzero. For our purposes, it is more convenient to specify orientations by this  $n$ -form.

Now consider a linear subspace of a vector space, say, of dimension  $k$ . One natural way to orient the  $k$ -subspace is to treat it as a vector space and choose a basis of  $k$  1-forms for the subspace. This is called an *inner orientation*. Another and different way to orient the linear subspace is to choose an ordered set of  $(n-k)$  1-forms, all of which pull back to zero on the subspace and which wedge together to form a nonzero  $(n-k)$ -form. This is called an *outer orientation* or a *transverse orientation*.

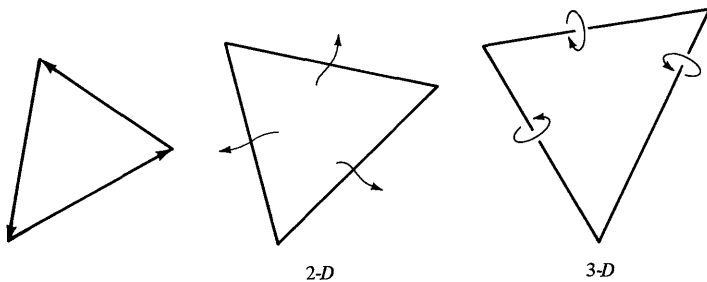
**Example:** The two transverse orientations for the line  $x = y = 0$  could be represented by the 2-forms  $dx dy$  and  $dy dx$ . The two inner orientations are represented by  $dz$  and  $-dz$ .

A transverse orientation may be deformed as long as its elements all pull back to zero and their product remains nonzero. A transverse orientation for the entire vector space is by convention just a choice of sign, plus or minus.

These two types of orientation, inner and outer, distinguish twisted



**Figure 28.1.** A twisted vector in two dimensions (left) and one in three dimensions (right).



**Figure 28.2.** The addition of ordinary vectors contrasted with the addition of twisted vectors.

geometric objects from ordinary ones. Look at an ordinary vector: it is a directed line segment of definite length. It has an outer orientation. On the other hand, a twisted vector is a line segment of definite length, but having an outer orientation instead of an inner orientation. In Figure 28.1, I sketch twisted vectors in two and three dimensions. These twisted vectors form a vector space. Their addition is compared with that for ordinary vectors in Figure 28.2. In each I show three vectors that add up to zero.

The usual representation of a 1-form as a pair of hyperplanes has an outer orientation relative to the hyperplanes given by the 1-form itself. The usual representation of a 2-form is a tube of specified area and alignment. The linear subspace here lies along the tube. The orientation is given by the 2-form itself.

A twisted 1-form will be an object in the vector space dual to the twisted vectors. A geometric picture of this operator in two dimensions is given in Figure 28.3, and in three dimensions in Figure 28.4. For ordinary 1-forms the level surfaces have an outer orientation; for twisted 1-forms they have an inner orientation. In Figure 28.5, I compare the addition of twisted 1-forms with the addition of ordinary 1-forms in two dimensions. The method will be explained in a moment.

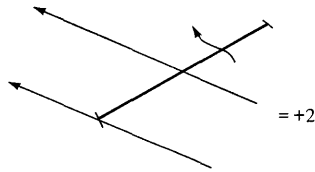


Figure 28.3. The duality between twisted 1-forms and twisted vectors in two dimensions.

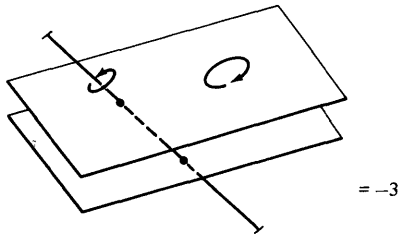


Figure 28.4. The duality between twisted 1-forms and twisted vectors in three dimensions.

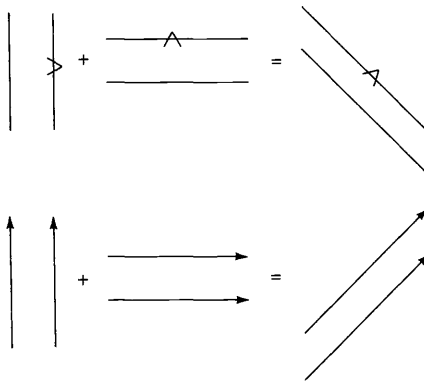
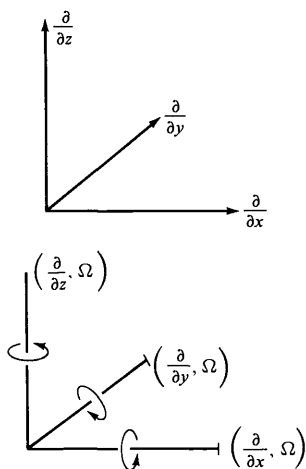
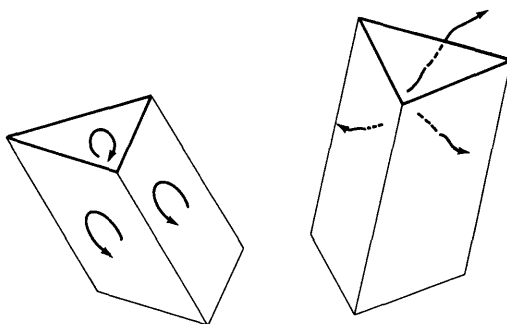


Figure 28.5. The addition of ordinary 1-forms in two dimensions contrasted with that of twisted 1-forms.

**Example:** In Figure 28.6 I draw a set of ordinary basis vectors and a set of twisted basis vectors in three dimensions. These twisted vectors were generated from the ordinary basis vectors by the right-hand rule.



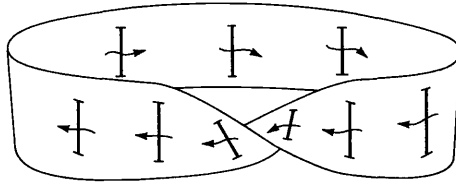
**Figure 28.6.** Basis vectors and twisted basis vectors in three dimensions.



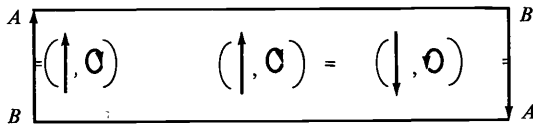
**Figure 28.7.** The addition of bivectors and twisted bivectors in three dimensions.

A bivector has a representation in the tangent space as a parallelogram of definite area with an inner orientation. A twisted bivector has instead an outer orientation. In Figure 28.7 I sketch the addition of both types of bivectors in three dimensions.

A twisted 2-form will be a linear operator on twisted bivectors. Again, it is related to an ordinary 2-form by a right-hand rule. Note that a twisted 2-form is the correct geometric object to represent the flow of electric charge. The tube lies along the flow of current, has a size that encloses unit current, and has an orientation given by the direction of the current flow.



**Figure 28.8.** A field of twisted vectors on the Moebius strip. The vectors are all transverse to the edge.



**Figure 28.9.** The twisted vector field of Figure 28.8, drawn using the representation in terms of ordinary vectors shown in Figure 28.10.

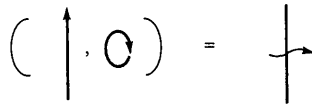
In a similar manner we can extend the idea of twisted tensors to multi-vectors and differential forms of any rank. In three dimensions a 3-form has a screw sense; a twisted 3-form has a sign, plus or minus.

**Example:** To see that there is a nontrivial distinction here, consider a Moebius strip. We cannot find a continuous vector field on it that is everywhere transverse to the edge. A twisted vector field satisfying these conditions is sketched in Figure 28.8.

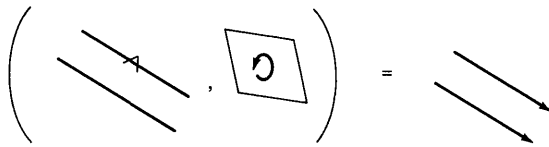
*Representation of twisted tensors*

To deal analytically with tensors of any sort, we choose a basis and represent the tensors as linear combinations of these basis elements. A given coordinate system leads naturally to a basis for ordinary vectors and 1-forms.

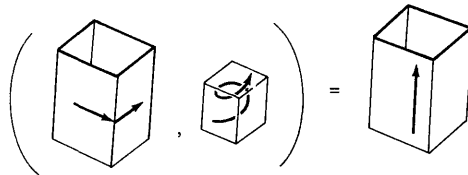
The basis of tangent vectors leads naturally to a basis for all the ordinary tensor spaces. When dealing with differential forms, we saw that it was economical to pick an orientation for the whole space and to use ordered summations to eliminate redundant terms. Choosing an orientation for the whole space also simplifies the representation of twisted tensors. We can formally define a twisted tensor as a pair, consisting of an ordinary tensor and an orientation, reduced by the equivalence relation



**Figure 28.10.** The representation of a twisted vector by an ordinary vector paired with an orientation.



**Figure 28.11.** The representation of a twisted 1-form in two dimensions.



**Figure 28.12.** The representation of a twisted 1-form in three dimensions.

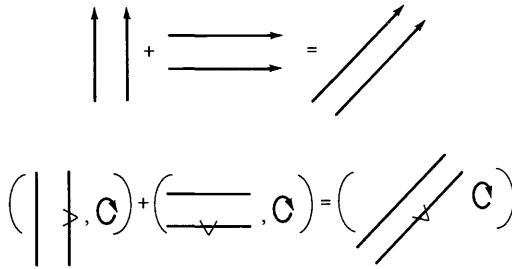
$$(\omega, \Omega) \equiv (-\omega, -\Omega),$$

where  $-\Omega$  is the orientation opposite to  $\Omega$ .

**Examples:** Let us use this representation for twisted vectors on Figure 28.8. A cut-apart Moebius strip is shown in Figure 28.9. The preceding equivalence is needed to have continuity across the join.

Figure 28.10 shows the representation scheme for vectors. The rule for 1-forms is shown in Figure 28.11 for two dimensions, and in Figure 28.12 for three dimensions. The rule will be given in equation 28.1.

Do not let the fact that twisted forms are being represented in terms of ordinary forms lead you to think that twisted forms are less fundamental.



**Figure 28.13.** The addition of twisted 1-forms shown both directly and by using the representation in terms of ordinary 1-forms paired with orientations.

The situation here is symmetric: we could define an ordinary vector as a pair consisting of a twisted vector and an orientation under the same equivalence relation.

**Example:** The addition of twisted 1-forms shown in Figure 28.5 can be demonstrated by using the representation in terms of ordinary forms, as shown in Figure 28.13.

To use a specific orientation  $\Omega$  of the whole space to represent twisted forms, we give the twisted form  $(\alpha, \Omega)$  an orientation  $\Omega_\alpha$  satisfying

$$\Omega_\alpha \wedge \alpha = \Omega. \tag{28.1}$$

Great care is needed in picking the conventions here. A careless choice will have peculiar signs all over the place.

**Examples:** The 1-form  $dx$  has an orientation specified by the 1-form  $dx$ . The twisted 1-form  $(dx, dx dy dz)$  must have the orientation specified by  $dy dz$ . This is the usual right-hand rule, pictured in Figure 28.14. The representation  $dx \mapsto (dx, \Omega)$  pictured depends on  $\Omega$  but not on the coordinates.

The twisted basis vectors in two dimensions,  $(dx, dx dy)$  and  $(dy, dx dy)$  are sketched in Figure 28.15.

A basis for twisted 2-forms in three dimensions is given by

$$j_x = (dy dz, dx dy dz)$$

and so on. Note that the orientation is being used only to give an unnatural  $y$ - $z$ -ish name to a purely  $x$ -ish object. See Figure 28.16.

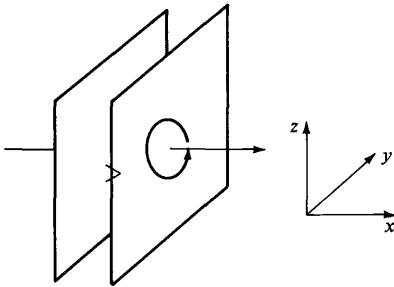


Figure 28.14. The representation of  $dx$  by the right-hand rule.

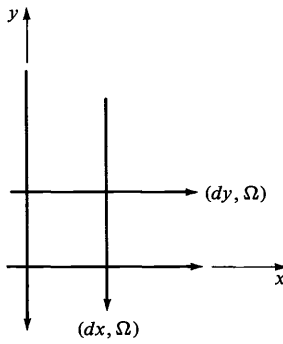


Figure 28.15. Twisted basis 1-forms in two dimensions.

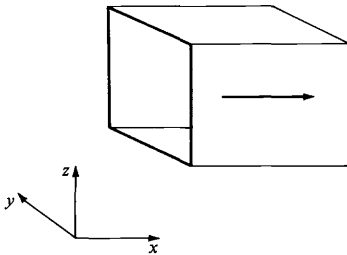


Figure 28.16. The basis twisted 2-form  $j_x$  in three dimensions.



Most of the time I will write the twisted form  $(\alpha, \Omega)$  as just  $\alpha$  unless confusion would result.

### *Operations on twisted forms*

We can extend the usual operations of the exterior calculus to twisted forms.

The wedge product of a twisted form and an ordinary form is itself twisted. We use the rule

$$(\alpha \wedge \beta, \Omega) = (\alpha, \Omega) \wedge \beta = \alpha \wedge (\beta, \Omega).$$

The orientation given by this satisfies

$$\{(\alpha \wedge (\beta, \Omega)), \{\alpha\}\} = \{(\beta, \Omega)\},$$

where I am writing  $\{\alpha\}$  for the differential form that specifies the orientation of  $\alpha$ , and so on. If  $\alpha$  is a  $p$ -form and  $\beta$  is a twisted  $q$ -form in  $n$ -space, then each side of the preceding equation is an  $(n - q)$ -form.

The exterior derivative of a twisted form is again twisted. We define it using

$$d(\alpha, \Omega) = (d\alpha, \Omega).$$

### *Pullback*

One reason why differential forms are so much more useful than vectors is their nice behavior under pullback. With twisted forms this behavior is not so simple. Twisted forms can be pulled back only onto surfaces that have a transverse orientation.

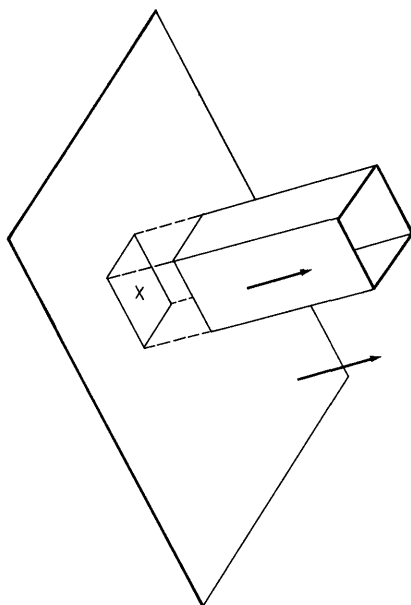
**Example:** In Figure 28.17 we pull back a twisted 2-form in  $\mathbb{R}^3$  onto a two-dimensional subspace. A twisted 2-form in two dimensions has a sign, here positive if the orientations of the subspace and the 2-form are equal.

If the transverse orientation is given by  $\eta$ , then the pullback  $\psi^* \cdot (\alpha, \Omega)$  has the orientation given by

$$\eta \wedge \{\psi^* \cdot (\alpha, \Omega)\} = \{(\alpha, \Omega)\}.$$

**Examples:** In  $\mathbb{R}^2$  we pull back  $(dy, dx dy)$  onto the surface  $x = 0$  with transverse orientation  $dx$ . We have

$$\{\psi^* \cdot (dy, dy dz)\} = +1.$$



**Figure 28.17.** A twisted 2-form being pulled back onto a surface that has a transverse orientation.

In  $\mathbb{R}^3$  we pull back  $(dy, dx dy dz)$  onto the surface  $x = 0$  with transverse orientation  $dx$ , finding

$$\{\psi^* \cdot (dy, dx dy dz)\} = -dz.$$

This pullback operation will be used in the next section to formulate the divergence theorem, and in Chapter VI to formulate the junction conditions of electrodynamics.

### *Hodge star*

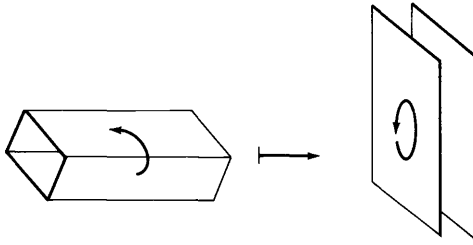
We can define a new star operator,  $(*, \Omega)$ , that maps  $p$ -forms to twisted  $(n-p)$ -forms in  $n$ -space. Despite its appearance,  $(*, \Omega)$  is independent of orientation, because  $*$  itself depends on orientation. A more rational notation would have used  $(*, \Omega)$  for the original star operator. We define  $(*, \Omega)$  so that

$$(*, \Omega)(\alpha, \Omega) = *\alpha$$

and

$$(*, \Omega)\alpha = (*\alpha, \Omega).$$

See Figure 28.18.



**Figure 28.18.** The action of a twisted Hodge star operator, mapping a 2-form into a twisted 1-form.

### Problems

- 28.1** Find a representation for twisted vectors  $(a, \Omega)$  compatible with that given for twisted 1-forms.
- 28.2** Develop a formalism for dealing with nonoriented differential forms, such as would be used to describe the density and alignment of unoriented lines.

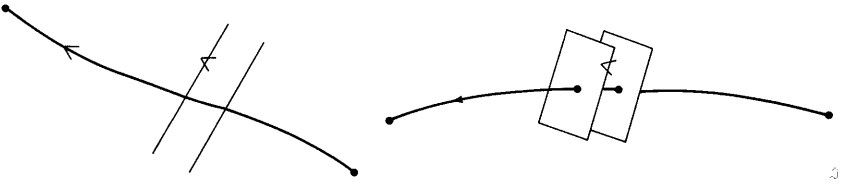
### 29. Integration

By now it will not surprise you to learn that in an integral like  $\int f dx$ , the integrand  $f dx$  is a 1-form. This is especially obvious for a line integral:

$$\int (f dx + g dy + h dz).$$

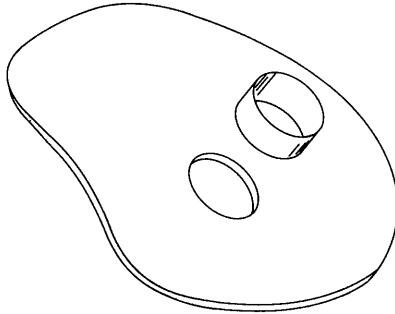
The geometric idea is shown in two dimensions in Figure 29.1, and in three dimensions in Figure 29.2. It is quite straightforward. A line integral through a field of 1-forms counts the number of unit contour lines crossed. When the 1-form is not constant, then what we count is the average number of lines crossed. Not all of the infinitesimal expressions used by physicists are 1-forms. For example, if we are finding the length of a line,  $dl$  is *not* a 1-form.

Similar considerations hold for multiple integrals. We integrate 2-forms over 2-surfaces, and so on. In the expression  $\int f dx dy$ , the integrand  $f dx dy$  is a 2-form. I think of the 2-surface as a piece of dough, and the 2-form as a cookie cutter, as shown in Figure 29.3. The integral tells me how many cookies I can cut out. Since this cookie cutter has an area but no shape, it cannot be reliably used to cut up spaghetti. Thus a 2-form cannot be integrated along a line.



**Figure 29.1** (left). The geometric ingredients for a line integral: a curve and a field of 1-forms.

**Figure 29.2** (right). The pattern for a line integral in three dimensions.



**Figure 29.3.** The pattern for a surface integral.

*Integration of differential forms*

We define these geometric integrals by reducing them to ordinary integrals. Let  $\Gamma: \mathbb{R}^k \rightarrow M$ .  $\Gamma$  defines a  $k$ -dimensional subspace of the manifold  $M$ . Let  $\Sigma$  be the region of  $\mathbb{R}^k$  such that  $\Gamma(\Sigma)$  is the region in  $M$  over which we want to integrate. We want to define

$$I = \int_{\Gamma(\Sigma)} \omega,$$

where  $\omega$  is a  $k$ -form on  $M$ . We will use  $x^1, x^2, \dots$ , for coordinates in  $\mathbb{R}^k$ . We define this integral to be the ordinary multiple integral in  $\mathbb{R}^k$ ,

$$I = \int_{\Sigma} \left[ \frac{\partial}{\partial x^k} \cdots \lrcorner \frac{\partial}{\partial x^1} \lrcorner (\Gamma^* \cdot \omega) \right] dx^1 dx^2 \cdots dx^k,$$

and we will usually write this

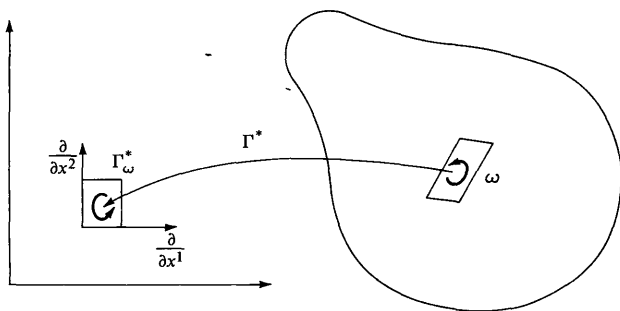


Figure 29.4. The pullback operation in the definition of a surface integral.

$$I = \int_{\Sigma} \Gamma^* \cdot \omega.$$

In the preceding definition, the quantity in square brackets is a number that gives the rate at which, for example, cookies are being cut out per unit area in  $\mathbb{R}^k$  (see Figure 29.4).

**Example:** Let  $\Gamma$  be the map

$$\Gamma: [0, 1] \rightarrow M$$

represented by the  $n$  functions  $\gamma^a(s)$ . The pullbacks of the basis forms are given by

$$\Gamma^* \cdot dx^a = \frac{d\gamma^a}{ds} ds,$$

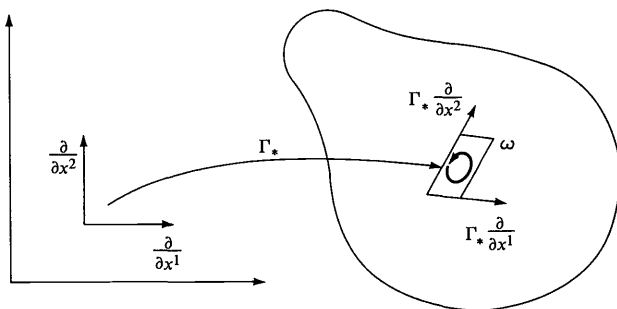
and the pullback of a 1-form  $\omega$  thus gives us

$$\Gamma^* \cdot (\omega_a dx^a) = \omega_a \frac{d\gamma^a}{ds} ds,$$

and our line integral is thus

$$\int \omega = \int_0^1 \omega_a \frac{d\gamma^a}{ds} ds.$$

Sometimes a complementary view of integration is useful. Instead of pulling back the differential form, break the surface in  $\mathbb{R}^k$  into little pieces, that is, into multivectors, and push them forward. Since these multivectors are dual to the differential form, each gives a number. Add



**Figure 29.5.** A dual view of a surface integral. The basis in  $\mathbb{R}^2$  is now being pushed forward.

them up. (See Figure 29.5 and compare with the preceding figure.) The region  $\Sigma$  in  $\mathbb{R}^k$  really plays a minor role in the integral. What counts is the interaction between the differential form and the oriented subspace. Note that the map  $\Gamma$  automatically assigns an inner orientation to the integration surface  $\Gamma(\Sigma)$  given by

$$\Gamma^* \cdot \Omega = dx^1 dx^2 \dots$$

It is easy to verify that these integrals are invariant under any change in parametrization of the submanifold  $\Gamma$  that preserves the orientation.

### *Integration of twisted forms*

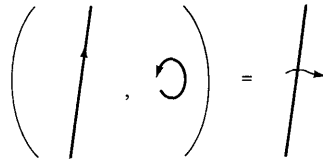
Whereas ordinary differential forms require an inner orientation for the region of integration, twisted forms require a transverse orientation. It is again convenient to define the integration of twisted forms in terms of ordinary forms. Bear in mind, however, that this again introduces a spurious dependence on orientation, which later cancels out. It is only a property of the representation.

We can represent the outer orientations of surfaces in terms of an orientation for the entire space as before, using the rule given in equation 28.1. This is shown for two dimensions in Figure 29.6, and for three dimensions in Figures 29.7 and 29.8.

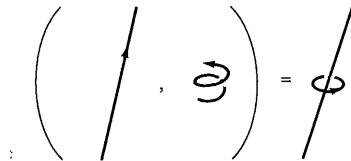
For twisted forms we define

$$\int_{(\Gamma, \Omega)} (\alpha, \Omega) = \int_{\Gamma} \alpha,$$

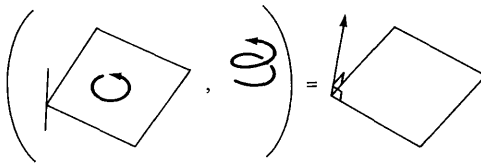
where here  $\Omega$  is an orientation for the whole space, and  $(\Gamma, \Omega)$  is a subspace that has an outer orientation.



**Figure 29.6.** Representing a twisted orientation for a surface, here a line, in two dimensions.



**Figure 29.7.** Representing a twisted orientation for a line in three dimensions.



**Figure 29.8.** A twisted orientation for a 2-surface in three dimensions. The vector on the right is coming up out of the surface.

**Example:** Suppose we have a Moebius strip embedded in  $\mathbb{R}^3$ . We cannot give the embedded Moebius strip either an inner orientation or an outer orientation. A 2-form in  $\mathbb{R}^3$  can be pulled back onto the Moebius strip, but, since the Moebius strip cannot be given a continuous inner orientation, the 2-form cannot be integrated over the Moebius strip. A twisted 2-form cannot even be pulled back onto the Moebius strip, since the Moebius strip also lacks an outer orientation.

We can define a twisted 2-form on the Moebius strip as a manifold itself. Since it is two-dimensional, we can just pick a signed area. Such a twisted 2-form could be integrated over the Moebius strip, to compute, for example, how much paper would be needed to make it.

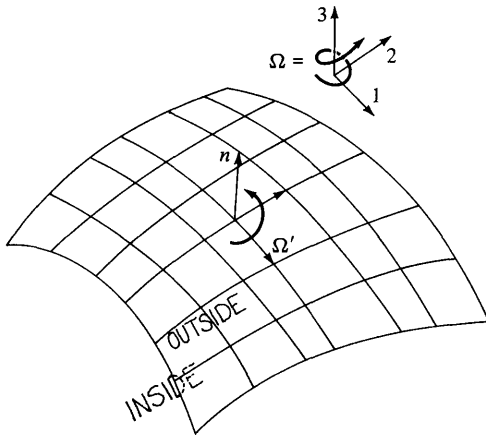


Figure 29.9. The pattern for Stokes' Theorem.

### Stokes' Theorem

In the exterior calculus the great variety of theorems that relate integration and differentiation is greatly simplified. If we write  $\partial\Gamma$  for the boundary of the region  $\Gamma$ , then we have *Stokes' Theorem*:

$$\int_{\Gamma} d\omega = \int_{\partial\Gamma} \omega.$$

The boundary  $\partial\Gamma$  has an inner orientation given by

$$d\phi \wedge \{\partial\Gamma\} = \{\Gamma\}$$

where  $\phi$  is a function increasing to the outside of the surface (see Figure 29.9).

Stokes' Theorem can also be applied to twisted forms. The boundary must now be given an outer orientation. The rule consistent with the preceding equation is

$$\{(\partial\Gamma, \Omega)\} = \{(\Gamma, \Omega)\} \wedge d\phi.$$

Since  $\Omega$  appears on both sides of the equality, this definition does not depend on  $\Omega$ . The orientation  $\Omega$  appears only as part of the representation.

Stokes' Theorem for twisted forms is

$$\int_{(\Gamma, \Omega)} d(\omega, \Omega) = \int_{(\partial\Gamma, \Omega)} (\omega, \Omega).$$



When  $\Gamma$  has the dimension of the manifold, this should really be called the divergence theorem. [For a proof of Stokes' Theorem and further discussion, see Spivak (1965) or Loomis and Sternberg (1968).] The generalization of integration by parts to multidimensional integrals is easy:

$$\int_{\Gamma} d\alpha \wedge \beta = (-)^{p+1} \int_{\Gamma} \alpha \wedge d\beta + \int_{\partial\Gamma} \alpha \wedge \beta,$$

where  $\alpha$  is a  $p$ -form.

To integrate on manifolds we may have to use several charts, and break up the integral into pieces by using a partition of unity. Each piece is then done in its own chart.

#### *The adjoint*

If we are given two  $p$ -form fields on a manifold without boundary that has a Hodge star operation defined on it, then we have a natural inner product on these fields:

$$\langle \alpha, \beta \rangle = \int_M \alpha \wedge * \beta.$$

The operator adjoint to  $d$  will be a differential operator  $\delta$  such that

$$\langle \alpha, d\gamma \rangle = \langle \delta\alpha, \gamma \rangle.$$

Clearly  $\delta$  must map  $p$ -forms to  $(p-1)$ -forms. Since there is no boundary, from Stokes' Theorem we have

$$*\delta = (-)^p d*.$$

The natural generalization of  $\nabla^2$  is the de Rham operator  $\Delta$ :

$$\Delta = d\delta + \delta d.$$

For the special case of functions (0-forms), we have

$$\nabla^2 \phi * 1 = d * d\phi,$$

where  $*1$  must be a properly oriented unit volume element.

#### *Green's Theorem*

Starting from

$$\begin{aligned} d(\phi \wedge *d\psi) &= d\phi \wedge *d\psi + \phi(d*d\psi) \\ &= d\phi \wedge *d\psi + \phi \nabla^2 \psi * 1, \end{aligned}$$

where  $\phi$  and  $\psi$  are scalar fields (0-forms), and using

$$d\phi \wedge *d\psi = d\psi \wedge *d\phi,$$

we have

$$d(\phi *d\psi - \psi *d\phi) = (\phi \nabla^2 \psi - \psi \nabla^2 \phi) *1.$$

In integral form this is the familiar Green's Theorem:

$$\int_V (\phi \nabla^2 \psi - \psi \nabla^2 \phi) *1 = \int_{\partial V} (\phi *d\psi - \psi *d\phi).$$

### *Deformations of integrals*

Look at the line integrals  $\int_{\Gamma_u} \omega$ , where  $\Gamma_u$  is now a 1-parameter family of curves. Suppose that  $\Gamma_0$  is the curve of interest. The  $\Gamma_u$  represents a deformation of that curve. The family  $\Gamma_u$  can be represented by the functions

$$\Gamma_u: \mathbb{R} \times \mathbb{R} \rightarrow M; (s, u) \mapsto X^\alpha(s, u).$$

The rate of change of the integral as we deform the curve will be given by

$$\begin{aligned} \frac{d}{du} \int_{\Gamma_u} \omega &= \frac{d}{du} \int_a^b \omega_\alpha \frac{\partial X^\alpha}{\partial s} ds \\ &= \int_a^b \left[ \omega_{\alpha, \beta} \frac{\partial X^\alpha}{\partial s} \frac{\partial X^\beta}{\partial u} + \omega_\alpha \frac{\partial^2 X^\alpha}{\partial u \partial s} \right] ds. \end{aligned}$$

Now the local deformation of  $\Gamma_0$  is represented by a vector field  $v$ . Of course,  $v$  need be defined only on the curve  $\Gamma_0$ . In components,

$$v = \frac{\partial X^\alpha}{\partial u} \frac{\partial}{\partial x^\alpha};$$

so we have

$$\frac{d}{du} \int_{\Gamma_u} \omega = \int_a^b [\omega_{\alpha, \beta} v^\beta + \omega_\beta v^{\beta, \alpha}] \frac{\partial X^\alpha}{\partial s} ds,$$

and the term in square brackets is just the Lie derivative. In fact, this is true not only for line integrals; we have in general

$$\mathfrak{L}_v \int_{\Gamma_u} \omega = \frac{d}{du} \int_{\Gamma_u} \omega = \int_{\Gamma_0} \mathfrak{L}_v \omega. \quad (29.1)$$

This result will be of use in the calculus of variations. A proof is easy if we use the normal form  $(\partial/\partial x^1)$  for  $v$ .

### 30. Cohomology

We often face the problem of finding potentials. That is, given a closed form  $\omega$ , one satisfying

$$d\omega = 0,$$

we want to find a differential form  $\alpha$  such that

$$\omega = d\alpha.$$

If an  $\alpha$  can be found, then  $\omega$  is called exact. Cohomology is the study of this problem. It turns out that there is a surprising connection between the conditions under which this problem is soluble and the topology of the manifold on which it is posed. The cohomology of a manifold is a topological invariant.

The material of this section and its later applications show the importance of existence proofs, which are often derided by "practical" physicists.

#### *The Poincaré Lemma*

Suppose first that the problem is posed in  $\mathbb{R}^n$  or in some other region that can be contracted to a single point, as sketched in Figure 30.1. On such a region, every closed form is exact. An explicit algorithm for computing  $\alpha$  can be given as follows. [This proof comes from Spivak (1965).] We write the  $r$ -form  $\omega$  in a basis

$$\omega = f_{|I|}(q) dq^I,$$

where the  $I$  indicates a block of indices and an ordered summation, as usual. We define the operator  $J$  by

$$J\omega = \left[ \int_0^1 t^{r-1} f_{|I|}(tq) dt \right] q \frac{\partial}{\partial q} \lrcorner dq^I.$$

In all these integrals,  $t^r$  is  $t$  raised to the  $r$ th power and  $(tq)$  is the argument of  $f_{|I|}$ . The points along the ray  $tq$  all lie in the region if it is contractible to the origin and the coordinates are properly chosen; the region is then called star-shaped with respect to the origin. The integral is therefore well-defined. We now show that for any  $r$ -form  $\omega$

$$\omega = J(d\omega) + d(J\omega).$$

For a closed form this gives us an explicit expression for  $\alpha$ :

$$\alpha = J\omega.$$

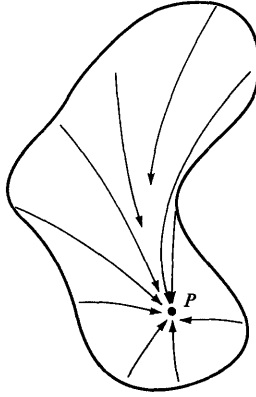


Figure 30.1. A region that can be contracted to a single point, here,  $P$ .

We compute

$$d\omega = \frac{\partial f_I}{\partial q^j} dq^j dq^I,$$

$$J(d\omega) = \left[ \int_0^1 t^r \frac{\partial f_I}{\partial q^j}(tq) dt \right] q \frac{\partial}{\partial q} \lrcorner (dq^j dq^I)$$

$$= \left[ \int_0^1 t^r \frac{\partial f_I}{\partial q^j}(tq) dt \right] \left[ q^j dq^I - dq^j \wedge \left( q \frac{\partial}{\partial q} \lrcorner dq^I \right) \right].$$

But

$$d(J\omega) = \left[ \int_0^1 t^r \frac{\partial f_I}{\partial q^j}(tq) dt \right] dq^j \wedge \left( q \frac{\partial}{\partial q} \lrcorner dq^I \right)$$

$$+ \left[ \int_0^1 t^{r-1} f_I(tq) dt \right] dq \frac{\partial}{\partial q} \lrcorner dq^I.$$

Now we have

$$dq \frac{\partial}{\partial q} \lrcorner dq^I = r dq^I.$$

Each nonzero term in the sum recreates the original term. Thus

$$Jd\omega + dJ\omega = \left[ \int_0^1 r t^{r-1} f_I(tq) dt + \int_0^1 q^j t^r \frac{\partial f_I}{\partial q^j}(tq) dt \right] dq^I,$$

and integrating by parts gives us the desired result. In practice  $\alpha$  can often be found by inspection, a process justified by the preceding explicit calculation.

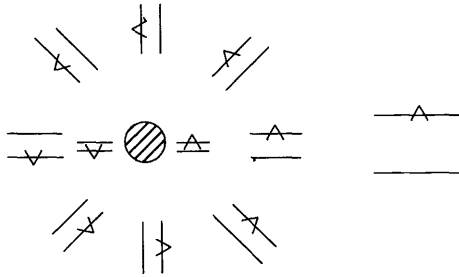


Figure 30.2. The closed form  $d\theta$ , not exact but smooth on  $\mathbb{R}^2$  with a disc removed.

### Cohomology

On a topologically more complicated manifold, such as  $\mathbb{R}^3$  with a ball removed around the origin, it may no longer be true that every closed form is exact. The obstruction to finding a potential is intrinsic to the manifold. This property of the manifold is represented by a series of vector spaces,  $H^r$ . Elements of  $H^r$  are closed  $r$ -forms on the manifold modulo the exact  $r$ -forms on the manifold. That is, two closed forms in  $H^r$  are equivalent if they differ by an exact  $r$ -form.

**Examples:** From Poincaré's Lemma, on  $\mathbb{R}^n$  there are no closed forms that are not exact; so  $H^r$  is zero for all  $r$ .

Look at  $\mathbb{R}^2$  minus a disc at the origin. Clearly  $d\theta$  gives us a closed 1-form that is not exact (see Figure 30.2). Not so easy to see is that every other 1-form differs from  $d\theta$  only by an exact form or some multiplicative factor.  $H^1$  here is one-dimensional.

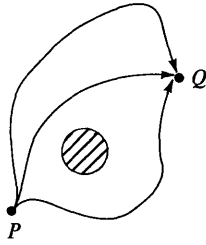
To see that every other closed form, say,  $\alpha$ , is "cohomologous" to  $d\theta$ , first integrate  $\alpha$  around the missing disc. Integrating  $d\theta$  around the disc gives us  $2\pi$ . The 1-form

$$2\pi\alpha - d\theta \int \alpha$$

will be exact. Write

$$2\pi\alpha - d\theta \int \alpha = 2\pi df,$$

where  $f(P)$  is computed from the line integral from some fixed



**Figure 30.3.** Different paths from  $P$  to  $Q$  must define the same differential form.

point  $Q$  to  $P$  along any path (see Figure 30.3). It is clearly path-independent. Thus

$$\alpha = \left[ \frac{1}{2\pi} \int \alpha \right] d\theta + df,$$

and so  $\alpha$  and  $d\theta$  are indeed equivalent.

### Topology

Some topological information about a manifold comes from a study of the boundary operator  $\partial$  acting on submanifolds. There is a surprising duality between the operator  $\partial$  and the exterior derivative  $d$ . Recall Stokes' Theorem and also the relation

$$\partial^2 = 0.$$

We can make a vector space out of oriented submanifolds by formally adding and scaling them, connecting these formal sums with reality by the relation

$$\int_{k_1\Gamma_1 + k_2\Gamma_2} \omega = k_1 \int_{\Gamma_1} \omega + k_2 \int_{\Gamma_2} \omega.$$

Such objects are called *chains*. Important topological information comes from studying the chains that have no boundary. These are called *cycles*. Some cycles are themselves boundaries; others are not. Cycles that are boundaries are trivial, in the sense that the preceding integral on closed forms vanishes for them:

$$\int_{\partial\Gamma} \omega = \int_{\Gamma} d\omega = 0.$$

The homology classes  $H_r$  are defined as the  $r$ -cycles on the manifold modulo those that are boundaries.

**Example:** For  $\mathbb{R}^2$  minus a disc,  $H_1$  is one-dimensional, and 1-cycles are closed curves. Those not encircling the disc are boundaries. Two different 1-cycles, each encircling the disc once, are together the boundary of an annular region. Cycles that encircle the disc  $k$  times are just  $k$  times a simple cycle. Thus the vector space  $H_1$  here is one-dimensional.

The important theorem here, far too hard for us to prove, is that the vector spaces  $H^r(M)$  and  $H_r(M)$  are isomorphic (i.e., have the same dimension).

**Example:**  $\mathbb{R}^2$  with  $k$  discs removed clearly has  $k$  cycles that are not boundaries and hence  $k$  independent closed 1-forms that are not exact.

The result of all this is the important theorem that a closed  $r$ -form is exact if and only if its integral over all  $r$ -cycles vanishes. [See Bott and Tu (1982) or Warner (1971) for proofs and discussion; a good introduction without proofs is in Guillemin and Pollack (1974).]

### Problems

**30.1** Show that every smooth, closed 1-form on the 2-torus can be written

$$\omega = A d\theta + B d\phi + df,$$

where  $A$  and  $B$  are numbers, and  $f$  is a smooth function.

**30.2** What is the relation between the cohomology of the  $n$ -spheres and their related projective spaces?

**30.3** Study the cohomology of  $\mathbb{R}^n$  for the case where the differential forms all have compact support; that is, they vanish outside of a compact region. For example, on the line the 1-form  $dx$  is closed but no longer exact, because the function  $x$  does not have compact support.

# V

## *Applications of the exterior calculus*

In this chapter we treat several short applications of the exterior calculus. The major uses, electrodynamics, mechanics, and gravitation, each have their own chapter. Here we will develop some of the remaining tools of the exterior calculus. One is the calculus of variations, which fits nicely into our language, including problems involving constraints. We will also study partial differential equations, representing them by an ideal of differential forms called an exterior differential system. Viewed in this way the distinction between dependent and independent variables disappears, and we can see clearly the relations between symmetries, conservation laws, and similarity solutions. This approach is among the most powerful known for nonlinear partial differential equations.

### **31. Diffusion equations**

Earlier we saw how naturally Laplace's equation and the wave equation fit into our geometric language. Here we extend that discussion to the heat equation and its generalizations.

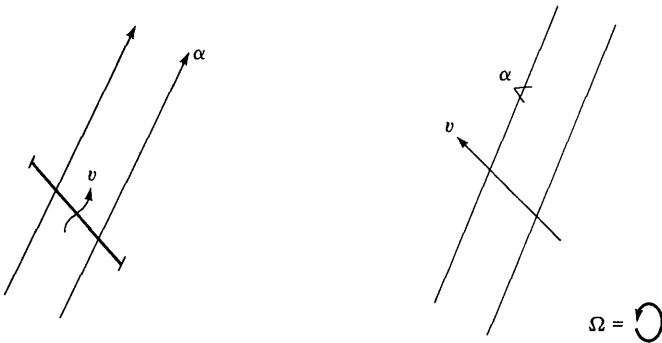
#### *The heat equation*

We start with the heat equation

$$\kappa \frac{\partial^2 \phi}{\partial x^2} = \frac{\partial \phi}{\partial t}.$$

Here  $\phi$  is the temperature, proportional to the density of heat, and  $\kappa$  is the specific thermal diffusivity. We write this as a first-order system,





**Figure 31.1** (left). The heat flux in two dimensions is described by a twisted 1-form  $\alpha$ . A piece of the transversely oriented surface is represented by a twisted vector.

**Figure 31.2** (right). The situation of Figure 31.1 represented using ordinary vectors and forms.

$$u = \kappa \frac{\partial \phi}{\partial x} \quad (31.1)$$

and

$$\frac{\partial u}{\partial x} - \frac{\partial \phi}{\partial t} = 0. \quad (31.2)$$

If we define a 1-form

$$\alpha = \phi dx + u dt,$$

then equation 31.2 becomes

$$d\alpha = 0.$$

The 1-form  $\alpha$  represents the spacetime flux of heat, and the preceding equation represents the conservation of heat (energy). A similar interpretation can be given to  $\alpha$  for other diffusion problems.

**Example:** The heat crossing a short line described by the vector  $v$  is given by  $v \lrcorner \alpha$ . The geometric picture is given in Figure 31.1. Note that both  $v$  and  $\alpha$  are twisted. In Figure 31.2 the less natural picture in terms of untwisted forms is shown, using the representation given in equation 28.1.

In either representation you can see that  $\phi$  is the local density of heat; it describes the flux of heat per unit length across a spacelike surface. Similarly,  $u$  describes the flux of heat past a given location per unit time.

Equation 31.1, on the other hand, is a material equation, one expressing a property of the material being considered. It says that the heat flux  $u$  is proportional to the temperature gradient,  $\phi_{,x}$ . We can geometrize the material equation most naturally by introducing a degenerate metric on  $(t, x)$  space. That this metric is degenerate should not bother you. The heat equation is only an approximate equation. The heat transport described by the heat equation can travel infinitely fast, corresponding to the limit  $c \rightarrow \infty$ . Thus it is not so surprising that a degenerate metric is useful here.

We can find a Hodge star operator even for a degenerate metric. For our applications this must be done in terms of the sharp operator:

$$\#dx = \frac{\partial}{\partial x}, \quad \#dt = 0.$$

This leads to a Hodge star operator

$$\begin{aligned} *1 &= dx dt, & *dt &= 0, \\ *dx &= dt, & *dx dt &= 0. \end{aligned}$$

We have

$$*\alpha = \phi dt \quad \text{and} \quad d*\alpha = \phi_{,x} dx dt.$$

This geometrizes the differential part of equation 31.1. The algebraic part can be written using the vector  $(\partial/\partial t)$  to find

$$u = \frac{\partial}{\partial t} \lrcorner \omega.$$

This is also geometrically natural. The vector  $(\partial/\partial t)$  represents the space-time motion of the material. Call this vector  $\lambda$ .

Thus we arrive at the geometric form of the heat equation

$$d\alpha = 0, \tag{31.3}$$

$$d\kappa * \alpha = * \lambda \lrcorner \alpha. \tag{31.4}$$

Although similar to the systems describing the wave equation and Laplace's equation, it involves one additional term. This is the only way the vector field  $\lambda$  can be used to make a 2-form from  $\alpha$ . For the rest of this section, I will set  $\kappa = 1$ ; I left it in thus far only to show that it appeared in the material equation, but not in the conservation law.

The 1-form field  $\alpha$  still describes a field of conserved flux lines, but the 1-form  $*\alpha$  is not the gradient of any function. These flux lines are not cut by any nice geometric family of orthogonal hypersurfaces.

*Symmetries of the heat equation*

Suppose we generate a new heat flux  $\alpha'$  according to

$$\alpha' = \mathfrak{L}_k \alpha,$$

where  $\alpha$  is any given solution to the heat-equation system. When will  $\alpha'$  also be a solution of the heat equation? Since  $d$  and  $\mathfrak{L}_k$  commute, we clearly have

$$d\alpha' = 0$$

for any vector field  $k$ .

If we pick  $k$  such that

$$[k, \lambda] = 0 \tag{31.5}$$

and

$$[\mathfrak{L}_k, *] = 0, \tag{31.6}$$

then taking the Lie derivative of equation 31.4 gives us

$$d * \mathfrak{L}_k \alpha = * \lambda \lrcorner \mathfrak{L}_k \alpha$$

and hence a new solution.

Vector fields that satisfy equations 31.5 and 31.6 are infinitesimal symmetries of the operators  $*$  and  $\lambda$ . We now proceed to find them. Look at equation 31.6. It is again sufficient to satisfy this equation for 0-forms and 1-forms. Since the metric is now degenerate, we must reexamine the earlier arguments to be sure that this is still true.

Write

$$k = T(t, x) \frac{\partial}{\partial t} + X(t, x) \frac{\partial}{\partial x}.$$

Then

$$\mathfrak{L}_k * 1 = \mathfrak{L}_k dx dt = dX dt + dx dT;$$

so we must have

$$X_{,x} - T_{,t} = 0.$$

Also we must have

$$\mathfrak{L}_k * dt - * \mathfrak{L}_k dt = 0,$$

which gives us

$$T_{,x} = 0.$$

The equation

$$[\mathfrak{L}_k, *] dx = 0$$

gives us nothing new. The isometries of  $*$  are the vector fields

$$k = A(t) \frac{\partial}{\partial t} + [A'(t)x + B(t)] \frac{\partial}{\partial x},$$

$A'$  being the derivative of  $A$ . Of these, those that commute with  $(\partial/\partial t)$  are just the vector fields

$$k = A \frac{\partial}{\partial t} + B \frac{\partial}{\partial x}$$

with  $A$  and  $B$  constant. Only the translations are symmetries of both  $*$  and  $\lambda$ . The Galilean transformations are not symmetries, because the material picks out a special worldline.

We have not necessarily found all the symmetries of the heat equation. The vector field  $k$  need not be separately a symmetry of  $*$  and  $\lambda$ . Also, it may be necessary to change the dependent variables in  $\alpha$  as well as  $x$  and  $t$ .

**Example:** The map

$$\phi(t, x) \mapsto \frac{e^{-x^2/4t}}{\sqrt{4\pi t}} \phi\left(\frac{x}{t}, -\frac{1}{t}\right)$$

takes solutions of the heat equation into other solutions. It is called the Appell transformation. Here  $\phi$  is changed rather than just transported.

The best framework for studying more general symmetries is one in which all the variables, dependent and independent, are coordinates. We will return to this question once we have developed the tools for handling partial differential equations in terms of exterior differential systems.

### *Diffusion in higher dimensions*

The same equations, 31.3 and 31.4, also describe diffusive transport in higher-dimensional spaces. For this we take  $\alpha$  to be a twisted  $(n-1)$ -form. Again  $\alpha$  describes the flux of a scalar quantity, with the same conservation law

$$d\alpha = 0 \tag{31.7}$$

and material law

$$d*\alpha = *(\lambda \lrcorner \alpha). \quad (31.8)$$

Here  $d*\alpha$  is a 2-form. If we are given  $\lambda$ ,  $\alpha$ , and  $*$ , the right-hand side is the only 2-form that we can write. Here we still take

$$\lambda = \frac{\partial}{\partial t}$$

and  $*$  is defined by

$$\#dt = 0, \quad \#dx = \frac{\partial}{\partial x},$$

and so on.

### *Diffusion in deformable materials*

The test of a good notation is the natural manner in which it generalizes. The diffusion may be taking place in a deformable material, such as a fluid. If the diffusing substance is still conserved, then the flux 1-form will remain closed. If heat is being described, then the heat generated by the deformation may have to be included. However, the material law must now use a velocity vector appropriate to the moving material. For nonrelativistic motions it will be sufficient to use a vector

$$\lambda = \frac{\partial}{\partial t} + v \frac{\partial}{\partial q},$$

where

$$v \frac{\partial}{\partial q}$$

is the spatial velocity of the material. Again, equations 31.7 and 31.8 describe the diffusion.

### *Telegraph equation*

It is tempting to continue the generalization, and to write a Lorentz-invariant transport equation by using the special-relativity star operator. Unfortunately, the preceding description fails already for speeds as large as that of sound. There will be delay terms in the transport law, and these will prevent whatever is diffusing from traveling infinitely fast.

For electrical signals on a lossy transmission line, the correct linear theory is straightforward, and the equation describing these signals is

$$\frac{\partial^2 I}{\partial x^2} - LC \frac{\partial^2 I}{\partial t^2} - (RC + GL) \frac{\partial I}{\partial t} - RGI = 0;$$

this is called the *telegraph equation*. Here  $R$  and  $G$  are resistances along and across the lines, respectively. The term involving  $I_{,tt}$  will limit the speed of a current pulse. To geometrize this equation for the case  $G=0$ , we use a metric

$$\#dx = \frac{\partial}{\partial x}, \quad \#dt = -CL \frac{\partial}{\partial t},$$

$$*1 = dt \, dx,$$

to find

$$\alpha = dI, \quad d\alpha = 0,$$

$$d*\alpha = RC*(\lambda \lrcorner \alpha),$$

and the effects of inductance do indeed fit neatly into the modification of the star operator. In Section 45 we will see how Maxwell's equations in a medium with linear conductivity also fall into this geometric form.

### Problems

- 31.1** Give a graphical justification for the appearance of a degenerate metric in this section.  
**31.2** Study the case of variable conductivity.  
**31.3** Are there any conformal symmetries of the heat equation?

### 32. First-order partial differential equations

The geometric structure generated by a single first-order partial differential equation (pde) is best represented by an *ideal* of differential forms. We are extending here the discussion of linear pdes given in Section 14.

In an algebra, an *ideal* is a subset  $I$  such that for any element  $a$  in the entire algebra, the product of  $a$  with any element of the ideal lies in the ideal. Multiples and sums of elements in the ideal must also be in the ideal. For differential forms, the proper algebra is the exterior algebra, and a set of forms constitutes an ideal if it is closed under exterior multiplication by any other differential form. The ideal generated by a form  $\alpha$  consists of all those forms that vanish when pulled back onto a surface for which  $\alpha$  itself pulls back to zero. An ideal is usually specified by giving a finite number of elements from which all others can be generated by multiplication and addition. A *differential ideal* is an ideal of the exterior algebra such that the exterior derivative of any element of the ideal is also in the ideal.

**Example:** A given contact-element field determines an ideal as follows. Put into the ideal all the 1-forms  $\omega$  such that

$$\omega \cdot v = 0$$

for all  $v$  in the contact element. Use these 1-forms to generate the ideal.

In Section 19 I mentioned the Frobenius Theorem, which gave the conditions under which a contact-element field has integral submanifolds. There is a dual version of the theorem; it states that an ideal of differential forms defines an integrable contact-element field if and only if the ideal is a differential ideal. The proof is an easy computation if we use equation 23.5 and its generalizations.

**Example:** The contact-element field given as an example in Section 19, and there generated by

$$\frac{\partial}{\partial y} \quad \text{and} \quad \frac{\partial}{\partial x} + y \frac{\partial}{\partial z},$$

has an ideal generated by the 1-form

$$\omega = y dx - dz.$$

It would be a differential ideal if the 2-form  $d\omega$  were also in the ideal:

$$d\omega = dy dx.$$

The 2-forms in the ideal are in the linear subspace spanned by

$$dx dz \quad \text{and} \quad y dx dy + dy dz,$$

plus polynomials times these. There is no way we can split off the  $dx dy$  terms. This is not a differential ideal, and again we see that the contact-element field is not integrable.

Consider now a partial differential equation for a single dependent variable  $u$ , with  $n$  independent variables  $x$ . Let  $p$  be the  $n$  partial derivatives

$$p = \frac{\partial u}{\partial x}.$$

In the contact space  $(x, u, p)$ , the pde reduces to the hypersurface given by some function

$$\phi(x, u, p) = 0.$$

A solution to the pde is a subsurface lying in this hypersurface on which the contact 1-form

$$\alpha = du - p dx$$

pulls back to zero.

If we form a differential ideal  $I$  from the 1-forms  $\alpha$  and  $d\phi$  and the 2-form  $d\alpha$ , then the graph of the solution in this contact space will be an integral manifold of the ideal  $I$ . An *integral manifold* is the natural generalization of an integral curve: all elements of  $I$  pull back to zero on an integral manifold.

The solutions to a single first-order pde can be found by what is called the method of characteristics. [The classical treatment of this is in Courant and Hilbert (1962).] We will construct the solution by finding a vector field  $v$  that lies in the graph of the solution without first having to find the solution. The integral curves of this vector field, called the characteristics of the pde, also lie in the graph. If we have boundary values on some  $(n-1)$ -dimensional surface, then these characteristics sweep out the  $n$ -dimensional graph containing that boundary information. We thus reduce our problem to one of, first, finding consistent boundary values, and, second, finding integral curves of a vector field. This second problem involves the solution of a system of ordinary differential equations with initial-value information. These are much easier to solve, especially numerically.

To say that the forms  $d\phi$ ,  $\alpha$ , and  $d\alpha$  all pull back to zero is equivalent to the requirements that

$$d\phi \cdot v = 0, \quad \alpha \cdot v = 0,$$

$$d\alpha \cdot (v, w) = 0,$$

for all vectors  $v$  and  $w$  in the graph. A characteristic vector  $v$  thus must satisfy three conditions. First, it must be tangent to the graph of  $\phi$ ,

$$v \lrcorner d\phi = 0.$$

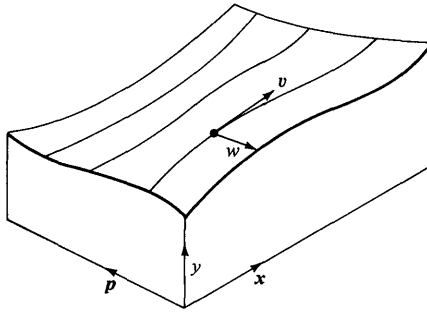
Second, to make the  $ps$  truly partial derivatives, we must have

$$v \lrcorner \alpha = 0.$$

Third, to form a surface it must be a differential ideal

$$w \lrcorner v \lrcorner d\alpha = 0,$$





**Figure 32.1.** The pattern for the differential ideal defining a solution to a partial differential equation.

for all vectors  $w$  that are also in the surface (see Figure 32.1). The vector  $w$  must satisfy

$$w \lrcorner \alpha = 0, \quad w \lrcorner d\phi = 0.$$

The third condition on  $v$  can easily be written by using two Lagrange multipliers,  $\lambda$  and  $\mu$ , in the form

$$v \lrcorner d\alpha = \lambda\alpha + \mu d\phi.$$

We have here  $2n + 3$  homogeneous equations for  $2n + 3$  unknowns. From these we can find  $v$  up to a scale.

Look at a quasilinear pde, that is, one linear in the derivatives:

$$au_{,x} + bu_{,y} = c,$$

where  $a$ ,  $b$ , and  $c$  may all be functions of  $x$ ,  $y$ , and  $u$ . The coordinates  $x$  are here  $x$  and  $y$ , the  $p$  are  $p$  and  $q$ . We have

$$\alpha = du - p dx - q dy,$$

$$d\alpha = dx dp + dy dq,$$

$$\phi = ap + bq - c,$$

$$d\phi = a dp + b dq + \phi_{,x} dx + \phi_{,y} dy + \phi_{,u} du.$$

A characteristic vector

$$v = U \frac{\partial}{\partial u} + X \frac{\partial}{\partial x} + \dots$$

must satisfy

$$u = pX + qY, \tag{32.1}$$

$$aP + bQ + \phi_{,x}X + \phi_{,y}Y + \phi_{,u}U = 0, \quad (32.2)$$

and finally

$$Xdp - Pdx + Ydq - Qdy = \lambda\alpha + \mu d\phi.$$

We can find the Lagrange multipliers by inspection. Pick the amount of  $d\phi$  to cancel the  $dp$  component, and the  $\alpha$  to cancel the  $du$ , leaving three essential equations:

$$bX = aY,$$

$$aP + X(\phi_{,x} + p\phi_{,u}) = 0,$$

$$bQ + Y(\phi_{,y} + q\phi_{,u}) = 0.$$

These equations are consistent with equations 32.1 and 32.2, and the five together have a solution proportional to

$$v = a \frac{\partial}{\partial x} + b \frac{\partial}{\partial y} + c \frac{\partial}{\partial u} - (\phi_{,x} + p\phi_{,u}) \frac{\partial}{\partial p} - (\phi_{,y} + q\phi_{,u}) \frac{\partial}{\partial q}. \quad (32.3)$$

Because of the quasilinearity it is possible to project this vector field down onto the base space  $(x, y, u)$  and ignore the  $(p, q)$  coordinates.

**Example:** Look at

$$uu_{,x} + u_{,y} = 1.$$

The projection of  $v$ , equation 32.3, onto the  $(x, y, u)$  3-space is

$$u \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial u},$$

with integral curves given by

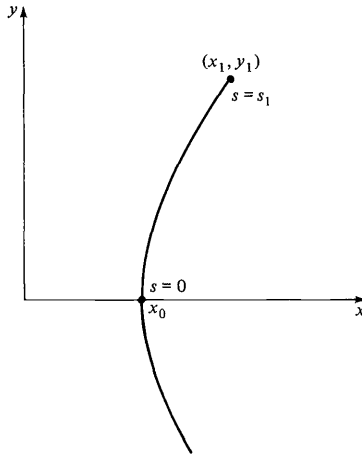
$$X = u, \quad Y = 1, \quad U = 1.$$

The solutions are

$$x = x_0 + u_0 s + \frac{1}{2} s^2,$$

$$y = y_0 + s, \quad u = u_0 + s.$$

These form a three-parameter family of parabolas, with parameters  $x_0$ ,  $y_0$ , and  $u_0$ . To find  $u$  at a given point  $(x_1, y_1)$ , we must find the parabola that passes through the given point, and trace that characteristic curve back to an intersection with the surface on which boundary-value data is specified (see Figure 32.2).



**Figure 32.2.** The characteristic curve, here a parabola, is extended until it crosses a surface on which boundary data is specified.

Suppose we knew  $u_0(x_0)$  on the surface  $y_0 = 0$ . We must have

$$x_0 + u_0 s_1 + \frac{1}{2} s_1^2 = x_1, \quad (32.4)$$

and

$$y_0 + s_1 = y_1.$$

Here  $s_1$  is a fourth unknown, the parameter value at the point  $(x_1, y_1)$ . The initial-value surface is given by  $y = 0$  and  $s = 0$ . Thus we must have

$$y_0 = 0, \quad \text{and} \quad s_1 = y_1.$$

From equation 32.4 we have

$$x_0 + u_0 y_1 + \frac{1}{2} y_1^2 = x_1,$$

and finally the boundary data give us a fourth equation  $u_0(x_0)$ .

Having found the four unknowns  $(x_0, y_0, u_0, s_1)$ , we find  $u$  from

$$u_1 = u_0 + s_1,$$

completing the solution.

We will extend the ideas of this section to systems of partial differential equations in Section 38.

**Problem**

**32.1** Solve the equation

$$uu_{,x} + u_{,y} = 1$$

for  $y \geq 0$  with initial data on  $y = 0$  of

$$u(x, 0) = -x.$$

How far into the  $y$  direction is your solution valid?

**33. Conservation laws**

An enduring feature of physical models of the world has been the idea that “you can’t get something for nothing.” Many things have a continuous existence and are not created spontaneously. Of course, there have been many wrong guesses; neither heat nor mass is truly conserved, although each is almost conserved in restricted circumstances. In addition to objects that are strictly conserved, there are those that appear and disappear in an orderly manner, such as electric flux lines, which appear and disappear on charges. These situations are described by a conservation law or a balance law, and these laws take a nice geometric form when we use the exterior calculus.

*Conserved particles*

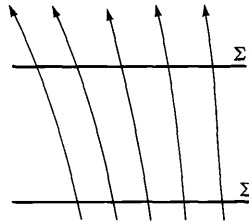
If we think of small conserved objects, which we will call particles, then conservation implies that their worldlines are continuous, and neither start nor end. The conservation laws are the “fluid” approximation to situations in which there are an enormous number of such particles in any volume of interest, all with the same velocity. The density in spacetime of a family of nonintersecting worldlines is given by a twisted 3-form, call it  $\alpha$ , pointing into the future. At any given time the number of particles in a spatial volume  $\Sigma$  is given by

$$N = \int_{\Sigma} \alpha,$$

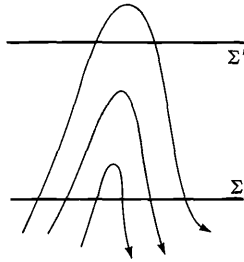
and for a situation in which the region of integration is so large that no worldlines cross at the edges, we have a conservation law

$$\int_{\Sigma} \alpha = \int_{\Sigma'} \alpha.$$

The spacetime configuration is shown in Figure 33.1. The twisted 3-form



**Figure 33.1.** A spacetime conservation law for particles having continuous existence.



**Figure 33.2.** Spacetime conservation for particles that can also move backward in time.

$\alpha$  can also represent worldlines moving backward in time; so the situation in Figure 33.2 also has a conservation law.

For small regions, some particles can flow out of the region of interest, and we must include the flux through the walls. If we call the side region  $\Sigma''$ , we have

$$\int_{\Sigma'} \alpha = \int_{\Sigma} \alpha + \int_{\Sigma''} \alpha,$$

and if we orient  $\Sigma'$  using the outward normal, we can write this

$$\int_{\partial\Gamma} \alpha = 0,$$

where we now have an integral over the boundary of some 4-volume  $\Gamma$ . Using the divergence theorem, we have

$$\int_{\Gamma} d\alpha = 0,$$

and if this is to be true for every region  $\Gamma$ , then we have the differential form of a conservation law

$$d\alpha = 0. \quad (33.1)$$

**Examples:** Suppose that the local density of particles is  $\rho$ , and their velocity is  $U$  in the  $x$  direction. Let us find  $\alpha$ . The most general possible 3-form is

$$\alpha = A dx dy dz + B dy dz dt + C dz dt dx + D dt dx dy.$$

The integration over a unit spatial volume  $\Sigma_1$ , must give  $\rho$ . This volume  $\Sigma_1$  is represented by the map

$$\Sigma_1: [0, 1] \times [0, 1] \times [0, 1] \rightarrow \mathbb{R}^4; (u, v, w) \mapsto (x, y, z, t) = (u, v, w, 0).$$

Since the pullback

$$\Sigma_1^* \cdot dt = 0,$$

this tells us that we must have  $A = \rho$ . No particles will cross a surface  $\Sigma_2$  given by

$$\Sigma_2: (u, v, w) \mapsto (u, 0, v, w)$$

or the surface  $\Sigma_3$  given by

$$\Sigma_3: (u, v, w) \mapsto (u, v, 0, w).$$

The first has

$$\Sigma_2^* \cdot dy = 0,$$

and tells us that  $C = 0$ ; and the second tells us that  $D = 0$ .

Finally, no particles cross the surface  $\Sigma_4$ , shown in Figure 33.3.

$$\Sigma_4: (u, v, w) \mapsto (u, v, w, Uu).$$

For  $\Sigma_4$  we have

$$\Sigma_4^* \cdot dx = du, \quad \Sigma_4^* \cdot dt = U du;$$

hence

$$\Sigma_4^* \cdot \alpha = (A + BU) du dv dw,$$

and so we must have

$$\alpha = \rho(dx - U dt) dy dz,$$

a useful result.

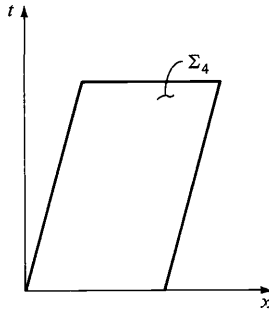


Figure 33.3.

In spacetime for general motion we have

$$\alpha = \rho[dx dy dz - U_x dy dz dt + U_y dz dt dx - U_z dt dx dy];$$

that is,

$$\alpha = \rho(\theta_t + U\theta)$$

where

$$\theta = dt dx dy dz$$

and missing indices range over  $x, y,$  and  $z$ . Our conservation law gives

$$\begin{aligned} \frac{\partial \rho}{\partial t} dt dx dy dz - \frac{\partial(\rho U_x)}{\partial x} dx dy dz dt \\ + \frac{\partial(\rho U_y)}{\partial y} dy dz dt dx - \frac{\partial(\rho U_z)}{\partial z} dt dx dy dz = 0, \end{aligned}$$

that is,

$$\frac{\partial \rho}{\partial t} \theta + d(\rho U)\theta = 0,$$

which is the analog of

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho U) = 0.$$

This shows how the usual vector notation can be incorporated into the exterior calculus. The differential form equation shows that the conservation is independent of the metric.

In addition to the fluid picture of conserved particles, there is also the kinetic-theory picture, where, at each spacetime point, particles are moving in all directions. The appropriate representation uses the line-element contact bundle  $C\mathbb{R}^4$ , which is seven-dimensional. The density of particles is represented by a twisted 6-form  $j$  that satisfies

$$j \wedge \alpha = 0,$$

where the  $\alpha$  are the contact forms

$$\alpha = dq - \dot{q} dt.$$

If the particles travel in straight lines, then we must have also

$$j \wedge d\dot{q} = 0.$$

Finally, the particles are conserved, and an argument similar to that given for equation 33.1 requires

$$dj = 0.$$

These are a total of seven conditions on the seven components of  $j$ .

If we write  $j$  as

$$j = \rho(dx dy dz d\dot{x} d\dot{y} d\dot{z} + \dot{x} dy dz d\dot{x} d\dot{y} d\dot{z} dt \\ - \dot{y} dx dz d\dot{x} d\dot{y} d\dot{z} dt + \dot{z} dx dy d\dot{x} d\dot{y} d\dot{z} dt),$$

then the six algebraic constraints are automatically satisfied. The remaining variable,  $\rho$ , must also satisfy

$$\frac{\partial \rho}{\partial t} + \dot{x} \frac{\partial \rho}{\partial x} + \dot{y} \frac{\partial \rho}{\partial y} + \dot{z} \frac{\partial \rho}{\partial z} = 0.$$

It is quite satisfying to see the convective derivative appear naturally in the context of differential forms.

The flow of cars on a single-lane road is a nice example of the utility of a conservation law. The local density of cars is represented by a 1-form  $j$ ,

$$j = \rho(dx - u dt),$$

where  $\rho$  is the spatial density and  $u$  the velocity. Conservation of cars requires that

$$dj = 0,$$



that is,

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u)}{\partial x} = 0.$$

To complete the theory, we need something like an equation of state. In the simplest models, this is an algebraic relation between the density  $\rho$  and the flux  $\rho u$ . This turns the conservation law into a first-order nonlinear partial differential equation for  $\rho$ , solvable by the method of characteristics of Section 32.

The geometric structure of conservation laws studied here is independent of the dimension of either the space or the subspaces. If at any instant we have lines, not particles in spacetime, in spacetime these lines will sweep out 2-surfaces. The density will be described by a 2-form, not a 3-form. If these lines exist for all time, then we have a conservation law

$$\int_{\partial\Gamma} \alpha = 0$$

where now  $\Gamma$  is a 3-volume and  $\partial\Gamma$  is a closed 2-surface.

We will see in Chapter VI that the electrical-charge density can be represented by a twisted 3-form  $\rho$ . The total charge is the integral of this 3-form over a volume. The electrostatic field will be represented by a twisted 2-form  $D$ , and half of Maxwell's equations can be written

$$dD = 4\pi\rho.$$

Charge conservation is an immediate consequence of this:

$$d\rho = 0.$$

In Section 44 we will find conservation laws associated with the electromagnetic field. The conserved currents there can spread to infinity, and the surface term may diverge to prevent us from turning the differential conservation law into a global integral conservation law.

Note that the conservation of energy and momentum does not fit into this geometric formalism. There is no way to write a conservation law for a symmetric 2-tensor, and in general relativity there is no global law of energy conservation. There do exist special conservation laws for isolated systems. These exploit asymptotic symmetries of the gravitational field, depend on very specific details of general relativity, and do not involve locally conserved fluxes.

### Problems

- 33.1** An interesting equation of state for traffic flow requires drivers to maintain a spacing between cars that is proportional to velocity and to obey a speed limit. Explore this equation of state, and in particular the effect of a finite length of the cars.
- 33.2** Suppose the plane is filled with round holes of radius  $a$ , randomly distributed with density  $\mu \ll 1/a^2$ . Random lines are drawn until the ends hit a hole. Describe the balance of lines in  $CR^2$  in the fluid approximation, where the lines and holes are described by an average density.
- 33.3** Repeat the preceding problem, but with lines now drawn between randomly chosen holes, one line ending at each hole. Again describe the balance of lines and compare with the previous problem.

### 34. Calculus of variations

Minimal principles are much beloved by physicists as a slick way to derive results already known to be true. Mathematicians tend to shy away from them, knowing too well the pitfalls and problems associated with them. We will give here a geometric picture of the calculus of variations, not pretending to any rigor that we do not have, and not particularly looking for trouble. Do realize that you must keep your common sense turned on here.

Our geometric approach to the calculus of variations will allow us to go beyond the standard treatment to be found in most mathematical physics books. In particular, we will be able to give a proof of the multiplier rule. We will pay particular attention to what spaces we are in. The usual discussions, say, in classical mechanics, jump between several different spaces with no fanfare or change in notation. This will occasionally make things a bit harder for us to prove. Easy proofs sometimes reflect the fact that there are many wrong ways to get the right answer. In this section the promise implicit in the dedication is kept.

#### *Hamilton's principle*

One formulation of the classical mechanics of unconstrained conservative systems asserts that there exists a Lagrangian function  $L$  of the state variables (such as position and velocity) and possibly time, and that the dynamical behavior of the mechanical system is such that the integral

$$\int L dt$$

is extremized. This integral is called the *action*. We will begin with this special problem to gain familiarity with our geometric approach. (I assume that you have already seen the standard approach.)

A line integral makes geometric sense only if its integrand is a 1-form. Is  $L dt$  a 1-form? Well, this is the wrong question. The correct question is: On what space is  $L dt$  a 1-form? It is *not* a 1-form on configuration space, the space of positions, because it can have a nonlinear dependence on the velocities. A 1-form must be a linear operator on the tangent vectors. The correct space for  $L dt$  is the line-element contact bundle of configuration space. This space has natural coordinates  $(t, q, \dot{q})$ , and on this space  $L dt$  is an honest 1-form. This space is a contact bundle, and only those curves on which the contact 1-forms

$$\alpha = dq - \dot{q} dt$$

pull back to zero represent possible motions of the system. This constrains the variations of the curves used to compute the action integral, and the constraints are anholonomic. That is, they allow you a different number of degrees of freedom of finite motions as opposed to infinitesimal motions, as in the car-parking example of Section 9.

We now vary the integral by pushing the curve around a little bit. We can describe this by a vector  $v$  defined along the original, undeformed curve

$$v = \underline{Q}(t) \frac{\partial}{\partial q} + \underline{\dot{Q}}(t) \frac{\partial}{\partial \dot{q}}. \quad (34.1)$$

For simplicity we consider variations that keep time fixed; this is usually written into Hamilton's principle. Only the curve through the space  $(t, q, \dot{q})$  makes physical sense, not its parametrization.

The variation of the action is given by

$$\int_{\Gamma} \mathfrak{L}_v(L dt),$$

(recall equation 29.1), and because we want variations with fixed end conditions, we can drop exact 1-forms to find

$$\begin{aligned} \int_{\Gamma} \mathfrak{L}_v(L dt) &= \int_{\Gamma} v \lrcorner (dL dt) \\ &= \int_{\Gamma} \left[ \underline{Q} \frac{\partial L}{\partial q} dt + \underline{\dot{Q}} \frac{\partial L}{\partial \dot{q}} dt \right]. \end{aligned}$$

Now, to limit the variations to those satisfying the constraint condition,

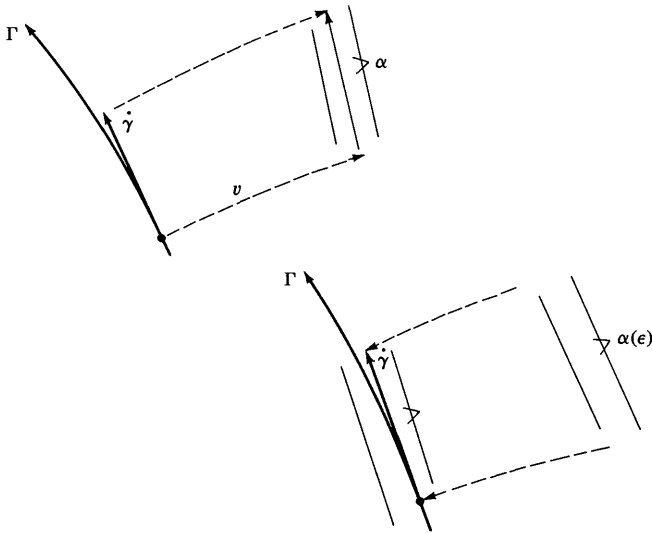


Figure 34.1. The deformation of a path constrained by the 1-form  $\alpha$ .

the vector field  $v$  must move the path into another path that is parallel to the  $\alpha$ , as shown in Figure 34.1. An easier view is to use the integral curves of  $v$  to pull the 1-forms  $\alpha$  back instead of pushing the path forward. The  $\alpha$  pulled back from a parameter change of  $\epsilon$  are

$$\alpha(\epsilon) = \alpha + \epsilon \mathcal{L}_v \alpha.$$

The condition

$$\dot{\gamma} \lrcorner \alpha(\epsilon) = 0$$

gives us

$$\dot{\gamma} \lrcorner \mathcal{L}_v \alpha = 0$$

as the condition that our variations must satisfy. If we take the  $v$  given in equation 34.1 and extend it smoothly to the neighborhood of the curve, we can write this

$$\dot{\gamma} \lrcorner (dQ - \dot{Q} dt) = 0. \tag{34.2}$$

The final result will be independent of this extension. We are only using it as a trick to avoid introducing some new notation.

We can now use this to simplify our integral. Recall that integration is equivalent to contracting the integrand with  $\dot{\gamma}$ . Thus we can use equation 34.2 under the integral sign to turn  $\dot{Q} dt$  into  $dQ$  to find

$$\mathfrak{L}_v \int_{\Gamma} L dt = \int_{\Gamma} \left[ Q \frac{\partial L}{\partial q} dt + \frac{\partial L}{\partial \dot{q}} dQ \right].$$

Now integrate the second term by parts and discard the perfect differential, to find

$$\mathfrak{L}_v \int_{\Gamma} L dt = \int_{\Gamma} Q \left[ \frac{\partial L}{\partial q} dt - d \left( \frac{\partial L}{\partial \dot{q}} \right) \right].$$

The functions  $Q(t)$  can be chosen freely, and so the integrand must vanish at each point. If  $\dot{\gamma}$  is the tangent to the path, then we must have

$$\dot{\gamma} \lrcorner \left[ \frac{\partial L}{\partial q} dt - d \left( \frac{\partial L}{\partial \dot{q}} \right) \right] = 0,$$

which together with the constraints

$$\dot{\gamma} \lrcorner (dq - \dot{q} dt) = 0$$

give us  $2n$  constraints on the  $2n$  components of the line element. These are the usual Euler-Lagrange equations. Note that only values of  $v$  along the curve appear.

Now look at what we have done. We have found conditions on the extremals of the line integral  $\int L dt$  subject to constraints  $d\dot{q} - \dot{q} dt$ . But we have not found the solution to the general problem of this type. The problem here is special. First, the integrand was not a general 1-form:  $L dt$  is special, since its exterior derivative  $dL dt$  is in the differential ideal generated by  $L dt$ .  $L dt$  satisfies the condition

$$L dt \wedge d(L dt) = 0.$$

Second, although the constraint 1-form is a general 1-form, to be sure, it is special because it is in canonical form in the *same* coordinates in which  $L dt$  is in canonical form. That is,  $d\alpha$  is also in the ideal generated by  $L dt$ . It is these two special properties that allow the Euler-Lagrange equations to be derived by the simple trick of integrating by parts. In general this will not work.

**Example:** In  $\mathbb{R}^3$ ,  $(t, q, \dot{q})$ , we cannot find the extrema of

$$(L dt + R dq)$$

subject to the constraint

$$\dot{\gamma} \lrcorner (dq - \dot{q} dt) = 0,$$

by integrating by parts. The term

$$\frac{\partial R}{\partial \dot{q}} \dot{Q} dq$$

is not reducible.

Fortunately, the general case is easily handled by the method of Lagrange multipliers. Few books bother to prove this, however. We will do so in the next section. Although I try to skip proofs that are easily accessible or uninteresting, this one is neither.

### *Unconstrained integrals*

For fun, let us look first at the case of an integral  $\int \omega$  subject to no constraints. Although this might seem to be a reasonable problem, it turns out to be surprisingly pathological.

Take a manifold  $M$  with a sufficiently smooth 1-form  $\omega$  defined on it. For given points  $A$  and  $B$  in  $M$ , we seek a curve  $\Gamma$  from  $A$  to  $B$  such that the integral

$$I = \int_{\Gamma} \omega,$$

does not vary in linear order as  $\Gamma$  is deformed. Let  $s$  be the parameter along the curve. Clearly  $I$  does not depend on the parametrization, and we will use the same range of parametrization for all the deformed curves as well. We will call the tangent vector to the curve  $\dot{\gamma}$ ,

$$\dot{\gamma} = \Gamma_* \cdot \left( \frac{\partial}{\partial s} \right),$$

the pushforward of the basis vector  $\partial/\partial s$ . Let  $v$  be a vector field defined along  $\Gamma$ , representing a possible deformation. From the results of Section 29 we must have

$$\int_{\Gamma} \mathcal{L}_v \omega = 0,$$

and we can write this

$$\int_{\Gamma} v \lrcorner d\omega + \int_{\partial\Gamma} v \lrcorner \omega = 0.$$

Since we demand that all deformed paths still go from  $A$  to  $B$ ,  $v$  must vanish at the endpoints, and so the boundary term vanishes. We have, from the definition of a line integral,

$$\int_{\Gamma} v \lrcorner d\omega = \int_a^b \Gamma^* \cdot (\dot{\gamma} \lrcorner v \lrcorner d\omega) ds = 0,$$

for all  $v$ . Since the vector  $v$  at different points can take arbitrary values, we must have, for all  $v$ ,

$$\dot{\gamma} \lrcorner v \lrcorner d\omega = 0,$$

and so we must have

$$\dot{\gamma} \lrcorner d\omega = 0. \tag{34.3}$$

This is an ordinary differential equation for  $\Gamma$ .

**Example:** In  $M = \mathbb{R}^3$  we consider the 1-form

$$\omega = (y^2 + z^2) dx$$

and seek a path  $\Gamma$  which extremizes  $\int \omega$ . The path must satisfy equation 34.3, and we calculate

$$d\omega = 2y dy dx + 2z dz dx,$$

$$\dot{\gamma} = \dot{x} \frac{\partial}{\partial x} + \dot{y} \frac{\partial}{\partial y} + \dot{z} \frac{\partial}{\partial z},$$

so that

$$\dot{\gamma} \lrcorner d\omega = 2y\dot{y} dx - 2y\dot{x} dy + 2z\dot{z} dx - 2z\dot{x} dz.$$

Setting all three components of this equal to zero, we find equations

$$y\dot{y} + z\dot{z} = 0, \quad 2y\dot{x} = 0, \quad 2z\dot{x} = 0. \tag{34.4}$$

In particular, the curve  $s \mapsto (s, 0, 0)$  is a trivial extremal. Others will be discussed below.

### *The geometry of the solutions*

The geometry of equation 34.3 was covered in Sections 3 and 22 when we discussed 2-forms. Odd- and even-dimension spaces behave quite differently. If  $d\omega$  is of maximal rank, then in an even-dimension space we have no solutions. The kernel of a linear operator is the subspace satisfying equation 34.3. Maximal rank means the kernel is as small as possible.

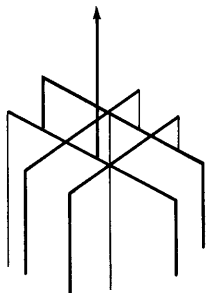


Figure 34.2.

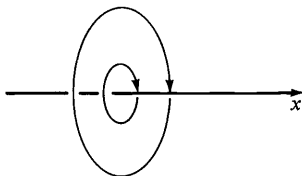


Figure 34.3.

Although at first puzzling, the condition  $d\omega \neq 0$  describes a situation with a nonzero curl, and the more we circulate around and around, the more (or less) our line integral picks up. On the other hand, in an odd-dimension space there is a single allowed direction at every point (see Figure 34.2).

**Example:** For the preceding example we have a  $d\omega$  of maximal rank. The kernel will be one-dimensional, given by equation 34.4. For  $yz \neq 0$  we have this kernel specified by the vector

$$\dot{\gamma} = z \frac{\partial}{\partial y} - y \frac{\partial}{\partial z},$$

which leads to circular curves around the  $x$ -axis. If both  $y$  and  $z$  vanish, then so does  $d\omega$ , and a singular solution along the  $x$ -axis appears (see Figure 34.3).

The usual two-point problem considered in the calculus of variations has no general solutions here. Spaces of even dimension are pathological.



In spaces of odd dimension, the initial-value problem is soluble, provided that just one end condition is imposed. Fortunately, these peculiarities disappear when constraints enter, as they will in the contact bundles of mechanics. [See Pars (1962), Section 3.2, for more on the peculiarities of the calculus of variations of line integrals.]

### Classical notation

To practice with our notation, let us solve this problem in the style of mathematical physics courses. We have an integral

$$I = \int \omega_\mu dx^\mu = \int \omega_\mu \dot{x}^\mu ds$$

and the Euler-Lagrange equations for this are

$$\omega_{\mu, \nu} \dot{x}^\mu = \frac{d}{ds} \omega_\nu,$$

where remember that  $\omega_\mu$  depends on the  $x$ . This equation becomes

$$\omega_{\mu, \nu} \dot{x}^\mu = \omega_{\nu, \mu} \dot{x}^\mu,$$

or

$$(\omega_{\mu, \nu} - \omega_{\nu, \mu}) \dot{x}^\mu = 0,$$

which is just equation 34.3, written out in components.

### Parameter-dependent integrals

Sometimes we need to deal with line integrals in which the parameter enters explicitly, say, in the form

$$\int [f(x, y, \dots, s)x + \dots] ds.$$

To deal with this case geometrically, we extend the manifold  $M$ , adding one more dimension for the parameter  $s$ . We need to do this because  $f dx + \dots$  is not a 1-form on  $M$ . It depends explicitly on the path. Call this larger space  $N$ . It has a natural projection on the original space  $M$ . If we use coordinates  $(q)$  for the space  $M$ , then we can use coordinates  $(s, q)$  for  $N$ . The natural projection here is  $\pi: M \rightarrow N; (s, q) \mapsto (q)$ . This map  $\pi$  lets you lift 1-forms from  $M$  to  $N$ . Lifting our 1-form up to  $N$

$$\hat{\omega} = \pi^* \omega,$$

we can consider the integral in  $N$ ,  $\int \hat{\omega}$ , and this integral has a value that

agrees with our original integral. On  $N$ , now,  $\hat{\omega}$  is a true 1-form. There  $s$  is just another coordinate, and we have a canonical problem on the space  $N$ .

### *Symmetries*

Suppose that we have an infinitesimal symmetry of our problem, a vector field  $k$  such that

$$\mathfrak{L}_k \omega = 0.$$

Then clearly

$$\int_{\Gamma} \mathfrak{L}_k \omega = 0.$$

Along any piece of a curve  $\Gamma$  that satisfies the Euler–Lagrange equations,

$$\dot{\gamma} \lrcorner d\omega = 0.$$

But

$$\int_{\Gamma} \mathfrak{L}_k \omega = \int_{\Gamma} d(k \lrcorner \omega) + \int_{\Gamma} k \lrcorner d\omega = \int_{\partial\Gamma} k \lrcorner \omega = 0.$$

Thus the quantity  $(k \lrcorner \omega)$  is constant along the solution curves. This is a simple example of a very general result called Noether's Theorem, relating infinitesimal symmetries and conservation laws.

## 35. Constrained variations

Since Hamilton's variational principle can be handled by *ad hoc* methods, the general problem of constraints in the calculus of variations is of physical interest mainly for the light it sheds on symmetries and conservation laws. Also, I present it here to show how it is handled by the use of differential forms, especially since a proof is very hard to find. It is called the problem of Lagrange. The answer is simple and pretty.

### *Holonomic constraints*

Here we have a constraint given by some function (or perhaps several)  $\phi: M \rightarrow R$ , such that allowed paths must lie in the  $\phi=0$  surface. Although we can in principle choose new coordinates such that we can define a smaller manifold, one in which there are no constraints, quite often that is not convenient. Again, we look at the integral

$$I = \int_{\Gamma} \omega$$

and demand that its variation vanish

$$\mathfrak{L}_v \int_{\Gamma} \omega = 0$$

for all deformations consistent with the constraint,

$$\mathfrak{L}_v \phi = 0;$$

that is,

$$v \lrcorner d\phi = 0.$$

This is a Lagrange multiplier problem in standard form. The variation at any point of the integral must vanish

$$v \lrcorner \dot{\gamma} \lrcorner d\omega = 0$$

for all  $v$  satisfying the condition

$$v \lrcorner d\phi = 0,$$

which implies that there exists a  $\lambda$ , a function of  $s$ , such that along the curve

$$\dot{\gamma} \lrcorner d\omega = \lambda d\phi,$$

and, of course, the curve must also satisfy

$$\phi = 0.$$

These equations determine both the curve and the Lagrange multiplier  $\lambda$ .

### *Anholonomic constraints*

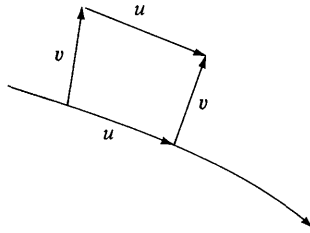
For this case the constraint has the form

$$\dot{\gamma} \lrcorner \alpha = 0 \tag{35.1}$$

for some given 1-form (or several)  $\alpha$ . This is a far trickier possibility. At any point there is no local constraint, such as the tempting condition

$$v \lrcorner \alpha = 0 \quad \text{[NO!]}$$

despite the proof offered in Goldstein (1959). Still, the constraint does really constrain. The correct condition is that equation 35.1 must be preserved by the deformation. The vector field  $\dot{\gamma}$  can be extended off the curve using the condition



**Figure 35.1.** Extending the tangent vector  $u$  off the curve by Lie transport.

$$\mathcal{L}_v \dot{\gamma} = 0$$

sketched in Figure 35.1. This amounts to choosing a simple parametrization for the deformed integrals. Since the value of the integral is parametrization-independent, this is allowed. The constraint condition is thus

$$\dot{\gamma} \lrcorner \mathcal{L}_v \alpha = \mathcal{L}_v (\dot{\gamma} \lrcorner \alpha) = 0. \quad (35.2)$$

The difficulty is that this expression involves not just derivatives along the curve (which we could get rid of if we integrated by parts), but transverse derivatives as well.

A Lagrange multiplier again does the trick, this time in the form

$$\dot{\gamma} \lrcorner d(\omega + \lambda\alpha) = 0,$$

and, along with

$$\dot{\gamma} \lrcorner \alpha = 0,$$

these again determine the path  $\Gamma$ . [An amazing number of good and careful books, such as Courant and Hilbert (1953) and Gelfand and Fomin (1963), decline to do more than to assert this result. I adapted the following proof from Akhiezer (1962). See also Pars (1962).] In outline, we must show that we can find a Lagrange multiplier  $\lambda$  such that for a constrained minimum of the original problem, we have an unconstrained minimum for the problem with 1-form  $\omega + \lambda\alpha$ .

The argument goes as follows. We first show that a vector field  $w$  can be chosen such that all deformations  $v$  (vector fields along  $\Gamma$ ) can be written in the form

$$v = v_c + \zeta w,$$

where the deformations  $v_c$  satisfy the constraint and  $\zeta$  is a scalar function that makes up for the degree of freedom lost in the constraint (Figure

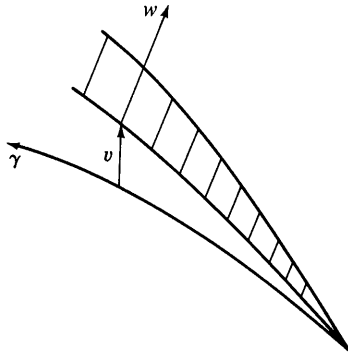


Figure 35.2. Motions along  $w$  restore the degree of freedom removed by the constraint.

35.2). Then we will show that for all deformations in this form, a  $\lambda$  can be found independent of  $\zeta$  and  $v$ .

The first result is shown as follows. Let  $v$  be any given deformation. Then the deformation  $v - \zeta w$  satisfies the constraint if we have

$$\dot{\gamma} \lrcorner \mathfrak{L}_{(v-\zeta w)} \alpha = 0.$$

To simplify matters we pick  $w$  to satisfy

$$w \lrcorner \alpha = 0, \tag{35.3}$$

a condition independent of  $v$ . Expanding, we have

$$\dot{\gamma} \lrcorner \mathfrak{L}_v \alpha = \zeta \dot{\gamma} \lrcorner w \lrcorner d\alpha,$$

and we have the explicit result that

$$\zeta = \frac{\dot{\gamma} \lrcorner \mathfrak{L}_v \alpha}{\dot{\gamma} \lrcorner w \lrcorner d\alpha}.$$

We must have

$$w \lrcorner d\alpha \neq 0,$$

and  $d\alpha$  cannot vanish or the constraint would be integrable.

Having shown that all deformations can be written as  $v_c + \zeta w$ , we now study the demand that the integral be an extremum for all deformations

$$\dot{\gamma} \lrcorner \mathfrak{L}_{(v_c+\zeta w)} (\omega + \lambda \alpha).$$

Since the original path is assumed to be extremal for deformations satis-

fying the constraint, we have

$$\dot{\gamma} \lrcorner \mathbb{F}_{v_c} \omega = 0.$$

Since  $v_c$  satisfies the constraint,

$$\lambda \dot{\gamma} \lrcorner \mathbb{F}_{v_c} \alpha = 0.$$

Finally

$$\dot{\gamma} \lrcorner \alpha(\mathbb{F}_{v_c} \lambda) = 0,$$

because the original path satisfies the constraint. Thus we have to satisfy only

$$\dot{\gamma} \lrcorner \mathbb{F}_{\zeta w}(\omega + \lambda \alpha) = 0.$$

We use the Cartan identity and discard the exact differential to find

$$\dot{\gamma} \lrcorner \zeta w \lrcorner d\omega + \lambda \dot{\gamma} \lrcorner \zeta w \lrcorner d\alpha + \dot{\gamma} \lrcorner \zeta w \lrcorner d\lambda \wedge \alpha = 0, \quad (35.4)$$

and  $\zeta$  factors out of the expression. The third term vanishes because either  $\dot{\gamma}$  or  $w$  must hit  $\alpha$ . Thus we have the explicit expression for  $\lambda$ , clearly independent of  $\zeta$ :

$$\lambda = - \frac{\dot{\gamma} \lrcorner w \lrcorner d\omega}{\dot{\gamma} \lrcorner w \lrcorner d\alpha}.$$

If we have several constraint 1-forms, we handle them by introducing several  $w$  fields and a corresponding number of Lagrange multipliers. The Lagrange multipliers may be found either from the preceding equation or directly from the new variational problem.

When the constraints  $\alpha$  involve higher derivatives, equation 35.3 cannot be satisfied. Our procedure is easily modified, however. The  $w$  representing first derivatives will satisfy 35.3, and their  $\zeta$  can be solved for directly. The second derivatives can then be solved for, and so on. The constraint equations do not separate, but they stratify. The only other change is in equation 35.4. The third term need not vanish, but will only involve derivatives along the original curve. The equations for the  $\lambda$  become differential equations rather than algebraic equations.

**Example:** The dynamics of a conservative system is given by a variational principle based on the functional

$$\int_{\Gamma} L(q, \dot{q}, t) dt$$

subject to the set of constraints

$$\dot{\gamma} \lrcorner (\dot{q} dt - dq) = 0.$$

A suitable set of vectors  $w$  is clearly  $\partial/\partial\dot{q}$ . We have

$$d\omega = \frac{\partial L}{\partial q} dq dt + \frac{\partial L}{\partial \dot{q}} d\dot{q} dt,$$

$$d\alpha = d\dot{q} dt,$$

$$\dot{\gamma} = Q \frac{\partial}{\partial q} + \dot{Q} \frac{\partial}{\partial \dot{q}} + \frac{\partial}{\partial t},$$

and so

$$\lambda = -\frac{\partial L}{\partial \dot{q}}.$$

The equivalent unconstrained problem is

$$\int_{\Gamma} \left[ L dt - \frac{\partial L}{\partial \dot{q}} (\dot{q} dt - dq) \right].$$

The Euler-Lagrange equations for this problem are

$$\dot{\gamma} \lrcorner \left[ dL dt - d \left( \frac{\partial L}{\partial \dot{q}} \right) (\dot{q} dt - dq) - \frac{\partial L}{\partial \dot{q}} d\dot{q} dt \right];$$

that is

$$\dot{\gamma} \lrcorner \left[ \frac{\partial L}{\partial q} dq dt - d \left( \frac{\partial L}{\partial \dot{q}} \right) (\dot{q} dt - dq) \right] = 0.$$

Now

$$\dot{\gamma} \lrcorner dq dt = \dot{\gamma} \lrcorner (dq - \dot{q} dt) dt = -(dq - \dot{q} dt) (\dot{\gamma} \lrcorner dt)$$

and so we have

$$\left\{ \dot{\gamma} \lrcorner \left[ \frac{\partial L}{\partial q} dt - d \left( \frac{\partial L}{\partial \dot{q}} \right) \right] \right\} \wedge (\dot{q} dt - dq) = 0.$$

The extremals are thus integral curves of the 1-forms

$$\frac{\partial L}{\partial q} dt - d \left( \frac{\partial L}{\partial \dot{q}} \right)$$

and

$$dq - \dot{q} dt.$$

*Symmetries*

Suppose that we have an infinitesimal symmetry as in Section 34

$$\mathfrak{L}_k \omega = 0, \quad \mathfrak{L}_k \alpha = 0.$$

It must be a symmetry of the constraints as well. This makes it a symmetry of the entire problem, including the Lagrange multiplier  $\lambda$ . The problem thus has the same form as that in the last section. It is only necessary that the Lie derivative of the  $\alpha$  be in the ideal generated by  $\alpha$ , because

$$\dot{\gamma} \lrcorner \alpha = 0$$

is true along the path. This lets us bring the  $\lambda$  inside the Lie differentiation as well, and we have

$$\mathfrak{L}_k(\omega + \lambda\alpha) \subset \mathcal{I}(\alpha).$$

From this we find

$$\int_{\Gamma} \{d[k \lrcorner (\omega + \lambda\alpha)] + k \lrcorner d(\omega + \lambda\alpha)\} = 0.$$

The second term is again just the Euler-Lagrange equation, and so vanishes, and we find

$$k \lrcorner (\omega + \lambda\alpha) = \text{constant}$$

along the solution curves. Again a symmetry leads to a conservation law. Note that it was necessary to put the variational problem into Lagrange-multiplier form for this.

**36. Variations of multiple integrals**

We now extend the calculus of variations from line integrals to integrals of any dimension. The geometric setting is a fiber bundle  $\pi: M \rightarrow N$ , with the objects of inquiry being sections of this bundle,  $\phi: N \rightarrow M$  (see Figure 36.1). The functional on these sections is given by an  $n$ -form  $\omega$  on  $M$ , where  $n$  is the dimension of  $N$ , integrated over some region  $\Gamma$  in  $N$ :

$$\int_{\phi(\Gamma)} \omega = \int_{\Gamma} \phi^* \cdot \omega.$$

The cross sections may be subject to constraints, the general anholonomic constraints being given by differential forms  $\alpha$  such that

$$\phi^* \cdot \alpha = 0.$$



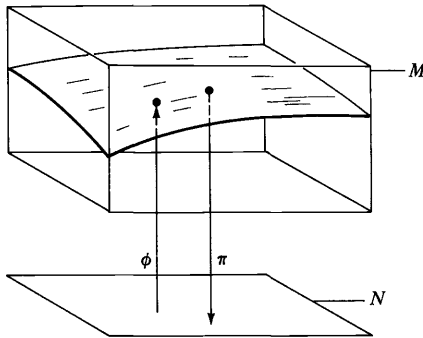


Figure 36.1. A field represented by a section of a fiber bundle.

**Example:** The description of a field in spacetime by a variational principle takes place in a bundle whose points consist of a point in spacetime, a value of the field variable  $\psi$ , and as many of its partial derivatives as are needed. For a field Lagrangian using only first derivatives, we need the four derivatives  $\psi_{,t}, \psi_{,x}, \psi_{,y}, \psi_{,z}$  as additional variables. The bundle is given by

$$\pi : (t, x, y, z, \psi, r, s, u, v) \rightarrow (t, x, y, z)$$

and cross sections by

$$\phi : (t, x, y, z) \rightarrow (t, x, y, z, \Psi, R, S, U, V)$$

where  $\Psi, R$ , etc., are all functions of  $t, x, y, z$ .

The cross sections of interest are integral submanifolds of the contact 1-form

$$\alpha = r dt + s dx + u dy + v dz - d\psi,$$

which forces  $R$  to be  $\psi_{,t}$  and so on. The action integral is given by

$$I(\phi) = \int_{\Gamma} \phi^* \cdot (L dt dx dy dz).$$

The variational problem is to make the integral

$$\int_{\phi(\Gamma)} \omega$$

stationary for all variations  $v$  that satisfy the constraint

$$\phi^* \cdot \alpha = 0.$$

For simplicity we deal with only a single constraint. If we have several constraints, we need only another layer of indices. In the bundle the constraint is

$$\mathbf{u} \lrcorner \mathfrak{L}_v \alpha = 0,$$

where the  $\mathbf{u}$  are  $n$  vectors that are pushforwards under  $\phi$  of a basis for  $N$ . They are the analogs of  $\dot{\gamma}$ .

We proceed as before. We factor deformations into the form

$$v = v_c + \zeta w,$$

where  $v_c$  satisfies the constraint, and the  $n$  vectors  $w$  satisfy

$$w \lrcorner \alpha = 0, \tag{36.1}$$

$$w \lrcorner \omega = 0. \tag{36.2}$$

Each constraint requires  $n$  vectors  $w$ . The modifications necessary when 36.1 cannot be satisfied were discussed in Section 35.

**Example:** For a section  $\mathbb{R}^2 \rightarrow \mathbb{R}^3$ , an arbitrary section requires two additional adjustments, the  $x$  and  $y$  derivatives, say, to satisfy the contact constraint

$$df - p dx - q dy.$$

The functions  $\zeta$  are now determined by the  $n$  simultaneous equations

$$\zeta u^a \lrcorner w \lrcorner d\alpha = u^a \lrcorner \mathfrak{L}_v \alpha,$$

and provided the 1-forms  $w \lrcorner d\alpha$  are nonsingular, these equations can be solved.

We now show that we can write this as an unconstrained problem by using Lagrange multipliers,

$$I'(\phi) = \int_{\phi(\Gamma)} (\omega + \lambda \wedge \alpha). \tag{36.3}$$

Since  $\phi^* \cdot \alpha = 0$ , this integrand has the same value for all constrained variations as our original integrand. Suppose  $\phi$  is an extremum of such constrained variations. We must show that we can pick  $\lambda$ , now an  $(n-1)$ -form, such that the functional in equation 36.3 is extremal for all variations.

As before, we write the deformation in terms of a part that satisfies the constraint plus a remainder, and we examine the requirement

$$U \lrcorner \mathfrak{L}_{(v_c - \zeta w)}(\omega + \lambda \wedge \alpha) = 0, \tag{36.4}$$

where

$$U = u_n \lrcorner u_{n-1} \lrcorner \cdots \lrcorner u_1$$

is a pushforward basis  $n$ -vector. Now

$$U \lrcorner \mathfrak{L}_{v_c} \omega = 0,$$

because  $v_c$  is a constrained variation, and the section  $\phi$  is assumed to be a solution of the original problem. Also

$$U \lrcorner \mathfrak{L}_{v_c}(\lambda \wedge \alpha) = U \lrcorner [(\mathfrak{L}_{v_c} \lambda) \wedge \alpha + \lambda \wedge \mathfrak{L}_{v_c} \alpha]. \tag{36.5}$$

The first term vanishes because

$$u \lrcorner \alpha = 0,$$

and one of the  $n$   $u$ s in  $U$  must hit  $\alpha$ . The second term in equation 36.5 vanishes because  $v_c$  satisfies the constraint

$$u \lrcorner \mathfrak{L}_{v_c} \alpha = 0.$$

Thus our  $\lambda$  will be independent of  $v_c$ .

Look at the remaining part of equation 36.4:

$$\begin{aligned} U \lrcorner \mathfrak{L}_{\zeta w}(\omega + \lambda \wedge \alpha) \\ = U \lrcorner \{ \zeta w \lrcorner (d\omega + d\lambda \wedge \alpha + (-)^{n-1} \lambda \wedge d\alpha) + d[\zeta w \lrcorner (\omega + \lambda \wedge \alpha)] \}, \end{aligned}$$

giving us  $n$  equations

$$U \lrcorner \{ w \lrcorner d\omega + \lambda \wedge (w \lrcorner d\alpha) + (-)^{n-1} (w \lrcorner \lambda) \wedge d\alpha \} = 0.$$

The  $(n-1)$ -form  $\lambda$  can be chosen to be the pullback of an  $(n-1)$ -form on  $N$ , and will then have exactly  $n$  unknown coefficients, for which we have  $n$  linear inhomogeneous equations. In general these can be solved, and the variational problem can then be written in Lagrange-multiplier form, as in equation 36.3.

**Example:** We continue the example of the field  $\psi$ . We extremize

$$\int L\theta,$$

where

$$\theta = dt \, dx \, dy \, dz.$$

The constraint is

$$\alpha = p dq - d\psi,$$

where the coordinates  $p$  are the gradients  $(r, s, u, v)$  used before. Vectors that satisfy equations 36.1 and 36.2 are

$$w = \frac{\partial}{\partial p}.$$

Since the Lagrange multiplier is a pullback from  $N$ , it satisfies

$$w \lrcorner \lambda = 0,$$

and so we have

$$U \lrcorner \{w \lrcorner d\omega + \lambda \wedge (w \lrcorner d\alpha)\} = 0.$$

Here

$$d\omega = \frac{\partial L}{\partial p} dp \wedge \theta + \frac{\partial L}{\partial \psi} d\psi \wedge \theta$$

and

$$d\alpha = dp dq$$

Thus

$$w \lrcorner d\omega = \frac{\partial}{\partial p} \lrcorner d\omega = \frac{\partial L}{\partial p} \theta.$$

Also,

$$w \lrcorner d\alpha = dq;$$

so we have

$$U \lrcorner \left( \frac{\partial L}{\partial p} \theta + \lambda \wedge dq \right) = 0.$$

The 3-form  $\lambda$  can be found by inspection.

$$\lambda = (-)^n \frac{\partial L}{\partial p} \theta,$$

where

$$\theta = \frac{\partial}{\partial q} \lrcorner \theta,$$

and we have used

$$dq^i \theta_j = \delta^i_j \theta$$

(equation 23.4). Thus the field-theory Lagrangian is equivalent to

the free variational problem that uses the functional

$$\int_{\phi(\Gamma)} \left[ L\theta - \frac{\partial L}{\partial \mathbf{p}} (p_i dq^i - d\psi) \wedge \theta \right].$$

The Euler–Lagrange equations follow as before. We must have

$$U \lrcorner \left[ dL\theta - d\left(\frac{\partial L}{\partial \mathbf{p}}\right)(p_i dq^i - d\psi) \wedge \theta - \frac{\partial L}{\partial \mathbf{p}} dp_i dq^i \theta \right] = 0,$$

and so

$$\begin{aligned} U \lrcorner \left[ \frac{\partial L}{\partial \psi} d\psi \wedge \theta + \frac{\partial L}{\partial \mathbf{p}} d\mathbf{p} \wedge \theta \right. \\ \left. - d\left(\frac{\partial L}{\partial \mathbf{p}}\right)(p_i dq^i - d\psi) \wedge \theta - \frac{\partial L}{\partial \mathbf{p}} d\mathbf{p} \wedge \theta \right] = 0. \end{aligned}$$

Use

$$d\psi \wedge \theta = (d\psi - \mathbf{p} dq) \wedge \theta$$

to find

$$\left\{ U \lrcorner \left[ \frac{\partial L}{\partial \psi} \theta - d\left(\frac{\partial L}{\partial \mathbf{p}}\right)\theta \right] \right\} \wedge (p_i dq^i - d\psi) = 0,$$

and so the solution is given by

$$\begin{aligned} U \lrcorner \left[ \frac{\partial L}{\partial \psi} \theta - d\left(\frac{\partial L}{\partial \mathbf{p}}\right)\theta \right] &= 0, \\ U \lrcorner (\mathbf{p} dq - \psi) &= 0, \end{aligned}$$

equivalent to the classical Euler–Lagrange equations.

### Conservation laws

Again the presence of a symmetry leads to a conservation law. Suppose that we have a vector  $k$  on  $M$  such that

$$\mathfrak{L}_k \omega = 0 \quad \text{and} \quad \mathfrak{L}_k \alpha = 0.$$

Since we have

$$\mathbf{u} \lrcorner \alpha = 0,$$

we also have

$$U \lrcorner \mathfrak{L}_k(\omega + \lambda \wedge \alpha) = 0.$$

To find the conservation law, integrate over an *arbitrary* region  $\Gamma$ , and, with  $\phi$  a solution, we have

$$\int_{\phi(\Gamma)} d[k \lrcorner (\omega + \lambda \wedge \alpha)] + \int_{\phi(\Gamma)} k \lrcorner d(\omega + \lambda \wedge \alpha) = 0.$$

The second term in the integrand is the Euler–Lagrange equation and so vanishes; and since the region is arbitrary, we must have

$$d[k \lrcorner (\omega + \lambda \wedge \alpha)] = 0.$$

This is a conservation law on  $N$  when pulled back using  $\phi$ . Note again how the Lagrange-multiplier form is essential.

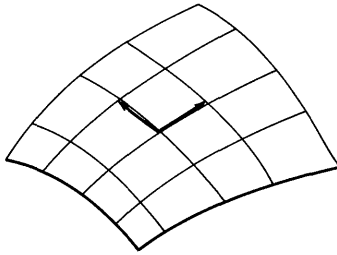
### 37. Holonomy and thermodynamics

In Section 19 we discussed how a set of vector fields could determine a submanifold by the requirement that the submanifold contain the vector fields (see Figure 37.1). The criterion that a set of vector fields must meet in order to define submanifolds was given by the Frobenius Theorem: the vector fields must be closed under the Lie bracket. We have also discussed submanifolds defined by ideals of differential forms, with the requirement that the forms pull back to zero on the submanifold (see Figure 37.2). There is a dual version of the Frobenius Theorem. To determine a submanifold, the ideal of differential forms generated by the given forms must be closed under exterior differentiation.

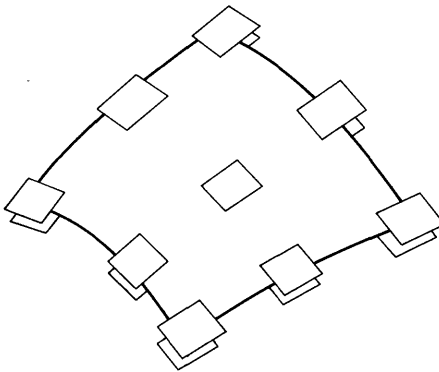
With the Frobenius Theorem and the tools of the exterior calculus, we can now finish our discussion of thermodynamics started in Section 16. We saw there that a thermodynamic system is described by an  $n$ -surface in a  $(2n+1)$ -dimensional contact bundle. This is the geometric content of the first law. The local properties of such a system are represented by the contact element to this surface, which in turn is described by  $n(n+1)$  partial derivatives. This local theory could be called differential thermodynamics. Half of these coefficients are constrained by the geometric structure of thermodynamics, and half reflect the particular properties of the system itself.

These constraints, called *Maxwell relations*, are often proven by using a Legendre transformation to change the coordinates so that the result is obvious from the symmetry of partial derivatives.

**Example:** Look at the energy 2-surface as the graph of a map  $\mathbb{R}^2 \rightarrow \mathbb{R}^3$ ;  $(S, V) \mapsto (U, S, V)$ , with partial derivatives defined by



**Figure 37.1.** A submanifold defined by a pair of vector fields.



**Figure 37.2.** A submanifold defined by an ideal of differential forms.

$$\frac{\partial U}{\partial S} = T, \quad \frac{\partial U}{\partial V} = -P.$$

From the symmetry of partial derivatives, we have also

$$\frac{\partial T}{\partial V} + \frac{\partial P}{\partial S} = 0.$$

These are thus three conditions that relate the six partial derivatives of  $U$ ,  $T$ , and  $P$ .

The exterior calculus gives us a nice geometric treatment of these relations in the more complicated situation where independent variables other than  $S$  and  $V$  are chosen. Look at systems having a single mechanical degree of freedom, with a contact structure given by

$$\alpha = dU - TdS + PdV.$$

Locally the system can be described by a two-dimensional surface element. Let us write this as the linear space spanned by the two vectors

$$A = a \frac{\partial}{\partial v} + \frac{\partial}{\partial T} + b \frac{\partial}{\partial S} + c \frac{\partial}{\partial P}$$

and

$$B = d \frac{\partial}{\partial U} + \frac{\partial}{\partial T} + e \frac{\partial}{\partial S} + f \frac{\partial}{\partial V}.$$

These vectors have been normalized to have unit  $T$  components, and have been oriented to be in a constant  $V$  and a constant  $P$  direction.

To be a piece of the surface describing a thermodynamic system, these vectors must lie in the zero surface of  $\alpha$ , and so we must have

$$A \lrcorner \alpha = \alpha \cdot A = 0, \quad B \lrcorner \alpha = \alpha \cdot B = 0.$$

These conditions give us

$$a = bT \quad \text{and} \quad d = eT - fP.$$

This reduces the number of independent coefficients from six to four.

The four remaining coefficients,  $b$ ,  $c$ ,  $e$ , and  $f$ , have simple physical interpretations. Let  $C_V$  be the heat capacity at constant volume. The vector  $A$  represents displacements at constant volume. The heat capacity is the ratio of  $T$  times the entropy change to the temperature change. Similarly, the thermal stress coefficient measures the pressure change at constant volume due to a changing temperature, which we write as  $(dP/dT)_V$ . The vector  $A$  can thus be written

$$A = \frac{\partial}{\partial T} + \frac{C_V}{T} \frac{\partial}{\partial S} + \left( \frac{dP}{dT} \right)_V \frac{\partial}{\partial P} + C_V \frac{\partial}{\partial U}.$$

Likewise,

$$B = \frac{\partial}{\partial T} + \frac{C_P}{T} \frac{\partial}{\partial S} + \left( \frac{dV}{dT} \right)_P \frac{\partial}{\partial V} + \left[ C_P - P \left( \frac{dV}{dT} \right)_P \right] \frac{\partial}{\partial U}.$$

There is one final thermodynamic requirement. The ideal must be a differential ideal, and we have to add  $d\alpha$  to the list of generators of the ideal. This imposes a final condition

$$B \lrcorner A \lrcorner d\alpha = 0.$$

We have

$$A \lrcorner d\alpha = A \lrcorner (-dTdS + dPdV) = -dS + (C_V/T) dT + (dP/dT)_V dV,$$



and so

$$(C_V - C_P)/T + \left(\frac{dV}{dT}\right)_P \left(\frac{dP}{dT}\right)_V = 0.$$

This is a thermodynamic identity, independent of the material comprising the system. A relation such as this is called a Maxwell relation. Because of the Maxwell relations, only three independent parameters are needed to describe the linear approximation to a thermodynamic system.

This completes our discussion of thermodynamics. The geometry is just that of an  $n$ -surface in  $(2n+1)$ -space, with its local structure described by the contact element to the surface. Note that all thermodynamic differentials have been turned into honest differential forms.

### 38. Exterior differential systems

We can now extend our treatment of single partial differential equations (pdes) (Section 32) to systems of partial differential equations. As with a single pde, the appropriate geometric structure is an ideal of differential forms closed under exterior multiplication and exterior differentiation. A space with such an ideal is called an *exterior differential system*.

**Examples:** A dynamical system can be specified in two ways. Either give a vector field, which determines parametrized curves as integral curves, or give a set of  $(n-1)$  1-forms, which have unparametrized curves as integral curves (Figure 38.1). The classical way of writing the latter systems is in the form

$$p dx : q dy : r dz,$$

which is equivalent to the system generated by the two 1-forms

$$p dy - q dx \quad \text{and} \quad r dy - q dz.$$

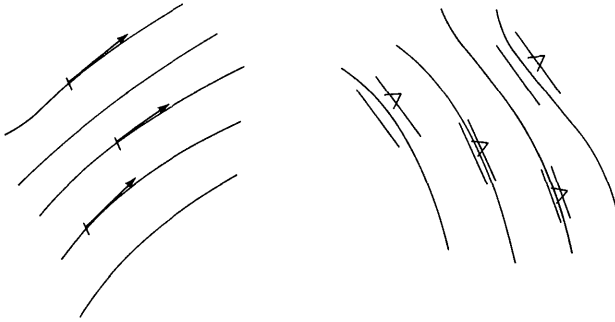
The example without peer for this is the heat equation

$$\frac{\partial \phi}{\partial t} = \kappa \frac{\partial^2 \phi}{\partial x^2}.$$

I use  $\phi$  for temperature, instead of  $T$ , in order to save  $T$  for the  $t$  component of vectors. As a first-order system, we have

$$u = \kappa \phi_{,x} \quad d[\phi dx + u dt] = 0. \quad (38.1)$$

Now look at the 4-space with coordinates  $(t, x, \phi, u)$  and the two



**Figure 38.1.** Two ways to specify a dynamical system: either as the parametrized integral curves of a vector field; or as integral submanifolds of an ideal of differential forms.

2-forms

$$\omega = d\phi dx + du dt, \quad \alpha = u dx dt - \kappa d\phi dt.$$

The second of these is generated, roughly, by “multiplying” by  $dx dt$ ; remember that exterior multiplication is anticommutative. The first is generated by exterior differentiation of equation 38.1. If we have a two-dimensional submanifold

$$\Phi: \mathbb{R}^2 \rightarrow \mathbb{R}^4; (t, x) \mapsto (t, x, \phi, u),$$

then the condition

$$\Phi^*(u dx dt - \kappa d\phi dt) = 0$$

is just

$$u = \kappa \phi_{,x}.$$

Thus the integral submanifolds of the ideal generated by the 2-forms  $\omega$  and  $\alpha$  are the solutions to the heat equation. It is convenient to study the heat equation by studying the geometry of this ideal.

Note that this ideal is closed under exterior differentiation:

$$d\alpha = du dx dt = -dx \wedge dt.$$

The wave equation

$$\frac{\partial^2 \psi}{\partial t^2} = \frac{\partial^2 \psi}{\partial x^2},$$

written as the system

$$\begin{aligned}\psi_{,x} &= u, & \psi_{,t} &= v, \\ u_{,t} &= v_{,x}, & u_{,x} &= v_{,t},\end{aligned}$$

translates into the ideal generated by the four 2-forms

$$\begin{aligned}d\psi dt - u dx dt, \\ d\psi dx + v dx dt, \\ du dx + dv dt, \\ du dt + dv dx.\end{aligned}$$

We now check that this is a differential ideal:

$$\begin{aligned}-du dx dt &= dx(du dt + dv dx) + dv dx dt \\ &= -dt(du dt + dv dx).\end{aligned}$$

Had we been awake we would have noticed that we are in a contact manifold, with contact 1-form

$$d\psi - u dx - v dt.$$

This 1-form generates the first two 2-forms and gives us instead the ideal

$$\begin{aligned}d\psi - u dx - v dt, & \quad du dx + dv dt, \\ & \quad du dt + dv dx.\end{aligned}$$

We will see in a bit that this is a better exterior differential system for representing the wave equation. If we were not interested in  $\psi$  itself, then the ideal consisting of the last two 2-forms could be studied in a space of only four dimensions.

### *Electrodynamics*

Maxwell's equations in units with  $c = 1$  (details will come in Section 42) are

$$\begin{aligned}\frac{\partial \mathbf{B}}{\partial t} &= -\nabla \times \mathbf{E}, \\ \frac{\partial \mathbf{E}}{\partial t} &= \nabla \times \mathbf{B} - 4\pi \mathbf{J}, \\ \nabla \cdot \mathbf{B} &= 0,\end{aligned}$$

$$\nabla \cdot E = 4\pi\rho.$$

To write these as an exterior differential system, we go to a ten-dimensional space with coordinates  $(t, x, \dots, E^x, \dots, B^x, \dots)$ . Let our missing indices run over just  $x, y,$  and  $z$ . The equation

$$\nabla \cdot B = 0$$

is equivalent to the 4-form

$$(dB^x dy dz + dB^y dz dx + dB^z dx dy) dt.$$

If we use the 3-space volume element (Section 23)

$$\theta = dx dy dz$$

and the contractions

$$\theta_x = \frac{\partial}{\partial x} \lrcorner \theta, \dots,$$

the preceding 4-form is

$$dB \theta dt.$$

Continuing in this way, we can represent Maxwell's equations by the ideal generated by the eight 4-forms

$$\begin{aligned} dB^i \theta + (dE dq) dt dq^i, \\ dE^i \theta - (dB dq) dt dq^i + 4\pi J^i dt \theta, \\ dB \theta dt, \\ dE \theta dt - 4\pi \rho \theta dt. \end{aligned} \tag{38.2}$$

Just as with the wave equation, we can find simpler generators for this idea. We play with the first three forms, using

$$dB^i \theta = dB^j \delta_j^i \theta = dB^j \theta_j dq^i$$

to write them

$$[(dB \theta) + (dE dq) dt] dq^i.$$

An ideal containing the single 3-form

$$dB \theta + dE dq dt$$

generates the first three 4-forms above. It also generates the one given in equation 38.2. Likewise the 3-form

$$dE\theta - dBdqdt - 4\pi J\theta dt - 4\pi\rho\theta$$

generates the other four 4-forms.

The new ideal generated by these two 3-forms is closed under exterior multiplication, provided that the functions  $J$  and  $\rho$  satisfy

$$dJ\theta dt + d\rho\theta = 0,$$

which is just the condition of charge conservation. Not only is this ideal with two 3-forms simpler, but there are defects in the original ideal that make it unusable. Try, for example, to show that the ideal generated by 4-forms is closed under exterior differentiation.

One advantage to using a differential ideal representation for a physical system rather than a system of pdes is that the distinction between dependent and independent variables can be easily forgotten. This leads to the possibility of simplifying a system of pdes by choosing different variables to be the independent variables. A given differential ideal thus represents a large family of different systems of pdes.

**Example:** Isentropic fluid flow in one space dimension is described by the conservation of mass

$$\rho_{,t} + \rho u_{,x} + u\rho_{,x} = 0$$

and conservation of momentum

$$\rho u_{,t} + \rho uu_{,x} + c^2\rho_{,x} = 0,$$

where  $c$  is the sound speed. These can be represented by the ideal

$$dx d\rho + \rho du dt + u d\rho dt,$$

$$\rho dx du + \rho u du dt + c^2 d\rho dt.$$

As far as the ideal is concerned, one could as well take  $\rho$  and  $u$  to be the independent variables, leading to equations

$$x_{,u} + \rho t_{,\rho} - ut_{,u} = 0,$$

$$\rho x_{,\rho} - \rho ut_{,\rho} + c^2 t_{,u} = 0.$$

Note that these are linear equations. Thus solutions of the original set cannot be superposed as functions of  $x$  and  $t$ , but they can be superposed as functions of  $\rho$  and  $u$ ! This is called a hodograph transformation. This trick works for any nonlinear system that is linear in the derivatives.

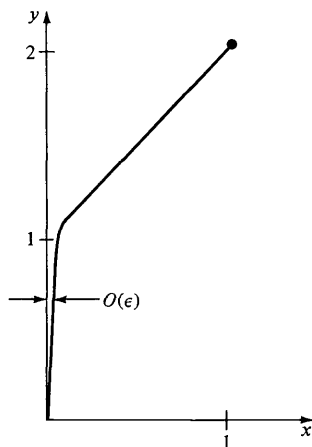


Figure 38.2.

A differential ideal is not precisely equivalent to a system of pdes. The ideal may allow singular solutions for which the original choice of independent variables is not appropriate. Far from being a defect, this is a natural way of including singular solutions, such as shock waves, in an orderly manner.

**Example:** The time-honored example of a system with a shock-wave solution is the ordinary differential equation

$$\epsilon y'' + y' - 1 = 0$$

with boundary conditions

$$y(0) = 0, \quad y(1) = 2,$$

in the limit  $\epsilon \rightarrow 0$ . The solution is sketched in Figure 38.2. Define

$$p = y';$$

then the equation can be represented by the ideal generated by

$$\begin{aligned} dy - p dx, \\ \epsilon dp + (p-1) dx. \end{aligned} \tag{38.3}$$

An alternative ideal, equivalent except for singular solutions, is to use not the slope, which blows up at a shock wave, but the angle  $\theta$

defined by replacing equation 38.3 with

$$\alpha = \sin \theta dx - \cos \theta dy.$$

The original differential equation then goes over to the 1-form generator

$$\beta = \epsilon d\theta + \cos \theta (\sin \theta - \cos \theta) dx.$$

When  $\epsilon = 0$ , this ideal has integral submanifolds of three types

$$y = x + C,$$

$$\theta = \frac{\pi}{2}, \quad x = \text{constant},$$

and

$$x = \text{constant}, \quad y = \text{constant}.$$

The  $\epsilon = 0$  shock-wave solution can be constructed from pieces of these.

### *Regular integral submanifolds*

Many different ideals can represent a physical system. Of particular utility are those for which the integral submanifolds are *regular*, by which we mean that the local construction of the integral submanifolds proceeds in an orderly fashion, as follows. Pick a curve satisfying the 1-forms in the ideal (with a great deal of freedom here, corresponding to different allowable boundary conditions), then extend this curve to a 2-surface that also satisfies the 2-forms in the ideal (and with less freedom), and so on up to the largest dimension, which is the dimension of the space of independent variables. Note the dual viewpoint here. Instead of pulling back the forms, we are pushing forward the tangent vectors. *Regular integral submanifolds* arise when this process ends with just the right number of conditions on the submanifold.

**Examples:** Look at Laplace's equation in two space dimensions. It is equivalent to the ideal generated by

$$d\phi - p dx - q dy, \quad (38.4)$$

$$dp dx + dq dy, \quad (38.5)$$

$$dp dy - dq dx. \quad (38.6)$$

Suppose we are given boundary conditions along the line  $y=0$ . That is, we want an integral submanifold that starts from the curve

$$\gamma: x \mapsto (x, y, p, q, \phi) = (x, 0, \Phi'(x), Q(x), \Phi(x)),$$

which contains two arbitrary functions. These functions allow us to match the boundary conditions. This curve is a one-dimensional integral submanifold, because

$$\gamma^* \cdot (d\phi - p dx - q dy) = 0.$$

Now, consider pushing out into the  $y$  direction. Since here we have two independent variables, we expect the extension into the  $y$  direction to be unique. Look at the point on the curve

$$(x, y, p, q, \phi) = (\bar{x}, 0, \Phi'(\bar{x}), Q(\bar{x}), \Phi(\bar{x})).$$

The tangent to the original curve at this point is

$$u = \frac{\partial}{\partial x} + \Phi''(\bar{x}) \frac{\partial}{\partial p} + Q'(\bar{x}) \frac{\partial}{\partial q} + \Phi'(\bar{x}) \frac{\partial}{\partial \phi}.$$

Take a vector in the  $y$  direction,

$$v = \frac{\partial}{\partial y} + A \frac{\partial}{\partial p} + B \frac{\partial}{\partial q} + C \frac{\partial}{\partial \phi}.$$

We have three conditions on  $v$ :

$$v \lrcorner (d\phi - p dx - q dy) = 0,$$

$$v \lrcorner u \lrcorner (dp dx + dq dy) = 0,$$

$$v \lrcorner u \lrcorner (dp dy - dq dx) = 0.$$

From these  $v$  is uniquely determined (up to an overall length, which we have fixed by picking the  $y$  component to be unity)

$$v = \frac{\partial}{\partial y} + Q'(\bar{x}) \frac{\partial}{\partial p} - \Phi''(\bar{x}) \frac{\partial}{\partial q} + Q(\bar{x}) \frac{\partial}{\partial \phi}.$$

The ideal given by equations 38.4, 38.5, and 38.6 thus has regular two-dimensional integral submanifolds.

Because  $v$  was uniquely determined, there are no three-dimensional integral submanifolds. If there were any, then another independent two-dimensional integral submanifold could be found. This contradicts the fact that  $v$  was uniquely determined.



The ideals for the wave equation and Maxwell's equations in their first forms in our examples do not have regular integral submanifolds.

Suppose we are given a 1-form  $\omega$  and want to find functions  $\alpha$  and  $\beta$  such that

$$\alpha d\beta = \omega.$$

If we take the exterior derivative of both sides, and then wedge it with the original equation, we see that the left-hand side vanishes identically. Thus a necessary condition for this problem to be soluble is

$$\omega \wedge d\omega = 0.$$

Likewise we must also have

$$d\omega \wedge d\omega = 0.$$

Look at the special case

$$\omega = x dy + 2y dx,$$

which satisfies the given conditions. This is equivalent to the exterior differential system generated by

$$\alpha d\beta - x dy - 2y dx$$

and

$$d\alpha d\beta + dx dy$$

in four dimensions. In order to make it a differential ideal, we had to explicitly include the exterior derivative of the first form in the generators. The original problem will be soluble, provided that this ideal has regular two-dimensional integral submanifolds.

To see that the ideal does indeed have regular integral 2-manifolds, we go to a typical point and try to pick two linearly independent vectors that will lie in the integral manifold, and see to what degree they are constrained by the ideal. Start with a vector

$$u = A \frac{\partial}{\partial \alpha} + B \frac{\partial}{\partial \beta} + C \frac{\partial}{\partial x} + D \frac{\partial}{\partial y}.$$

It must satisfy

$$\alpha B - xD - 2yC = 0.$$

The length of  $u$  is arbitrary; let us normalize it by taking  $C = 1$ . Then we have two arbitrary components, say,  $A$  and  $D$ , with  $B$  given by

$$B = \frac{xD + 2y}{\alpha}.$$

Now, pick a second vector,

$$v = E \frac{\partial}{\partial \alpha} + F \frac{\partial}{\partial \beta} + G \frac{\partial}{\partial x} + H \frac{\partial}{\partial y}.$$

It must satisfy

$$\alpha F - xH - 2yG = 0,$$

and, from the 2-form generator, it must satisfy

$$AF - EB + CH - GD = 0.$$

Again the length is arbitrary; we set  $H = 1$ . Any vector proportional to  $u$  will be a trivial solution; we rule this out by setting  $G = 0$ . This leaves us two equations:

$$\alpha F - x = 0,$$

$$AF - \frac{xD + 2y}{\alpha} E + 1 = 0.$$

These have a unique solution

$$F = \frac{x}{\alpha},$$

$$E = -\frac{xA + \alpha}{xD + 2y}.$$

Thus our ideal does have regular integral 2-manifolds; so our original problem is soluble in the neighborhood of a typical point. In fact, we have

$$x^3 d(yx^{-2}) = x dy - 2y dx.$$

This is a special case of the normal form for 1-forms discussed in Section 26.

### Problems

**38.1** Show that the ideal of 4-forms for Maxwell's equations does not have regular integral submanifolds.

**38.2** Consider a set of 1-forms (called a Pfaffian system) that satisfy

$$\omega^i \wedge d\omega^j = 0.$$

Show that these form a differential ideal.

**38.3** Generalize the last example, first to arbitrary forms in two dimensions, and then to higher dimensions.

### 39. Symmetries and similarity solutions

One of the main uses of the representation of a system of pdes as an ideal of differential forms is to study the symmetries of the system. We will concentrate here on continuous families of symmetries. An infinitesimal transformation will be represented by a vector field on the space where the differential ideal lives. Using the ideal of differential forms will make it easy for us to study even those symmetries that mix up dependent and independent variables. Also, by using the ideal to represent the system, we automatically achieve a treatment of its symmetries that is independent of any special choice of coordinates. [The basic reference for this is Harrison and Estabrook (1971).]

#### *Isovectors*

A transformation is a symmetry if it takes one solution into itself or another solution. This means that the ideal representing the exterior differential system must be invariant under the transformation. If  $S$  is a vector field that is an infinitesimal symmetry, then we must have, for any differential form  $\alpha$  in  $\mathcal{I}$ ,

$$\mathcal{L}_S \alpha \in \mathcal{I}.$$

Such vector fields are called *isovectors*.

**Examples:** The vector field

$$S = 2t \frac{\partial}{\partial t} + x \frac{\partial}{\partial x} + \phi \frac{\partial}{\partial \phi}$$

is an isovector of the heat-equation ideal,

$$\theta = u dx dt - d\phi dt,$$

$$\alpha = d\phi dx + du dt,$$

taking  $\kappa = 1$  for simplicity. We have, in fact

$$\mathcal{L}_S \theta = 3u dx dt - 3 d\phi dt = 3\theta,$$

$$\mathcal{L}_S \alpha = 2 d\phi dx + 2 du dt = 2\alpha.$$

Not all symmetries are obvious. Another isovector for the heat-equation ideal is

$$S = -2t \frac{\partial}{\partial x} + x\phi \frac{\partial}{\partial \phi} + (\phi + xu) \frac{\partial}{\partial u}.$$

We expect that the fluid-flow ideal of Section 38,

$$\alpha = dx d\rho + \rho du dt + u d\rho dt,$$

$$\beta = \rho dx du + \rho u du dt + c^2 d\rho dt,$$

will have a Galilean symmetry corresponding to adding a uniform velocity to everything. This shears worldlines by an operator  $t(\partial/\partial x)$ , translates velocities by an operator  $\partial/\partial u$ , and leaves densities unchanged. Thus we expect

$$S = t \frac{\partial}{\partial x} + \frac{\partial}{\partial u}$$

to be an isovector, and it is.

We can search systematically and exhaustively for isovectors by methods similar to those used to search for Killing vector fields. We find a set of linear pdes for the components of the isovector. The set is overdetermined, and will have only a few, special solutions, if any. It is tedious, but usually routine to find its solutions.

**Example:** For the heat-equation ideal given earlier we look for an isovector

$$S = T \frac{\partial}{\partial t} + X \frac{\partial}{\partial x} + U \frac{\partial}{\partial u} + \Phi \frac{\partial}{\partial \phi}$$

that contains four unknown functions  $T$ ,  $X$ ,  $U$ , and  $\Phi$ . These functions must satisfy the linear pdes

$$T_{,u} = 0, \tag{39.1}$$

$$\begin{aligned}
 T_{,x} + \Phi_{,u} &= 0, & (39.2) \\
 T_{,\phi} - X_{,u} &= 0, & T_{,x} - \Phi_{,u} + u(X_{,u} + T_{,\phi}) = 0, \\
 \Phi_{,\phi} + X_{,x} - U_{,u} - T_{,t} &= 0, \\
 \Phi_{,t} - U_{,x} - u(X_{,t} + U_{,\phi}) &= 0, \\
 U - \Phi_{,x} + u(X_{,x} - \Phi_{,\phi}) + u^2 X_{,\phi} &= 0,
 \end{aligned}$$

which we find from the equations

$$\begin{aligned}
 \mathfrak{L}_S(u \, dx \, dt - d\phi \, dt) &= 0 \text{ mod } \mathfrak{I}, \\
 \mathfrak{L}_S(d\phi \, dx + du \, dt) &= 0 \text{ mod } \mathfrak{I},
 \end{aligned}$$

by using the ideal to eliminate the 2-forms  $dt \, d\phi$  and  $dx \, d\phi$ . Several pages of work are needed to get these pdes. We solve these equations by replacing the dependent variables with more variables that depend on fewer independent variables. For example, from equations 39.1 and 39.2 we have

$$\Phi_{,uu} = 0;$$

so we can write  $\Phi$  in terms of new variables  $A$  and  $B$  as

$$\Phi(t, x, u, \phi) = A(t, x, \phi) + uB(t, x, \phi).$$

The process continues until the new variables introduced depend on nothing at all. Then they are just what we call constants of integration. In this way we can find a complete list of isovectors for the heat equation:

$$\begin{aligned}
 S_1 &= \frac{\partial}{\partial t}, & S_2 &= \frac{\partial}{\partial x}, \\
 S_3 &= \phi \frac{\partial}{\partial \phi} + u \frac{\partial}{\partial u}, & S_4 &= 2t \frac{\partial}{\partial t} + x \frac{\partial}{\partial x} - u \frac{\partial}{\partial u}, \\
 S_5 &= -2t \frac{\partial}{\partial x} + x\phi \frac{\partial}{\partial \phi} - (\phi + ux) \frac{\partial}{\partial u}, \\
 S_6 &= 2t^2 \frac{\partial}{\partial t} + 2xt \frac{\partial}{\partial x} + (\frac{1}{2}x^2 + t) \frac{\partial}{\partial \phi} - (\frac{1}{2}x^2u + 3ut + x\phi) \frac{\partial}{\partial u},
 \end{aligned}$$

and

$$S_7 = f(x, t) \frac{\partial}{\partial \phi} + \frac{\partial f}{\partial x} \frac{\partial}{\partial u},$$

for any function  $f$  satisfying

$$f_{,t} = f_{,xx}.$$

Any linear combination of these is also an isovector. It would have been difficult to find these symmetries even in the geometric language of Section 31.

In the preceding example we see that the heat equation has an isovector that depends on an arbitrary function. This reflects the special symmetry of superposition. Any equation that has a superposition principle, not necessarily in the variables originally used, will have such an isovector. Recall the nonobvious superposition principle for the fluid flow ideal given earlier.

### *Similarity solutions*

Given a symmetry, we can also search for special solutions that are themselves invariant under the transformation. That is, the transformation takes the solution into itself, not just into another solution. The solutions that satisfy this extra condition are integral submanifolds of the ideal generated by adding to the original ideal all the forms found by contracting the original ideal with  $S'$ . It is clear that any similarity solution must annihilate these forms, because it will contain the isovector  $S$  in its tangent space. Think of pushing forward the tangent space rather than pulling back the forms. The new ideal is automatically closed under exterior differentiation if and only if  $S$  is an isovector.

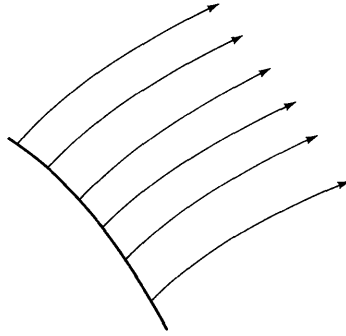
A vector field that satisfies

$$V \lrcorner \mathcal{I} \subset \mathcal{I}$$

is called a *Cauchy characteristic vector* of the ideal. Its integral curves are characteristics in the sense of Section 32. The flow generated by a Cauchy characteristic vector field sweeps an  $m$ -dimensional integral manifold into an  $(m+1)$ -dimensional integral manifold (see Figure 39.1). The augmented ideal constructed earlier automatically has the isovector  $S$  as a Cauchy characteristic. Thus the problem of finding similarity solutions for a single isovector involves one fewer dimension than the original problem.

**Example:** Look at the one-parameter family of isovectors to the heat equation,

$$S_\lambda = 2t \frac{\partial}{\partial t} + x \frac{\partial}{\partial x} - \lambda \phi \frac{\partial}{\partial \phi} - (\lambda + 1)u \frac{\partial}{\partial u}.$$



**Figure 39.1.** Cauchy characteristics used to sweep compatible boundary data into an integral submanifold with one more dimension.

Remember, the equation for isovectors is linear, and they can be superposed. The contractions of this vector with the original ideal give us the 1-forms

$$2t \, d\phi - 2tu \, dx + (ux - \lambda\phi) \, dt$$

and

$$\lambda\phi \, dx + x \, d\phi + (\lambda + 1)u \, dt + 2t \, du.$$

Now, we first find a one-dimensional integral manifold, say, lying along the  $t = 1$  line. We have then  $\phi$  and  $u$ , which are just functions of  $x$ , satisfying

$$\frac{d\phi}{dx} = u$$

and

$$2 \frac{du}{dx} + x \frac{d\phi}{dx} + \lambda\phi = 0.$$

The latter equation is an eigenvalue equation. The solutions vanishing at infinity occur for  $\lambda = 1, 2, \dots$ , and are

$$\phi = e^{-x^2/4} H_{\lambda+1} \left( \frac{x}{2} \right),$$

where the  $H_n$  are Hermite polynomials ( $H_0 = 1$ ,  $H_1 = 2x$ ,  $H_2 = 4x^2 - 2$ ).

Look at the lowest solution,  $\lambda = 1$ ,

$$\phi(1, x) = e^{-x^2/4}.$$

We extend this solution by using the vector field

$$2t \frac{\partial}{\partial t} + x \frac{\partial}{\partial x} - 2u \frac{\partial}{\partial u} - \phi \frac{\partial}{\partial \phi},$$

which has integral curves

$$s \mapsto (t, x, u, \phi) = (e^{2s}, Ae^s, Be^{-2s}, Ce^{-s}).$$

To find the field  $\phi$  at a point  $(t, x)$ , we must find which of these integral curves passes through that point. Clearly it has

$$A = \frac{x}{\sqrt{t}}.$$

This integral curve crosses the  $t=1$  line at an  $x$  value of

$$x_0 = A = \frac{x}{\sqrt{t}},$$

and there  $\phi$  has the value

$$\phi(1, x_0) = e^{-x^2/4t},$$

which give us

$$C = e^{-x^2/4t}$$

and a value of  $\phi$  at  $(t, x)$  of

$$\phi = \frac{1}{\sqrt{t}} e^{-x^2/4t}.$$

This is the familiar impulse solution of the heat equation.

Note that the sum of the similarity solutions that correspond to two isovectors is not a similarity solution of their sum. It may not even satisfy the original equation unless it is linear.

The classical similarity solutions are found by dimensional analysis. These correspond to isovectors of the form

$$S = At \frac{\partial}{\partial t} + Bx \frac{\partial}{\partial x} + \dots$$

with  $A, B, \dots$  being constants. A systematic search for isovectors in this special form is very easy.



### Problems

- 39.1** What happened to the solution of the heat equation involving the error functions?
- 39.2** Show that the only isovector for the fluid-flow ideal of Section 38 that has the form

$$S = vt \frac{\partial}{\partial x} + U(x, t) \frac{\partial}{\partial u} + P(x, t) \frac{\partial}{\partial \rho}$$

is the Galilean symmetry given in the example.

### 40. Variational principles and conservation laws

It is a bit of physics lore that to every symmetry there corresponds a conservation law and vice versa. We saw earlier that every isovector generates a similarity solution. Sometimes an isovector leads to one or more conservation laws, but not always. A conservation law does always lead to a potential, however.

#### Conservation laws

A conservation law for an exterior differential system is a differential form whose restriction to the integral manifolds is closed. To be non-trivial it should not be identically closed. Any closed generator of the ideal leads to a conservation law.

**Example:** The heat-equation ideal,

$$\omega = d\phi dx + du dt,$$

$$\alpha = u dx dt - d\phi dt,$$

has  $\omega$  closed. The 1-form

$$j = S \lrcorner \omega$$

satisfies

$$dj = \mathcal{L}_S \omega - S \lrcorner d\omega$$

and, because  $d\omega$  pulls back to zero, this vanishes for every isovector  $S$ . For some closed generators this leads to trivial conservation laws,

$$\frac{\partial}{\partial t} \lrcorner (d\phi dx - du dt) = du,$$

or well-known ones,

$$\left( \phi \frac{\partial}{\partial \phi} + u \frac{\partial}{\partial u} \right) \lrcorner \omega = \phi dx - u dt,$$

here the conservation of heat. Some surprising conservation laws can also appear. The isovector called  $S_5$  in the preceding section leads to a conserved current,

$$j = 2t du - u\phi dx + (\phi + ux) dt.$$

Subtracting the exact form  $2d(ut)$ , we find an equivalent current

$$j = -u\phi dx + (\phi + ux - 2u) dt.$$

Some “stuff” with density  $(-u\phi)$  is conserved, and can be assigned a flux  $(\phi + ux - 2u)$ . I know of no use or physical interpretation for this. It could be used to check on a numerical integration.

### Variational principles

We now take the view that the ideal is the given, and ask when can it be derived from a variational principle. That is, when does a differential form  $\zeta$  exist such that

$$v \lrcorner d\zeta \in \mathcal{I}$$

for all vector fields  $v$ ? The form  $\zeta$  is called a *Cartan form*. It should generate in the preceding fashion all of the ideal, except, perhaps, low-rank forms, such as contact 1-forms. Even these turn up as common factors of the preceding forms. We will here study only complete variational principles where all of the ideal can be found from the Cartan form. We call such a form a *variational principle*, because the integral of  $\zeta$  over any piece of an integral manifold is unchanged by any infinitesimal deformation of the piece that keeps the boundary fixed. The Cartan form is the integrand of the variational problem in Lagrange-multiplier form.

The Cartan form  $\zeta$  is unique only up to the addition of any exact form. This complicates the discussion of symmetries and their associated conservation laws. The definition of a symmetry used in Noether’s Theorem

$$\mathcal{L}_S \zeta = 0$$

is not unchanged if we add an exact form to  $\zeta$ . Suppose we tried to define a symmetry of the variational problem as any  $S$  for which there exists a  $\phi$  such that

$$\mathcal{L}_S(\zeta + d\phi) \in \mathcal{I}.$$

We can show that for every vector field  $S$ , we can find a  $\phi$  to satisfy this equation. Just pick  $\phi$  to satisfy

$$\mathfrak{L}_S \phi = -S \lrcorner \zeta.$$

There are no obstructions to solving this equation, as was shown in Section 26. For the “symmetries” found this way, the current

$$j = S \lrcorner (\zeta + d\phi)$$

is identically conserved and not interesting.

A nontrivial conservation law comes from a vector field  $S$  that satisfies

$$\mathfrak{L}_S d\zeta = 0. \quad (40.1)$$

All Noetherian symmetries are included here, and it is clearly independent of any exact form added to  $\zeta$ . We have

$$d(S \lrcorner d\zeta) = 0,$$

and hence can find a  $j$  such that

$$dj = S \lrcorner d\zeta \in \mathfrak{G}.$$

This  $j$  is a nontrivial conservation law, provided  $S \lrcorner d\zeta$  is not zero.

**Example:** The Lagrangian dynamics of a particle with one degree of freedom is described by the ideal

$$dq - v dt,$$

$$dL_v - L_q dt,$$

where  $L = L(q, v)$  is the usual Lagrangian function.

A Cartan form for this system is

$$\zeta = L dt + L_v (dq - v dt).$$

For a free particle

$$L = \frac{1}{2} v^2,$$

$$\zeta = v dq - \frac{1}{2} v^2 dt,$$

$$d\zeta = dv dq - v dv dt;$$

contracting in  $\partial/\partial q$  gives us  $dv$ , and contracting in  $\partial/\partial v$  gives us  $dq - v dt$ . This Cartan form reproduces the whole ideal.

The vector field  $\partial/\partial q$  is even a Noetherian symmetry:

$$\mathfrak{L}_{\partial/\partial q} \zeta = 0,$$

and the conserved current here can be written down explicitly as

$$j = \frac{\partial}{\partial q} \lrcorner \zeta = v.$$

For a complete variational principle, every symmetry of the Cartan form, equation 40.1, is an isovector. Every generator can be written in the form  $v \lrcorner d\zeta$  for some  $v$ , and we have

$$\mathfrak{L}_S v \lrcorner d\zeta = [S, v] \lrcorner d\zeta.$$

The converse is not true in general.

**Example:** A vector field  $S$  that satisfies

$$\mathfrak{L}_S d\zeta = \lambda d\zeta$$

is not a symmetry of the Cartan form, but is an isovector, because

$$\mathfrak{L}_S v \lrcorner d\zeta = [S, v] \lrcorner d\zeta + \lambda v \lrcorner d\zeta,$$

assuming a complete variational principle.

### Potentials

Whenever we have a conservation law for an exterior differential system, a form  $j$  with  $dj$  in the ideal, then we can automatically introduce potentials. We add new coordinates to the space to represent the potential,  $y$ , and relate it to the other variables by adding to the ideal another generator

$$dy - j.$$

Since  $dj$  is in the original ideal, the new ideal is still closed under exterior differentiation.

**Example:** We saw in Section 38 that electrodynamics can be represented by the ideal

$$\begin{aligned} dB\theta + dE dq dt, \\ dE\theta - dB dq dt - 4\pi J\theta dt - 4\pi\rho\theta. \end{aligned}$$

One conservation law for this ideal is the 2-form

$$j = B\theta + E dq dt.$$

This conservation law is what is usually called Faraday's law of induction. The potential associated with this law is a 1-form

$A dq - V dt$ . We enlarge the space, introduce four new coordinates  $A$  and  $V$ , and add another generator

$$-V dt + dA dq - B\theta - E dq dt$$

to the ideal. These are the usual vector and scalar potentials.

It is also possible to introduce potentials in the form

$$dy - j$$

with  $dj$  not in the original ideal but in the new, larger ideal. These are called pseudopotentials. Unfortunately, no systematic method for finding them is known.

### Problem

**40.1** What electrodynamic conservation law comes from the field equation with source terms?

### 41. When not to use forms

It is time to correct the impression I may have given that differential forms are the solution to all mathematical problems. I tend to be overzealous. The formalism of differential forms and the exterior calculus is a highly structured language. This structure is both a strength and a limitation. In this language there are things we cannot say. This leads to an advantage: what can be said usually makes sense (i.e., is well-defined and coordinate-independent). This structure is also a disadvantage. Sometimes a calculation or a proof involves steps in which terms are split into pieces that are separately not coordinate-independent. Such calculations are either impossible in differential-forms language, or else involve arduous circumlocutions. I must admit that in several places in this book I first had to work things out in “old tensor.”

The limitation of most relevance to physics involves metrics. The Hodge star operator allows the algebraic entry of the metric into exterior calculus. Combined with the Lie derivative, this leads to a natural and efficient discussion of isometries. What does not fit in so naturally is the notion of a connection, such as a covariant derivative based on a metric. Even then, the situation is not hopeless, as we will see when we come to the method of moving frames in Section 66.

Such problems arise in the study of conservation laws. Some conserved objects are intrinsically geometric: flux lines, fly specks, etc. Yet some

conserved objects are not: momentum, for example. Momentum is vector-like stuff. Putting momentum into a box and sending it somewhere is one realization of the mathematical structure of a connection on the tangent bundle. We cannot usually describe momentum balance naturally and efficiently by using differential forms. Electromagnetism appears to be the outstanding exception to this.

As a concrete example of the difficulty that the exterior calculus has with momentum conservation, let us consider the hypothetical case of a scalar field and its associated stress-momentum-energy tensor.

If we consider the field by itself, and ignore the complications of sources, we have a field equation

$$d * d\phi = 0,$$

which falls naturally into the language of differential forms. The bind comes in trying to construct a stress-momentum-energy tensor for the field. This tensor is a linear map from Killing vectors to conserved currents.

If we use the first derivative of  $\phi$  and the metric tensor, there are only two possible terms quadratic in  $\phi$ :

$$\phi_{,\mu} \phi_{,\nu} \quad \text{and} \quad \phi_{,\alpha} \phi_{,\beta} g^{\alpha\beta} g_{\mu\nu}.$$

The combination

$$T_{\mu\nu} = \phi_{,\mu} \phi_{,\nu} - \frac{1}{2} \phi_{,\alpha} \phi_{,\beta} g^{\alpha\beta} g_{\mu\nu}$$

has zero divergence as a consequence of the field equations. The calculation is easily done if we use covariant derivatives:

$$T_{\mu\nu};\sigma g^{\nu\sigma} = [\phi_{;\mu\sigma} \phi_{,\nu} + \phi_{,\mu} \phi_{;\nu\sigma} - \phi_{;\alpha\sigma} \phi_{,\beta} g^{\alpha\beta} g_{\mu\nu}] g^{\nu\sigma}.$$

Since the covariant derivative of  $g_{\mu\nu}$  vanishes, and two covariant derivatives of a scalar field commute, we have

$$T_{\mu\nu};\sigma g^{\nu\sigma} = 0,$$

and from this in general we find conserved currents  $T_{\mu\nu} k^\nu$  for every Killing vector. Well, we don't yet know about covariant derivatives. They will only appear in Chapter VII. Can't we do this by using differential forms and the Hodge star operator?

Well, yes, we can, but it is surprisingly difficult and unnatural. The conserved current associated with the Killing vector  $k$  is

$$j = *[(k \lrcorner d\phi) d\phi - \frac{1}{2}(u \lrcorner d\phi)\alpha],$$

where

$$u = \#d\phi, \quad k = \#\alpha,$$

and the sharp operator denotes the use of the metric to map a 1-form into a tangent vector. Is it not routine to show that  $dj = 0$ ? Definitely not. Any serious student of differential forms should try to prove this for himself or herself before reading the following hints.

Start by expanding

$$\begin{aligned} dj &= (k \lrcorner \phi) d * d\phi + d(k \lrcorner d\phi) \wedge * d\phi \\ &\quad - \frac{1}{2} d(u \lrcorner d\phi) \wedge (k \lrcorner *1) - \frac{1}{2} (u \lrcorner d\phi) d(k \lrcorner *1), \end{aligned}$$

where

$$*\alpha = *(1 \wedge \alpha) = k \lrcorner *1,$$

was used. The first term drops out because of the field equations. Now use equations 23.5, 18.1, and the Leibniz rule for Lie differentiation, and in a dozen lines the last three terms will cancel.

Only if you try to complete the proof (good practice with forms) will you appreciate how clumsy it is. For a few days I was even convinced that a proof did not exist within the set of relations that I had given for forms. The moral: Use the right tool for the job, even if it is encrusted with indices.

# VI

## *Classical electrodynamics*

The test of a mathematical formalism is shown in the applications. Although I have long been convinced of this, it was emphasized to me again when I decided to teach a graduate electrodynamics course using differential forms instead of the usual vector notation. I expected only modest gains, but in fact it made a tremendous improvement. The mathematics became “transparent” and the underlying physical structures became visible. The clarity and economy of the differential-forms language for electrodynamics led me to include a separate chapter on it.

**Example:** Electrostatics in 2-space using differential forms is pictured in Figure VI.1, showing a potential difference of +4 between two conductors. The Hodge star construction of Section 24 is being used here.

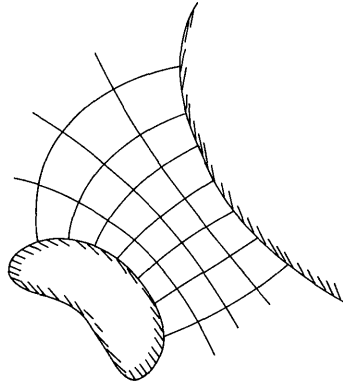
Nor was it all to the advantage of the physics. Actually *using* the formalism led me to an important theorem for the star operator, and to a definition of the star operator that is more direct and useful than the conventional one. It also led to the characterization of Killing vectors in the beautiful and useful expression

$$\mathfrak{L}_k * = * \mathfrak{L}_k.$$

It will be a while before anyone can write a standard textbook on electrodynamics using differential forms. I hope that this chapter shows you that it is inevitable. The exterior calculus is to multivariate calculus what del, cross, and dot are to three-dimensional Euclidean calculus.

Every classical field theory must have a force law. Electrodynamics is a very special classical field theory. Force is geometrically a 1-form. Think





**Figure VI.1.** The pattern of electrostatics, showing field lines and equipotentials.

Similar to Figure 24.6.

of it either as the rate of change of momentum or as the operator taking displacements into energy changes. In order to keep the rest masses of charges unchanged, the 4-force  $f$  must satisfy

$$u \lrcorner f = 0,$$

where  $u$  is the 4-velocity of the charge. A natural way to ensure this is to use a field tensor  $F$ , which is a 2-form, and a force law

$$f = u \lrcorner F.$$

A 2-form in spacetime has six degrees of freedom, the three components of  $E$  and the three components of  $B$ . There is a natural match between electrodynamics and differential forms, and they do more for electrodynamics than, say, for elasticity or fluid mechanics.

Both the spacetime theory and the 3+1 space and time theory are useful, and both appear in this chapter. I will use here  $*$  for the 3-space Hodge star and  $\star$  for the Hodge star in spacetime. For  $d$  you will have to keep awake. Also, most of the differential forms used here are twisted. I will mention this when they are introduced, but I will not use any special notation for them. For charge density I will keep  $\rho$  a scalar, and write the charge-density twisted 3-form  $\rho\theta$ .

## 42. Electrodynamics and differential forms

Undoubtedly the most useful electrodynamic formalism lives in 3-space, with time treated as a parameter. This 3+1 splitting reflects the behavior

of physical systems, and also the structure of the equations themselves, for which the initial-value problem is of particular importance. The section after this one will develop the spacetime formalism and discuss Lorentz invariance.

Before we develop the geometric aspects of electrodynamics, we must clear up the always troublesome question of units. It is not a problem that we can lightly push aside; I intend to use different types of geometric objects for the various electric and magnetic fields: twisted differential forms for  $D$  and  $H$ , and ordinary differential forms for  $E$  and  $B$ . Thus I need a notation that keeps them distinct. Gaussian units fit the bill except for their cumbersome dual units, esu and emu. This problem disappears, however, if you set  $c=1$ ; then no distinction is needed. On the other hand, nearly all textbooks today use SI units. I will use here unrationalized  $c=1$  Gaussian units. Use equations from Jackson (1962) by setting  $c=1$ ; use SI like Panofsky and Phillips (1955) by setting  $4\pi\epsilon_0 = \mu_0/4\pi = 1$ . Maxwell's equations in these units are

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E},$$

$$\frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{B} - 4\pi \mathbf{J},$$

$$\nabla \cdot \mathbf{B} = 0, \tag{42.1}$$

$$\nabla \cdot \mathbf{E} = 4\pi \rho, \tag{42.2}$$

and

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}).$$

The geometric unit of length is the light-second,

$$1 \text{ sec} = 2.998 \times 10^8 \text{ meters}$$

(light travels 1 sec in 1 sec). The geometric unit of mass is

$$1 \text{ sec} = 4.04 \times 10^{35} \text{ kg}$$

(at a distance of 1 sec from a mass of 1 sec the gravitational acceleration will be  $1 \text{ sec}^{-1}$ ) and that of force is

$$1 = 1.211 \times 10^{44} \text{ newtons}$$

(unit force gives a mass of 1 sec an acceleration of  $1 \text{ sec}^{-1}$ ). We add electromagnetic units of charge

$$1 \text{ sec} = 3.480 \times 10^{25} \text{ coulombs}$$

(two unit charges 1 sec apart have unit electrical force between them), current density

$$1 \text{ sec}^{-2} = 3.872 \times 10^8 \text{ amp/meter}^2,$$

electric field

$$1 \text{ sec}^{-1} = 3.5 \times 10^{18} \text{ volts/meter},$$

(unit charge in unit electric field feels unit force), and magnetic field

$$1 \text{ sec}^{-1} = 1.16 \times 10^{10} \text{ tesla}.$$

### Differential forms

We start in Cartesian coordinates. Once we have translated electrodynamics into differential forms, we will have a theory that can easily be used in curvilinear coordinates. Consider first the electric field. From the force law we see that the electric field gives directly the force on a particle. Since force is geometrically a 1-form, we define the electric field 1-form

$$E = E_x dx + E_y dy + E_z dz = E dq,$$

using the ordinary Cartesian components of the electric field.

The dynamical equation for the magnetic field requires the curl of the electric field. If we look at the exterior derivative of  $E$ , we find

$$\begin{aligned} dE &= \left( \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right) dx dy \\ &\quad + \left( \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \right) dy dz + \left( \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} \right) dz dx. \end{aligned}$$

The operator  $d$  here is the 3-space exterior derivative. Time is just a parameter. If we introduce the magnetic field 2-form

$$B = B_z dx dy + B_x dy dz + B_y dz dx = B \theta,$$

then the dynamical equation for  $B$  becomes

$$\frac{\partial B}{\partial t} = -dE.$$

What about  $dB$ ? We calculate

$$dB = \left( \frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} \right) dx dy dz$$

and so equation 42.1 for  $B$  can be written

$$dB = 0.$$

How will we get the curl of  $B$  and the divergence of  $E$ ? A divergence comes from an exterior derivative acting on a 2-form; so we use the star operator,

$$*E = E_x dy dz + E_y dz dx + E_z dx dy = E\theta,$$

and then Poisson's equation becomes

$$d*B = 4\pi\rho\theta,$$

provided we introduce the charge density 3-form

$$\rho dx dy dz = \rho\theta.$$

Here  $\rho$  is the usual scalar charge density, and  $\theta$  is the 3-space volume element. Similarly, we use

$$*B = B_x dx + B_y dy + B_z dz = B dq$$

and the current 2-form

$$J = J_x dy dz + J_y dz dx + J_z dx dy = J\theta.$$

Thus we have Maxwell's equation in vacuum:

$$\frac{\partial B}{\partial t} = -dE, \quad (42.3)$$

$$\frac{\partial *E}{\partial t} = d*B - 4\pi J, \quad (42.4)$$

$$dB = 0, \quad d*B = 4\pi\rho\theta.$$

In all of these,  $d$  is the 3-space exterior derivative. The force law translates into

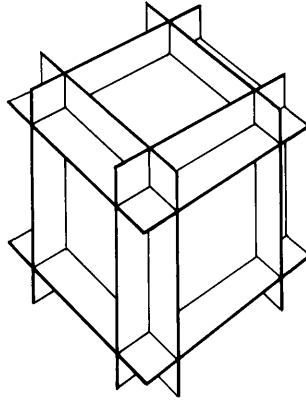
$$F = q(E - v \lrcorner B),$$

where  $v$ , the 3-velocity, is a tangent vector. When we consider materials in Section 45, we will introduce a  $D$  field for  $*E$  and an  $H$  field for  $*B$ .

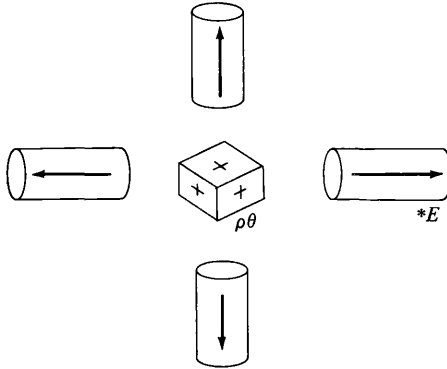
### Geometric structure

The preceding translation was motivated by the algebraic structure of the equations. We could have used instead the geometric structure, most apparent in the integral form of the equations.

Charge density is naturally a twisted 3-form. In the egg-crate picture, Figure 42.1, the 3-form basic cell contains exactly one unit of charge. The zero 3-form is represented by an infinite box, and that is the correct description of the size of box needed when the charge density goes to zero.

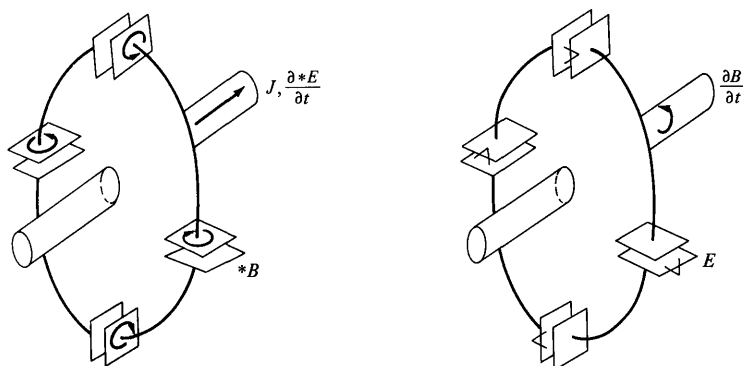


**Figure 42.1.** Charge density represented by a 3-form; the cell encloses a unit amount of charge in the limit.



**Figure 42.2.** The relation between electrostatic flux tubes described by the twisted 2-form  $*E$  and their source in the twisted 3-form representing charge density.

The integral form of equation 42.2 relates a surface integral of  $*E$  to the volume integral of  $\rho\theta$ . Only a 3-form can be integrated over a volume, and only a 2-form over an area. Thus we need to represent the electric field here as a twisted 2-form  $*E$ . The pictorial representation of  $*E$  is a flux tube oriented to point away from positive charges, as sketched in Figure 42.2. In regions where there is no charge, flux is conserved and the flux tubes are continuous.



**Figure 42.3** (left). The relation between the magnetic intensity twisted 1-form  $*B$  and its source in the twisted 2-form representing current density.

**Figure 42.4** (right). The relation between the electric field intensity 1-form  $E$  and its source in a changing magnetic flux represented by the 2-form  $(\partial B/\partial t)$ .

Currents are also represented by twisted 2-forms. The relation between currents (and displacement current) and the magnetic-field twisted 1-form is shown in Figure 42.3. Note how naturally the signs go, with no need for awkward right-hand rules.

Finally, we sketch the law of induction in Figure 42.4. This relates the  $B$ -field 2-form to the electric-field 1-form.

### *The initial-value problem*

Equations 42.3 and 42.4 provide a complete description of the time evolution of the field quantities. For consistency, the equations

$$\begin{aligned} dB &= 0, \\ d*E &= 4\pi\rho\theta, \end{aligned}$$

must be preserved by this dynamics. To check this we compute their time derivatives. First,

$$\frac{\partial}{\partial t}(dB) = d\left(\frac{\partial B}{\partial t}\right) = -dE = 0.$$

Thus if the initial  $B$  field satisfies  $dB = 0$ , then the dynamics preserves this condition. Next we compute

$$\frac{\partial}{\partial t}(d*E - 4\pi\rho\theta) = d(d*E - 4\pi J) - 4\pi \frac{\partial \rho}{\partial t} \theta,$$

and this derivative vanishes only if the sources satisfy

$$\frac{\partial \rho}{\partial t} + dJ = 0.$$

An equation arising this way is called an integrability condition. The equation here expresses the conservation of charge. If you are asked to solve Maxwell's equations for sources that do not satisfy charge conservation, you should refuse. They will have no consistent solution.

### *Curvilinear coordinates*

The operations  $d$ , wedge, and angle can be performed in any coordinate system whatsoever. Only star is at all tricky. To find the star operator in curvilinear coordinates, pick an orthonormal basis of 1-forms. Star acts on this orthonormal basis just as it does on  $dx$ ,  $dy$ , and  $dz$ . Note that the curvilinear coordinates create only an algebraic complication. The exterior differentiations are all straightforward.

**Example:** Let us write out Poisson's equation in spherical polar coordinates. That is, write out

$$d * dV = -4\pi\rho\theta$$

for a space with metric tensor

$$\mathcal{G} = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2.$$

Picking an orthonormal basis (this is not unique, of course) is easily done by just factoring the metric. The 1-forms  $dr$ ,  $r d\theta$ , and  $r \sin \theta d\phi$  clearly form an orthonormal basis. We have

$$dV = \frac{\partial V}{\partial r} dr + \frac{\partial V}{\partial \theta} d\theta + \frac{\partial V}{\partial \phi} d\phi$$

and in terms of our orthonormal basis we have

$$dV = \frac{\partial V}{\partial r} dr + \frac{1}{r} \frac{\partial V}{\partial \theta} r d\theta + \frac{1}{r \sin \theta} \frac{\partial V}{\partial \phi} r \sin \theta d\phi,$$

and so

$$*dV = \frac{\partial V}{\partial r} r^2 \sin \theta d\theta d\phi + \frac{1}{r} \frac{\partial V}{\partial \theta} r \sin \theta d\phi dr + \frac{1}{r \sin \theta} \frac{\partial V}{\partial \phi} r dr d\theta.$$

From this

$$*dV = r^2 \sin \theta \frac{\partial V}{\partial r} d\theta d\phi + \sin \theta \frac{\partial V}{\partial \theta} d\phi dr + \frac{1}{\sin \theta} \frac{\partial V}{\partial \phi} dr d\theta$$

and so

$$d*dV = \left[ \frac{\partial}{\partial r} \left( r^2 \sin \theta \frac{\partial V}{\partial r} \right) + \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial V}{\partial \theta} \right) + \frac{\partial}{\partial \phi} \left( \frac{1}{\sin \theta} \frac{\partial V}{\partial \phi} \right) \right] dr d\theta d\phi,$$

from which Poisson's equation can be read off.

Note the mechanical nature of this example. Although dull, such routine calculational algorithms are a safe way to get the correct answers. A calculation using the rotations of a field of unit vectors is pretty tricky, and unlikely to lead to the correct answer without careful debugging.

### *Junction conditions*

In addition to considering a charge continuously distributed throughout a volume, electrodynamics considers charge and current confined to a surface of negligible thickness. Because electrodynamics is a linear theory, we can find a description of these that ignores the internal details of the surface. We replace Maxwell's differential equations with algebraic conditions on how solutions on the two sides of the surface are to be joined together. These conditions have a natural description in terms of the pullback of ordinary and twisted differential forms.

I will use square brackets here to denote the operator that computes the discontinuity in a twisted differential form across a surface:  $[F]$  is the sum of the pullbacks from either side, using the appropriate vector pointing away from the surface on each side. Note that we do not need to orient the surface nor do we need to choose an arbitrary normal direction.

Look at Gauss's law

$$\int_{\partial\Gamma} *E = \int_{\Gamma} 4\pi\rho\theta$$

for a region  $\Gamma$  that is a tuna-fish can of negligible height, shown in Figure 42.5. For a surface distribution of charge  $\lambda$ , the right-hand integral is equivalent to a surface integral

$$4\pi \int_{\Gamma} \rho\theta = 4\pi \int_{\Sigma} \lambda,$$



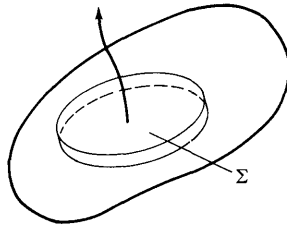


Figure 42.5. The geometry for the junction condition on  $*E$ .

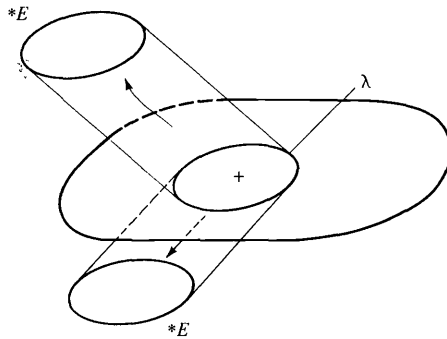


Figure 42.6. The junction condition on the electric flux generated by a surface charge density.

where  $\Sigma$  is the part of the surface inside the tuna-fish can. Here  $\lambda$  is a twisted 2-form; its egg crate encloses unit charge, and it has a plus sign for positive charge.

Now, note that the vector *away* from  $\Sigma$  is an outward normal vector for  $\partial\Gamma$  on both sides. Thus, letting the can shrink in all dimensions, we find

$$[*E] = 4\pi\lambda$$

for junction conditions across any surface  $\Sigma$  with surface charge density  $\lambda$ . See Figure 42.6 for this.

A similar argument can be applied to the equation

$$\frac{\partial *E}{\partial t} = d*B - r\pi J.$$

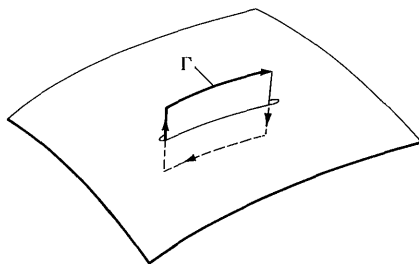
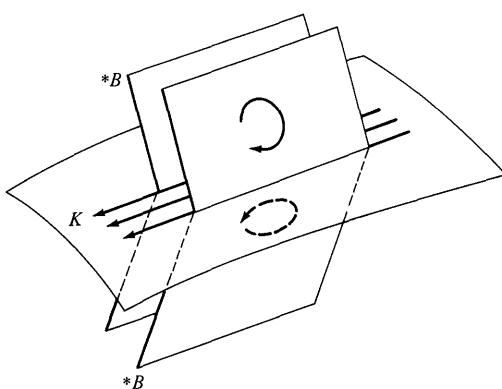


Figure 42.7.



**Figure 42.8.** The junction condition on magnetic intensity generated by a surface current density. Note that the pullbacks of the fields on the two sides are equal.

For fields that are finite but possibly discontinuous, we must have

$$\int_{\partial\Gamma} *B = 4\pi \int_{\Gamma} J$$

for  $\Gamma$  (shown in Figure 42.7). Again, we can pull back the twisted forms from the two sides to find

$$[*B] = 4\pi K,$$

where  $K$  is a twisted 1-form on the surface giving the distribution of the surface current. The result is shown in Figure 42.8.

The two homogeneous Maxwell equations involve the pullback of ordinary differential forms. Since the relations are homogeneous, again no

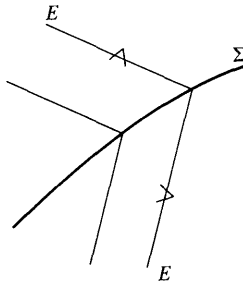


Figure 42.9. The junction condition on electric field intensity.

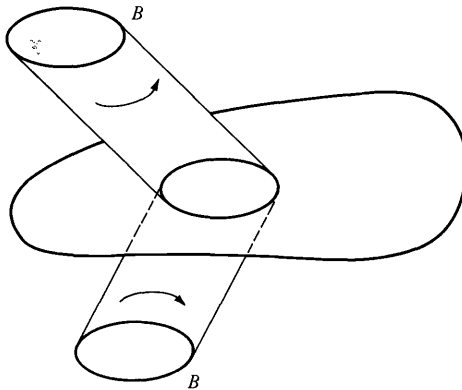


Figure 42.10. The junction condition on magnetic flux density.

arbitrary sign need be chosen. The pullback of both  $E$  and  $B$  into the surface from the two sides must be equal. Examples of this are shown in Figures 42.9 and 42.10. Note how important it is that the inhomogeneous equations involve twisted forms.

### 43. Electrodynamics in spacetime

The 3-space theory of the preceding section, although useful for the initial-value problem and many explicit problems, can be considerably simplified by a spacetime representation. In this section we use the exterior derivative of spacetime unless specifically noted otherwise; we retain the  $dq$ ,  $\theta$ , and  $\theta$  of 2-space. This hybrid notation is very useful. We have, then, for a scalar  $f$ ,

$$df = \frac{\partial f}{\partial \mathbf{q}} d\mathbf{q} + \frac{\partial f}{\partial t} dt.$$

This leads us to combine the homogeneous equations into the form

$$d_3 \mathbf{B}\theta + \frac{\partial \mathbf{B}}{\partial t} dt \theta + dt d_3 \mathbf{E} d\mathbf{q} = 0,$$

which collapses neatly into

$$d(\mathbf{B}\theta + \mathbf{E} d\mathbf{q} dt) = 0.$$

Thus we define the four-dimensional field 2-form

$$F = \mathbf{B}\theta + \mathbf{E} d\mathbf{q} dt = \mathbf{B} + \mathbf{E} dt,$$

satisfying

$$dF = 0.$$

Repeat this trick on the inhomogeneous equation

$$d_3 \mathbf{E}\theta + \frac{\partial \mathbf{E}}{\partial t} dt \theta - dt d_3 \mathbf{B} d\mathbf{q} = -4\pi \mathbf{J}\theta dt + 4\pi \rho \theta,$$

and it collapses to

$$d(\mathbf{E}\theta - \mathbf{B} d\mathbf{q} dt) = 4\pi(\rho \theta - \mathbf{J}\theta dt).$$

The 4-space dual of  $F$  is easily found from the results in Section 24,

$$\star F = \mathbf{E}\theta - \mathbf{B} d\mathbf{q} dt = \star \mathbf{E} - \star \mathbf{B} dt,$$

and using a spacetime current 3-form

$$j = \rho \theta - \mathbf{J}\theta dt = \star \rho - \mathbf{J} dt$$

we have

$$d\star F = 4\pi j. \tag{43.1}$$

Current conservation is just

$$dj = 0.$$

Some of the symmetries of electrodynamics are now obvious: the symmetries of  $\star$ .

### *Potentials*

Whenever we have a conservation law, we can reduce the problem by introducing potentials, using the Poincaré Lemma of Section 30. Here we have

$$d(\mathbf{B}\theta + E dq dt) = 0,$$

and hence we know that we can find a 1-form, which we write as

$$\phi = A dq - V dt,$$

satisfying

$$d(A dq - V dt) = \mathbf{B}\theta + E dq dt.$$

The spatial part gives us

$$\mathbf{B}\theta = dA dq,$$

and the spacetime part gives

$$E dq dt = -dV dt - \frac{\partial A}{\partial t} dq dt.$$

You can't factor out the  $dt$  unless you change  $dV$  to the 3-space  $d$  operating on  $V$ .

Now, this vector potential is not unique, and any closed form can be added to it without changing  $E$  or  $B$ . We can find a closed form  $\psi$  such that

$$d\star(A dq - V dt - d\psi) = 0,$$

since this just requires solving

$$d\star d\psi = d\star(A dq - V dt),$$

which is the ordinary wave equation (Section 24). This condition is the familiar Lorentz gauge condition.

From equation 43.1 we have

$$\star d\star d\phi = 4\pi\star j.$$

Now, recall that the vector wave equation needs an additional term,  $d\star d\star$ . This can be added here, because in Lorentz gauges it vanishes. Thus in Lorentz gauges the potentials satisfy the wave equation.

Well, we can go on. The vector potential itself is not free, but satisfies in Lorentz gauges

$$d\star\phi = 0$$

and we can introduce a potential for this as well. This route leads to the Hertz potentials.

### *Fields seen by a local observer*

An observer moving with spacetime velocity  $u$  sees a current density 2-form

$$J = -u \lrcorner j$$

and a charge density 3-form

$$\rho\theta = -u \lrcorner (\star j)\theta.$$

He sees an electric field

$$E = -u \lrcorner F$$

and a magnetic field

$$B = *(u \lrcorner \star F).$$

These relations are obvious in a frame at rest, and being tensors are true in general. From these we can find the usual result that a moving observer sees an electric field even when an observer at rest sees a pure magnetic field.

### Problem

**43.1** Study the Hertz potentials. Do they fit into the scheme of Section 40? How do we get wave equations for the Hertz potentials (Section 38)?

## 44. Laws of conservation and balance

The electromagnetic field deposits energy and momentum on charges and currents. If energy and momentum are to be conserved, then this energy and momentum must come from the electromagnetic field. Although this may seem like only a bookkeeping trick, it turns out that just as charge conservation is essential for the consistency of electrodynamics, so too energy and momentum conservation are essential for the consistency of general relativity.

### *Electrostatic body force*

We need here the continuum version of the Lorentz force law. Now, this cannot strictly be derived from the particle force law, nor is the reverse possible. There is always the possibility that a charge that is truly a point charge does not behave like the limit of a continuous charge distribution. We will here guess a plausible continuum force, called a *body force*, and show that in the limit it is consistent with the Lorentz force law. This body force is verified by experiments. There will be several unusual features of our development, all of them visible in the electrostatic case, and we do it first.

We will work in a 3+1 formalism, with  $*$  and  $d$  being the 3-space  $*$  and  $d$ . We expect a body-force operator to be one that takes in a small volume element and gives out a force 1-form. We are going to interpret the force 1-form a bit differently here. It will not be the operator that maps displacements into work done. Instead, it will map Killing vectors into the rate at which the corresponding momentum appears on the body. Viewing them in this way, we can treat all ten conserved quantities uniformly.

We will take the  $x$  component of the force on the charge in a unit volume to be

$$f = \left( \frac{\partial}{\partial x} \lrcorner E \right) \rho \theta;$$

in terms of how the Killing vector  $k = \partial/\partial x$  generates  $x$ -momentum, this is

$$\lrcorner f_k = (k \lrcorner E) \rho \theta = (k \lrcorner \rho \theta) \wedge E. \quad (44.1)$$

The last equality follows from

$$\theta \wedge E = 0,$$

because a purely spatial 4-form must vanish. We will assume that this is true for all Killing vectors, not just those representing linear momentum. This is a definite physical generalization. Equation 44.1 lets us calculate both the forces and the torques on a static charge distribution. No work can be done on a *fixed* charge distribution, nor does the momentum of its center of mass change; so these four conserved quantities are trivially constant.

Let us recover the static Lorentz force law from this body force. We consider a charge distribution contained in a small volume, in the limit as the volume shrinks to zero. The net body force on the entire body  $B$  will be

$$\int_B f_k = \int_B (k \lrcorner \rho \theta) \wedge E.$$

The  $E$  field here has two parts, one due to the charges themselves and another due to charges external to the body. Since the electrostatic force is a central force, equal and opposite on pairs of charges, the self-force on the body should vanish. To verify this without multiple integrals over Green functions requires a useful bit of trickery. The body force is a spatial 3-form and hence necessarily exact. More important, however, it can be written as the exterior derivative of field quantities. We have, in fact,

$$(k \lrcorner \rho \theta) \wedge E = \frac{1}{8\pi} d[(k \lrcorner E) \wedge *E - (k \lrcorner *E) \wedge E] = d\kappa, \quad (44.2)$$

provided that the fields satisfy Maxwell's equations and that  $k$  is a Killing vector. It will be useful to work this out for any  $p$ -form  $E$  such that  $dE=0$  in a space of  $p+q$  dimension. The general identity is

$$d[(-)^{pq}(k \lrcorner E) \wedge *E - (k \lrcorner *E) \wedge E] = (2k \lrcorner (d*E)) \wedge E.$$

To derive this identity we bring the  $d$  inside, using

$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-)^p \alpha \wedge d\beta$$

and

$$\mathfrak{L}_k \alpha = d(k \lrcorner \alpha) + k \lrcorner d\alpha,$$

to arrive at

$$\begin{aligned} d[ ] &= (k \lrcorner d*E) \wedge E + (-)^q (d*E) \wedge (k \lrcorner E) \\ &\quad + (-)^{pq} \mathfrak{L}_k E \wedge *E - (\mathfrak{L}_k *E) \wedge E. \end{aligned}$$

The first two terms are equal, because

$$(d*E) \wedge E = 0.$$

Now we use the properties of  $*$ ,

$$\alpha \wedge * \beta = \beta \wedge * \alpha$$

for  $\alpha, \beta$  forms of the same rank, and of the Lie derivative with respect to a Killing vector,

$$\mathfrak{L}_k * = * \mathfrak{L}_k,$$

to eliminate the last term. We have then a balance law for  $k$ -momentum,

$$d\kappa = (k \lrcorner \rho \theta) \wedge E$$

with the flux  $\kappa$  given by equation 44.2.

Back to the total force on a body,

$$\int_B f_k = \int_B (k \lrcorner \rho \theta) \wedge (E_{\text{int}} + E_{\text{ext}}),$$

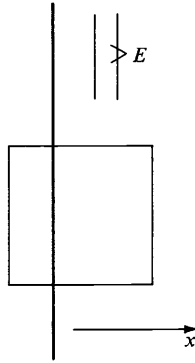
where  $E_{\text{int}}$  is the electric field due to the charge itself, and  $E_{\text{ext}}$  is the field due to all the other charges.

Write the first term, using our conservation law, as

$$\frac{1}{8\pi} \int_{\partial B} [(k \lrcorner E) \wedge *E - (k \lrcorner *E) \wedge E]$$

and since  $\rho=0$  outside the body, the surface  $\partial B$  can be pushed off to infinity. Since both  $E$  and  $*E$  die off like  $1/r^2$ , and  $k$  grows only as fast as  $r$  (for angular momentum), this surface integral must vanish.





**Figure 44.1.** The generation of an electric field by the surface charge on a conductor.

The body force comes only from the external field. To calculate this, write

$$\int_B f_k = \int_B (k \lrcorner E_{\text{ext}}) \rho \theta = (k \lrcorner E_{\text{ext}}) \int_B \rho \theta = Qk \lrcorner E_{\text{ext}}$$

where, as the body shrinks, both  $k$  and  $E_{\text{ext}}$  are essentially constant over the body.

Thus we have recovered the Lorentz force law, and in addition, picked up the balance law for  $k$ -momentum,

$$d\kappa = (k \lrcorner \rho \theta) \wedge E,$$

with

$$8\pi\kappa = (k \lrcorner E) \wedge *E - (k \lrcorner *E) \wedge E.$$

It is a physical fact that these continuum balance laws work. This law can be used to compute the total force on a body in terms of surface integrals surrounding it.

**Examples:** We compute the stress on the surface of a charged conductor. Here  $\rho$  is complicated and depends on the details of the surface of the conductor. Instead, we evaluate the rate at which linear momentum is deposited in this surface layer by using the conservation law to find the linear momentum flowing into a box surrounding the surface layer. If the surface is along the plane  $x = 0$ , and the surface charge density is given by  $\lambda$ , then we have an electric field

$$E = 4\pi\lambda dx,$$

$$*E = 4\pi\lambda dy dz,$$

as shown in Figure 44.1. The net momentum flux into the region is given by surface integrals over the six faces of the box, which we take to be a unit cube. The stress ( $x$ -momentum flux density) is

$$\kappa = 2\pi\lambda^2 dy dz,$$

and only integrals over the surfaces  $x = \text{constant}$  contribute. Of these one is inside the conductor where the field is zero. The other has an orientation  $dy dz$ , and so there is a stress on the surface of strength  $2\pi\lambda^2$ . On an area  $A$  of the surface,  $2\pi\lambda^2 A$  worth of outward momentum appears per unit time.

To see that a charge exerts the same force on a conductor as it would on its image charge, just surround the conductor with a surface and note that the same stress would be calculated for either situation.

#### *Spacetime conservation laws*

The equations of electrostatics in terms of differential forms are the same as those for electrodynamics in spacetime, with the correspondence  $F$  for  $E$ ,  $j$  for  $\rho\theta$ ,  $\star$  for  $*$ , and  $d$  for  $d_3$ . A spacetime body-force density

$$f_k = -(k \lrcorner j) \wedge F$$

completes the analogy. The balance of  $k$ -momentum

$$8\pi\kappa = (k \lrcorner \star F) \wedge F - (k \lrcorner F) \wedge \star F$$

is given by

$$d\kappa = -(k \lrcorner j) \wedge F.$$

Does this lead to the usual Lorentz force law? Well, yes and no. There are some fine points here. For an accelerating charge, the radiation fields drop off only like  $1/r$ , and this is not fast enough to remove self-force terms. Here is where a charge that is truly a point charge can behave differently from the continuum limit. Indeed, a controversy about runaway charges has gone on for years. [For my opinion, see Burke (1970).]

For a charge moving with uniform velocity, everything works out as before. Look, then, at a ball of charge *and* current moving with constant velocity  $v$  in our reference frame. We compute the  $k$ -momentum

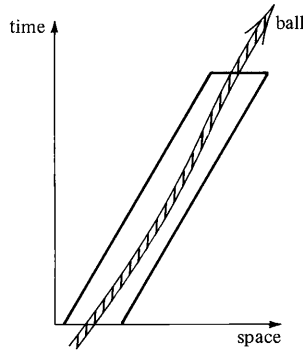


Figure 44.2.

deposited in unit time by integrating over a spacetime region of unit extent in time and over a spatial region following the ball, as shown in Figure 44.2. We have, first of all,

$$-\int (k \lrcorner j) \wedge F = + \int (k \lrcorner F) \wedge j$$

coming from

$$F \wedge j = 0,$$

because all 5-forms in spacetime vanish. Now, in the limit as the ball shrinks to zero, the 1-form  $(k \lrcorner F)$  is sensibly constant. Write it

$$(k \lrcorner F) = d\alpha.$$

Then we have, since  $dj = 0$ ,

$$\int_{\Gamma} (k \lrcorner j) \wedge F = \int_{\Gamma} d\alpha \wedge j = \int_{\Gamma} d(\alpha j) = \int_{\partial\Gamma} \alpha j.$$

Because  $j$  vanishes outside the ball, the only contributions to the surface integral come from the two  $t = \text{constant}$  ends.

$$\int_{\partial\Gamma} \alpha j = \int_{t=t_0+1} \alpha \rho \theta - \int_{t=t_0} \alpha \rho \theta = \Delta\alpha \int \rho \theta = \Delta\alpha Q.$$

The change in  $\alpha$  is given by

$$\begin{aligned}
\Delta\alpha &= \left(\frac{\partial}{\partial t} + v\right) \lrcorner d\alpha \\
&= \left(\frac{\partial}{\partial t} + v\right) \lrcorner k \lrcorner F \\
&= -k \lrcorner \left(\frac{\partial}{\partial t} + v\right) \lrcorner F \\
&= -k \lrcorner \left(\frac{\partial}{\partial t} + v\right) \lrcorner (B - dt \wedge E) \\
&= +k \lrcorner (E - v \lrcorner B),
\end{aligned}$$

and so the force is given by

$$Qk \lrcorner (E - v \lrcorner B).$$

#### *The Poynting–Heaviside Theorem*

Look at the balance law generated by  $k = \partial/\partial t$ . We have, in 3 + 1 language,

$$8\pi\kappa = *B \wedge B + *E \wedge E + 2 dt \wedge *B \wedge E.$$

Thus the energy density is

$$(*B \wedge B + *E \wedge E)/8\pi,$$

and the flux density is

$$(E \wedge *B)/4\pi.$$

The 3 + 1 version of the conservation law is

$$\frac{dt}{8\pi} \wedge \left[ \frac{\partial}{\partial t} (*B \wedge B + *E \wedge E) + 2 d(*B \wedge E) \right] = -dt \wedge J \wedge E.$$

We know that this conservation law is connected with energy, not because a  $\partial/\partial t$  symmetry always generates energy conservation, but because for this  $k$  we have

$$d\kappa = -dt \wedge J \wedge E,$$

and the right-hand side here is indeed the rate at which energy is appearing on the currents. Not every mathematical balance law represents a physically interesting quantity. Also, remember that this conservation law is unique only up to the addition of any closed 3-form.

Look also at the linear momentum-conservation law. The density  $\mu$  of  $k$ -momentum is given by the projection operator  $(\partial/\partial t) \lrcorner \star$  acting on the  $k$ -momentum 3-form

$$\mu = \frac{1}{8\pi} \frac{\partial}{\partial t} \lrcorner \star [(k \lrcorner \star F) \wedge F - (k \lrcorner F) \wedge \star F].$$

We expect that this will equal the energy flux in the  $k$ -direction, which is given by

$$\mu' = \frac{1}{8\pi} k \lrcorner \star \left[ \left( \frac{\partial}{\partial t} \lrcorner \star F \right) \wedge F - \left( \frac{\partial}{\partial t} \lrcorner F \right) \wedge \star F \right]. \quad (44.3)$$

To verify this symmetry, look at the first term in equation 44.3, and use a general vector  $v$  instead of  $\partial/\partial t$ :

$$k \lrcorner \star [(v \lrcorner \star F) \wedge F] = \star [(v \lrcorner \star F) \wedge F \wedge \phi],$$

where the 1-form  $\phi$  satisfies

$$k = \# \phi.$$

Here  $\#$  is the sharp operator defined in Section 24, and the metric is being used to convert a 1-form into a vector. Thus the first term of 44.3 equals

$$\star [\star (F \wedge v) \wedge (F \wedge \phi)].$$

Here  $v$  is similarly associated with  $\nu$ . Since  $F \wedge v$  and  $F \wedge \phi$  have the same rank, it also equals

$$\star [\star (F \wedge \phi) \wedge (F \wedge \nu)],$$

and now reversing all the steps demonstrates the symmetry of this first term. The second term is also symmetric by the same argument. The symmetry tells us that the momentum density is the same as the energy-flux density.

### Uniqueness

One use of the law of energy conservation is to show that Maxwell's equations have a unique solution. This is an important result, for it allows us to guess solutions and to verify them by merely checking Maxwell's equations. To prove uniqueness, we show that two solutions with the same charge and current distributions that agree on an initial surface and on the boundaries must be identical.

Look at the difference of two such solutions. It satisfies homogeneous Maxwell's equations and hence the energy-conservation law with

no sources. Its energy density on the initial surface is zero, because the two solutions agree there by assumption. Similarly, there is no energy flux into the region, because the two solutions agree on the boundaries, the difference field vanishes there, and the energy flux is bilinear in these fields. The total energy is thus always zero. Finally, since the energy is positive definite, the difference field must vanish everywhere. Of crucial importance here is the restriction to fixed currents and charges. Solutions are not unique if the currents, say, are allowed to vary, as in the dynamo theory of the Earth's magnetic field.

### Problem

**44.1** Where is the angular momentum in a circularly polarized plane electromagnetic wave? (See remarks at the end of Section 33.)

## 45. Macroscopic electrodynamics

So far we have set up a theory of electrodynamics that discusses the electric and magnetic fields of charges and currents in vacuum. This is called microscopic electrodynamics. We need to extend the theory to include some simple properties of materials. Much of the behavior is independent of the detailed structure of the matter. The detailed structure is the proper realm of solid-state physics or plasma physics. The universal behavior common to much matter in bulk is called macroscopic electrodynamics.

### Conductors

The simplest type of matter is the conductor, in which an electric field causes currents to flow according to Ohm's law,

$$J = \sigma * E,$$

which is true only in the rest frame of the conductor. This restriction to the rest frame is important, and a correct description of moving conductors will be given in the next section. As stated, Ohm's law neglects the dynamical aspects of the current carriers, and applies in the limit of low frequencies. A more general conductivity law would be

$$J(t) = \int_{-\infty}^t s(t-t') * E(t') dt',$$

corresponding to a frequency-dependent conductivity.

**Example:** Let us repeat the commonly given argument that shows an exponential decay for free charge inside a conductor. Starting from

$$\frac{\partial \rho}{\partial t} \theta + dJ = 0$$

and using Ohm's law

$$\frac{\partial \rho}{\partial t} \theta + \sigma d*E = 0,$$

we have

$$\frac{\partial \rho}{\partial t} + 4\pi\sigma\rho = 0.$$

This shows exponential decay of the charge density on a time scale of  $(4\pi\sigma)^{-1}$ .

For materials considered very good conductors, this time scale is, in fact, too short for Ohm's law to be valid. [See Jackson (1962), problem 7.7]

Inside a conductor, Maxwell's equations can be written

$$dF = 0, \quad d\star F = -4\pi\sigma\star(\lambda \lrcorner F),$$

as promised in Section 31.

### *Polarizable materials*

In addition to conductivity, matter can also respond to an applied electric field by a rearrangement of its charges rather than with a steady current flow. Such a body is called *polarizable*, and the resulting charge distribution is called *polarization charge* or *bound charge*. We will study the nearly static limit of this behavior. Here both  $*$  and  $d$  refer to 3-space. We will first show that for *any* rearrangement of the charge, the bound charge can be represented by a 2-form  $P$  such that

$$\rho\theta = -dP,$$

with the important property that  $P$  vanishes outside the body. For small displacements of the charge, a Taylor's Series expansion is useful, and the leading term  $P$  is then the induced dipole moment per unit volume.

To see that any deformation of a charge distribution of zero total charge can be written in terms of a polarization 2-form with  $P = 0$  outside the body, we proceed as follows. We have

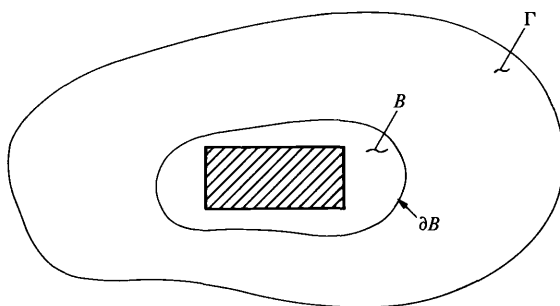


Figure 45.1.

$$d(\rho\theta) = 0$$

everywhere; hence we can write

$$\rho\theta = d\alpha$$

for some  $\alpha$  not necessarily vanishing outside the body. Outside the body we have

$$d\alpha = 0$$

and so

$$\alpha = d\beta$$

provided that we have

$$\int \alpha = 0$$

over every 2-cycle in the region  $\Gamma$  outside the body, as shown in Figure 45.1. The 2-cycle  $\partial B$  is the only one of interest, and we have

$$\int_{\partial B} \alpha = \int_B d\alpha = \int_B \rho\theta = 0,$$

because our requirement was that the total charge vanish. Thus we can write

$$\alpha = d\beta$$

outside the body. Extend  $\beta$  inside the body in any smooth fashion. Define the polarization



$$P = \alpha - d\beta,$$

which clearly satisfies our conditions. The spacetime proof proceeds in a similar fashion. Note the usefulness of cohomology for regions that are not simply connected.

### *D and H*

A useful trick in polarizable materials is to define a 2-form  $D$  field. We have

$$d*E = 4\pi\rho_F\theta - 4\pi dP,$$

where  $\rho_F$  is the charge not accounted for as a bound charge, called *free charge*. If we define

$$D = *E + 4\pi P,$$

then we have

$$dD = 4\pi\rho_F\theta.$$

In vacuum this agrees with our previous use of  $*E$ . The bound charge must move as the field changes, and this leads to a current as well. The bound-charge current, as well as other currents induced by magnetic fields, can be written

$$j_B = -d\gamma,$$

where the spacetime 2-form  $\gamma$  vanishes outside the body. As with conductivity, this applies only to matter at rest. Any 2-form in spacetime can be decomposed

$$\gamma = P + dt \wedge M,$$

where the 2-form  $P$  and the 1-form  $M$  are purely spatial. We have, then,

$$\rho_B = -dP, \quad J_B = dM + \frac{\partial P}{\partial t}.$$

The 1-form  $M$  describes possible persistent currents that can be set up in the material by a magnetic field, and it is called the magnetization.

Much as we did for  $D$ , we redefine  $H$  so that the equation

$$\frac{\partial}{\partial t}(*E) = d(*B) - 4\pi J$$

reads

$$\frac{\partial D}{\partial t} = d(*B) - 4\pi J_F - 4\pi dM,$$

and taking

$$H = *B - 4\pi M,$$

we have the macroscopic equations

$$\frac{\partial B}{\partial t} = -dE, \quad \frac{\partial D}{\partial t} = dH - 4\pi J_F,$$

$$dB = 0, \quad dD = 4\pi \rho_F,$$

whose sources are only the free charge and the free current.

These equations are useful only if we understand enough of the properties of matter to find relations between  $P$  and  $M$  and the applied fields  $E$  and  $B$ . A great many materials are modeled by the linear, isotropic relations

$$D = \epsilon * E, \quad B = \mu * H,$$

where  $\epsilon$  and  $\mu$  are scalar parameters characterizing the matter, just like  $\sigma$ . These equations hold only in an inertial frame in which the matter is at rest.

### *Smoothing*

The macroscopic fields ignore the detailed structure of matter and the complex fields inside the body. They represent only the smooth properties of the fields. For a cloud of electrons, the charge density is highly irregular. For a great many questions, this highly irregular charge density is equivalent to a smooth averaged charge density. These questions must involve a length scale much larger than that typical of the irregularities. A smoothed function will be an equivalence class of functions all of which agree when integrated with smooth integrands over macroscopic regions. This is very similar to delta functions.

A smoothed function is best represented by finding the equivalent smooth function. This process involves averaging the microscopically rough function over regions intermediate in size between microscopic and macroscopic regions. To see that Maxwell's equations still hold for the smooth functions representing classes of smoothed functions, look at just a two-point average. Now, the averaging is supposed to be done over a length scale small compared with the scale of the smooth functions, but large compared with the microscopic fields that we are averaging. Thus we cannot represent the changes in the fields by a Taylor's Series or a Lie derivative.

It is clear from the linearity of Maxwell's equations that if we average together two solutions, we will get another solution, provided that the points of the average are separated by a rigid displacement in spacetime. Smoothed functions cannot be multiplied, nor can delta functions. What, then, are we to do with the force law

$$(k \lrcorner j) \wedge F?$$

Well, if you look at a volume of intermediate size, then the  $F$  field from external sources will be smooth over that volume. The internal field will, however, not be smooth. The smoothed force law allows us to compute the average force (momentum deposition) on a volume, but leaves the internal stresses undetermined. A correct treatment of internal stresses requires a microscopic theory. Nor will this smoothed theory apply to radiation fields with wavelengths on the scale of atomic spacings, but we already know that classical electrodynamics is not the theory to use for that.

#### 46. Electrodynamics of moving bodies

The usual constitutive relations

$$\mathbf{J} = \sigma \mathbf{E}, \quad \mathbf{D} = \epsilon \mathbf{E}, \quad \mathbf{B} = \mu \mathbf{H}.$$

we write as

$$j = \sigma * E, \quad D = \epsilon * E, \quad B = \mu * H.$$

They are true only in the rest frame of the material. In our notation this is obvious, because  $\star$  is Lorentz-invariant, but  $*$  is not. In this section we will find the correct extension of these laws for moving bodies, assuming that the preceding relations remain true for small fluid elements in a frame in which they are momentarily at rest.

#### *Conductivity*

One approach to this problem would be to transform the fields from the lab frame to the fluid frame, use the standard laws, and then transform back. An equivalent but better approach is to write the basic laws in a form that makes no reference to the frame whatsoever, a manifestly Lorentz-invariant form.

In the fluid rest frame, we have

$$\frac{\partial}{\partial t} \lrcorner F = -E,$$

$$\star\left(\frac{\partial}{\partial t}\lrcorner F\right) = dt \wedge \star E,$$

$$\sigma\star\left(\frac{\partial}{\partial t}\lrcorner F\right) = dt \wedge J.$$

Unlike the situation with a solid conductor, some free charge may be convected along with the body; let us call its density  $\phi \star 1$ . The spacetime current 2-form is thus

$$j = -\sigma\star(u \lrcorner F) + \phi(u \lrcorner \star 1), \quad (46.1)$$

where  $u$  is the fluid 4-velocity. In the rest frame

$$u = \frac{\partial}{\partial t}.$$

The preceding expression for the spacetime current involves only Lorentz-invariant quantities and so can be used in any frame, not just the fluid frame.

The 3+1 splitting of this equation will be useful. In general we have

$$u = \gamma\left(\frac{\partial}{\partial t} + v\right)$$

where  $v$  is a purely spatial vector, and  $\gamma$  is the usual Lorentz factor. We have

$$u \lrcorner F = \gamma[-E + u \lrcorner B + dt(u \lrcorner E)]$$

and so

$$J = \sigma\gamma\star(E - u \lrcorner B) + \gamma\phi(u \lrcorner \star 1),$$

$$\rho = \gamma[\sigma(u \lrcorner E) + \phi]\theta.$$

Another condition needed for the electrodynamics of conducting bodies is that for finite conductivity there can be no surface currents in the rest frame of the material; otherwise the electric field would be infinite. There can be a delta-function surface charge density, and for a moving conductor this will be convected along with the conductor and lead to delta-function surface currents.

#### *Free charge in a conductor*

In equation 46.1 we allowed for the possibility of free charge convecting with the moving conductor. The dynamics of this free charge is specified by the condition  $dj = 0$ . This gives us

$$-d[\sigma\star(u \lrcorner F)] + d\phi \wedge (u \lrcorner \star 1) + \phi d(u \lrcorner \star 1) = 0. \quad (46.2)$$

The second term here can be written

$$(u \lrcorner \star d\phi) \star 1 = (\mathfrak{L}_u \phi) \star 1$$

and this is the convective derivative of the density of free charge. The third term here involves the divergence of the velocity field. For slow motions, with

$$u = \frac{\partial}{\partial t} + v,$$

we have

$$u \lrcorner \star 1 = \theta - dt \wedge (v \lrcorner \theta)$$

and so

$$\begin{aligned} d(u \lrcorner \star 1) &= dt \wedge d(v \lrcorner \theta) \\ &= dt \wedge \mathfrak{L}_v \theta \\ &= (\operatorname{div} v) \star 1 \end{aligned}$$

and this vanishes for solids and incompressible fluids.

The first term in equation 46.2 can be simplified if we use the 1-form  $\nu$  obtained from  $u$  by using the metric; the definition of star gives us

$$\star(u \lrcorner F) = -\star(F \wedge \nu),$$

using the fact that on 2-forms in four dimension  $\star\star = -1$ . Thus

$$d\star(u \lrcorner F) = 4\pi j \wedge \nu - \star F \wedge d\nu$$

and recycling Ohm's law gives us

$$d\star(u \lrcorner F) = 4\pi\sigma\star(u \lrcorner F) \wedge \nu - 4\pi\phi(u \lrcorner \star 1) \wedge \nu.$$

The first term here vanishes because

$$\star[\star(u \lrcorner F) \wedge \nu] = -u \lrcorner u \lrcorner F = 0,$$

since  $F$  is antisymmetric. For the second term here, we have

$$(u \lrcorner \star 1) \wedge \nu = -(u \lrcorner \nu) \star 1 = \star 1.$$

Remember that  $u$  is a timelike vector, with  $u \cdot u = u \lrcorner \nu = -1$ . Thus the dynamics of the free charge in a conductor is specified by

$$[\mathfrak{L}_u \phi + 4\pi\sigma\phi + (\operatorname{div} v)\phi] \star 1 + \sigma \star F \wedge d\nu = 0.$$

Inside a solid conductor this gives us

$$\mathcal{E}_u \phi + 4\pi\sigma\phi = 0$$

and the free charge decays in a time  $T = 1/4\pi\sigma$ .

### *Frozen flux*

Suppose we consider the problem of a conducting fluid moving in a magnetic field. Let  $L$  be a typical spatial dimension, and  $T$  be a typical time scale. One dimensionless number characterizing the problem is

$$\delta = L/T.$$

The conductivity is another parameter. From it a second dimensionless number can be formed,

$$Rm = \sigma L^2/T.$$

$Rm$  is called the magnetic Reynold's number. We will consider only slow motions,  $\delta \ll 1$ .

If the electric field is generated only by induction, it will be of size  $LB/T$ . With these sizes taken into account, we have Maxwell's equations

$$\begin{aligned} \frac{\partial B}{\partial \left(\frac{t}{T}\right)} &= -L d \left(\frac{ET}{L}\right), \\ \delta^2 \frac{\partial \left(*\frac{ET}{L}\right)}{\partial \left(\frac{t}{T}\right)} &= L d * B - 4\pi \left(\frac{\sigma L^2}{T}\right) \left(\frac{ET}{L} - \frac{vT}{L} \lrcorner B\right) \\ &\quad - 4\pi \delta^2 \phi T \left(\frac{vT}{L}\right) \lrcorner \theta. \end{aligned}$$

The sizes are now explicit, and our slow-motion approximation will ignore the  $\delta^2$  terms here, just as it ignored the  $\delta^2$  factors in Ohm's law. Note that the basis 1-forms are dimensionless, there being one length in  $x$  and an inverse length in the  $d$ . The size of the charge density came from Maxwell's equations. We see that we have a singular dynamical system, such as we studied in Section 25, since our small parameter multiplies one of the time derivatives. An initial situation will have a rapid flow of charge until the condition

$$d * B = 4\pi\sigma * (E - v \lrcorner B)$$

is established. The dynamics then evolves to preserve this condition and

$$\frac{\partial B}{\partial t} = -dE. \tag{46.3}$$

An initial unrelaxed state will snap to this state in a time  $\delta^2 T$ .

For a fluid with good conductivity,

$$1 \ll Rm \ll \frac{1}{\delta},$$

which in terms of the skin depth  $l$  is

$$l \ll L \ll T,$$

we must have

$$E = v \lrcorner B. \quad (46.4)$$

**Example:** This  $E$  field gives us a charge density

$$4\pi\rho\theta = d*(v \lrcorner B).$$

This does not necessarily vanish, and this distinguishes the conducting fluid in differential motion from the stationary conductor. For example, if

$$B = dx \, dy \quad \text{and} \quad v = y \frac{\partial}{\partial x},$$

representing a fluid with shear in a constant magnetic field, then we have a nonzero density of free charge.

We now repeat the Lie-derivative calculation of Section 18. Using equation 46.4 in equation 46.3 give us

$$\frac{\partial B}{\partial t} + d(v \lrcorner B) = 0,$$

and since  $dB = 0$  this can be written

$$\frac{\partial B}{\partial t} + \mathfrak{L}_v B = 0.$$

For steady motions the  $B$  field satisfies

$$\mathfrak{L}_v B = 0,$$

which is commonly said to imply that the magnetic field is frozen into the fluid. This statement requires caution in uniform magnetic fields, since this equation does allow a uniform field to slide through a conductor.

For time-dependent motions we have

$$\mathfrak{L}_{(\partial/\partial t + v)} B = 0,$$

and again the flux is frozen into the fluid.

*Polarization and magnetization*

The spacetime version of the macroscopic equations involves

$$F = B - dt \wedge E \quad \text{and} \quad G = D + dt \wedge H.$$

Maxwell's equations are

$$dF = 0, \quad dG = 4\pi j.$$

In the rest frame of the material, we have, using  $u = \partial/\partial t$ ,

$$u \lrcorner F = -E, \quad u \lrcorner \star F = \star B,$$

$$u \lrcorner G = H, \quad u \lrcorner \star G = \star D.$$

These allow us to write the constitutive equations for a linear, isotropic material in the invariant form

$$u \lrcorner \star G = -\epsilon u \lrcorner F, \quad u \lrcorner \star F = \mu u \lrcorner G.$$

Let us write out these equations in a frame where the fluid velocity is small. Then again we have

$$u = \frac{\partial}{\partial t} + v$$

and

$$u \lrcorner F = v \lrcorner B - E + dt(v \lrcorner E),$$

$$u \lrcorner \star F = v \lrcorner \star E + \star B - dt(v \lrcorner \star B),$$

$$u \lrcorner G = v \lrcorner D + H - dt(v \lrcorner H),$$

$$u \lrcorner \star G = -v \lrcorner \star H + \star D - dt(v \lrcorner \star D).$$

Thus our equations become

$$\star D - v \lrcorner \star H = \epsilon(E - v \lrcorner B),$$

$$(v \lrcorner \star D) = \epsilon(v \lrcorner E),$$

$$\star B + v \lrcorner \star E = \mu(H + v \lrcorner D),$$

$$(v \lrcorner B) = \mu(v \lrcorner \star H).$$

To solve these, contract them with  $v$  and discard terms quadratic in velocity, to find

$$\star D = \epsilon E + (1 - \epsilon\mu)v \lrcorner B,$$

$$H = \star B + (1 - \epsilon\mu)(v \lrcorner \star E).$$



These clearly have the correct limits for  $v = 0$  and for  $\epsilon = \mu = 1$  (moving nonexistent material).

### **Problem**

- 46.1** Repeat the discussion of the conducting fluid in a magnetic field, Section 18, using differential forms.

# VII

## *Dynamics of particles and fields*

Classical mechanics displays the advantages of differential forms as well as electrodynamics does. The development of mechanics that I give here is slightly unusual, in that it takes place in the special contact bundles called jet bundles rather than in the tangent and cotangent bundles. Only for time-independent systems are the tangent and cotangent bundles the proper ones. Contact bundles are not usually mentioned, perhaps because they lack natural coordinates. This lack forces us to use generators for the contact ideal that appear to single out the  $t$  direction. Once the lack of symmetry in the generators is understood, however, the contact bundle  $CM$  is seen to be much simpler than the tangent bundle  $TM$ . The geometric structure of both  $TM$  and  $CM$  consists in the fact that there are special curves in the bundles, and these, when projected down to the base space, have tangent vectors that are truly tangent or contact elements truly in contact. In  $CM$  such curves are characterized as integral submanifolds of the contact ideal.  $TM$  has no such simple characterization. In fact, the structure of  $TM$  is best described by adding one more dimension and making it into a contact bundle.

Of course, analytical dynamics has been able to deal with time-dependent Hamiltonians, but only by adding a spurious extra dimension, going to the cotangent bundle of spacetime. This extra dimension is in the way, and fouls up the geometry of kinetic theory, for example. The phase-space density in kinetic theory should be a 6-form, but that represents the density of lines only in 7-space, not in 8-space.

I will start with a careful development of mechanics from  $F=ma$  particle mechanics. This will lead naturally into the study of families of solutions and the Lagrangian and Hamiltonian formulations. I will also give a deductive picture of particle dynamics as a special case of wave-packet

dynamics in the high-frequency limit. This gives us a much needed physical picture of the  $ps$  in Hamilton's equations.

#### 47. Lagrangian mechanics of conservative systems

Let us start from Newtonian particle mechanics and see how far we can go. Start by considering  $n$  particles described by  $3n$  rectangular coordinates  $q$  moving under the influence of forces derivable from a potential  $V(q, t)$ . If  $m_i$  is the mass appropriate to the  $i$ th coordinates, we have

$$m_i \frac{d^2 q^i}{dt^2} = - \frac{\partial V}{\partial q^i} \quad (\text{no sum}),$$

with  $m_i$  constant over index triples. Our first task is to recast this into proper geometric language.

##### *Mechanics and differential forms*

The first step, always, is to get into the right space. The state of the mechanical system is best described by giving its position and velocity as a function of time. Thus we are naturally led to the line-element contact bundle

$$\pi: CQ \rightarrow Q: (t, q, \dot{q}) \mapsto (t, q)$$

with contact structure given by the ideal generated by the  $3n$  1-forms

$$dq - \dot{q} dt.$$

In fact, the  $t$  coordinate is special, and we are dealing with the special contact bundles called jet bundles.

The dynamics on this space is described by the vector field

$$\frac{\partial}{\partial t} + \dot{q} \frac{\partial}{\partial q} - \sum_i \frac{1}{m_i} \frac{\partial V}{\partial q^i} \frac{\partial}{\partial \dot{q}^i},$$

which is everywhere tangent to the path of the system through the contact bundle  $CQ$ . The length of this vector is irrelevant. I have chosen a convenient normalization.

The proper geometric structure is really a line element tangent to the preceding vector. This can be described by the ideal of differential forms that annul the line element. By inspection, these are the  $3n$  contact forms and also the  $3n$  1-forms

$$m_i d\dot{q}^i + \frac{\partial V}{\partial q^i} dt \quad (\text{no sum}),$$

both sets pulling back to zero on the given vector. There are  $3n$  forms in each set, and since this contact bundle is  $(6n+1)$ -dimensional, these determine the line element completely.

We can put this into better geometric shape by introducing the function on  $CQ$  called the Lagrangian,

$$L: (t, \mathbf{q}, \dot{\mathbf{q}}) \mapsto \sum_i \frac{1}{2} m_i \dot{q}^i \dot{q}^i - V(t, \mathbf{q}).$$

Now we can describe the mechanical system as the exterior differential system generated by

$$\alpha = d\mathbf{q} - \dot{\mathbf{q}} dt,$$

and

$$\beta = d\left(\frac{\partial L}{\partial \dot{\mathbf{q}}}\right) - \frac{\partial L}{\partial \mathbf{q}} dt. \quad (47.1)$$

We can now use the summation convention. Pragmatically, this calculation was driven by the urge to get the indices properly tucked away. The geometric view of this result is that a mechanical system moves along a path in  $CQ$  that is an integral curve of the ideal of 1-forms generated by  $\alpha$  and  $\beta$ .

As a bonus we recognize this ideal as that belonging to a variational problem on the contact bundle for the integral

$$\int L dt.$$

This interpretation is called Hamilton's principle. This makes it obvious that the preceding ideal depends only on  $L$  and the contact structure, not on the coordinates. Recall the 1-form appearing in the variational principle, called the Cartan form

$$\omega = L dt + \frac{\partial L}{\partial \dot{\mathbf{q}}} \alpha.$$

We have the useful identity

$$d\omega = \alpha \wedge \beta.$$

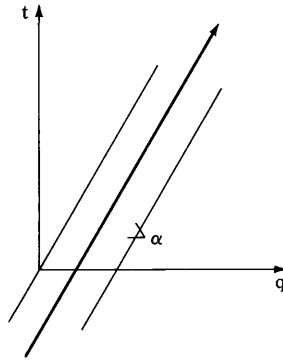
The conditions that

$$\dot{\gamma} \lrcorner d\omega = 0, \quad \dot{\gamma} \lrcorner \alpha = 0,$$

easily give us

$$\dot{\gamma} \lrcorner \beta = 0,$$

using the preceding identity.



**Figure 47.1.** The system path is an integral curve of the 1-form  $\alpha$ .

**Example:** For a free particle in one dimension, we have a Lagrangian

$$L: (t, q, \dot{q}) \mapsto \frac{1}{2} \dot{q}^2,$$

and so the system path is an integral curve of the ideal generated by

$$\alpha = dq - \dot{q} dt, \quad \beta = d\dot{q}.$$

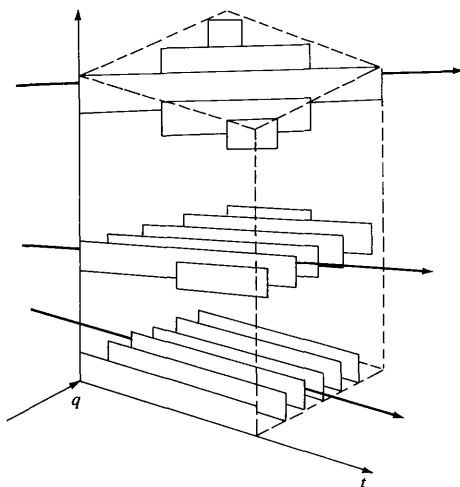
The path stays in a plane of given  $q$  and moves along the contours of  $\alpha$  as shown in Figure 47.1. A three-dimensional view is sketched in Figure 47.2. You could make a model of this by gluing together rafts of soda straws, twisting the various layers progressively.

### *Generalized coordinates*

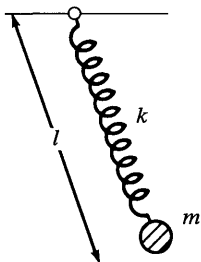
Thus far we have only restated  $F = ma$  in fancy notation. Now comes the beautiful move. Consider stiff systems, systems that have very different time scales for the dynamics of different parts of the system.

**Example:** Look at the pendulum shown in Figure 47.3, in which a mass  $m$  is hung on a spring of strength  $k$  and length  $l$ , and swings freely in a gravitational field of strength  $g$ . If the swinging frequency  $(g/l)^{1/2}$  is much less than the vibrational frequency  $(k/m)^{1/2}$  then the system is stiff.

If we have a stiff system, then for some situations we can just ignore the stiff degrees of freedom and consider a reduced system. This requires not



**Figure 47.2.** The kinematics of the system in the bundle  $CR^2$ .



**Figure 47.3.** The springy pendulum, a system with two independent time scales.

only a high frequency for the stiff degrees of freedom, but also low forcing frequencies and sufficient damping to keep the stiff degrees of freedom from being excited.

**Examples:** For the stiff pendulum in the preceding example, with initial conditions

$$\dot{x} = \dot{y} = v_0,$$

we cannot ignore the stiff mode until it has damped out. Nor is it

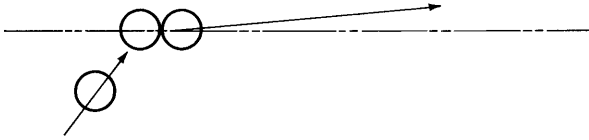


Figure 47.4.

generally obvious what are the correct initial conditions on the reduced system. Here angular momentum conservation gives it to you. In general you are faced with a singular dynamical system in the sense of Section 14, and must look closely at the path by which it snaps onto the slow manifold.

Although billiard balls are the staple of mechanics metaphors, they are stiff systems that are not easily reducible. Every mechanics student should contemplate the marvels of the pool table. In particular, try to explain the “throw” of frozen balls, shown in Figure 47.4. [See Martin and Reeves (1977).]

When a stiff system can be reduced, then we use the coordinate invariance of our exterior differential system to write the system in a form where the stiff modes are described by separate coordinates. To reduce the system just set them equal to their fixed values. The remaining coordinates are called generalized coordinates.

**Example:** For our stiff pendulum the angle is a generalized coordinate.

The task of writing the Lagrangian function  $L$  in generalized coordinates is simplified if we note that the term  $\sum_i \frac{1}{2} m_i \dot{q}^i \dot{q}^i$  is just the total kinetic energy. The Lagrangian is just the kinetic energy minus the potential energy. We can usually write down the Lagrangian for a physical system by inspection, but try writing one for a bicycle.

### *Symmetries*

Suppose we have a vector field  $k$  such that

$$\mathcal{L}_k L dt = 0,$$

$$\mathcal{L}_k \alpha = \mathcal{L}_k (d\dot{q} - \dot{q} dt) \in \mathcal{G}(\alpha).$$

The symmetry  $k$  leads to a conserved quantity using the Cartan 1-form

$$\omega = L dt + \frac{\partial L}{\partial \dot{q}} \alpha.$$

Since

$$\mathcal{L}_k \omega \in \mathfrak{G}(\alpha),$$

we have along the solution curve

$$\mathcal{L}_k \omega = 0,$$

and if we write this out, one term is the Euler-Lagrange equation and the other gives

$$\int d(k \lrcorner \omega) = 0;$$

that is,  $k \lrcorner \omega$  is constant.

In fact, the question of conserved quantities for discrete mechanical systems is pathological. Along every trajectory the initial conditions are conserved. Thus every such system has the full number of conservation laws allowed. Of course, these trivial ones are not useful. The useful ones are those that can be found without first solving the equations of motion.

#### 48. Lagrange's equations for general systems

We are now ready to extend the equations of motion given in the last section, which were based on the Lagrangian function, to systems that can have nonconservative forces, perhaps velocity-dependent forces, and constraints that may be anholonomic. There is no new physics here. These are all still consequences of  $F = ma$ .

##### *Generalized forces*

Not all forces come from the gradients of potentials. In general, we only have a force 1-form telling us how much it costs to "push the system" in different directions. Look back at the last section to see how the potential entered into our equation of motion. The Lagrangian is given by

$$L = T - V,$$

where  $V$  is the potential generating the forces. This shows up as a term in  $d\omega$ ,

$$-\frac{\partial V}{\partial q} dq dt,$$



and, after the contraction  $\dot{\gamma} \lrcorner d\omega$ , becomes

$$\left(\frac{\partial V}{\partial q}\right)[Q' dt - T' dq],$$

and this is the value, along the path, of the 1-form on the contact bundle

$$\frac{\partial V}{\partial q}(\dot{q} dt - dq).$$

Thus, if we had not wanted to include this force in  $L$ , we would have written the correct equations if we had included it as a separate force term

$$f\alpha = \left(-\frac{\partial V}{\partial q}\right)(dq - \dot{q} dt) = \left(-\frac{\partial V}{\partial q}\right)\alpha,$$

and used equations of motion

$$\dot{\gamma} \lrcorner d\omega = -f\alpha.$$

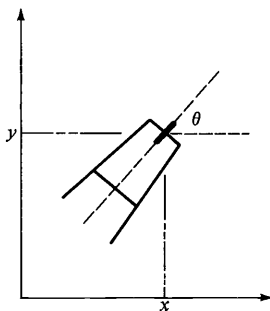
Any force that is not derivable from a potential can be treated this way. If the force in configuration space is given by  $f dq$ , then we lift it up to a force on the contact bundle of  $f\alpha$ , and find the system motion from the preceding exterior differential system, equation 47.1. This leads to what are called Lagrange's equations. They are the most general dynamical equations. Lagrangian mechanics based only on the Lagrangian function cannot handle velocity-dependent forces. Likewise, the Hamiltonian description to be covered in Section 50 is also less general.

### *Anholonomic constraints*

Return to the idea of physical systems with forces on them so strong that they dominate the dynamics. Such forces are called constraints, and the constraint is presumed to be able to apply whatever force is needed to keep the constraint satisfied. A common form for such constraints involves local conditions on the derivatives. These are called *anholonomic constraints*.

**Example:** Consider a wheelbarrow, free to move on the  $(x, y)$  plane. Three coordinates are needed to specify its position, say, the  $x$  and  $y$  coordinates of the wheel and an angle for the direction of the wheelbarrow. We cannot change the wheelbarrow's state arbitrarily, though. The wheel cannot be dragged sideways. The condition for this is that

$$\dot{y} = \dot{x} \tan \theta,$$



**Figure 48.1.** The wheelbarrow as an example of a system with an anholonomic constraint.

where  $\theta$  gives the angle between the  $x$  axis and the direction that the wheelbarrow is pointing. In terms of 1-forms this condition is

$$\dot{\gamma} \lrcorner (dy - \tan \theta dx) = 0.$$

See Figure 48.1.

For a constraint condition

$$\dot{\gamma} \lrcorner \phi = 0,$$

with  $\phi$  a 1-form, we have a system able to produce sufficiently strong forces in the  $\phi$  direction to keep the constraint condition satisfied. Putting these (as yet) unknown forces in with a strength parameter  $\mu(s)$ , we have a system

$$(\dot{\gamma} \lrcorner \beta) dq = f dq + \mu \phi,$$

$$\dot{\gamma} \lrcorner \alpha = 0, \quad \dot{\gamma} \lrcorner \phi = 0.$$

The unknown strength  $\mu(s)$  is very much like a Lagrange multiplier. You no doubt expect that the variational formalism given earlier will apply to anholonomic mechanical constraints. Although that puts you in good company, it is not true. [I gather from some remarks in Sommerfeld (1952) that the argument over this dates back to Hertz at least. Goldstein (1959) is wrong; Whittaker (1965) has an honest but confusing treatment; Saletan and Cromer (1971) have an adequate algebraic discussion; and Pars (1965) has a clear, lucid discussion and gives a useful example.]

The trouble is that if we take the action to be extremal over variations that are consistent with the constraint, then when the constraint is anholonomic, we get the wrong equations of motion. Rather than abandon

a variational approach, some authors redefine the idea of a variation. For mechanical constraints they make variations satisfying the condition

$$v \lrcorner \phi = 0$$

instead of the correct condition, given in equation 35.2. The claim in Goldstein (1959) that this follows from

$$\dot{\gamma} \lrcorner \phi = 0$$

is incorrect. The problem appears to be one of taking a symbolic notation too seriously. For anholonomic constraints it is not true that their operators  $d$  and  $\delta$  commute. The modified variations, which need to be introduced in order to yield the correct equations of motion, do not move the path into another acceptable path. It seems to be a rather forced way to produce the equations. One goal of this book is to eliminate all such funny deltas.

### Problems

- 48.1** Write Lagrange's equations for a charged particle moving in a magnetic field.
- 48.2** Show that the procedure used to lift the force 1-form up into the jet bundle is really coordinate-independent. Try to find an elegant argument rather than just grinding it out.

## 49. Lagrangian field theory

We can easily extend our discussion of dynamics to continuum systems and field theory. The proper geometric setting is again a contact manifold. For a scalar field this is the manifold of the independent variables, the dependent variable, and its derivatives.

There is a temptation to view continuum systems as systems with an infinite number of degrees of freedom, and so jump into spaces of infinite dimension, Sobolev spaces or worse. That leads into the morass of functional analysis. Powerful theorems come from this approach, but it is outside the scope of this book and not congenial to my style. Here I will reap what geometric advantages I can from the finite-dimensional setting. For some systems, notably electrodynamics, the entire discussion can even be given in spacetime itself.

The continuum limit that we arrive at here is special, in that it does not include materials with intrinsic spin. The angular-momentum density here all comes from bulk motions. To include intrinsic spin, we must

develop continuum mechanics directly from the ten conservation laws and the appropriate constitutive relations.

### *Hamilton's principle*

We will ignore systems with anholonomic constraints and base our discussion on a variational principle. The Lagrangian for a distributed system involves an integration over spatial variables as well as the time integration demanded by Hamilton's principle.

**Example:** The usual linearized stretched elastic string has a kinetic energy

$$T = \int_0^L \frac{1}{2} \rho \left( \frac{\partial y}{\partial t} \right)^2 dx,$$

where  $x$  is a coordinate labeling points on the string and  $y(x, t)$  is the transverse displacement of the point  $x$  at time  $t$ . The potential energy for small slopes and perfect elasticity is

$$V = \int_0^L \frac{1}{2} T \left( \frac{\partial y}{\partial x} \right)^2 dx.$$

Thus Hamilton's principle demands that the double integral

$$\int_{t_1}^{t_2} \int_0^L \frac{1}{2} \left[ \rho \left( \frac{\partial y}{\partial t} \right)^2 - T \left( \frac{\partial y}{\partial x} \right)^2 \right] dx dt$$

be stationary for variations of the function  $y(x, t)$ . In the contact manifold described by coordinates  $(x, t, y, u, v)$  with contact 1-form

$$\alpha = dy - u dt - v dx, \quad (49.1)$$

this integral is

$$\int_{\Gamma} \frac{1}{2} (\rho u^2 - T v^2) \theta,$$

where the volume-element 2-form is

$$\theta = dx dt.$$

Note that  $u$  and  $v$  are defined by equation 49.1.

As was true in particle mechanics, this is a degenerate form of the calculus of variations, and the Euler-Lagrange equations can be found by the integration-by-parts trick. The integral is to be stationary for infinitesimal variations given by the vector field

$$w = Y \frac{\partial}{\partial y} + P \frac{\partial}{\partial p},$$

in which the graph of the field variable  $y$  and its partial derivatives  $p$  are deformed while keeping the graph an integral submanifold of  $\alpha$ . This condition constrains  $w$  so that

$$u \lrcorner \mathfrak{L}_w \alpha = 0$$

for any vector  $u$  in the graph. Now,

$$\mathfrak{L}_w \alpha = -w \lrcorner dp dq + dY,$$

and so we must have

$$u \lrcorner (dY - P dq) = 0.$$

The deformation of the integral is

$$\begin{aligned} \mathfrak{L}_w \int_{\Gamma} L \theta &= \int_{\Gamma} \mathfrak{L}_w (L \theta) = \int_{\Gamma} w \lrcorner (dL \wedge \theta) \\ &= \int_{\Gamma} \left( \frac{\partial L}{\partial y} \theta + P \frac{\partial L}{\partial p} \theta \right). \end{aligned}$$

Look at the second term. Restoring the indices, we have

$$\frac{\partial L}{\partial p_i} P_i \theta = \frac{\partial L}{\partial p_i} P_j \delta_i^j \theta = \frac{\partial L}{\partial p_i} P_j dq^j \theta_i,$$

using the identity

$$dq^j \theta_i = \delta_i^j \theta.$$

Now we can use the contact form to find

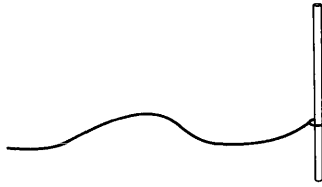
$$\int_{\Gamma} P \frac{\partial L}{\partial p} \theta = \int_{\Gamma} \frac{\partial L}{\partial p} dY \theta = \int_{\Gamma} \frac{\partial L}{\partial p} d(Y \theta),$$

and integrating by parts we find

$$\mathfrak{L}_w \int_{\Gamma} L \theta = \int_{\Gamma} Y \left[ \frac{\partial L}{\partial y} \theta - d \left( \frac{\partial L}{\partial p} \right) \theta \right] + \int_{\partial \Gamma} Y \frac{\partial L}{\partial p} \theta.$$

From the volume integral we find the Euler-Lagrange equations as the exterior differential system

$$\begin{aligned} d \left( \frac{\partial L}{\partial p} \right) \theta - \frac{\partial L}{\partial y} \theta, \\ dy - p dq. \end{aligned}$$



**Figure 49.1.** An elastic string with an end free to move up and down.

The surface integral does not vanish identically. Hamilton's principle demands that  $Y$  vanish at the initial and final times, but not necessarily at intermediate times. The boundary conditions determine the surface term. Conversely, if the boundary conditions are not specified, the requirement that the surface terms vanish provides natural boundary conditions.

**Example:** Suppose that for our elastic string the ends are not fixed, but free to slide up and down (Figure 49.1). The surface term is

$$\int_{\partial\Gamma} Y \frac{\partial L}{\partial p} \theta = \int_{\partial\Gamma} Y \left[ \frac{\partial L}{\partial u} \frac{\partial}{\partial t} \lrcorner dt dx + \frac{\partial L}{\partial v} \frac{\partial}{\partial x} \lrcorner dt dx \right],$$

giving us

$$\int_{\partial\Gamma} Y [\rho u dx + Tv dt] = 0.$$

This is satisfied for arbitrary  $Y$  on the sides only if

$$v = \frac{\partial y}{\partial x} = 0.$$

This is indeed the correct boundary condition for the elastic string with free ends.

### *Elastic beam*

A nice illustration of all this is the simple, linear, elastic beam. The kinetic energy involves only the transverse velocity, and the potential energy involves only bending. More complicated beam models take into account rotational kinetic energy and other factors. They are routine extensions of this problem.

Some care is needed to pick the right contact manifold. It should have all the necessary variables and no more. The dependent variable  $y$  and

the independent variables  $x$  and  $t$  are surely needed. Both first derivatives will also be needed; let us call them  $u$  and  $v$ , and constrain them using the contact 1-form

$$\alpha = dy - u dt - v dx.$$

The solution must be an integral submanifold of this 1-form. Because the potential energy involves the curvature, we need to introduce second derivatives as well, but only in the spatial direction. Call the second spatial derivative  $r$ , and introduce the additional contact 2-form

$$\beta = (dv - r dx) dt.$$

The  $dt$  factor is necessary to kill off a  $v_{,t}$  term. If we require that our solution be an integral submanifold of this form as well, then  $r$  will be  $y_{,xx}$ .

We write Hamilton's principle

$$\mathcal{L}_w \int_{\Gamma} \frac{1}{2}(u^2 - r^2) dt dx = 0.$$

There are also material parameters. For clarity I will keep them constant; and I have absorbed them into the length and time variables.

The deformation of the integral is given by the vector field

$$w = Y \frac{\partial}{\partial y} + U \frac{\partial}{\partial u} + V \frac{\partial}{\partial v} + R \frac{\partial}{\partial r},$$

and the functions  $Y$ ,  $U$ ,  $V$ ,  $R$  are constrained by

$$\sigma \lrcorner \mathcal{L}_w \alpha = 0, \quad \sigma \lrcorner \mathcal{L}_w \beta = 0,$$

where  $\sigma$  is any surface element. This implies that under the integral we have

$$\int (dY - U dt - V dx) = 0$$

and

$$\int (dV - R dx) dt = 0.$$

We proceed in the now familiar fashion. The deformation of the integral is given by

$$\mathcal{L}_w \int_{\Gamma} L \theta = \int_{\Gamma} (uU - rR) dt dx.$$

Here  $u$  is the value of  $u$  on the submanifold, whereas  $U$  is the  $u$  component

of the deformation. The first term reduces as usual:

$$\begin{aligned} -\int_{\Gamma} u dx U dt &= -\int_{\Gamma} u dx (dY - V dx) = \int_{\Gamma} u dY dx \\ &= -\int_{\Gamma} Y du dx + \int_{\partial\Gamma} Yu dx. \end{aligned}$$

The surface integral vanishes identically;  $Y$  must vanish at the initial and final times, and  $dx$  pulls back to zero on the ends.

The second term is more challenging. We have

$$-\int_{\Gamma} rR dt dx = \int_{\Gamma} r dV dt = -\int_{\Gamma} V dr dt + \int_{\partial\Gamma} rV dt. \quad (49.2)$$

It seems that we can proceed no further. The problem is that we need to do another integration by parts to change  $V$  into  $Y$ . This involves another derivative, but we lack a variable for it. We introduce  $s = y_{,xxx}$ , using the 2-form

$$\gamma = (dr - s dx) dt,$$

which is suggested strongly by equation 49.2. None of the earlier work is changed by this enlarging of our space. Now we can proceed

$$-\int_{\Gamma} V dr dt = -\int_{\Gamma} Vs dx dt = -\int_{\Gamma} s dY dt$$

and integrating by parts again

$$= +\int_{\Gamma} Y ds dt - \int_{\partial\Gamma} sY dt.$$

Collecting all this, we have

$$\mathcal{L}_w \int_{\Gamma} L\theta = \int_{\Gamma} Y [dx du - ds dt] + \int_{\partial\Gamma} (rV - sY) dt.$$

At long last we find the Euler-Lagrange exterior differential system

$$ds dt + dx du, \quad dy - u dt - v dx,$$

$$du dt + dv dx,$$

$$dv dt - r dx dt, \quad dr dt - s dx dt.$$

The surface integrals provide natural boundary conditions

$$sY = 0, \quad rV = 0.$$



If the ends are free to move, then  $s$  must vanish there. If they are free to rotate, then  $r$  must vanish there.

### Problem

**49.1** How much less will a bookshelf sag if the ends of the shelf are built in and not free to rotate, compared with a shelf whose ends are simply supported?

## 50. Hamiltonian systems

Classical mechanics is a highly evolved subject, and many obscure moves have been explored and found to be useful. The transformation of a dynamical system into Hamiltonian form is one of these. It is difficult to motivate this transformation. Perhaps it was easier when everyone had a thorough background in thermodynamics. Then a Legendre transformation would come naturally to mind. Furthermore, many of the reasons given in textbooks to motivate this transformation are wrong. It is often said that Lagrange's equations are  $n$  second-order equations, whereas Hamilton's equations are  $2n$  first-order equations. This is nonsense, as the exterior differential form of Lagrange's equations shows.

The true advantage of the Hamiltonian form is in the contact and symplectic structures that arise. This will be clear when the applications are studied. Thus when we describe optics as a Hamiltonian system in Section 52, we will find it to be an obvious result that no lens can increase the brightness of an object. With this apology, let us now wander up, down, and across a veritable forest of contact bundles.

### *Lagrangian systems in the jet bundle*

Use enough dimensions. This seems to be a good general rule. Suppose we look at the bundle  $CQ \times \mathbb{R}$ , where  $CQ$  is the line-element contact bundle to configuration space. We will use the extra coordinate to graph the Lagrangian function. In this space a given Lagrangian system is described by a submanifold, the graph of  $L$ . Call this extra coordinate  $l$ . We have a natural projection

$$(t, q, \dot{q}, l) \mapsto (t, q, \dot{q})$$

and using this we pull the basis forms  $dq$ ,  $d\dot{q}$ , and  $dt$  up to the larger bundle.

The 1-jet of this new bundle has additional coordinates for all the partial derivatives of  $L$ . A useful mnemonic labeling for these coordinates is

$l_q$ ,  $l_{\dot{q}}$ , and  $l_t$ . The contact structure of this 1-jet is given by

$$dl - l_q dq - l_{\dot{q}} d\dot{q} - l_t dt.$$

A Lagrangian dynamical system on  $CQ$  can be lifted to a system on  $CQ \times \mathbb{R}$  by pulling up the contact form

$$d\dot{q} - \dot{q} dt,$$

and the Lagrange equations

$$d\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} dt.$$

We impose the requirement that the motion remain in the graph of  $L$

$$l = L(q, \dot{q}, t),$$

and this implies

$$dl = dL. \tag{50.1}$$

Continue upward, and pull these forms up into the 1-jet bundle, to find the exterior differential system

$$\begin{aligned} dl - l_q dq - l_{\dot{q}} d\dot{q} - l_t dt, \\ dq - \dot{q} dt, \quad dl_{\dot{q}} - l_q dt, \\ dl - dL. \end{aligned} \tag{50.2}$$

We have used equation 50.1 to simplify the Lagrange equations. Now they involve only coordinates and basis vectors, not functions. This simplification is the whole reason for this construction. Note the clarity that results from a notation that separates the independent variable  $l$  and the function  $L$ . As an exterior system, the one given in the preceding is not closed under  $d$ . This does not matter, because the maximal integral submanifolds are only one-dimensional; 2-forms do not matter. This exterior differential system defines a line-element field on the jet bundle.

### *The Hamiltonian bundle*

If we start instead with the hypersurface contact-element bundle  $C^*Q$ , with coordinates  $(t, q, p)$  and contact structure

$$dt - p dq,$$

we can generate the same tree of bundles. We now have the graph of a function called the Hamiltonian. We look at the 1-jet of the bundle

$$(t, q, p, h) \mapsto (t, q, p)$$

with contact structure

$$dh - h_q dq - h_p dp - h_t dt. \quad (50.3)$$

Again, this equation should be read as a convenient packaging of the definitions of the new coordinates. This jet bundle and also the Lagrangian jet bundle both have hypersurface-element contact structures, equations 50.2 and 50.3. Contrast this with the different contact structures of  $CQ$  and  $C^*Q$ . We can find a contact transformation between the jet bundles that preserves their contact structures; this is the Legendre transformation:

$$(q, \dot{q}, t, l, l_q, l_{\dot{q}}, l_t) \mapsto (q, p, t, h, h_q, h_p, h_t) = (q, l_{\dot{q}}, t, l_{\dot{q}} \dot{q} - l, -l_q, \dot{q}, -l_t).$$

Note that this involves only the coordinates, not the functions  $L$  or  $H$ . They enter the problem only by defining the submanifolds on which the system moves. The inverse transformation is

$$\begin{aligned} (q, p, t, h, h_q, h_p, h_t) &\mapsto (q, \dot{q}, t, l, l_q, l_{\dot{q}}, l_t) \\ &= (q, h_p, t, h_p p - h, -h_q, p, -h_t). \end{aligned} \quad (50.4)$$

These transformations map integral submanifolds of 50.2 into integral submanifolds of the 1-form in equation 15.3 and vice versa.

### *Hamilton's equations*

Using the inverse transformation, equation 50.4, we can pull our Lagrangian system over to the Hamiltonian jet bundle. Because it consists of only basis 1-forms and coordinates, this is easy. We find the system

$$\begin{aligned} dh - h_q dq - h_p dp - h_t dt, \\ dq - h_p dt, \quad dp + h_q dt, \\ dh - dH, \end{aligned}$$

where the hypersurface  $l = L$  is mapped onto the hypersurface  $h = H$ . The Hamiltonian function  $H$  is found by inverting

$$p = \frac{\partial L}{\partial \dot{q}}(q, \dot{Q}(q, p, t), t)$$

for the functions  $\dot{Q}(t, q, p)$ . Then we have

$$H = p\dot{Q} - L(q, \dot{Q}, t).$$

Again the exterior system defines a line-element field on the jet bundle.

Now to descend. The jet bundles are used only for the Legendre transformation. The Hamiltonian exterior system is the pullback of the exterior system on  $C^*Q$  generated by

$$dq - \frac{\partial H}{\partial p} dt, \quad dp + \frac{\partial H}{\partial q} dt.$$

A direct map between  $CQ$  and  $C^*Q$  is very difficult. (Try it!) A partial derivative such as  $\partial L/\partial q$  does not transform simply. The function part goes against the map, and the derivation goes with it. The trick was to go to a space where the partial derivatives are just coordinates.

A Cauchy characteristic vector for the exterior system on  $C^*Q$  is

$$u = \frac{\partial H}{\partial p} \frac{\partial}{\partial q} - \frac{\partial H}{\partial q} \frac{\partial}{\partial p} + \frac{\partial}{\partial t}.$$

This vector field yields the usual set of Hamilton's equations.

### Integral invariants

We now jump from considering a single integral curve to studying properties of families of them. Here there is additional structure not apparent in any single integral curve.

Look at a 1-parameter family of integral curves that have the topology of the circle, that is, a 1-cycle (see Figure 50.1). The integral curves of our system sweep this into a cylinder in  $C^*Q$ . The integral

$$\int_{\partial\Gamma} (p dq - H dt)$$

is called the *action* for this ensemble. The action is the same for any 1-cycle  $\partial\Gamma$  around this cylinder. Different parts of the 1-cycle  $\partial\Gamma$  can be pushed forward along the integral curves at different rates.

I will show this invariance in the next section. There are also higher forms of these conservation laws. For any two-dimensional ensemble, we have an invariant integral

$$\int_{\Gamma} (dp dq - dH dt)$$

provided  $\Gamma$  is carried along by the integral curves (Figure 50.2). Remember that  $H$  is a function on  $C^*Q$ . For an  $n$ -dimensional ensemble, wedge the above integrand with itself, and so on, up to  $n$  factors. The last invariance is called Liouville's Theorem. After  $n$  factors they all vanish identically.

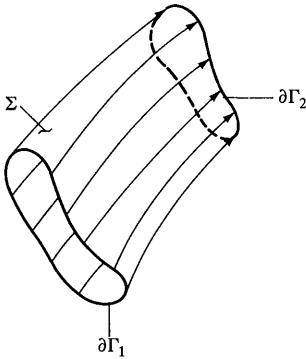


Figure 50.1.

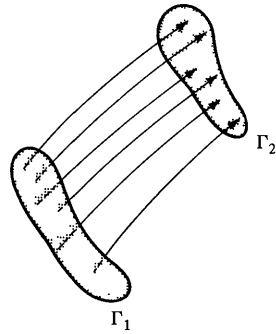


Figure 50.2.

*Hamilton's principle*

Associated with any Hamiltonian system is a variational principle. This variational principle is a useful tool for guessing Hamiltonian functions. Look at the action, the line integral in  $C^*Q$

$$S = \int (p dq - H dt).$$

Hamilton's principle states that the motion of the system is along an extremum of this integral, that is

$$\delta \int (p dq - dH dt) = 0.$$

Along a solution curve

$$s \mapsto (Q(s), P(s), s),$$

we must have

$$P' dq - Q' dp + \frac{\partial H}{\partial q} dq + \frac{\partial H}{\partial p} dp + Q' \frac{\partial H}{\partial q} dt + P' \frac{\partial H}{\partial p} dt = 0,$$

where

$$Q' = \frac{dQ}{ds}$$

and so on. These lead to the usual Hamiltonian equations

$$\frac{dQ}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dP}{dt} = -\frac{\partial H}{\partial q}$$

and also

$$\frac{dH}{dt} = \frac{\partial H}{\partial t},$$

which follows from the first two.

### Problem

**50.1** What is the relation between Hamiltonian mechanics and the exterior differential system generated by  $dp \wedge dq$  and  $dH$ ?

## 51. Symplectic geometry

We now specialize to the situation in which the Hamiltonian function does not depend on one coordinate, usually time. Here the special geometry of the cotangent bundle plays a central role. This special situation is so much studied that you could mistakenly think that it is all of the Hamiltonian dynamics. In Section 12 we saw that the structure of the cotangent bundle manifested itself in the canonical 1-form

$$\theta = p dq,$$

and the 2-form

$$\Omega = d\theta = dp dq. \quad (51.1)$$

This 2-form satisfies

$$d\Omega = 0$$

and is of maximum rank. Such a 2-form generates a mathematical structure called *symplectic geometry*. A manifold of even dimension with such a 2-form is called a *symplectic manifold*. Manifolds with odd dimension can have the closely related contact structure.

In Section 26, on normal forms, we saw that any closed 2-form can be written in some finite region in the manner of equation 51.1. Thus all symplectic geometries are locally the same, and are thus quite different from metric geometries. A finite piece of a sphere is intrinsically different from a finite piece of the plane.

### *Geometric significance of $\Omega$*

The 2-form  $\Omega$  determines special submanifolds  $(N, \psi)$  that are integral submanifolds

$$\psi^* \cdot \Omega = 0.$$

An integral submanifold of  $\Omega$  of half the dimension of the space, which is the largest possible, is called a *Lagrangian submanifold*.

Suppose we have a 1-form field, a section  $\nu: M \rightarrow T^*M$ . If we have

$$d\nu = 0, \quad (51.2)$$

then this section is a Lagrangian submanifold. Locally the reverse is true, but there can be bad points where the Lagrangian submanifold is vertical and is not the graph of any 1-form field.

To prove equation 51.2, represent the map  $\nu: M \rightarrow T^*M$  by functions  $W$

$$\nu: q \mapsto (q, p) = (q, W(q)).$$

Then

$$\nu^* \cdot dq = dq$$

and

$$\nu^* \cdot dp = \frac{\partial W}{\partial q^\alpha} dq^\alpha$$

and so

$$\nu^* \cdot (dp_\alpha dq^\alpha) = \frac{\partial W_\alpha}{\partial q^\beta} dq^\beta dq^\alpha.$$

But

$$d(W_\alpha dq^\alpha) = \frac{\partial W_\alpha}{\partial q^\beta} dq^\beta dq^\alpha,$$

hence

$$\nu^* \cdot (dp dq) = d\nu,$$

proving the assertion.

### *Symplectic gradient*

The 2-form  $\Omega$  has many of the properties of a metric. We can use it to transfer vectors to the cotangent space and vice versa, using the relation

$$u \lrcorner \Omega = \nu. \quad (51.3)$$

If  $\Omega$  is nondegenerate, that is,  $u \lrcorner \Omega$  never vanishes unless  $u$  itself does, then equation 51.3 can be solved for  $u$ . The gradient of a function is a covector. Using the preceding operation, we can turn it into a vector, called the *symplectic gradient*.

### *Hamiltonian dynamical systems*

A dynamical system is a vector field on a manifold. Some dynamical systems have more structure; here we will look at the special systems that

have a symplectic structure as well. These special systems are most familiar to physicists from classical mechanics. These dynamical systems can be derived from a Hamiltonian function that is independent of time.

If  $t$  does not appear in  $H$ , then we can reduce the dimension of the representation by one, and relegate  $t$  to being only the parameter along the integral curves. Thus the parametrization becomes significant, and we replace the exterior differential system by the vector field

$$\frac{\partial H}{\partial p} \frac{\partial}{\partial q} - \frac{\partial H}{\partial q} \frac{\partial}{\partial p}.$$

Geometrically we have a function  $H: T^*(M) \rightarrow \mathbb{R}$ , called the Hamiltonian function. We can use the 2-form  $\Omega$  to convert the gradient of  $H$  into a vector field

$$v \lrcorner \Omega = -dH, \quad (51.4)$$

going from  $H$  to a dynamical system by using symplectic geometry.

**Example:** Let us show that this produces the usual Hamiltonian equations of classical mechanics. Write

$$v = \dot{q} \frac{\partial}{\partial q} + \dot{p} \frac{\partial}{\partial p};$$

then

$$v \lrcorner \Omega = \dot{p} dq - \dot{q} dp,$$

and so from equation 51.4 we have

$$\dot{p} dq - \dot{q} dp = -\frac{\partial H}{\partial q} dq - \frac{\partial H}{\partial p} dp;$$

that is

$$\dot{q} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial q}.$$

Where do these Hamiltonian systems come from? For a few physical systems there are informal algorithms for producing a suitable Hamiltonian function to describe the system. In these situations the Hamiltonian is a quick way to produce the equations of motion. In many situations we have the equations already; so we search around and find a suitable Hamiltonian function. Why do that if we already have the equations of motion? Because Hamiltonian systems are special, and they have a great deal of useful structure. Also, we can use the connection between Lagrangian systems and Hamiltonian systems. There are quite precise algorithms



for the Lagrangian description of a mechanical system. Finally, Hamiltonian systems also arise in the high-frequency limit of dispersive wave propagation.

### *The Hamiltonian flow*

In Euclidean geometry the divergence of a vector field measures the extent to which a fluid carried along with the flow of the vector field is compressed or expanded. The idea can be carried over to a manifold provided that we are given some way to compare densities at different points. Such a standard density is available if we are given some particular  $n$ -form ( $n$  being the dimension of the manifold) at every point. The extent to which the flow of the vector field  $v$  expands relative to the given density is described by the divergence ( $\operatorname{div} v$ ), defined by

$$\mathcal{L}_v \omega = (\operatorname{div} v) \omega.$$

Remember that the  $n$ -forms are a one-dimensional vector space. Since  $\mathcal{L}_v \omega$  is an  $n$ -form, it must be proportional to  $\omega$ ; ( $\operatorname{div} v$ ) is the constant of proportionality.

**Example:** The familiar Euclidean divergence uses the 3-form

$$\theta = dx dy dz.$$

For a vector

$$v = a \frac{\partial}{\partial x} + b \frac{\partial}{\partial y} + c \frac{\partial}{\partial z},$$

we have

$$\begin{aligned} \mathcal{L}_v \theta &= d(v \lrcorner \theta) \\ &= d(a dy dz + b dz dx + c dx dy) \\ &= \left( \frac{\partial a}{\partial x} + \frac{\partial b}{\partial y} + \frac{\partial c}{\partial z} \right) dx dy dz \end{aligned}$$

and so we have the familiar result

$$\operatorname{div} v = \frac{\partial a}{\partial x} + \frac{\partial b}{\partial y} + \frac{\partial c}{\partial z}.$$

For a density

$$\omega = a dx^1 dx^2 \dots$$

we have

$$\mathfrak{L}_v \omega = \frac{\partial(av^j)}{\partial x^j} dx^1 dx^2 \dots$$

and so

$$\operatorname{div} v = \frac{1}{a} \frac{\partial(av^j)}{\partial x^j}. \quad (51.5)$$

There is a natural density on any cotangent bundle  $T^*M$

$$\omega = \Omega \wedge \Omega \wedge \dots \wedge \Omega$$

where, if  $M$  is  $n$ -dimensional, there are  $n$  factors of  $\Omega$  on the right. Since we have for our Hamiltonian vector field  $v$

$$\mathfrak{L}_v \Omega = d(v \lrcorner \Omega) + v \lrcorner d\Omega = 0,$$

we also have

$$\mathfrak{L}_v \omega = 0$$

and so

$$\operatorname{div} v = 0.$$

Thus the flow of any Hamiltonian vector field has no divergence and is similar to the flow of an incompressible fluid.

Systems with damping cannot be Hamiltonian systems. If all initial states collect at a single point, then this implies a negative divergence no matter what  $\omega$  is chosen. Of course, in small, local, regions any system can be written as a Hamiltonian system. Recall that any vector field can be transformed into a constant vector field in some small region. These limited transformations are usually unnatural and of little interest, since the systems usually run out of the local regions before anything interesting happens. [See Santilli (1978) for more on this.]

### *Integral invariants*

The representation of systems in the cotangent bundle is very well-suited to a study of ensembles of systems. Consider a 1-parameter family of systems, each moving along an integral curve of some given Hamiltonian  $H$ . Suppose further that they have the topology of a circle. The action integral is defined to be

$$S = \int_{\partial\Gamma} p dq,$$

and we want to see under what conditions  $S$  has the same value for any

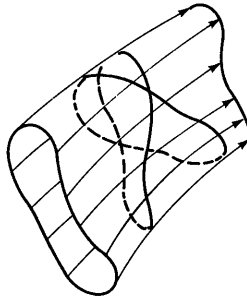


Figure 51.1.

cycle  $\partial\Gamma$  that goes around the family of integral curves, not necessarily at constant time (see Figure 51.1).

We will calculate the effect of pushing the curve  $\partial\Gamma$  along the integral curves, not necessarily with the same rate for all of its parts. The change in  $S$  is given by

$$\mathcal{L}_{\phi v} \int_{\partial\Gamma} \mathbf{p} \, dq,$$

where  $v$  is the Hamiltonian vector field satisfying equation 51.4, and  $\phi$  is an arbitrary function. We have the change given by

$$\int_{\partial\Gamma} \mathcal{L}_{\phi v}(\mathbf{p} \, dq) = \int_{\partial\Gamma} \phi v \lrcorner d\mathbf{p} \, dq = - \int_{\partial\Gamma} \phi \, dH.$$

Now, integrating by parts we find that this equals

$$\int_{\partial\Gamma} H \, d\phi.$$

If the value of  $H$  is the same on all of  $\partial\Gamma$ , then  $H$  can be factored out of the integral and the integral vanishes. If  $H$  is not constant, then we must take  $\phi$  constant, and  $\partial\Gamma$  must be carried along at the uniform rate of the flow.

Using Stokes' Theorem, we can write this as a surface integral

$$\int_{\Gamma} \Omega = \int_{\Gamma} d\mathbf{p} \, dq = \int_{\partial\Gamma} \mathbf{p} \, dq$$

invariant under the same conditions. Using the Leibniz rule we see that products like

$$\int_{\Gamma} \Omega \wedge \Omega \wedge \cdots \wedge \Omega$$

are also invariant.

### *Parameter invariance*

Suppose we compare a Hamiltonian system with Hamiltonian function  $H$  with one with Hamiltonian function  $H' = f(H)$ . From

$$v' \lrcorner \Omega = -f'(H) dH,$$

we have

$$v' = f'(H)v,$$

and the same curves in  $T^*M$  are traced out, but at a different rate.

For some Hamiltonian systems the states of interest do not fill the cotangent bundle. It is common to have systems confined to a hypersurface  $H = \text{constant}$ . For these systems an even more drastic reparametrization is possible:  $H$  can be replaced by any other function that preserves the  $H = \text{constant}$  hypersurface. This allows the Hamiltonian vector to be rescaled at every point, tracing out the same curves but at a different rate. Since the states of interest do not fill  $T^*M$ , we can go to a smaller space to describe the dynamics. This may or may not be more efficient.

**Example:** A free particle in special relativity is described by a Hamiltonian

$$H = g^{\alpha\beta} p_\alpha p_\beta.$$

With this Hamiltonian the parameter turns out to be proper time. Exactly the same Hamiltonian describes free particles in general relativity as well. The value of the Hamiltonian is the square of the mass of the particle. An ensemble of particles all of the same mass will not fill  $T^*\mathbb{R}^4$ .

### *Time-dependent systems*

Although it is clear that time-independent systems are a special case of time-dependent ones, that the converse is true seems surprising. Still, it is so. By enlarging the space, we find that a time-dependent system on a space  $(t, \mathbf{q}, \mathbf{p})$  of odd dimension leads to a time-independent Hamiltonian system on the even-dimensional space  $(\mathbf{q}, t, \mathbf{p}, r)$ , with new Hamiltonian function

$$W = H(t, \mathbf{q}, \mathbf{p}) + r.$$

The symplectic structure is given by  $d\mathbf{p} d\mathbf{q} + dr dt$ . The parameter along the integral curves should be called something other than  $t$ .

Hamilton's equations for the preceding Hamiltonian are

$$\begin{aligned} \frac{d\mathbf{q}}{ds} &= \frac{\partial H}{\partial \mathbf{p}}, & \frac{dt}{ds} &= 1, \\ \frac{d\mathbf{p}}{ds} &= -\frac{\partial H}{\partial \mathbf{q}}, & \frac{dr}{ds} &= -\frac{\partial H}{\partial t}. \end{aligned}$$

From these we see that the curve parameter is indeed time, and that the momentum  $r$  is just minus the value of the Hamiltonian.

The utility of this approach is that we can use directly the proof of the integral invariants for time-dependent systems. This is much simpler than a direct discussion. The action is

$$\int_{\partial\Gamma} \mathbf{p} d\mathbf{q} + r dt = \int_{\partial\Gamma} \mathbf{p} d\mathbf{q} - H dt.$$

Now, in lifting a given state  $(t, \mathbf{q}, \mathbf{p})$  up into  $(\mathbf{q}, t, \mathbf{p}, r)$  space, we must give some rule for assigning the momentum  $r$ . A constant shift in  $r$  makes no change in other variables, and we can pick  $r$  such that the Hamiltonian  $W$  vanishes identically. This is therefore a special situation in which the action can be computed for any 1-cycle enclosing the integral curve, not necessarily at constant time or constant parameter increment.

This leads us to augment our earlier discussion for time-independent systems. We have the integral in the symplectic bundle

$$\int_{\partial\Gamma} \mathbf{p} d\mathbf{q}$$

unchanged by uniform displacement along the integral curves. The integral in the contact bundle

$$\int_{\partial\Gamma} \mathbf{p} d\mathbf{q} - H dt,$$

on the other hand, is unchanged by any displacements whatsoever along the integral curves. Thus we see that even time-independent systems display their full structure only in the contact bundle rather than in the symplectic bundle.

## 52. Hamiltonian optics

Rather than repeat the familiar examples of elementary mechanics, let me illustrate the ideas of Hamiltonian systems with the interesting and useful application to optics. It will be instructive to see an example that is not a mechanical system. We will find the results useful when we discuss gravitational lenses in Section 65. Historically, the applications in optics preceded those in mechanics. In this section,  $z$ , the distance along the optical axis of the system, plays the role of time.

The description of a light ray requires two spatial coordinates and two directional ones. A Lagrangian description would take place on  $C\mathbb{R}^3$  with coordinates  $(z, x, y, \dot{x}, \dot{y})$  and a contact structure generated by the 1-forms

$$\dot{x} dz - dx \quad \text{and} \quad \dot{y} dz - dy.$$

The Hamiltonian description takes place in  $C^*\mathbb{R}^3$  with coordinates  $(z, x, y, \alpha, \beta)$  and a contact structure generated by

$$dz - \alpha dx - \beta dy.$$

In early drafts I used the six-dimensional symplectic manifold  $T^*\mathbb{R}^3$ , but this turns out to be remarkably clumsy. The system properly lives in the five-dimensional contact bundle.

### *The Hamiltonian*

To begin, we need the Hamiltonian function. We start with Snell's law, which says that if the index of refraction  $n$  is independent of  $x$ , then

$$n \sin \theta = \text{constant}.$$

As a Hamiltonian system, this is a system with a translational symmetry and hence a constant momentum. Thus we expect a momentum variable

$$\alpha = n \sin \theta.$$

We must write this in the proper variables, of course:

$$\alpha = n\dot{x}/\sqrt{1+(\dot{x})^2}.$$

Solving for  $\dot{x}$  we have

$$\dot{x} = \alpha/\sqrt{n^2 - \alpha^2}.$$

This should be one of Hamilton's equations, and so we must have

$$\frac{\partial H}{\partial \alpha} = \frac{\alpha}{\sqrt{n^2 - \alpha^2}}$$

and this leads us to guess the Hamiltonian

$$H = -\sqrt{n^2 - \alpha^2}$$

or in 3-space

$$H = -\sqrt{n^2 - \alpha^2 - \beta^2}.$$

### Variational principle

From Section 50 we know that the line integral

$$S = \int (\mathbf{p} \, d\mathbf{q} - H \, dt)$$

is stationary along light rays. Here this integral is (using two dimensions for clarity)

$$S = \int \alpha \, dx + \sqrt{n^2 - \alpha^2} \, dz.$$

Using Hamilton's equations gives us Fermat's principle, that

$$S = \int \frac{n \, dz}{\cos \theta}$$

is stationary. You might have expected me to start from Fermat's principle. That would be circular reasoning, however. What we really know about light rays is Snell's law. Here the actual inductive step is clearly displayed.

### Brightness

Turn now to study bundles of rays. There is a four-parameter family of rays near any given ray.

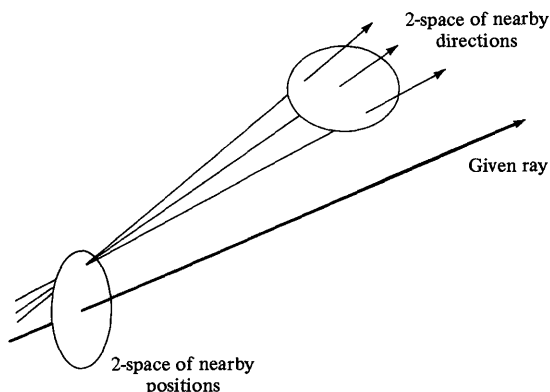
**Example:** The four-parameter family of rays around the ray

$$s \mapsto (x, y, z) = (0, 0, s)$$

is given by

$$(u, v, m, n) \mapsto [s \mapsto (x, y, z) = (u + ms, v + ns, s)].$$

See Figure 52.1.



**Figure 52.1.** The four-parameter family of rays in the neighborhood of a given ray.

Ray bundles are simple to discuss in the five-dimensional phase space  $C^*\mathbb{R}^3$ . Whereas in  $\mathbb{R}^3$  the rays crisscross everywhere, in  $C^*\mathbb{R}^3$  they form a smooth congruence, with only one ray passing through any given point.

The local density of rays is called *brightness*. The density of rays in five dimensions is represented by a 4-form aligned along the rays, with a cross-sectional area enclosing a single ray. The direction of the rays is specified by the Hamiltonian vector field

$$\frac{\partial H}{\partial \alpha} \frac{\partial}{\partial x} + \frac{\partial H}{\partial \beta} \frac{\partial}{\partial y} - \frac{\partial H}{\partial x} \frac{\partial}{\partial \alpha} - \frac{\partial H}{\partial y} \frac{\partial}{\partial \beta} + \frac{\partial}{\partial z}.$$

We can find a 2-form aligned along this direction

$$d\alpha dx + d\beta dy - \left( \frac{\partial H}{\partial \alpha} d\alpha + \frac{\partial H}{\partial \beta} d\beta + \frac{\partial H}{\partial x} dx + \frac{\partial H}{\partial y} dy \right) \wedge dz.$$

This is just our old friend

$$\Omega = dp dq - dH dz,$$

and we clearly have

$$v \lrcorner \Omega = 0.$$

A 4-form aligned along the rays will be  $\Omega \wedge \Omega$ .

The phase-space density of a four-parameter set of rays is called *brightness*. The scalar brightness  $B$  is such that the 4-form describing the density of rays is  $B\Omega \wedge \Omega$ . If rays are conserved, then we must have

$$d(B\Omega \wedge \Omega) = 0.$$



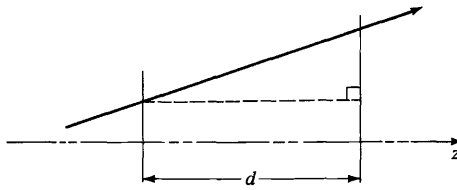


Figure 52.2.

We have

$$(\mathcal{L}_v B)\Omega \wedge \Omega = \mathcal{L}_v(B\Omega \wedge \Omega) = v \lrcorner d(B\Omega \wedge \Omega) + d(v \lrcorner B\Omega \wedge \Omega) = 0.$$

The scalar brightness  $B$  is constant along a ray.

Associated with this brightness is an integral

$$E = \int_{\Sigma} B\Omega \wedge \Omega,$$

where the surface  $\Sigma$  describes some four-parameter family of rays. If the surface  $\Sigma$  is carried along with the rays, the value of the integral does not change. If  $\Sigma$  contains all the rays passing through an optical instrument, then  $E$  is an invariant of the instrument called the *etendue* or throughput. The information-carrying capacity of an instrument depends on the *etendue*, there being  $E/\lambda^2$  picture elements, a number often called the space-bandwidth of the instrument. For an instrument with an oblique pupil, we would need the  $dH dz$  terms. These are often left out.

### Gaussian optics

Look at light spreading in vacuum from a small source. The propagation of a ray a distance  $d$  along the  $z$  axis is described by the transformation

$$(x, y, \alpha, \beta) \mapsto \left( x + \frac{\alpha d}{\sqrt{1 - \alpha^2 - \beta^2}}, y + \frac{\beta d}{\sqrt{1 - \alpha^2 - \beta^2}}, \alpha, \beta \right)$$

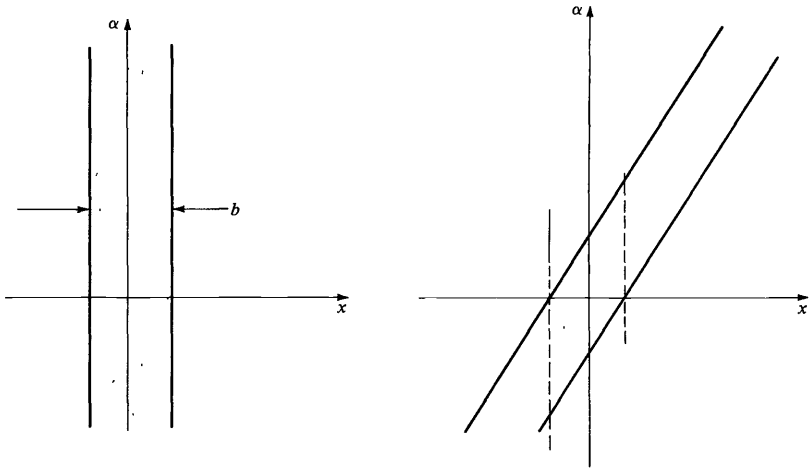
(see Figure 52.2). For small angles this is just the shear

$$(x, y, \alpha, \beta) \mapsto (x + \alpha d, y + \beta d, \alpha, \beta).$$

Look next at the passage of light through a thin lens in the limit of small angles. Again we have a shear

$$(x, y, \alpha, \beta) \mapsto (x, y, \alpha - x/f, \beta - y/f).$$

Both of these are obviously symplectic transformations; that is, they pre-



**Figure 52.3** (left). The rays emitted by a source of width  $b$ .

**Figure 52.4** (right). The rays intercepted by a receiver of finite width after the rays of Figure 52.3 have propagated some distance.

serve  $d\alpha dx + d\beta dy$ . The theory of optical instruments based on these approximate transformations is called Gaussian optics.

**Examples:** To see that the parameter  $f$  in the preceding equation is the focal length, take a two-parameter family of parallel rays

$$(u, v) \mapsto (x, y, \alpha, \beta) = (u, v, 0, 0).$$

After the rays pass through the lens, we have the family

$$(u, v, -u/f, -v/f).$$

Let the resulting beam propagate a distance  $d$ ; then we find

$$(u - ud/f, v - vd/f, -u/f, -v/f).$$

Indeed, after a distance  $d = f$ , this parallel beam is focused on the  $z$  axis.

To see how conservation of brightness is compatible with the inverse-square law, look at a beam emitted from a source of small size  $b$ . The rays now form a four-parameter family. For simplicity let us ignore the  $y$  direction, and derive a  $1/r$  law. The rays originally lie in the strip of width  $b$  shown in Figure 52.3. After the rays propagate a distance  $r$ , this strip is sheared, as shown in Figure 52.4. The

energy collected by a detector of fixed size in  $x$  is contained in the parallelogram shown. Since brightness is constant, the energy collected will be just the area of this parallelogram. And indeed, we see that this is proportional to  $1/r$ .

### *Perfect lens*

Our final application of Hamiltonian optics will be to show that a perfect lens is impossible, even allowing for a continuously variable index of refraction. We take from geometrical optics the result that any imaging system that maps straight lines into straight lines must obey exactly the thin-lens law

$$\frac{1}{f} = \frac{1}{s} + \frac{1}{d},$$

perhaps with allowance for thickness of the lens. This leads to a transformation

$$(x, \alpha) \mapsto (x, \alpha')$$

with  $\alpha'$  given by

$$\begin{aligned} \sin \theta &= \alpha, \\ \tan \theta' &= \tan \theta - x/f, \\ \alpha' &= \sin \theta'. \end{aligned}$$

Such a transformation is not a symplectic transformation. To preserve areas it must be a simple shear in  $\alpha$ , which the preceding is not. Thus the geometrical properties of an imaging system are incompatible with the behavior of bundles of light rays.

### **Problem**

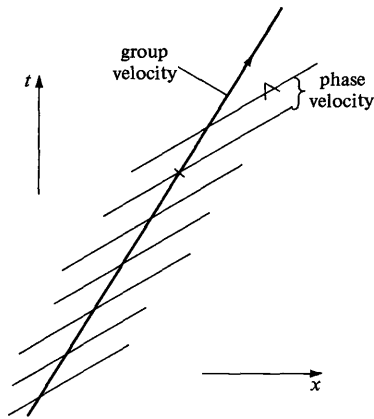
**52.1** Are there conservation laws for the intensity of point sources?

### **53. Dynamics of wave packets**

In this section I will discuss the high-frequency (WKB) limit of wave motion. This is a metatheory for Hamiltonian dynamics. It contains Hamiltonian dynamics as the frequency tends to infinity.

#### *Wave packets*

We start with a linear partial differential equation, or perhaps a set of them. We consider high-frequency solutions, where the phase of the wave



**Figure 53.1.** The geometric representation of a wave packet of dispersive waves in spacetime. The light lines are wavecrests, drawn only where the wave packet has a significant amplitude.

train changes rapidly compared with all other variables. We can make up lumps of these wave trains, lumps containing many waves and with slowly varying amplitude and frequency. These lumps I call wave packets. We will see that they behave very much like particles, but with the additional physical phenomenon of interference.

These wave packets are described geometrically by sections of the tangent and cotangent bundles of spacetime. The wave packet has a 1-form defined at every point, the gradient of its phase, which describes the alignment and spacing of its wavecrests. This 1-form is the geometric representation of phase velocity. The wave packet also has a vector showing its motion through spacetime, the group velocity. Many elementary books try to treat phase velocity as a vector. This is possible only in two-dimensional spacetime. Phase velocity is really a 1-form. Our view of a wave packet is sketched in Figure 53.1.

### *Dispersion relation*

The original partial differential equation constrains the phase velocity of the wave packet to a hypersurface in the cotangent space.

**Example:** Look at the ordinary wave equation

$$\frac{\partial^2 \psi}{\partial t^2} = \frac{\partial^2 \psi}{\partial x^2};$$

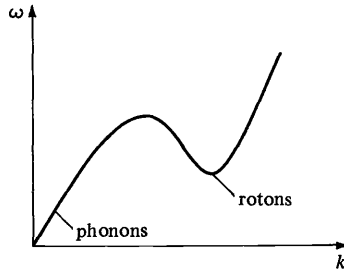


Figure 53.2. The dispersion relation for waves in liquid helium.

and consider solutions of the form

$$\psi \sim A \exp[i(\omega t + kx)],$$

where  $A$ ,  $k$ , and  $\omega$  can all be slowly varying functions. The derivatives of  $\psi$  will have the phase gradients as their largest terms, here

$$(k^2 - \omega^2)A \exp[i(\omega t + kx)],$$

and the phase gradient 1-form

$$P = P dq = \omega dt + k dx$$

must satisfy

$$k^2 - \omega^2 = 0.$$

The wave number  $k$  used here has a sign opposite to that usually given. I do this so that  $x$  and  $t$  can be treated uniformly. This constraint on the phase gradient is a function  $W: T^*M \rightarrow \mathbb{R}$  called the dispersion relation. Wave equations can be characterized by their dispersion relations. Deep-water waves on still water satisfy

$$\omega^4 - g^2 k^2 = 0.$$

Elastic waves on a beam with both tension and stiffness satisfy

$$\omega^2 - k^2 + k^4 = 0$$

(I have left out material constants here). Figure 53.2 shows the dispersion relation for some quantum waves in liquid helium. Quantum wave packets are called quasiparticles, here, phonons and rotons.

### Dynamics

To find one of these high-frequency solutions to a wave equation, we must solve the following geometric problem. Find a section  $P$  of the

cotangent bundle such that it satisfies the dispersion relation  $W$ ,

$$W \circ P = 0$$

such that

$$dP = 0. \quad (53.1)$$

The latter condition comes from the fact that  $P$  is already the gradient of the phase

$$P = d\phi.$$

This geometric problem can readily be translated into an exterior differential system. The section  $P$  must be an integral submanifold of the ideal generated by

$$dp \, dq \quad \text{and} \quad dW. \quad (53.2)$$

The relation between equations 53.1 and 53.2 was shown in Section 51. The section  $P$  is a Lagrangian submanifold.

This exterior system has a Cauchy characteristic vector

$$u = \frac{\partial W}{\partial p} \frac{\partial}{\partial q} - \frac{\partial W}{\partial q} \frac{\partial}{\partial p} \quad (53.3)$$

satisfying

$$u \lrcorner dp \, dq = 0, \quad u \lrcorner dW = 0.$$

Given initial conditions on the wave packet, this vector generates the section, i.e., the dynamics of the wave packet. The projection of  $u$  onto spacetime is the group velocity vector.

We see from this that these wave packets behave like particles following Hamiltonian dynamics. The relation to our classical dynamics picture is clear. The group velocity has no particular scale, and is really a line element. This is the Lagrangian picture in  $CQ$ . The Hamiltonian picture does not see the 1-form, the frequency having gone to infinity. This is the contact element in  $C^*Q$ . Thus we finally see a physical interpretation for the  $P$ s in Hamilton's equations; they are the residue of the phase velocity as the frequency goes to infinity.

#### *Wave diagram*

So far the group velocity has been specified by only a line element. We can use a vector to represent the group velocity in spacetime if we pick a normalization. We will take

$$u \lrcorner P \, dq = 1. \quad (53.4)$$

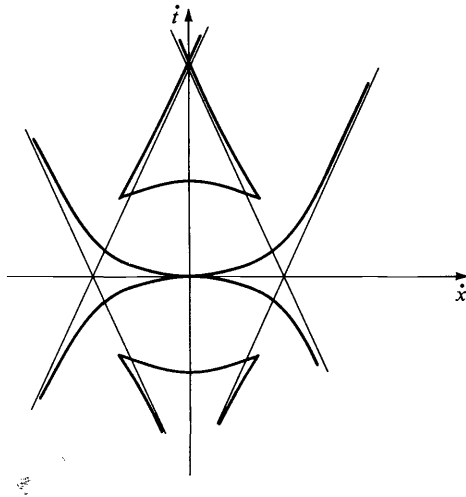


Figure 53.3. Wave diagram for waves on an elastic beam.

This leads to a group-velocity vector

$$u = \frac{\partial W}{\partial p_i} \frac{\partial}{\partial q_i} \bigg|_p \frac{\partial W}{\partial p}.$$

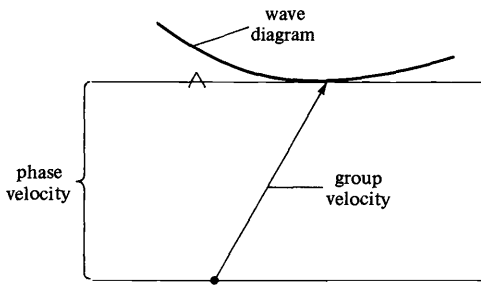
The dispersion relation is a hypersurface in the cotangent space. Corresponding to this there is the set of group-velocity vectors in the tangent space, each phase velocity being associated with a unique group-velocity vector. This set I call the *wave diagram*. Although the dispersion relation is a fairly tame surface, being the zero contour of a smooth function, the corresponding surface in the tangent space is not. It generally has cusps, asymptotes, and other interesting peculiarities.

**Example:** The wave diagram for an elastic beam is shown in Figure 53.3. The beam has bending stiffness, a compressive load, and transverse elastic support. The dispersion relation is

$$\omega^2 = k^4 - k^2 + 1.$$

Fortunately, the peculiarities of the wave diagram are all important physically, and so it is good that the wave diagram brings them to our attention.

**Examples:** An asymptote corresponds to a wave packet with no dispersion. The group velocity then lies in the constant-phase sur-



**Figure 53.4.** The relation between group velocity and phase velocity provided by the wave diagram.

faces. Although the general dispersive wave pulse will spread with time into a nearly harmonic wavetrain, this is not true for modes around such nondispersive points.

A cusp corresponds to a turning point in velocity. In the preceding wave diagram, it was a local maximum velocity. If wave energy is smoothly distributed over frequencies, then this turning point will be very prominent. This occurs in rainbows.

From equation 53.3 it follows that these wave packets follow Hamiltonian dynamics. Using the normalization given in 53.4 gives us Hamilton's equations in the form

$$\frac{dq^i}{dt} = \frac{\partial W}{\partial p_i} \Big|_p \frac{\partial W}{\partial p},$$

$$\frac{dp_i}{dt} = - \frac{\partial W}{\partial q^i} \Big|_p \frac{\partial W}{\partial p}.$$

The group velocity and the phase velocity are related by the wave diagram as shown in Figure 53.4. The phase-velocity 1-form is the tangent to the wave diagram at the group-velocity vector.

**Example:** In Figure 53.5 I sketch the turning-point wave train for the elastic beam shown in Figure 53.3. Observe that the wavecrests are moving to the left, but the energy is flowing to the right.

In fact, the dispersion relation and the wave diagram are related by a Legendre transformation. In general this is singular at cusps and cannot be inverted blindly.



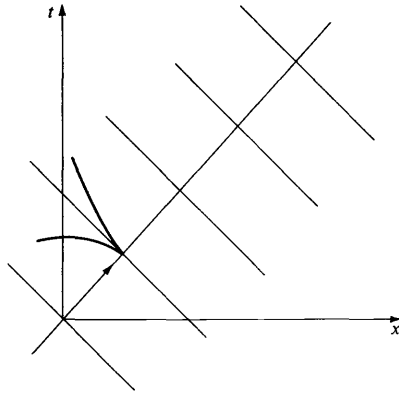


Figure 53.5. The wave train for the elastic beam at the turning point (cusp).

To see that the phase velocity is tangent to the wave diagram, we can argue directly, rather than making the Legendre transform. From the normalization in equation 53.4, we see that the unit contour of  $Pdq$  passes through the group-velocity vector. It will be tangent provided

$$(u + \Delta u) \lrcorner Pdq = 1$$

for a small change in group velocity  $\Delta u$ . This is equivalent to showing that

$$u \lrcorner \Delta Pdq = 0$$

for  $P + \Delta P$  a nearby phase velocity. We have

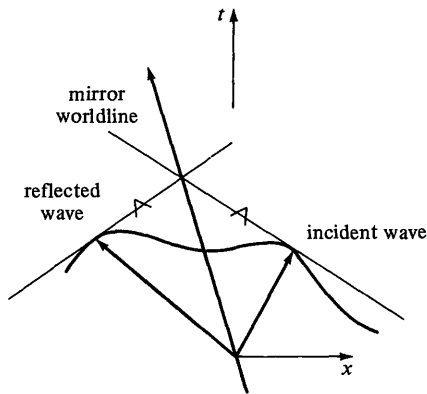
$$u \lrcorner \Delta P = \frac{(\partial W / \partial P_i) \Delta p_i}{p(\partial W / \partial p)}.$$

The numerator is the change in  $W$  along the dispersion relation. Since the dispersion relation is in fact specified by  $W = 0$ , this vanishes.

### Interactions

When a wave packet hits a mirror or interacts linearly with a perturbation or another wave packet, we must have the various disturbances all in phase over the many cycles that make up the wave packet. This leads to kinematic constraints on the interactions.

**Example:** For waves described by the squiggle shown in Figure 53.6, an incident wave train can only reflect off the moving mirror



**Figure 53.6.** Phase-velocity relations at a moving mirror.

as shown. The mirror must see the same local frequency on its surface (pullback of the phase-velocity 1-form) from each wave train.

Some further examples of wave-packet dynamics, including a discussion of relativity and the twin paradox for water waves, can be found in Burke (1980).

### Problem

**53.1** Derive the Hamiltonian for geometric optics from a dispersion relation.



# VIII

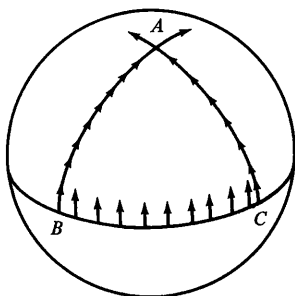
## *Calculus on fiber bundles*

Physical theories today are predominantly field theories. The geometric view of fields is that they are sections of fiber bundles. Handling the local relations (partial differential equations) between fields requires some care when we are working on manifolds, and the difficulty comes from an unexpected direction. We can do ordinary calculus with no trouble. The local affine approximations to these sections will be contact elements. This calculus is done in the jet bundle. The difficulty lies in deciding which contact element represents a constant field. A similar problem exists, say, for angles. Until we are given the geometric structure of a metric, the idea of an angle on a manifold does not exist.

**Example:** There is no problem in deciding what a constant real-valued function is. Thus it is easy to physically measure the temperature everywhere in a building, and mathematically easy to recognize when it is constant.

Not so for a field variable that is a vector. Because of this we are unable to define globally the state of rest on spacetime, and thus are not able to define, for example, the total energy of an expanding universe. Does the kinetic energy of the expansion count or not? This is not a problem of inadequate mechanics. There is no physical meaning to the question.

If there is to be physical meaning to the idea of a constant field, then it must be added to the bundle, and we need some mathematical structure to represent it. The appropriate mathematical structure is called a *connection*, and the calculus that it leads to is called (for shaky historical reasons) *covariant differentiation*. That name can be misleading. The



**Figure VIII.1.** A simple illustration that a 2-sphere does not allow a constant vector field to be consistently defined.

differentials of fields are well-defined and coordinate-independent. They lack only a reference value, the constant differential.

Note the preceding restriction to a *local* idea of a constant field. The useful concept here is not the idea of a globally constant field, but only that of a field with vanishing first derivative. In fact, a field constant over a loop of finite area may not exist.

**Example:** The embedding of the 2-sphere in Euclidean 3-space lets us define locally constant vector fields. In Figure VIII.1, I sketch a failed attempt to find a constant vector field along the (truly) right triangle  $ABC$ .

Constant vectors can be thought of as parallel. To pick vectors along a curve that are parallel is called *parallel transport*. The preceding example shows that a parallel transport may very well be path-dependent. Such connections are called *anholonomic*. One useful physical operation that is modeled by using the mathematical idea of parallel transport is the propagation of a wave packet of dispersive waves. The wave packet gives us a phase-gradient 1-form along the path of the wave packet. We can represent this physical structure using a connection, and this connection is the one used in general relativity.

In this chapter we will develop the theory of connections in the tangent bundle and its associated tensor bundles. All these bundles have a linear structure on their fibers, and the connection will be compatible with this linear structure. These connections are usually called affine connections, although they could better be called linear connections.

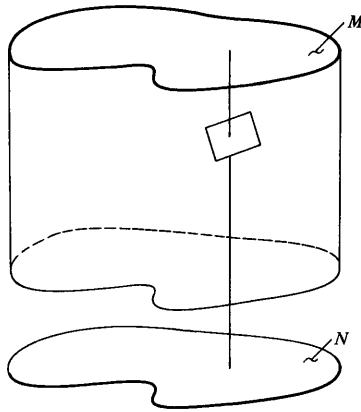


Figure 54.1. The geometric pattern of a connection.

## 54. Connections

A connection is the geometric structure used to answer the question: What is the constant field in the neighborhood of a given point? The geometric picture is shown schematically in Figure 54.1. The field is represented by a section of a fiber bundle. The local affine approximation to a section is a contact element. Taken literally, Figure 54.1 shows a field with one degree of freedom defined over a two-dimensional base space. If the field is a vector field or a 1-form field, then both the fibers and the base space have the same dimension. In Section 1 we discussed this geometric situation and found that it had an affine structure. Given two contact elements,  $b$  and  $c$ , we can find a unique parametrized line of contact elements passing through  $b$  and  $c$ . The situation does not have a linear structure, however. Given a contact element  $b$ , we have no way to single out a contact element  $k$  times  $b$ . An affine space can be made into a linear space only if a particular element is singled out as the origin. This is what a connection does. A connection is a choice at each point in the bundle of a contact element representing there the idea of a constant field. This contact element is often called horizontal. A section tangent to this contact element has no first derivative at the point. A bundle can have many different connections defined on it. A connection is a geometric structure, and different ones are used to model different physical situations, just as a vector field is a geometric structure used to model, for example, dynamical systems.

*Compatibility*

Our bundle may have other geometric structures as well, and usually we will want our connections to be compatible with these structures. The tangent bundle has a linear structure. A connection there should respect this linear structure. The sum of two constant fields should also be constant. Addition there is to be done fiber by fiber. Other structures that can arise are groups, group actions, and metric structure. In this chapter we will be concerned only with connections in the tangent bundle and its associated tensor bundles. By some miscarriage these are called affine connections, even though they preserve not just an affine structure on the fibers but also a linear structure.

**Example:** A physical realization of a connection on the tangent bundle to a curved 2-manifold such as an automobile fender is given by a sheet of graph paper. Pasted onto the surface it locally defines constant vector fields, such as the vectors on the sides of the squares. It only works in the neighborhood of the point, however. Over finite distances it wrinkles, and there is no unique way to squash it onto the surface.

*Representation of a connection*

There are two ways to represent a contact element. We can specify a linear subspace of the tangent space of the bundle by giving a set of tangent vectors that span the contact element. Alternatively, we can give an ideal of 1-forms that have the contact element as an integral submanifold. The number of tangent vectors is the dimension of the base space. The number of 1-forms is the dimension of the fiber.

**Example:** For a connection in the tangent bundle, it takes either  $n$  vectors or  $n$  1-forms to specify the connection.

If we take coordinates  $y^\mu$  for the fibers and  $x^i$  for the base space, then any contact element can be specified by the vectors

$$X_i = \frac{\partial}{\partial x^i} - f_i \frac{\partial}{\partial y}$$

or by the 1-forms

$$\alpha = dy + f_i dx^i.$$

The coefficients  $f_i(x, y)$  represent the connection.

For a linear connection these  $f_i$  coefficients must satisfy certain compatibility conditions. Suppose the  $y$  are linear coordinates for the fibers. Let  $x \mapsto (x, Y(x))$  and  $x \mapsto (x, Y'(x))$  be two sections, each constant at  $x = a$ . The sum of these is the section  $x \mapsto (x, Y + Y')$ . For the sections to be constant at  $x = a$ , we must have the pullbacks of the  $\alpha$  vanishing. Thus if we have

$$\left[ \frac{\partial Y}{\partial x^i} + f_i(Y) \right] dx^i = 0$$

and

$$\left[ \frac{\partial Y'}{\partial x^i} + f_i(Y') \right] dx^i = 0,$$

then for a linear connection we must have

$$f_i(Y) + f_i(Y') = f_i(Y + Y').$$

This plus scaling implies that the  $f_i$  can be written in terms of the linear coordinates in the form

$$f_i(x, Y) = \Gamma_{i\mu} Y^\mu,$$

and the connection can be specified by either the 1-forms

$$\alpha^\mu = dy^\mu + \Gamma^\mu_{j\nu} y^\nu dx^j, \quad (54.1)$$

or the vectors

$$X_i = \frac{\partial}{\partial x^i} - \Gamma^\mu_{i\nu} y^\nu \frac{\partial}{\partial y^\mu}.$$

The coefficients  $\Gamma^\mu_{j\nu}$  depend only on  $x$ . If the bundle is the tangent bundle, then they are the usual affine-connection components of tensor analysis.

Despite the fact that the  $\Gamma^\mu_{j\nu}$  are festooned with indices in a tensorial manner, they are *not* tensors. We can demonstrate this by calculating how the  $\Gamma^\mu_{j\nu}$  transform under a change in coordinates.

**Example:** In the tangent bundle, the  $y$  coordinates are usually denoted by  $\dot{x}$ . If we transform to new curvilinear coordinates

$$z^i = Z^i(x),$$

then the new fiber coordinates are given by

$$\dot{z}^i = \frac{\partial Z^i}{\partial x^j} \dot{x}^j.$$



The basis 1-forms transform according to

$$dz^i = \frac{\partial Z^i}{\partial x^j} d\dot{x}^j + \frac{\partial^2 Z^i}{\partial x^k \partial x^j} \dot{x}^j dx^k$$

and the new connection 1-forms can be written

$$\begin{aligned} dz^i + \bar{\Gamma}^i_{jk} \dot{z}^k dz^j \\ = \frac{\partial Z^i}{\partial x^j} dx^j + \bar{\Gamma}^i_{jk} \dot{x}^l dx^m \frac{\partial Z^k}{\partial x^l} \frac{\partial Z^j}{\partial x^m} + \frac{\partial^2 Z^i}{\partial x^k \partial x^j} \dot{x}^j dx^k. \end{aligned} \quad (54.2)$$

The tensorial nature of the  $\Gamma^i_{jk}$  is ruined by the last term on the right-hand side, which prevents the new components from being linear functions of the old components. Note that it is an affine function. Perhaps this is the historical reason for calling it an affine connection rather than a linear connection. The old-timers were obsessed with these transformation properties. In Section 56 we will see the use of these transformation equations.

Since for a linear connection the  $\Gamma^i_{jk}$  depend only on  $x$ , the 1-forms

$$\omega^i_k = \Gamma^i_{jk} dx^j$$

are the pullbacks of 1-forms on the base space. They do depend on the choice of the coordinates  $x$  and  $y$ . They are called connection 1-forms, and often are an efficient representation of a linear connection. The entire set of them is, of course, coordinate-independent.

#### *The linear connection as a derivation*

A linear connection on the tangent bundle can be introduced abstractly as an operator mapping pairs of vector fields to another vector field,

$$A \times B \mapsto C = \nabla_A B,$$

satisfying, for smooth functions  $f$  and  $g$

$$\nabla_{(fA+gB)} C = f\nabla_A C + g\nabla_B C, \quad (54.3)$$

$$\nabla_A (B+C) = \nabla_A B + \nabla_A C,$$

and

$$\nabla_A (fB) = f\nabla_A B + (df \cdot A)B. \quad (54.4)$$

The last equation forces the operator  $\nabla_A \cdot$  to recognize only linear terms in a Taylor's Series; it is a form of Liebniz rule.

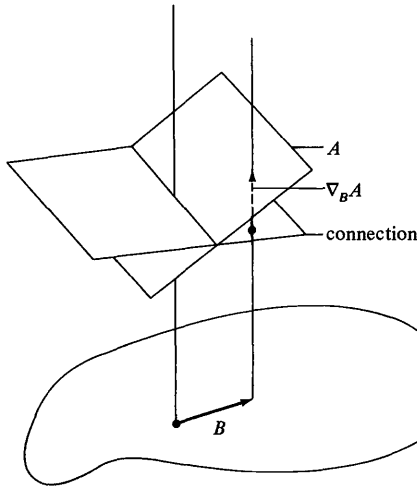


Figure 54.2. The relation between a connection and its associated covariant derivative.

Note that the operator

$$\nabla: A: B \mapsto \nabla_B A$$

is a tensor, but  $\nabla_A \cdot$  is not. The geometric picture of  $\nabla \cdot A$  is shown in Figure 54.2. Clearly the piece cut off from the vector spaces by the two planes will be linear. This also shows that the difference of two connections is a tensor.

We can represent  $\nabla$  by giving its action on basis vector fields. The coefficients  $\Gamma^k_{ij}$  in the expansion

$$\nabla_{\partial/\partial x^i} \frac{\partial}{\partial x^j} = \Gamma^k_{ij} \frac{\partial}{\partial x^k}$$

are sufficient to compute the action of  $\nabla$  on any vector field. For a locally constant vector field,  $Y^j(\partial/\partial x^j)$  we have

$$\nabla_{\partial/\partial x^i} Y^j \frac{\partial}{\partial x^j} = \left[ \frac{\partial Y^j}{\partial x^i} + \Gamma^j_{ik} Y^k \right] \frac{\partial}{\partial x^j} = 0$$

and comparing this with equation 54.1 we see that these are the same  $\Gamma^i_{jk}$  defined earlier. On the tangent bundle the  $Y$  could be called  $\dot{x}$ .

The abstract definition of a connection is efficient, but we need the insight of the geometric definition to fully appreciate the structure that is being introduced.

## 55. Parallel transport

Given a connection on a fiber bundle and a curve in the base space, we can find fields that are constant along the curve. If we are in the tangent bundle, then the field is a field of vectors. If we are given one vector on the curve, all the rest are determined. This operation is called *parallel transport*.

The geometric picture is sketched in Figure 55.1 for a space with one-dimensional fibers. For a given curve  $u \mapsto X(u)$ , and a bundle  $\pi: (x, y) \mapsto (x)$  whose fibers are  $n$ -dimensional, we look in the sheet of fibers over the curve, with coordinates  $(u, y)$ . If we pull back the connection 1-forms

$$\omega = dy + f_i dx^i$$

onto this sheet, we have  $n$  1-forms in  $n+1$  dimensions. This is a line-element system that can be integrated by the method of characteristics. From any starting value, the solutions to this line-element system determine a field that is constant along the given curve. We have

$$(u, y) \mapsto (X(u), y)$$

and so the pullback is

$$dy + f_i \frac{dX^i}{du} du$$

and the characteristic equations are

$$\frac{dy}{du} + f_i \frac{dX^i}{du} = 0.$$

For a linear connection we have

$$\frac{dy^\mu}{du} + \Gamma^\mu_{i\nu} y^\nu \frac{dX^i}{du} = 0.$$

The operation of parallel transport is often the physical operation actually modeled by a connection. An inertial navigation system uses gyroscopes to parallel-transport a reference frame along the worldline of a plane or missile. Often the operation of parallel transport can be used to find the components of the connection.

If we have a connection on the tangent bundle, then we can define special curves called *geodesics* for which the tangent vector is parallel-transported:

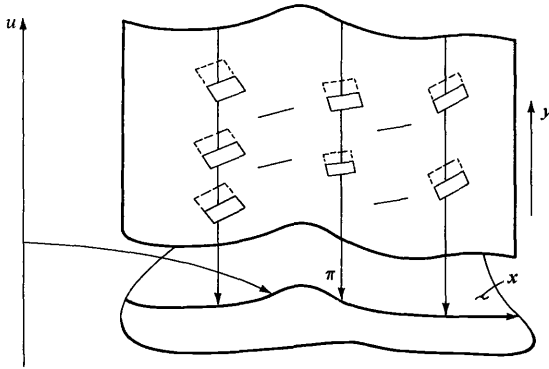


Figure 55.1. Parallel transport defined by a connection.

$$\frac{d^2 X^i}{du^2} + \Gamma^i_{jk} \frac{dX^j}{du} \frac{dX^k}{du} = 0.$$

This equation also singles out an affine family of parameters for the curve.

A connection is fully specified by the law of parallel transport. At a given point in the bundle, lift all the curves passing through the base point, using parallel transport. They have a unique contact element, which is the horizontal subspace of the connection.

#### *The metric connection*

If a manifold has metric structure, then this defines geodesics and a law of parallel transport. We can work out explicitly the linear parallel transport on a 2-sphere embedded in Euclidean 3-space. Use the tangent cylindrical strip to define parallel transport. Along any meridian we have

$$\nabla_{\partial/\partial\theta} \frac{\partial}{\partial\theta} = 0.$$

There is no reason for the vector to deviate to one side in preference to the other. This is really the original idea of a geodesic: it divides the sphere into two equal parts. Going away from the pole, the vector  $\partial/\partial\phi$  increases its length. The constant vector in its direction is given by

$$\nabla_{\partial/\partial\theta} \left( \frac{1}{\sin\theta} \frac{\partial}{\partial\phi} \right) = 0;$$

that is,

$$\nabla_{\partial/\partial\theta} \frac{\partial}{\partial\phi} = \frac{\cos\theta}{\sin\theta} \frac{\partial}{\partial\phi}.$$

The arguments in the  $\phi$  direction are more complicated. The results are

$$\nabla_{\partial/\partial\phi} \frac{\partial}{\partial\phi} = -\sin\theta \cos\theta \frac{\partial}{\partial\phi}$$

and

$$\nabla_{\partial/\partial\phi} \frac{\partial}{\partial\theta} = \frac{\cos\theta}{\sin\theta} \frac{\partial}{\partial\phi}.$$

The equations of parallel transport are

$$\frac{dy^\theta}{du} - \sin\theta \cos\theta y^\phi \frac{d\phi}{du} = 0 \quad (55.1)$$

and

$$\frac{dy^\phi}{du} + \frac{\cos\theta}{\sin\theta} \left( y^\theta \frac{d\phi}{du} + y^\phi \frac{d\theta}{du} \right) = 0. \quad (55.2)$$

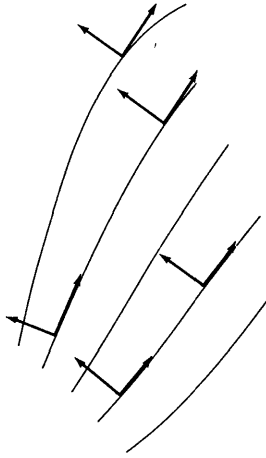
Note that parallel transport on  $S^2$  is path-dependent. In Section 60 we will find an easier and safer way to find these connection components.

### *The geodesy connection*

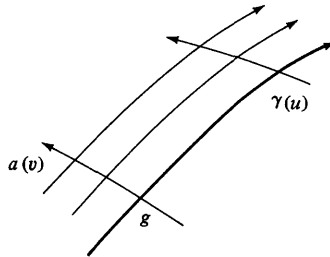
An instructive linear connection can be defined as follows. Take  $\mathbb{R}^2$  with its standard Euclidean metric. Take any smooth congruence of curves filling  $\mathbb{R}^2$ , one curve through every point. At every point form a reference frame by pointing one vector along the curve, and give it unit Euclidean length. Take the second vector at a Euclidean right angle to the first (Figure 55.2). Let these vectors define parallel transport by asserting that they are parallel-transported. For a linear connection this specifies the parallel transport. In Grossman (1974) this is used to model long-distance surveying, where the Euclidean geometry is used only to describe the construction of the sighting instruments (transits, etc.). The curved lines model astronomical observations of a star through a refracting atmosphere. Laser surveying has made this mathematical model less useful, but it is a fine example of a connection. It was discussed earlier in Schouten (1954). Note that parallel transport here is path-independent.

### *Lie-group connections*

At any point in a Lie group, there are two natural definitions of parallel transport. Let us transport the vector  $A$  from the point  $g$  along the curve  $\gamma(u)$ . Do it as follows. Pick any curve having  $A$  as a tangent vector, say,



**Figure 55.2.** A connection representing some features of geodetic surveying.



**Figure 55.3.** One of the two natural connections on a Lie group.

$a(v)$ , with  $a(0) = g$ . The curve

$$v \mapsto \gamma(u)g^{-1}a(v)$$

defines a tangent vector at  $\gamma(u)$  that is called the left translation of  $A$ . The curve

$$v \mapsto a(v)g^{-1}\gamma(u)$$

defines a vector at  $\gamma(u)$  called the right translation of  $A$ . Juxtaposition here indicates group multiplication (see Figure 55.3). Either of these laws of transport can be used to define a connection. These will not be linear connections, but rather ones compatible with the group structure.

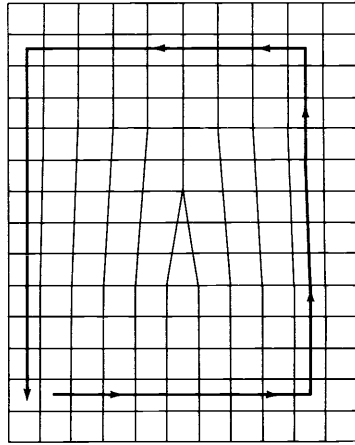


Figure 55.4. A connection representing a dislocation in a crystal.

### *The crystal connection*

We can use the regular lattice structure of a crystal to define a law of parallel transport. Let us consider a limit where the lattice spacing is far smaller than our regions of interest, and where the lattice has a density of dislocation defects that is small compared with the lattice size (nearly perfect crystal), but large enough so that over regions of interest the dislocations can be described by a continuous density.

Look at a circuit surrounding a single dislocation (Figure 55.4). The instructions, “Go  $k$  steps, turn, and repeat three times,” generate not a closed path, but one that has a gap of exactly one lattice spacing. A large circuit will have an error of closure proportional to the number of dislocations that it contains, that is, proportional to its area. In the next section we will see how to model this property of a connection. [See Bilby (1955) for more on this.]

### **Problem**

**55.1** Study the construction called Schild’s Ladder in Misner, Thorne, and Wheeler (1973) and criticize it.

## **56. Curvature and torsion**

The geometric structure of a connection can be represented by the ideal of differential forms

$$dy + f_i dx^i.$$

For any representation there arises the problem of equivalence. When are two different representations really representing the same geometric structure? Here different representations of a connection could arise from different base-space or fiber coordinates, different generators for the ideal, and so on. One approach to the equivalence problem is the study of invariants, properties that are independent of the representation. For connections the most important invariant is its curvature. In general it is the only local invariant. For connections on the tangent bundle there is another local invariant called torsion.

### *Curvature*

The curvature of a connection reflects the path-dependence of its parallel transport. If the parallel transport law of a connection is path-independent, then the horizontal contact elements line up to form integral submanifolds of the same dimension as the base space. Equivalently, the ideal of connection 1-forms would then be closed under exterior differentiation.

**Example:** The metric connection on  $S^2$  given in the last section has connection forms that can be directly read off from equations 55.1 and 55.2. Calling the fiber coordinates  $\theta$  and  $\phi$ , we have generators

$$d\dot{\theta} - \sin \theta \cos \theta \dot{\phi} d\phi$$

and

$$d\dot{\phi} + \frac{\cos \theta}{\sin \theta} (\dot{\theta} d\phi + \dot{\phi} d\theta).$$

To see that here parallel transport is path-dependent, just take the exterior derivative of the first generator

$$(\sin^2 \theta - \cos^2 \theta) \dot{\phi} d\theta d\phi - \sin \theta \cos \theta d\dot{\phi} d\phi.$$

To see that this is not in the ideal generated by the connection forms, use them to eliminate  $d\dot{\theta}$  and  $d\dot{\phi}$ . This leads to

$$\sin^2 \theta \dot{\phi} d\theta d\phi,$$

which is clearly not in the ideal. Any further use of the ideal brings in a  $d\dot{\theta}$  or a  $d\dot{\phi}$  term.

I am going to discuss curvature only for connections on the tangent bundle. The general framework, suitable for gauge theories, would require the



momentary introduction of principal bundles, associated bundles, and a great deal of formalism. The pedestrian approach that I give has the virtue of simplicity and may even seem reasonably motivated.

We will study parallel transport around infinitesimal paths, following closely the discussion of the Lie bracket in Section 9. The infinitesimal circuit will be similar. The sides will now be geodesics rather than integral curves. To keep opposite sides as parallel as possible, both tangent vectors must be parallel-transported along the sides. (You can see how we might reasonably want to introduce a bundle whose fibers were frames at each point.) The gap in the diagram must be at the origin, of course. Other diagrams are meaningless. We will discuss parallel transport around this circuit in two stages. First, we will discuss the closure of the circuit; then we will discuss the twists undergone by an arbitrary vector parallel-transported around the circuit. The circuit is of linear size  $\sqrt{\epsilon}$ , and the effects that we calculate are  $O(\epsilon)$ . The discussion will be simplified greatly by a careful choice of coordinates. We first discuss these special coordinates.

### *Gaussian coordinates*

Pick a particular point  $P$  in a manifold  $M$ , and a representation for the tangent space at  $P$ , a basis  $\partial/\partial q$ . If the space has a connection on its tangent bundle, then we can find particularly simple coordinates for the neighborhood of  $P$  as follows. For each vector  $\dot{q}(\partial/\partial q)$  in the tangent space at  $P$ , find the geodesic through  $P$  with that vector as tangent vector. Follow that geodesic for one unit of parameter change, to some point in  $M$ . This is a map  $T_P(M) \rightarrow M$ , called the exponential map. Until the geodesics start to cross, the numbers  $\dot{q}$  can be used as coordinates. These are called Gaussian or normal coordinates.

In these coordinates, a geodesic through  $P$  has the simple form

$$s \mapsto q = \dot{q}s.$$

A glance at the geodesic equation

$$\frac{d^2 q^i}{du^2} + \Gamma^i_{jk} \frac{dq^j}{du} \frac{dq^k}{du} = 0$$

shows that at  $P$  the symmetric part of  $\Gamma$  must vanish,

$$\Gamma^i_{jk} + \Gamma^i_{kj} = 0.$$

The derivatives at  $P$  will not necessarily have this symmetry, nor can it be imposed simultaneously at every point of  $M$ .

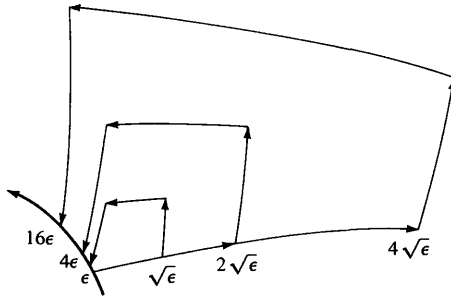


Figure 56.1. The failure of a geodesic square to close to terms linear in the area indicates torsion in a connection.

We will use these special coordinates to simplify proofs. First we will prove the desired relation in Gaussian coordinates. Then we will express the result in terms of tensors or operations that are manifestly coordinate-independent. Such a result must hold in every coordinates system if it holds in any one. For safety, I will usually use a modified equals sign, one with a superposed \*, to indicate relations true only in Gaussian coordinates.

*Torsion*

The failure of a circuit to close to linear terms in  $\epsilon$  is called *torsion*. Going around circuits of different sizes generates a curve of points (see Figure 56.1), whose tangent vector represents the linear error of closure. We expect to represent torsion by an operator  $T(X, Y)$  that takes in two vectors  $X$  and  $Y$ , representing the sides of the circuit, and yields the mismatch vector.

We describe the circuit in Gaussian coordinates. The first step is to move in the  $X$  direction for a parameter increment of  $+\sqrt{\epsilon}$ . Let the coordinates of the starting point be zero for simplicity. Then we have this leg of the circuit given by the curve

$$s \mapsto \sqrt{s} X^i + \frac{1}{2} s \Gamma^i_{jk} X^j X^k.$$

Using Gaussian coordinates will eliminate the term linear in  $s$ . Next we parallel-transport  $Y$  out to the end of this first leg, to find a vector with components

$$Y^i + \sqrt{\epsilon} \Gamma^i_{jk} Y^k X^j. \tag{56.1}$$

We will need these vectors only to a precision of  $\sqrt{\epsilon}$ .

Now proceed in the direction of the vector in equation 56.1 for a parameter increment of  $+\sqrt{\epsilon}$ . This moves us to the point

$$\sqrt{\epsilon} X^i + \sqrt{\epsilon} [Y^i + \sqrt{\epsilon} \Gamma_{jk}^i Y^k X^j].$$

Even though the symmetric part of the connection vanishes only at a single point, things are still simplified because the symmetric part is of order  $\sqrt{\epsilon}$ .

Now we must parallel-transport  $X$  out to the end of the second leg. To order  $\sqrt{\epsilon}$  this is

$$X^i + \sqrt{\epsilon} \Gamma_{jk}^i X^k Y^j.$$

We now move back a distance  $\sqrt{\epsilon}$  along a geodesic in a direction opposite to this. This takes us to the point

$$\sqrt{\epsilon} Y^i + \epsilon \Gamma_{jk}^i Y^k X^j - \epsilon \Gamma_{jk}^i X^k Y^j.$$

Now we need to carry  $Y^i$  out to this point. We find that it is just  $Y^i$ . The final point of the circuit is found by a retreat for a parameter increment  $\sqrt{\epsilon}$  in the  $-Y$  direction, to the point

$$\epsilon(\Gamma_{jk}^i - \Gamma_{kj}^i) Y^k X^j.$$

As advertised, there is a mismatch that is linear in  $\epsilon$ . It is due to the anti-symmetric part of the connection. A connection without torsion is symmetric, and is often called a symmetric connection.

To complete our argument we need to show that this operation is a tensor. To do this write this operation in the following tricky fashion. Start with

$$T(X, Y) \doteq (\Gamma_{jk}^i - \Gamma_{kj}^i) Y^k X^j \frac{\partial}{\partial X^i}. \quad (56.2)$$

The modified equals sign is used to show that this result is proven so far only for Gaussian coordinates. Now use

$$\nabla_X Y = Y^i_{,j} X^j + \Gamma_{jk}^i Y^k X^j$$

and

$$[X, Y] = Y^i_{,j} X^j - X^i_{,j} Y^j$$

to write this

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]. \quad (56.3)$$

So far, so good; the operator  $T$  is coordinate-independent, but is it a tensor? The right-hand side is defined only for  $X$  and  $Y$ , which are vector

fields. If  $T$  is a tensor, then any vector fields that agree with  $X$  and  $Y$  at the point of interest will do. It must depend only on the values of  $X$  and  $Y$  at a point. This will be true provided that

$$T(fX, Y) = fT(X, Y)$$

for  $f$  any function. This is similar to the argument in Section 9. The result is not obvious, since several derivatives of  $f$  occur. In fact, it is an easy consequence of equations 54.4 and 9.1, and  $T$  is indeed a tensor, of type  $\binom{1}{2}$ . Since the antisymmetric components of the connection do form a tensor, no coordinate change can eliminate them if they are nonzero.

Thus both our original description in terms of geodesics and our final result in terms of the tensor  $T(X, Y)$  make no reference to the Gaussian coordinates used to simplify the calculation. Thus the result is true in general. In fact, using equation 56.3 we can show that equation 56.2 holds in all coordinate systems, not just in Gaussian coordinates. We can also see directly from the transformation law given in equation 54.2 that it is a linear transformation law for the antisymmetric part of  $\Gamma$ ; hence that the antisymmetric part is a tensor.

**Examples:** The geodesy connection of the last section is a connection with nonzero torsion. To see this, let  $X$  and  $Y$  be the two constant vector fields defined by the special curves. Thus we have

$$\nabla_X Y = 0, \quad \nabla_Y X = 0,$$

but unless the curves are parallel, we will not have

$$[X, Y] = 0$$

and so  $T(X, Y)$  is nonzero.

Every dispersive wave system has its wave packets tracing out the geodesics of a linear connection with no torsion. The system of geodesics is equivalent to equations

$$\frac{d^2 x^i}{du^2} + \Gamma^i_{jk} \frac{dx^j}{du} \frac{dx^k}{du} = 0,$$

and these are insensitive to the torsion, which can be set equal to zero. Remember, to find torsion it was necessary to parallel-transport a pair of directions around the circuit, as well as follow geodesics.

The crystal connection of the last section has a torsion tensor proportional to the density of dislocations.

Torsion does not play a large role in our applications. Still, we need to know what it is in order to assert that it is zero. Also, it is good practice for the curvature calculation to follow.

### Curvature tensor

Proceed now to the second phase of the study of the infinitesimal circuit, and discuss the change in a vector parallel-transported around the  $\sqrt{\epsilon}$  circuit defined previously. The operation will be represented by a map  $R(\cdot, \cdot)\cdot$  and if  $X$  and  $Y$  define the circuit, and  $Z$  is the vector transported around the circuit, then  $R(X, Y)Z$  is the linear rate of change of the transported vector. For a linear connection on the tangent bundle, all of these vectors are tangent vectors. Unlike torsion, curvature can be defined for any linear connection in this same fashion, although now the field variable  $Z$  and its rate of change are not in the tangent space with  $X$  and  $Y$ .

The calculation proceeds just as for torsion, although more precision must be retained in the transport of  $Z$ . The calculation involves more tedious expansions. The result in special coordinates is

$$R(X, Y)\cdot Z \cong (\Gamma^i_{jk,l} - \Gamma^i_{lk,j})Z^k X^l Y^j. \quad (56.4)$$

Just as with torsion, this can be written in a manifestly covariant form

$$R(X, Y)\cdot Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X, Y]} Z. \quad (56.5)$$

In this form we can check that  $R(\cdot, \cdot)\cdot$  is a linear operator on *all* of its arguments. This tensor is called the *Riemann curvature tensor*, of type  $\binom{1}{3}$ . Its components are given by

$$R(X, Y)\cdot Z = R^i_{jkl} Z^j X^k Y^l$$

with

$$R^i_{jkl} = \Gamma^i_{lj,k} - \Gamma^i_{kj,l} + \Gamma^i_{km} \Gamma^m_{lj} - \Gamma^i_{lm} \Gamma^m_{kj}. \quad (56.6)$$

Be careful. There are several different sign conventions for the Riemann tensor and its contractions. I am following Misner, Thorne, and Wheeler (1973).

### Problems

- 56.1** Give a direct treatment of the curvature of 2-manifolds that are surfaces of revolution.
- 56.2** Compute the torsion tensor for the crystal connection of Figure 55.4.

## 57. Covariant differentiation

In the tangent bundle the operation

$$\nabla: A: B \mapsto \nabla_B A$$

is a tensor of type  $\binom{1}{1}$ . This follows from equation 54.3. This  $A$ -dependent tensor is called the *covariant derivative* of  $A$ . We define its components by

$$\nabla_B A = A^i{}_{;j} B^j \frac{\partial}{\partial x^i}.$$

It depends on the connection, and we could as well call it *a* covariant derivative. We have

$$A^i{}_{;j} = A^i{}_{,j} + \Gamma^i{}_{jk} A^k. \quad (57.1)$$

A geometric picture of this was given in Figure 54.2.

*Connection on  $T^*M$* 

Once we are given a linear connection on  $TM$ , we can find a natural connection on all of the associated tensor bundles. We demand compatibility of all these connections with the linear structure, and this is sufficient to determine the connection uniquely. Thus we will demand that the covariant derivative of a 1-form be compatible with the linear operation of that 1-form on any tangent vector. Using the same semicolon notation, we demand a Leibniz rule

$$(a^i \omega_i)_{;j} = a^i{}_{;j} \omega_i + a^i \omega_{i;j},$$

and compatibility of covariant differentiation and partial differentiation for scalar functions  $f$  (a scalar is a tensor of type  $\binom{0}{0}$ ),

$$f_{;i} = f_{,i}.$$

From these we deduce that

$$a^i \omega_{i;j} = a^i \omega_{i,j} - a^k \Gamma^i{}_{jk} \omega_i$$

for all  $a^i$ . Relabel the second term, and cancel the  $a^i$  to find

$$\omega_{i;j} = \omega_{i,j} - \Gamma^k{}_{ji} \omega_k. \quad (57.2)$$

In a similar fashion a connection can be put on any of the tensor bundles associated with a given vector bundle. The index-intensive expressions are easy to write down. There will be a correction term like that

in equation 57.1 for each contravariant index, and a correction term like that in equation 57.2 for each covariant index.

**Example:** For a second rank, covariant tensor we have

$$h_{ii;k} = h_{ij,k} - \Gamma^l_{ki} h_{lj} - \Gamma^l_{kj} h_{il}.$$

### *Symmetric connections*

We will deal almost exclusively with torsion-free connections on the tangent bundle. These are called symmetric connections, because their coefficients satisfy

$$\Gamma^i_{jk} = \Gamma^i_{kj}. \quad (57.3)$$

From equation 56.3 we therefore have

$$\nabla_X Y - \nabla_Y X = [X, Y].$$

Considering the close relation of the bracket operation to Lie derivatives, it is not too surprising that a similar simplification occurs also for Lie derivatives. In equation 18.3, all the partial derivatives can be replaced by covariant derivatives without changing the result. Similarly for exterior derivatives; we have, for example,

$$\omega_{i,j} - \omega_{j,i} = \omega_{i;j} - \omega_{j;i},$$

which follows easily from equation 57.3.

### *Commutation of covariant derivatives*

The covariant derivative  $a^i_{;j}$  is a tensor and so a further covariant derivative of it can be taken. As with repeated partial derivatives, we abbreviate this

$$a^i_{;jk} = (a^i_{;j})_{;k}.$$

The symmetry of partial derivatives

$$a^i_{;jk} = a^i_{;kj}$$

suggest that there should be some similar relation for repeated covariant derivatives.

If we write out equation 56.5 in covariant-derivative notation, we have

$$\begin{aligned} R^i_{jkl} Z^j X^k Y^l \\ = (Z^i_{;j} Y^j)_{;k} X^k - (Z^i_{;k} X^k)_{;j} Y^j - Z^i_{;j} Y^j_{;k} X^k + Z^i_{;k} X^k_{;j} Y^j \end{aligned}$$

and expanding the covariant derivatives using Leibniz's rule, we have

$$R^i{}_{jkl}Z^jX^kY^l = (Z^i{}_{;lk} - Z^i{}_{;kl})Y^lX^k.$$

Since this is true for all  $X$  and  $Y$ , we have

$$Z^i{}_{;kl} - Z^i{}_{;lk} = R^i{}_{jlk}Z^j. \quad (57.4)$$

### Bianchi identity

Since equation 56.6 is quadratic in the undifferentiated terms, Gaussian coordinates also simplify the first derivatives

$$R^i{}_{jkl;m} \stackrel{*}{=} \Gamma^i{}_{jl,km} - \Gamma^i{}_{jk,lm}.$$

Using the symmetry of partial derivatives, we have

$$R^i{}_{jkl;m} + R^i{}_{jmk;l} + R^i{}_{jlm;k} = 0,$$

which is true in all coordinates, because it involves tensors. This is called the *Bianchi identity*.

## 58. Metric connections

A linear connection compatible with a metric on the fibers is called a *metric connection*. Compatibility with a metric determines a unique symmetric connection.

Compatibility demands that the dot product of two vectors that are parallel-transported be constant. This in turn requires that the covariant derivative of the metric

$$\mathcal{G} = g_{ij} dx^i \otimes dx^j$$

vanish:

$$g_{ij,k} - \Gamma^l{}_{ki}g_{lj} - \Gamma^l{}_{jk}g_{il} = 0. \quad (58.1)$$

We can relabel this equation

$$g_{jk,i} - \Gamma^l{}_{ij}g_{lk} - \Gamma^l{}_{ki}g_{jl} = 0,$$

and again

$$g_{ki,j} - \Gamma^l{}_{jk}g_{li} - \Gamma^l{}_{ij}g_{kl} = 0.$$

If we add the first and third of these, subtract the second, and use the symmetry of the  $g_{ij}$  and the  $\Gamma^i{}_{jk}$ , we find

$$\Gamma^i{}_{jk} = \frac{1}{2}g^{il}(g_{lj,k} + g_{lk,j} - g_{jk,l}). \quad (58.2)$$

Thus our conditions do determine a unique symmetric connection.



*Killing vectors*

The infinitesimal symmetries of a metric are given by Killing vectors, which are vectors  $k$  such that

$$\mathfrak{L}_k \mathfrak{G} = 0.$$

In coordinate components this is

$$g_{ij,l}k^l + g_{il}k^l{}_{,j} + g_{lj}k^l{}_{,i} = 0. \tag{58.3}$$

As we observed earlier, we can change all these derivatives to covariant derivatives. If we define the associated 1-form

$$k_i = g_{ij}k^j,$$

then we have Killing's equation

$$k_{i;j} + k_{j;i} = 0. \tag{58.4}$$

Despite the apparent simplicity of equation 58.4, the expression in equation 58.3 is easier to use in explicit calculations. Equation 58.4 is useful primarily in proofs and formal calculations.

*Symmetries of the Riemann tensor*

These are easily discussed in Gaussian coordinates. For a metric connection we have at the point of interest

$$\Gamma^i{}_{jk} \stackrel{*}{=} 0$$

and hence

$$g_{ij,k} \stackrel{*}{=} 0.$$

Combining this with the definition, equation 56.6, we can write

$$g_{im}R^m{}_{jkl} = R_{ijkl},$$

and find that

$$2R_{ijkl} \stackrel{*}{=} g_{il,jk} - g_{jl,ik} - g_{ik,jl} + g_{jk,il}$$

and in this form we can see the following symmetries:

$$\begin{aligned} R_{ijkl} &= -R_{ijlk}, & R_{ijkl} &= -R_{jikl}, \\ R_{ijkl} &= R_{klij}, \end{aligned} \tag{58.5}$$

and

$$R_{ijkl} + R_{iklj} + R_{iljk} = 0. \tag{58.6}$$

Note that these are all tensor equations, and the conditional equality has been replaced with general equality.

**Example:** These symmetries greatly restrict the number of independent components of the Riemann tensor. In four dimensions there are only 20 independent components. Suspend the summation convention for this discussion. There are no nonzero components of the form  $R_{aaaa}$ ,  $R_{aaab}$ ,  $R_{aaba}$ ,  $R_{abaa}$ , or  $R_{baaa}$ . Components of the form  $R_{abab}$  need not vanish, and there are six independent possible pairs of indices. Similarly we can have terms such as  $R_{abac}$ . There are four choices for the  $a$  index. The symmetry in equation 58.5 shows that the order of  $b$  and  $c$  does not matter. There are three pairs for each choice of  $a$ , and so a further 12 components. Finally, we have terms with all indices distinct. We can arrange each pair in alphabetical order, and also put the first index in the first block. This leaves components  $R_{abcd}$ ,  $R_{acbd}$ , and  $R_{adbcd}$ , which are independent under the first three symmetries. Equation 58.6 gives one relation between these. Thus we have in four dimensions  $6 + 12 + 2 = 20$  independent components for the Riemann tensor of a metric connection.

### Divergence

A manifold with a metric has a natural density defined on it

$$\omega = \sqrt{|g|} \, dx^1 \wedge dx^2 \cdots$$

This is the tensor  $\eta_A$  defined in Section 27. Here  $g$  is the determinant of the covariant components of the metric tensor. The divergence of a vector field  $v$  with respect to this natural density is given by

$$\mathcal{L}_v \omega = (\operatorname{div} v) \omega$$

and

$$\operatorname{div} v = \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x^i} (\sqrt{|g|} v^i).$$

See equation 51.5. We can differentiate this determinant to find

$$\operatorname{div} v = v^i_{,i} + \frac{1}{2} v^i \frac{1}{|g|} \frac{\partial}{\partial x^i} |g|$$

and this is

$$\operatorname{div} v = v^i_{,i} + \frac{1}{2} v^i g^{jk} g_{jk,i},$$

which can be written

$$\operatorname{div} v = v^i{}_{,i} + \frac{1}{2}v^i g^{jk}(g_{jk,i} + g_{ik,j} - g_{ij,k}).$$

This collapses into a covariant derivative

$$\operatorname{div} v = v^i{}_{;i}.$$

# IX

## *Gravitation*

Differential geometry originally sneaked into theoretical physics through Einstein's theory of general relativity. Never before had tensors played such a fundamental role in a physical theory, nor had so much emphasis been placed before on the transformation properties of physical quantities under coordinate change. Indeed, the class from which this book sprang was originally a general-relativity class. By now you have grasped the message of this book: differential geometry is useful everywhere in physics. That feeling, plus the number of good books specializing in general relativity, led me to include very little about gravitation here, just enough to fill out the geometric message of the book. Also, since general relativity is a metric theory, it could lead you into the bad habit of failing to distinguish between tangent vectors and 1-forms. By now you realize that that is a poor style of thinking.

In Chapter VI I argued that electrodynamics was a special classical field theory, arguing from the need of the force law to preserve rest masses. Since in electrodynamics like charges repel, no theory like electrodynamics can explain gravity. How, then, is gravity to be described? Give up the constancy of rest mass. Settle instead for the constancy only of mass ratios. This allows us to include one additional classical field theory, one not satisfying  $u \lrcorner f = 0$ . This field theory must be universal, with the same coupling to all matter. We can finish by rescaling lengths and times so that masses appear constant. The price paid is the introduction of a curved spacetime.

Now this new theory has a symmetry. The original flat background is not observable, and cannot be recovered uniquely from the curved spacetime. We finish the generalization by saying boldly that the original flat background does not exist. This is very similar to what happened in special

relativity. There Lorentz symmetry made absolute rest unobservable, and we boldly said that therefore absolute rest does not exist. Of course, neither special relativity nor general relativity was discovered this way. Still it is very instructive to consider this mock-historical approach.

The loss of the flat-background spacetime with its global affine structure leads to a surprising loss. No longer can the total energy of a system be defined. To do so requires a global idea of rest; otherwise the accounting of kinetic energy is not possible. General relativity is a complete theory, and all of its local pieces and fields go about their business of obeying the laws of physics, but without any conservation of total energy. There is now no longer a ban on the free lunch, and speculation that the entire universe is nothing but a vacuum fluctuation can now be sensibly entertained.

### **59. General relativity**

A nice understanding of general relativity results if we pretend that it is being invented in a straightforward, logical manner, in response to observational evidence that was in fact barely available at the time. In such a mock-historical approach, we can view general relativity as a theory invented to resolve two crises. The first is the incompatibility of Newtonian gravity with special relativity. The second is its lack of compatibility with an infinite universe. The resolution of the first of these crises leads to theories in which influences can spread no faster than the speed of light. This leads to the existence of gravitational waves as the influence of changing gravitational fields spreads out at a finite speed. The resolution of the second crisis leads to the expanding universe, and to an explanation of the mystery of the dark night sky, called Olber's paradox.

General relativity as a theory of gravity involves a very interesting paradigm shift. In special relativity the free particles were identified as a realization of an affine structure of spacetime. In general relativity this affine structure is changed to that of a Riemannian spacetime. This by itself is a fairly natural curvilinearization. The striking feature of general relativity is that a physically different set of particles are identified as the free particles. This is really a paradigm shift. It is rare that there should be two different physical realizations of a mathematical structure. General relativity switches the class of free particles with no forces acting on them to those particles with no forces acting except gravity. Viewed operationally this is reasonable. Gravity is a universal force. There is no known possibility of shielding a system from gravity. Such a universal force is quite

naturally built right into the very language of our physical theory. The inability to identify “truly free” particles is a reflection of our inability to identify the background affine structure mentioned in the Introduction.

Why should gravitation be a universal force? Experimentally this is known to be true. The experiment is known as the Eötvös experiment. Should it have been expected theoretically? Philosophically? Surprisingly, the answer here is yes.

Consider any continuum field theory. Energy conservation demands that the field itself carry energy. Only if this energy is positive will the theory have a stable vacuum. Now in electrodynamics this leads to the following reasonable thought experiment.

Take two like charges very far apart. Each has a certain amount of field energy. Now push the charges together. The field strength has now doubled everywhere, and since the field energy is quadratic, there is now four times as much of it as there was around a single charge, twice as much as we started with. This reflects the fact that work needed to be done to force the two like charges together.

For a field theory of gravitation, nothing changes except that we extract work from the system as the two masses are brought together. How can that be? Here we see what makes electrodynamics such a special theory. A charge moved through an electrodynamic potential difference does not change its strength or its mass. This cannot be true for a theory where likes attract, such as gravitation. The energy deficit in this scenario must be made up by a decrease in the rest masses of the particles.

Now this can lead to all sorts of interesting problems. For matter to maintain its properties, all masses had better vary by the same factor. Thus we are led to expect that any theory with likes attracting must be universal. If masses can change with position, then so too will the sizes of atoms, meter sticks, and physics professors. Thus, without even asking for it, the metric tensor describing physical measurements will describe a curved spacetime. Even with the best of intentions we are forced into a geometric theory of gravity.

Actually, the situation in general relativity is even more complicated. There are consistency conditions similar to those for current conservation in electrodynamics. These consistency conditions constrain the motion of sources; no additional assumption, such as about geodesic motion, is needed. This was not realized at first, and geodesic motion was first taken to be an additional postulate. To derive this result from general relativity is rather delicate, especially for bodies moving slowly enough for gravitation to be important. The approximations seem by now reasonably

reliable, and the result that small bodies move along the geodesics of a background pseudo-Riemannian metric is nearly a theorem of general relativity.

Not only is general relativity set in the context of pseudo-Riemannian geometry, but this geometry is also a dynamical variable, comparable to the electrodynamic field. The geometry is constrained by the stress, energy, and momentum present in spacetime. Einstein picked the simplest intrinsic equations to represent this constraint. There is no observational evidence that anything more complicated is needed.

There are solutions to these field equations that represent both the outer regions around spherically symmetric bodies and, if properly continued, singularities of the geometry with no source at all. These singular solutions are quite romantic and go under the general name of black holes. Probably no one has escaped the vivid popular accounts of these objects. There are also solutions known that describe both finite and infinite models of the universe.

In the first few sections to follow we will take the simplest spherically symmetric solution as a given spacetime and study it carefully. We follow the behavior of clocks, test particles, and light rays in its vicinity. Only in this fashion can we discover the intrinsic properties of the spacetime. Remember, the coordinates in which the solution is written do not necessarily have any observational significance.

### Problem

- 59.1** Study explicitly the energy balance as two scalar charges attract one another. The energy-momentum tensor was given in Section 41.

### 60. Geodesics

In a manifold  $M$  with a linear connection on its tangent bundle, we can define geodesics as the curves  $\gamma: \mathbb{R} \rightarrow M; u \mapsto X(u)$  satisfying

$$\nabla_{\dot{\gamma}} \dot{\gamma} = 0. \quad (60.1)$$

This determines both a curve and also a special parametrization on the curve unique up to affine transformations. Such a parameter is called a *special affine parameter*. Such curves are straight; the condition here says that they do not turn in any direction relative to a parallel-transported frame of reference.

When the connection can be derived from a metric, then these same curves are also extrema for the integrals

$$L = \int [g_{ij} \dot{X}^i \dot{X}^j]^{1/2} du,$$

and

$$E = \int g_{ij} \dot{X}^i \dot{X}^j du. \quad (60.2)$$

The first of these says that the length of the curve is extremal. It is clearly independent of the parameter  $u$ . The second is a stronger condition and easier to work with. It is not independent of the parametrization, and forces the parameter  $u$  to be the *special affine parameter* for which equation 60.1 holds. For indefinite metrics a minus sign will have to be inserted under the square root for some curves, and curves with null tangents will take some special care. We now discuss these two integrals.

#### *Extremal length*

We first discuss curves of extremal length. This is a simple problem in the calculus of variations. The extrema of

$$L = \int \mathcal{L}(x, \dot{x}) du$$

satisfy the Euler-Lagrange equations

$$\frac{d}{du} \left( \frac{\partial \mathcal{L}}{\partial \dot{x}} \right) = \frac{\partial \mathcal{L}}{\partial x}. \quad (60.3)$$

We are interested only in paths for which  $u$  is a special affine parameter, but the variations need not satisfy this condition. We can set

$$|g_{ij} \dot{X}^i \dot{X}^j| = 1$$

only after we have done the differentiations off the curve. From equation 60.3 we have

$$g_{ij,k} \dot{X}^i \dot{X}^j = \frac{d}{du} (2g_{kj} \dot{X}^j) \quad (60.4)$$

and expanding this we have

$$\frac{d\dot{X}^i}{du} + \frac{1}{2} g^{ij} (g_{jk,l} + g_{jl,k} - g_{kl,j}) \dot{X}^k \dot{X}^l = 0 \quad (60.5)$$

which is precisely equation 60.1. This shows that straight curves and



extremal curves (shortest curves in some cases, longest curves in space-time) coincide.

### *Extremal energy*

The variation given in equation 60.2 is easier, because it lacks the square root. Its extrema satisfy the same equation, 60.4, and so it leads to the same extremal curves. The parameter is now forced to be a special affine parameter. The variations of the energy integral are well-behaved even for curves with null tangent vectors, and it is the most comprehensive variational principle.

### *Calculation of connection components*

The energy extremal is the basis of an efficient scheme for calculating the connection components. It is most efficient when many of the components are zero.

**Example:** The connection on the 2-sphere given as an example in Section 54 comes from a metric

$$\mathcal{G} = d\theta^2 + \sin^2 \theta d\phi^2.$$

The energy integral is

$$E = \int (\dot{\theta}^2 + \sin^2 \theta \dot{\phi}^2) du.$$

The Euler-Lagrange equations are

$$\frac{d}{du}(2\dot{\theta}) = 2 \sin \theta \cos \theta \dot{\phi}^2,$$

$$\frac{d}{du}(2 \sin^2 \theta \dot{\phi}) = 0,$$

and if we put these in standard form

$$\ddot{\theta} - \sin \theta \cos \theta \dot{\phi}^2 = 0,$$

$$\ddot{\phi} + \frac{\cos \theta}{\sin \theta} (\dot{\theta} \dot{\phi} + \dot{\phi} \dot{\theta}) = 0,$$

we can read off the nonzero components of the connection. The equations must have the form

$$\ddot{x}^i + \Gamma^i_{jk} \dot{x}^j \dot{x}^k = 0$$

and so we find

$$\Gamma^{\theta}_{\phi\phi} = -\sin\theta \cos\theta,$$

$$\Gamma^{\phi}_{\phi\theta} = \cos\theta/\sin\theta = \Gamma^{\phi}_{\theta\phi},$$

all others vanishing.

This method is easier and far safer than a direct computation using equation 58.2. In Section 66 we will find another scheme that is even better if we want the Riemann tensor components as well.

### *Hamiltonian system for geodesics*

The geodesic equations can be put into Hamiltonian form. The Hamiltonian

$$H(\mathbf{q}, \mathbf{p}) = g^{ij} p_i p_j$$

leads to equations of motion

$$\frac{dq^i}{du} = g^{ij} p_k,$$

$$\frac{dp_i}{du} = -\frac{1}{2} g^{jk}{}_{,i} p_j p_k.$$

Here  $g^{ij}$  is the inverse of the usual metric tensor, defined by

$$g^{ij} g_{jk} = \delta_k^i.$$

It is a routine exercise to eliminate the  $p$ s from these equations and recover equation 60.4.

This Hamiltonian system does not involve the parameter  $s$  explicitly, and so generates a symplectic geometry in the cotangent bundle based on the 2-form

$$dp_i dq^i.$$

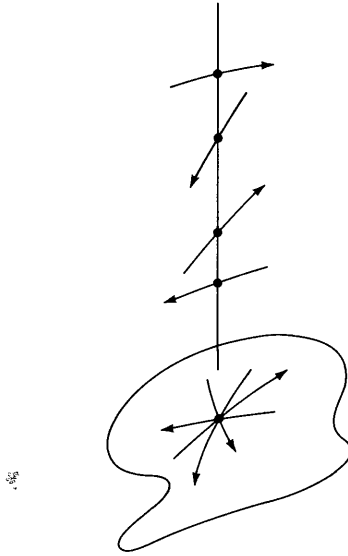
All the usual integral invariants follow from this.

### **Problem**

**60.1** Show that the geodesic equations do indeed form a Hamiltonian system.

## **61. Geodesic deviation**

Every space with a connection on its tangent bundle can be covered with geodesics. Through every point they pass in all possible directions. A



**Figure 61.1.** Simplifying a set of geodesics by lifting them into the tangent bundle.

natural question to ask now is about the behavior of nearby geodesics. We could approach this “calculus of geodesics” either in the manifold itself, or by lifting all the geodesics up into the tangent bundle. The latter simplifies things somewhat, because there the geodesics generally form a well-behaved congruence, with one and only one geodesic through each point (see Figure 61.1). The tangent-bundle discussion is no more than the local behavior of integral curves of a vector field. The description in the manifold is simpler, but involves second-order equations. We have considerable intuition about these second-order equations, and we will follow here this Lagrangian-like approach. The Riemann tensor describes the behavior of nearby geodesics, and this provides us with another way – not really all that different, to be sure – of grasping the physical significance of the Riemann tensor.

### *Deformations of a geodesic*

We start with a given geodesic  $\gamma: u \mapsto \gamma(u)$ , with tangent vector  $\dot{\gamma} = \lambda$ . A deformation of this curve will be a 1-parameter family of curves given by a map  $\Lambda: \mathbb{R}^2 \rightarrow M$  such that the curves

$$\Lambda_v: \mathbb{R} \rightarrow M; u \mapsto \Lambda(u, v)$$

are all geodesics with special parameter  $u$ , and with  $\Lambda_0$  identical with  $\gamma$ . We are interested in the behavior of  $\Lambda$  only up to linear terms in  $v$ . This nearby behavior of curves can be represented by a vector  $\eta(u)$  defined along the curve  $\gamma$ . This vector  $\eta$ , called the connection vector, is the tangent to the curves

$$\Lambda_u : v \mapsto \Lambda(u, v).$$

To specify a nearby geodesic, we need to know one point on the geodesic, say,  $\eta(0)$ , and also the direction at that point, given by  $\nabla_\lambda \eta(0)$ . These should be enough to specify the geodesic completely; so we do not expect to be able to freely choose the second derivative of  $\eta$ .

### Equation of geodesic deviation

Let us compute this second derivative. We are given that the curves  $\Lambda_u$  and  $\Lambda_v$  fit together to form coordinate lines on a 2-surface; hence we have

$$[\lambda, \eta] = 0.$$

Also, the curves are all geodesics with a special parameter, so that

$$\nabla_\lambda \lambda = 0.$$

We will consider only zero torsion; so we have

$$\nabla_\lambda \eta - \nabla_\eta \lambda = [\lambda, \eta] = 0.$$

We take the second derivative, and use the preceding equation to find

$$\nabla_\lambda \nabla_\lambda \eta = \nabla_\lambda \nabla_\eta \lambda$$

and now interchange the order of differentiation, using equation 57.4,

$$R(\lambda, \eta)\lambda = \nabla_\lambda \nabla_\eta \lambda - \nabla_\eta \nabla_\lambda \lambda = \nabla_\lambda \nabla_\eta \lambda$$

to find the equation of geodesic deviation

$$\nabla_\lambda \nabla_\lambda \eta - R(\lambda, \eta)\lambda = 0.$$

This is a second-order equation for the connection vector  $\eta$ , as promised. The derivatives are covariant derivatives, however, and the interpretation of this is not exactly  $F = ma$ , as we shall see. A vector field along a geodesic satisfying the preceding equation is called a *Jacobi vector field*.

### Tides

The most natural interpretation of the Riemann tensor is in terms of tidal forces, that is, gravitational field *gradients*.

**Example:** If we consider geodesics to be the worldlines of freely falling particles, then we can use the equation of geodesic deviation to find some of the components of the corresponding Riemann tensor. Consider geodesics that are near a free particle that is falling straight down toward the center of a spherical gravitating body. The path of this fiducial particle is well-described by Newtonian gravitation, and satisfies

$$\frac{d^2r}{dt^2} = -\frac{M}{r^2}.$$

I am using units with  $G=1$  here. They were defined in Section 42. We will compute the Riemann tensor at the point where the velocity of this particle is zero. A parallel-propagated frame of vectors along this curve is easily found, working only to terms linear in  $M$ . If the particle is on the equator, then

$$e_x = \frac{1}{r} \frac{\partial}{\partial \theta} \quad \text{and} \quad e_y = \frac{1}{r} \frac{\partial}{\partial \phi}$$

will be two such parallel propagated vectors, and

$$e_z = \frac{\partial}{\partial r}$$

a third. Near points of nonzero velocity, this  $\partial/\partial r$  vector must be tipped into the  $\partial/\partial t$  direction as well.

Look at a neighboring geodesic a distance  $\zeta$  away in the  $\partial/\partial r$  direction. We have

$$\frac{d^2(r+\zeta)}{dt^2} = -\frac{M}{(r+\zeta)^2};$$

that is,

$$\frac{d^2\zeta}{dt^2} = \frac{2M\zeta}{r^3}.$$

From this we find the Riemann tensor components

$$R^z{}_{tzt} = -2M/r^3$$

(see Figure 61.2).

We can also put the particles side by side, as in Figure 61.3. The motion is linear in  $r$ , but quadratic in time

$$\zeta = Ar = A(r_0 - Mt^2/2r_0^2).$$

Hence

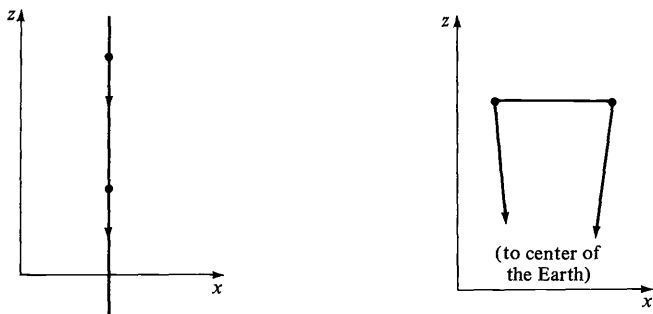


Figure 61.2 (left). Two particles falling along the same vertical line.

Figure 61.3 (right). Two particles falling side by side.

$$\frac{d\zeta}{dt} = -\frac{AM}{r^2} = -\frac{M\zeta}{r^3}.$$

This gives us the tidal-force matrix

$$R_{t,t} = \frac{M}{r^3} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix},$$

compressive in two directions and extensional in the third.

### Computations

The equation of geodesic deviation is split into two covariant pieces. This split is analytically useful, but not efficient for the actual computation of geodesics. Many of the derivatives of the  $\Gamma$ s appearing in the  $\nabla\nabla$  term are canceled by the curvature term. The situation is reminiscent of Killing's equation, where equation 21.1 is far easier to use than 58.4, which contains covariant derivatives. The efficient equation here is

$$\frac{d^2\eta^i}{ds^2} + 2\Gamma^i_{jk} \frac{dx^j}{ds} \frac{d\eta^k}{ds} + \Gamma^i_{jk,l} \frac{dx^j}{ds} \frac{dx^k}{ds} \eta^l = 0.$$

[For a proof and discussion, see Faulkner and Flannery (1978).]

### Problem

**61.1** Use the equation of geodesic deviation to compute the flattening of the Sun at sunset.

## 62. Symmetries and conserved quantities

We consider here a manifold with a metric (not necessarily positive definite) and its associated symmetric connection. Important properties of such a manifold are its symmetries. Most manifolds have none, but the manifolds that arise in applications are special and do often have symmetries. Every symmetry eases the problem of solving the geodesic equation.

### *Conserved quantities*

We suppose that we have a geodesic with special parameter

$$u \mapsto \gamma(u), \quad \nabla_{\dot{\gamma}} \dot{\gamma} = 0, \quad (62.1)$$

and a Killing vector  $k$

$$k_{i;j} + k_{j;i} = 0.$$

Here the quantity

$$E = k \cdot \dot{\gamma}$$

is constant along the geodesic.

A direct calculation gives

$$\nabla_{\dot{\gamma}}(k \cdot \dot{\gamma}) = \nabla_{\dot{\gamma}} k \cdot \dot{\gamma}$$

and in components

$$k_{i;j} \dot{\gamma}^i \dot{\gamma}^j = \frac{1}{2}(k_{i;j} + k_{j;i}) \dot{\gamma}^i \dot{\gamma}^j = 0.$$

**Examples:** Minkowski spacetime has ten Killing vectors and hence ten conserved quantities. The time translation  $\partial/\partial t$  generates the conserved quantity called energy. The three space translations, such as  $\partial/\partial x$ , give us momentum conservation. The rotations such as  $x(\partial/\partial y) - y(\partial/\partial x)$  give us angular-momentum conservation.

What about the Lorentz transformations, such as  $t(\partial/\partial x) + x(\partial/\partial t)$ ? For a single particle these lead to conserved quantities proportional to  $x - vt$ , and are trivial. For composite systems we will see that these are nontrivial, and force the center of mass of a system to move with constant velocity. These last three conservation laws prevent us from making a rocketlike engine that does not throw momentum away in its exhaust.

Look at geodesics on a surface of revolution. Useful coordinates are the angle  $\phi$  and proper distance  $s$  along the surface at constant

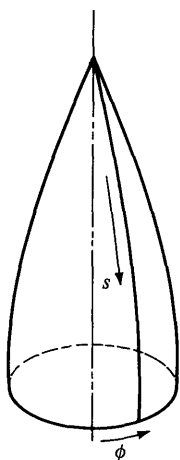


Figure 62.1.

$\phi$  (see Figure 62.1). The metric is

$$\mathcal{G} = ds^2 + R^2(s) d\phi^2.$$

The vector field  $\partial/\partial\phi$  is a Killing vector field, since  $\phi$  does not appear in the metric.

Look at a geodesic with unit tangent vector

$$\lambda = a \frac{\partial}{\partial s} + b \frac{\partial}{\partial \phi},$$

where

$$a^2 + R^2 b^2 = 1.$$

Our symmetry gives us a conserved quantity

$$L = k_\phi \lambda^\phi = bR^2,$$

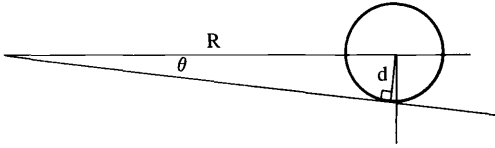
and we have

$$a^2 = 1 - L^2/R^2.$$

If the surface of revolution becomes flat for large  $s$ , then the geodesic becomes a straight line and the constant of motion  $L$  is related to the apparent impact parameter  $d$  by Euclidean geometry, as shown in Figure 62.2,

$$\sin \theta = L/R.$$





**Figure 62.2.** The apparent impact parameter.

Also

$$\sin \theta = d/R;$$

hence

$$d = L.$$

Now, no matter how the surface is curved, the geodesic will have a radial turning point at the radius where  $a = 0$ :

$$L = R = d.$$

Thus we have a remarkable result: The closest approach to the axis is independent of the shape of the surface! John Faulkner devised this pretty example.

### *Continuum systems*

The mechanical description of a continuum system is given by a stress-energy-momentum tensor  $T^{ij}$  subject to the Cauchy equations

$$T^{ij}{}_{;j} = 0.$$

General relativity extends the laws to curved spacetime by replacing partial derivatives with covariant derivatives.

The total energy, momentum, etc., of a system will be given by an integral over the system. Now, recall that an integral over an  $m$ -space can be defined only for an  $m$ -form. The metric of spacetime provides us with an oriented volume element

$$\eta_{ijkl} = \epsilon_{ijkl} \sqrt{-g},$$

(recall Section 58). This volume element can be written as the correctly ordered product of any orthonormal basis. If we parallel-transport the orthonormal basis along any curve, it will define this same volume element. Thus the covariant derivative of the volume element is zero.

Given only a stress-energy tensor and a volume element, we cannot produce a 3-form. Thus we cannot produce any conserved quantities, and so we cannot define what we could mean by, for example, the total energy of a system at some particular time. The problem is insoluble without additional geometric structure. A Killing vector will do. For each symmetry we can write down the 3-form

$$j_{klm} = \eta_{klmn} T^{np} k_p.$$

This 3-form gives us a conservation law if it is closed. Define the vector

$$v^i = T^{ij} k_j.$$

Now we have

$$dj = \mathcal{L}_v \eta = (\operatorname{div} v)$$

and we will have a conservation law provided that

$$(T^{ij} k_j)_{;i} = 0.$$

Now  $T^{ij}$  is symmetric and satisfies

$$T^{ij}_{;j} = 0;$$

so we have

$$(T^{ij} k_j)_{;i} = T^{ij} k_{j;i} = \frac{1}{2} T^{ij} (k_{j;i} + k_{i;j}) = 0.$$

Every Killing vector leads to a conserved current  $T^{ij} k_j$ . If we integrate over any volume, then the conservation law says that

$$\int_{\partial V} j = 0,$$

and for a confined system we have

$$\int_{S_1} j = \int_{S_2} j,$$

for the situation sketched in Figure 62.3.

**Examples:** In Minkowski spacetime the Killing vector

$$k = t \frac{\partial}{\partial x} + x \frac{\partial}{\partial t}$$

leads to a conserved quantity

$$\int (xT^{tt} + tT^{tx}) dx dy dz.$$

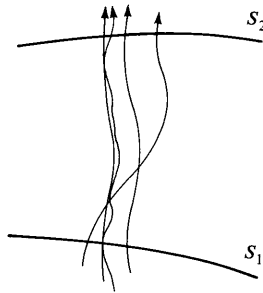


Figure 62.3.

The first term is the  $x$  coordinate of the center of mass; the second is the total momentum  $P$ , separately conserved because of the Killing vector  $\partial/\partial x$ . Thus the center of mass moves with a constant velocity related to the total momentum of the system.

The 2-cylinder has two Killing vector fields that are continuous over the entire cylinder,  $\partial/\partial\theta$  and  $\partial/\partial\phi$ . However, the vector field

$$k = \theta \frac{\partial}{\partial z} - z \frac{\partial}{\partial \theta}$$

is not continuous all around the cylinder. The angular momentum defined by this Killing vector will be conserved only for systems that do not stretch all the way around the cylinder. A spider living on the surface of a slippery cylinder could stop himself from spinning around only by encircling the cylinder with a strand of spider web.

**Problems**

62.1 Show that Cauchy's equations

$$T^{uv}{}_{;v} = 0$$

are equivalent to

$$d(*k \lrcorner T) = 0,$$

where  $k$  is almost a Killing vector, satisfying at that point

$$[\mathcal{L}_k, *] = 0.$$

Is this good for anything?

**62.2** On a spacetime  $\mathbb{R} \times S^3$  there are only seven, not ten, Killing vectors. No longer are there constraints on the motion of the center of mass of a mechanical system. Construct a mechanical device that can move from rest an arbitrarily large distance, then come to rest again in its original state (reactionless rocket).

### 63. Schwarzschild orbit problem

We illustrate now the effectiveness of symmetry in simplifying a geodesic problem. The motion of particles of negligible mass near a spherically symmetric gravitating body is described in general relativity by the geodesics of the spacetime metric,

$$\mathcal{G} = -(1 - 2m/r) dt^2 + dr^2/(1 - 2m/r) + r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

This representation is valid only for  $r > 2m$ . We can find a wealth of detail about this region of spacetime by exploiting the symmetries and without integrating any differential equations.

#### *Symmetries*

The spacetime was constructed to have the full set of Killing vectors representing spherical symmetry. Geodesics will be confined to a plane, and without loss of generality we can take that plane to be the  $\theta = \pi/2$  equatorial plane. In that plane we have a Killing vector  $\partial/\partial\phi$ , and hence a conserved angular momentum

$$L = \lambda_\phi,$$

where  $\lambda$  is the 4-velocity of the geodesic. In physical terms,  $L$  is the angular momentum per unit mass. We have also the time-shift symmetry  $\partial/\partial t$  and its conserved quantity

$$E = -\lambda_t,$$

which is physically the energy (including rest-mass energy) per unit mass. Finally, since  $\lambda$  is a 4-velocity, we have the algebraic relation

$$\mathcal{G}(\lambda, \lambda) = -\kappa.$$

Here  $\kappa = 1$  for particles with nonzero rest mass and  $\kappa = 0$  for light rays. Physically this says that the rest mass of a particle is constant. With three equations for the three nonzero components of the 4-velocity, we can solve immediately for the 4-velocity as a function of radius. If we write out the preceding equation we find

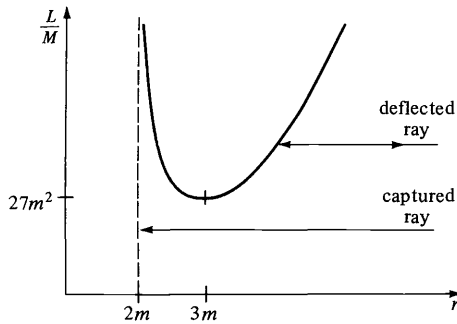


Figure 63.1. Light rays in the Schwarzschild metric.

$$\frac{L^2}{r^2} - \frac{E^2}{(1-2m/r)} + \kappa + \frac{(\lambda')^2}{(1-2m/r)} = 0,$$

and dropping the positive  $(\lambda')^2$  term, we have the inequality

$$L^2 - r^2(E^2 - \kappa) - 2mr\kappa \leq 2mL^2/r. \quad (63.1)$$

For comparison the Newtonian expression is

$$L^2 - 2r^2\varepsilon - 2mr \leq 0.$$

Here  $\varepsilon$  is the kinetic energy. The relativistic energy  $E$  includes the rest mass. The other differences are that equation 63.1 also applies to photons, and the extra potential on the right-hand side of equation 63.1 contains modifications of Newtonian gravity coming from general relativity.

#### Light rays

For light rays the two parameters  $E$  and  $L$  are not independent. This reflects the fact that all light signals travel with a unique velocity. We have

$$L^2 - r^2E^2 \leq 2mL^2/r,$$

and this can be written

$$r^3/(r-2m) \geq (L/E)^2. \quad (63.2)$$

This inequality is sketched in Figure 63.1. An incoming light ray must turn when  $\lambda' = 0$ , which corresponds to equality in equation 63.1. Light rays with

$$(L/E)^2 < 27m^2$$

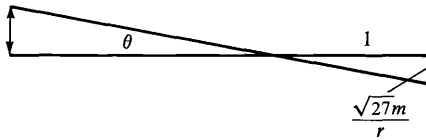


Figure 63.2. The impact parameter of the critical ray.

have no turning point anywhere in the region  $r > 2m$ . A global analysis shows that in fact it has none in any region that we can describe with general-relativity physics. This one-way behavior is why such spacetimes are called black holes. What about the critical ray, with  $(L/E)^2 = 27m^2$ ? Far from the gravitating body, spacetime becomes flat, and an orthonormal basis is given by the vectors  $\partial/\partial t$ ,  $\partial/\partial r$ , and  $r^{-1}(\partial/\partial\phi)$ . The 4-velocity of the light ray in that region is given by

$$\lambda = \frac{\partial}{\partial t} + \frac{\sqrt{27}m}{r^2} \frac{\partial}{\partial\phi} + \frac{\partial}{\partial r}.$$

The angle this ray makes with the radial direction can be found from the components in the orthonormal basis (see Figure 63.2)

$$\theta = \sqrt{27}m/r.$$

A drop in angular size like  $1/r$  is characteristic of Euclidean geometry. The rays that are trapped are those that have an impact parameter at infinity less than  $\sqrt{27}m$ . The optical cross section is thus  $27\pi m^2$ .

#### *Particles with nonzero rest mass*

For particles with rest mass we put  $\kappa = 1$  and sketch the four possibilities in Figures 63.3 to 63.6. Particles with  $E > 1$  are unbound and can escape to infinity. Particles with  $E < 1$  are bound and cannot get out beyond some finite radius. The contacts between pairs of possibilities correspond to two turning points on top of one another. These are circular orbits. There are two types of circular orbits, shown in Figures 63.7 and 63.8.

We can easily discuss the stability of these circular orbits. We consider perturbations of  $E$ ,  $L$ , and  $r$ , which are physically allowed by our inequality. For the situation shown in Figure 63.8 there are perturbations of  $E$  and  $L$  that give us Figure 63.9, and no matter how you perturb  $r$ , the particle will run away a finite distance. These circular orbits are unstable. The ones shown in Figure 63.7 are stable. The boundary between stable

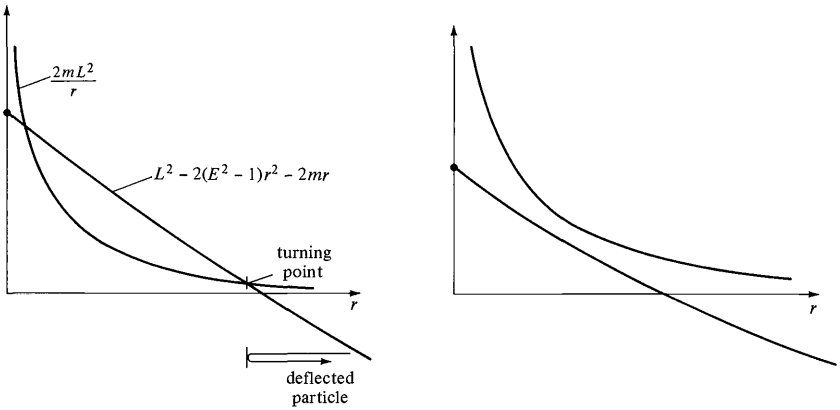


Figure 63.3 (left). Particles with  $E > 1$ , unbound, with a turning point.  
 Figure 63.4 (right). Another unbound particle, this one captured.

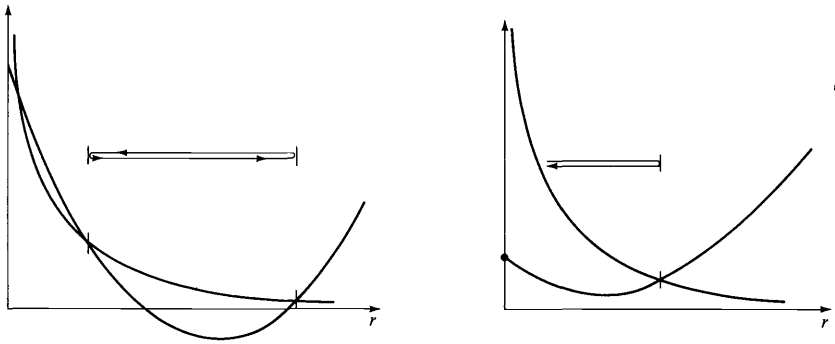


Figure 63.5 (left). A bound particle,  $E < 1$ , trapped in a noncircular orbit.  
 Figure 63.6 (right). A bound particle with no stable orbit.

and unstable circular orbits comes when there is triple contact. There we have contact up to the second derivative.

Zero-order contact, intersection, at a radius  $a$  requires

$$L^2 - a^2(E^2 - 1) - 2ma = 2mL^2/a, \tag{63.3}$$

and first-derivative contact there requires

$$-2a(E^2 - 1) - 2m = -2mL^2/a^2.$$

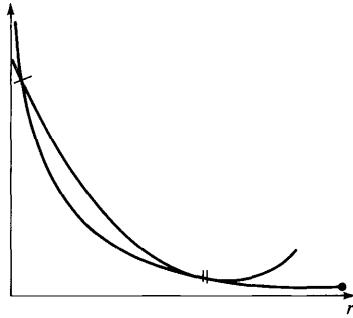


Figure 63.7. A double contact of the two outside points leads to a stable circular orbit.

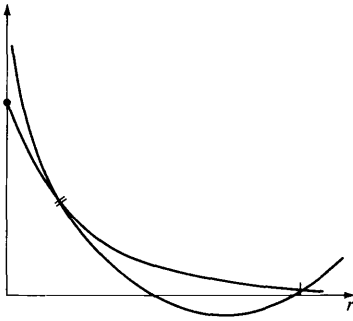


Figure 63.8 (left). A double contact of the two inside points leads to an unstable circular orbit, as shown in Figure 63.9.

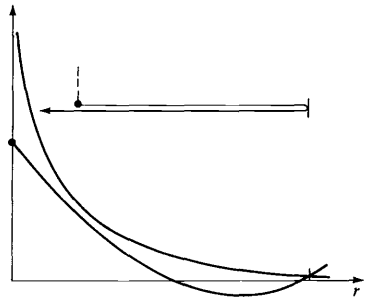


Figure 63.9 (right). A deformation of the preceding figure. The allowed orbits all fall into the center.

From these we can find the equations of circular orbits

$$E = (a - 2m)/\sqrt{a(a - 3m)}, \quad L = a\sqrt{m/(a - 3m)}. \quad (63.4)$$

Matching the second derivatives as well gives us

$$-2(E^2 - 1) = 4mL^2/a^3,$$

and together with the other conditions this gives us  $r = 6m$  as the radius of the smallest stable circular orbit.

Suppose that a particle is in a circular orbit and by some frictional process, energy and angular momentum are extracted and sent off to



infinity. This process can continue slowly only until the orbit gets to  $r = 6m$ . Beyond this the particle falls out of orbit and plunges into the black hole. The Killing vector  $\partial/\partial t$  ensures that the entire process must conserve energy. The energy that appears at infinity is just the energy  $E$  lost by the body. The maximum energy loss can be found from equation 63.4. At most a fraction  $1 - (2\sqrt{2}/3) \cong 6\%$  of the rest-mass energy of the particle can be extracted by such a process.

### *Capture of slow particles*

As another example, let us compute the capture cross section for slow particles, ones having  $E$  just a little longer than unity. The critical situation involves an unstable circular orbit from which particles can fall into the origin. From equation 63.4 we find a radius of  $4m$  for the orbit having  $E = 1$ , and from equation 63.3 we see that for this orbit we have  $L = 4m$ .

As seen in the Euclidean geometry at infinity, the angular momentum is just the velocity times the impact parameter, and so the cross section for slow particles will be equal to

$$16\pi m^2/u^2,$$

where  $u$  ( $\ll 1$ ) is the velocity of the particle at infinity.

### *Circular orbits*

We can find the periods of these circular orbits without detailed integration. The coordinate  $r$  is not directly measurable, and it is good style to state this law in terms of observable quantities. Let us calculate the period seen by an observer at infinity for a circular orbit of apparent radius  $b$ , as measured by its angular size as seen from infinity.

To relate the apparent size to the  $r$ -coordinate size  $a$ , we follow a grazing light ray out to infinity. If it has a turning point on the circular orbit, then it has

$$a^3/(a-2m) = (L/E)^2.$$

The 4-velocity of this light ray at infinity is

$$\lambda = E \frac{\partial}{\partial t} + \frac{L}{r} \left( \frac{1}{r} \frac{\partial}{\partial \phi} \right) + \sqrt{E^2 - \frac{L^2}{r^2}} \frac{\partial}{\partial r},$$

and its apparent angular size is

$$\tan \theta = L/\sqrt{E^2 r^2 - L^2}.$$

As  $r$  goes to infinity,

$$\theta \approx (b/r) \approx (L/Er)$$

and so

$$b = [a^3/(a-2m)]^{1/2}. \quad (63.5)$$

The coordinate-time period of a particle in this circular orbit as measured by any time-invariant process will just be the coordinate increment  $\Delta t$ , since  $\partial/\partial t$  generates the time-shift symmetry. We have

$$\frac{d\phi}{dt} = \frac{\lambda^\phi}{\lambda^t},$$

and so we have

$$\frac{d\phi}{dt} = \frac{a-2m}{a^3} \frac{L}{E};$$

so the period is

$$P = 2\pi a^3/m.$$

This is related to the observable apparent size  $b$  by equation 63.5.

#### 64. Light deflection

One of the earliest confirmations of Einstein's theory of gravitation was the observation that light passing the Sun is deflected by an amount

$$\theta = 4m/b,$$

where  $b$  is the distance of closest approach. This expression neglects corrections of order  $(m/b)^2$ .

It seems that everything we try to calculate in general relativity turns into a big mess. Here there is a conflict between spherical coordinates, in which the spacetime metric is simple, and rectangular coordinates, in which the orbit is simple. Writing the Euler-Lagrange equations in either coordinates is very messy.

You might think to exploit the spherical symmetry, as we did in the last section. From this we can find the orbit equation

$$\frac{d\phi}{dr} = \frac{\lambda^\phi}{\lambda^r}.$$

And from here the calculation of the deflection is straightforward. [The details of this approach can be found in Weinberg (1972).] A clever trick

here is to convert the orbit equation into a second-order equation. The perturbation is more easily done in this form. [For details see Robertson and Noonan (1968).]

The straightforward calculations are so involved that they conceal the interesting features of the calculation. I am going to give a different calculation of the light deflection. It will seem at first like the long way around the barn, but it exploits coordinate invariance and leads to some insight into what is going on and why the deflection is twice the Newtonian result.

### *Spacetime with two metrics*

The light deflection is the difference between the flat-spacetime orbit, a simple straight line, and the true curved orbit in Schwarzschild spacetime. Flat spacetime describes the Schwarzschild spacetime far from the mass, where the angles are actually measured. We can exploit this differential nature of the problem and use the fact that the difference of two affine connections is a tensor.

If we write the curved-spacetime metric

$$g_{ij} = \eta_{ij} + h_{ij},$$

with  $\eta_{ij}$  the flat-spacetime metric, then the geodesic equation is

$$\sigma^i{}_{,j} \sigma^j + g \Gamma^i{}_{jk} \sigma^j \sigma^k = 0.$$

Here  $\sigma^i{}_{,j} \sigma^j$  is a way of writing a derivative along the curve without introducing new notation. As so often happens, there is an apparent dependence on the values of  $\sigma^i$  off the curve, but no real dependence. Here  $g \Gamma$  is the affine connection of the Schwarzschild metric. If we write this

$$\sigma^i{}_{,j} \sigma^j + \partial \Gamma^i{}_{jk} \sigma^j \sigma^k + (g \Gamma^i{}_{jk} - \partial \Gamma^i{}_{jk}) \sigma^j \sigma^k = 0$$

then it collapses into the form

$$\sigma^i{}_{;j} \sigma^j + \Lambda^i{}_{jk} \sigma^j \sigma^k = 0, \quad (64.1)$$

where the semicolon is a flat-spacetime covariant derivative (curvilinear coordinates, perhaps), and  $\Lambda^i{}_{jk}$  is the tensor given to first order in  $h$  by

$$\Lambda^i{}_{jk} = \frac{1}{2} \eta^{im} (h_{mj;k} + h_{mk;j} - h_{kj;m}). \quad (64.2)$$

Note that each term in equations 64.1 and 64.2 is a tensor by itself. Equation 64.1 looks like a force law. The corresponding equation in electrodynamics is

$$\lambda^i{}_{;j}\lambda^j + (e/m)F^i{}_j\lambda^j = 0, \quad (64.3)$$

using  $\lambda^i$  for 4-velocity, and reserving  $\sigma^i$  for null vectors.

### *Light deflection by a point-mass*

The weak-field approximation to equation 62.1 is

$$h = (2m/r)(dt^2 + dr^2).$$

Let us work in rectangular components, where the unperturbed photon orbit is just

$$u \mapsto (t, x, y, z) = (u, u, b, 0).$$

The deflection will be bound by integrating the  $y$  component of the force along the unperturbed orbit.

We need three components of the force

$$\begin{aligned} \Lambda^y{}_{tt} &= my/r^3, & \Lambda^y{}_{tx} &= 0, \\ \Lambda^y{}_{xx} &= 2my/r^3 - 3mx^2y/r^5. \end{aligned}$$

The  $\Lambda^y{}_{tt}$  term is the analog of the Coulomb force in electrodynamics. The  $\Lambda^y{}_{xt}$  and  $\Lambda^y{}_{xx}$  are velocity-dependent forces analogous to magnetism and a double magnetism. Note the two velocity terms in equation 64.1 compared to the single one in equation 64.3.

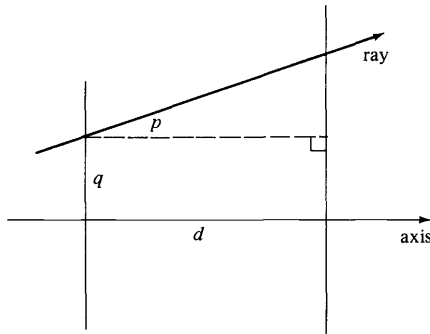
The deflection angle is thus

$$\theta = \int_{-\infty}^{\infty} \left( \frac{3mb}{r^3} - \frac{3mx^2b}{r^5} \right) dx = \frac{4m}{b}.$$

If you fill in the details, you will see that the advantage of this calculation method is in being able to work in rectangular components. In the next section we will apply this result to cosmology.

## **65. Gravitational lenses**

The gravitational deflection of light calculated in the last section and well-verified by observations is an important astrophysical effect. For these applications it is necessary to properly take into account both the curved spacetime of the universe and the Hubble expansion. Naive attempts to ignore these run afoul of aberration, which is important because it is a first-order effect. Details of cosmological models can be found in Burke (1980). The development here uses the Hamiltonian-optics formalism of Section 52.



**Figure 65.1.** A single ray described as a Hamiltonian system, propagating a distance  $d$ . The slope  $p$  remains constant, and the coordinate  $q$  increases linearly.

### Minkowski spacetime

First we study flat spacetime. Using the small-angles approximation, we have a given ray described by two transverse coordinates and two conjugate momenta, the small angles. For simplicity we look only at rays in a plane. This leads to a two-dimensional phase space that we can sketch, and think about easily. Because of symmetry, the full problem is just the direct sum of these plane problems.

Propagation through a distance  $d$  is represented by the symplectic transformation

$$(q, p) \mapsto (q + pd, p).$$

The situation in space is shown in Figure 65.1, and in phase space in Figure 65.2. The gravitational lens will be thin in this approximation, and so the lens is represented by the symplectic transformation

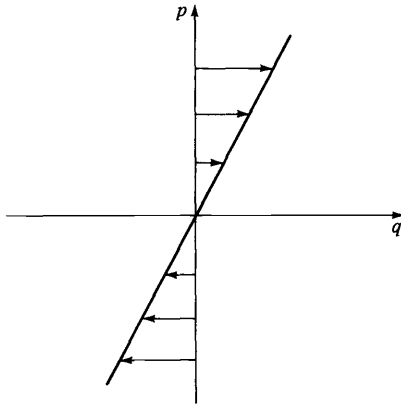
$$(q, p) \mapsto (q, p - 4m/q).$$

We have here a theorem from Newtonian gravity. A ring of matter deflects light passing outside the ring as if its mass were all concentrated at the center. Light passing through the ring is not deflected at all. We can use potential theory to prove this. The deflection  $E$  of a point-mass can be derived from a potential

$$E = \frac{1}{r} = \nabla\phi$$

and satisfies

$$\nabla \cdot E = 0.$$



**Figure 65.2.** In phase space we can view the behavior of all rays at once. The transformation of phase space is a shear along the  $q$  axis.

Now the analog of Gauss's law holds, and from it the preceding results. Be careful. The result holds for cylindrical shells, not spherical shells.

### *Gravitational lenses*

Rather than treat just the point-mass, let us discuss an extended, transparent object. In the center the cylindrical mass increases like  $r^2$  and the deflection increases like  $r$ . Beyond the mass the deflection drops off like  $1/r$ . The central part acts like an ordinary lens of focal length

$$f = (4\pi\rho_c)^{-1},$$

where  $\rho_c$  is the cylindrical density.

**Examples:** The focal length of the Sun is given by

$$\rho_c = 2 \int_0^\infty \rho \, dr,$$

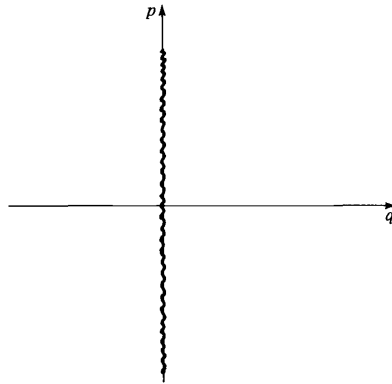
where  $\rho$  is the ordinary density. For the Sun we have, roughly,

$$\rho_c = 20 \text{ g cm}^{-3} R.$$

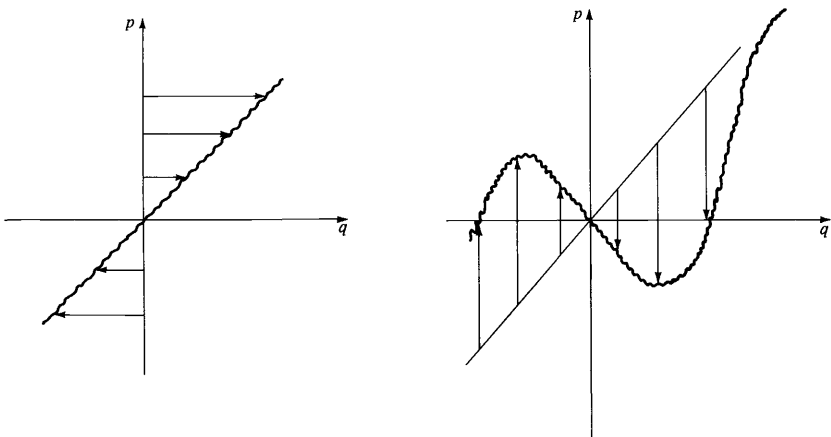
and this leads to a focal length

$$f = 25 \text{ AU} = 1.3 \times 10^4 \text{ s}.$$

This applies only to things like neutrinos and gravitational waves that pass right through the Sun.



**Figure 65.3.** Phase-space picture of the one-parameter family of rays emitted by a point-source.



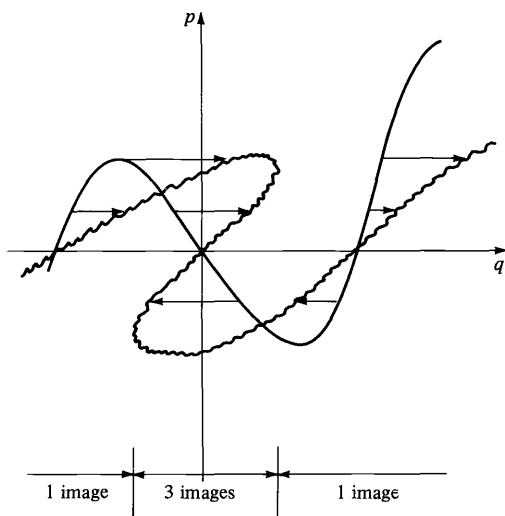
**Figure 65.4 (left).** The rays after they have propagated to the vicinity of the deflector.

**Figure 65.5 (right).** The rays just after they have passed the deflector.

The propagation of light from a point source at  $q = 0$  past an extended deflector is sketched in Figures 65.3 to 65.6. Note that beyond the conjugate focus there are three images, three  $ps$  for a given observer's position  $q$ .

### *Cosmological spacetimes*

I give here a quick review of the relevant formulas. The spacetime describing a homogeneous isotropic universe filled with matter having zero



**Figure 65.6.** The rays after they have propagated beyond the focal length of the central region. Observers not too far off the axis see three images.

pressure is

$$\mathcal{G} = R^2(\eta) [-d\eta^2 + d\chi^2 + S^2(\chi)(d\theta^2 + \sin^2\theta d\phi^2)]$$

where the function  $S(\chi)$  can be either  $\sin \chi$ ,  $\chi$ , or  $\sinh \chi$ . If it is  $\sin \chi$ , the spacetime looks like  $\mathbb{R} \times S^3$ , and these are called closed spacetimes. The spatial volume is finite, although the spacetime is unbounded in space. For  $\chi$  or  $\sinh \chi$ , the spacetime looks like time crossed into either Euclidean space or a pseudosphere. The coordinates  $(\chi, \theta, \phi)$  are comoving with the matter. The timelike coordinate  $\eta$  is called arc time. During an arc-time interval  $\Delta\eta$ , a light signal travels  $\Delta\eta$  radians around the universe.

The function  $R(\eta)$  is given by

$$R(\eta) = \begin{cases} \frac{1}{2}a(1 - \cos \eta), \\ a\eta^2/4, \\ \frac{1}{2}a(\cosh \eta - 1) \end{cases}$$

for the three functions of  $\chi$ . Here  $a$  is a scale factor for the entire spacetime. The parameter  $R$  converts proper lengths measured in light-seconds into arc length on  $S^3$ .

The redshift of light emitted at arc time  $\eta$  and received at  $\eta_0$  is

$$(1+z) = R(\eta_0)/R(\eta).$$



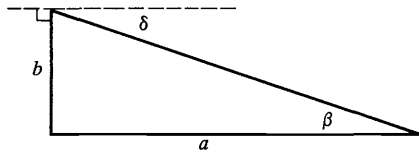


Figure 65.7. A long, skinny spherical triangle.

A measure of the local expansion rate of the universe is given by the Hubble constant

$$H_0 = 2 \sin \eta_0 / [a(1 - \cos \eta_0)^2].$$

Numerically, a Hubble constant

$$50 \text{ km/sec Mpc} = (6.18 \times 10^{17} \text{ sec})^{-1}$$

agrees with observations, although some argue for a value twice as large.

The use of arc distance  $\chi$  and arc time  $\eta$  allows us to discuss the properties of cosmological models using intrinsic non-Euclidean geometry rather than coordinate-intensive analytic geometry. For example, the closed models live for an arc time of  $\pi$ . A photon gets to travel halfway around the universe during the lifetime of the universe.

### *Cosmological light propagation*

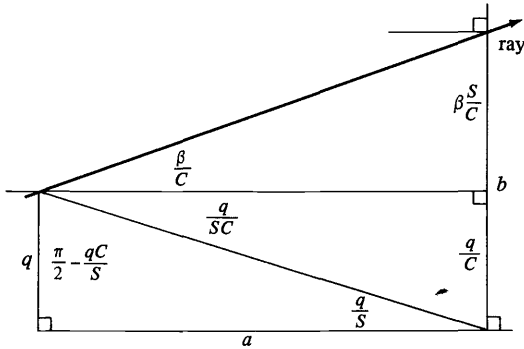
What change in the propagator must be made to take into account curved expanding spacetime? Both curvature and expansion are important, the latter because of aberration. This is most easily studied in comoving coordinates. The orbits of light rays are just geodesics in either the 3-sphere or the 3-pseudosphere. In both arc length provides a natural dimensionless measure of distance, one unchanged by the expansion of the universe.

This calculation is a nice problem in spherical and pseudospherical trigonometry. We need the laws for long, thin triangles. On the sphere we have

$$b = \beta \sin a, \quad \delta = \beta \cos a,$$

for the angles and lengths sketched in Figure 65.7. On the pseudosphere these become just  $\sinh$  and  $\cosh$ .

The geometry of a paraxial ray is surprisingly involved (Figure 65.8). Remember that the geometry is not Euclidean. The dotted construction line in the figure takes some thought. There are right angles only where



**Figure 65.8.** The cosmological analog of Figure 65.1, defining the state variables for the ray. Be careful, there are right angles only where indicated.  $S$  and  $C$  stand for trigonometric or hyperbolic functions.

explicitly indicated. Tracing through everything leads to the transformations

$$(q, p) \mapsto (q \cos a + p \sin a, -q \sin a + p \cos a)$$

and

$$(q, p) \mapsto (q \cosh a + p \sinh a, q \sinh a + p \cosh a).$$

The transformations of phase space are no longer shears (Galilean transformations), but are now either rotations or Lorentz transformations. Both  $q$  and  $p$  now have the dimensions of angle.

**Examples:** In a closed cosmological model the light from a point source focuses again after an arc distance of  $\pi$  without the need for any intervening lens.

In a closed cosmological model, an observer at an arc time  $\pi/2$  wants to place a gravitational lens to focus light from a distance  $z = 10$  on himself. Where should he place a lens to use the least mass? What mass density is needed?

The arc distance to  $z = 10$  comes out to be  $\chi = 1.14$  radians. To be most effective the lens should be placed halfway in arc distance, at an arc time of 1.0 radian, a redshift of  $z = 1.18$ .

To find the mass of the deflector, note that it must be the shear in phase space

$$(q, p) \mapsto (q, p - 3.85q).$$

In our dimensionless phase space, we have a mass density  $\rho_c$  leading to a shear

$$(q, p) \mapsto (q, p - 4\pi\rho_c qR),$$

where  $R$  is the scale factor of the universe at the epoch of deflection. Thus we must have

$$\rho_c = 2.3 \times 10^3 \frac{M_\odot}{pc^2} \left( \frac{H_0}{50 \text{ km sec}^{-1} \text{ Mpc}^{-1}} \right).$$

Since  $H_0$  is so poorly known, current practice is to keep explicit the way that the answer depends on it.

### 66. Moving frames

If you actually try<sup>8</sup> to compute curvature tensors from the definition in equation 56.6, you will find it a lot of work. As with most calculations in differential geometry, trickery is rewarded. Relativists do not spend most of their time grinding out Riemann tensors. Just as the energy Lagrangian for geodesics allows us to compute the affine connection easily, there is an efficient means for computing curvature tensors, called the method of moving frames. The basic idea is an old one that is extremely versatile and useful in many geometric calculations. Cartan systematically exploited this idea, and the name *repere mobile* is his.

Simply put, the idea of moving frames is to represent geometric objects not in a basis that is derived from the coordinates, but in one that is tied to some of the intrinsic geometric features of the problem. To compute Riemann curvature tensors, we pick a set of basis vectors in which the Riemannian metric is simply represented. In a problem with fluid flow, we might want to align one basis vector along the flow. In a radiation problem, we can simplify things by picking a basis of null vectors aligned with the propagation of the radiation.

Two further ideas are needed to make these moving frames into an efficient computational tool. The first is to use an explicit notation for vectors and tensors rather than lists of components. This is the source of the efficiency of the energy-Lagrangian method for connections. We waste no time computing components that vanish. The second idea, more ingenious, is to solve certain crucial equations by guesswork. For spacetimes of reasonable simplicity, this is often quite easy, and only for reasonably simple spacetimes will you be able to use the curvature tensor for anything.

*The moving frame*

In an  $n$ -dimensional manifold we pick  $n$  smooth vector fields to be used as a basis. You should not expect to cover the entire manifold with one such set, of course. Write these basis vector fields  $E_i$ , where the index  $i$  labels different vector fields, not the components of a single vector field. These vectors can be represented in a coordinate basis in the usual fashion

$$E_i = E_i^j \frac{\partial}{\partial x^j},$$

where the numbers  $E_i^j$  are the coordinate components. Note the aesthetic placement of the indices here. We will rarely ever need the coordinate components of these frame vectors; so we will not bother to invent a special notation to distinguish frame indices from coordinate-component indices. We will deal here with a frame adapted to a given Riemannian metric. We will pick our frame of basis vectors so that the metric inner products of the frame vectors

$$\eta_{ij} = \mathcal{G}(E_i, E_j),$$

are constant. For most spacetime uses, the vectors will be chosen so that the matrix  $\eta_{ij}$  will be diagonal, with three plus ones and a minus one.

In practice it is easier to pick a basis of 1-form fields such that

$$\mathcal{G} = \eta_{ij} \omega^i \otimes \omega^j,$$

where each  $\omega^i$  is a smooth 1-form field, dual to the  $E_i$ s

$$\omega^i \cdot E_j = \delta_j^i.$$

Any 1-form field can be written in terms of such a basis

$$\alpha = \alpha_i \omega^i,$$

and the frame components  $\alpha_i$  can be found from

$$\alpha_i = \alpha \cdot E_i.$$

Suitable  $\omega^i$  are easily picked. They are not unique.

**Example:** The spacetime metric tensor

$$\mathcal{G} = -dt^2 + R^2(t)[d\chi^2 + \sin^2 \chi(d\theta^2 + \sin^2 \theta d\phi^2)]$$

can be written

$$\mathcal{G} = -dt^2 + (R d\chi)^2 + (R \sin \chi d\theta)^2 + (R \sin \chi \sin \theta d\phi)^2$$

and an obvious choice of a frame is

$$\begin{aligned}\omega^t &= dt, & \omega^\chi &= R d\chi, \\ \omega^\theta &= R \sin \chi d\theta, & \omega^\phi &= R \sin \chi \sin \theta d\phi.\end{aligned}$$

This spacetime is called Robertson-Walker spacetime. We continue with it in what follows.

The  $\omega^i$  themselves must be represented in terms of a coordinate basis, which after all has some advantages. Here  $d\omega^i$  does not necessarily vanish.

### Affine connection

The torsion-free affine connection associated with the metric  $\mathcal{G}$  can be represented by the coefficients  $F^k_{ij}$  in the expansion

$$\bar{\nabla}_{E_i} E_j = F^k_{ij} E_k.$$

The bar is added here to prevent confusion with the coordinate-based coefficients. The  $F$  are not symmetric even when there is no torsion. To use an explicit notation (first trick), we introduce instead the 1-forms

$$\omega^i_j = F^i_{jk} \omega^k. \quad (66.1)$$

Our index placement will follow the usual convention of one-up-one-down in summations. The metric enters via the use of the  $\eta_{ij}$  to raise and lower indices:

$$\omega_{ij} = \eta_{ik} \omega^k_j.$$

The 1-forms  $\omega_{ij}$  are called connection 1-forms. Remember that they are 1-forms even though they have two indices. These indices are just numerical labels. Raising and lowering the indices requires some sign changes; so we do not use the boldface missing-index convention here. The connection 1-forms satisfy

$$\omega_{ij} + \omega_{ji} = 0, \quad (66.2)$$

and

$$d\omega_i + \omega_{ij} \omega^j = 0. \quad (66.3)$$

In a moment we will derive these equations. The second clever trick will be to find the  $\omega_{ij}$  not from equation 66.1, but by guessing solutions to equation 66.3.

**Example:** The connection 1-forms for the Robertson-Walker spacetime are

$$\omega_{\chi t} = R' d\chi = (R'/R)\omega^\chi,$$

$$\omega_{\theta t} = (R'/R)\omega^\theta,$$

$$\omega_{\phi t} = (R'/R)\omega^\phi,$$

$$\omega_{\theta\chi} = \cos \chi d\theta,$$

$$\omega_{\phi\chi} = \cos \chi \sin \theta d\phi,$$

$$\omega_{\phi\theta} = \cos \theta d\phi.$$

The  $\omega^i_j$  contain all the information about the affine connection. It is easy to show that we have

$$\nabla_A B = [(\omega^j_k \cdot A)(\omega^k \cdot B) + \nabla_A(\omega^j \cdot B)]E_j,$$

and so a knowledge of the connection 1-forms is equivalent to a knowledge of the connection.

### Proofs

To show that the  $\omega_{ij}$  are skew, we start from the constancy of the  $\eta_{ij}$ ,

$$\nabla_{E_l} \eta_{ij} = 0,$$

and write this

$$\nabla_{E_l} \mathcal{G}(E_i, E_j) = 0;$$

since the connection is derived from this metric,

$$\mathcal{G}(\nabla_{E_l} E_i, E_j) + \mathcal{G}(E_i, \nabla_{E_l} E_j) = 0.$$

Now, just chase around the definitions to find

$$\eta_{jk} \mathbf{F}^k_{li} + \eta_{ik} \mathbf{F}^k_{lj} = 0$$

and contract with  $\omega^l$  to arrive at equation 66.2.

To derive equation 66.3, we start from the identity

$$d\omega(A, B) = \nabla_A(\omega \cdot B) - \nabla_B(\omega \cdot A) - \omega \cdot [A, B]$$

and evaluate it on our frame

$$d\omega_i \cdot (E_k, E_l) = \nabla_{E_k}(\omega_i \cdot E_l) - \nabla_{E_l}(\omega_i \cdot E_k) - \omega_i \cdot [E_k, E_l].$$

Since we have

$$\omega_i \cdot E_l = \eta_{il},$$

which are constants, the first two terms drop out. For zero torsion

$$\begin{aligned} [E_k, E_l] &= \nabla_{E_k} E_l - \nabla_{E_l} E_k \\ &= (\mathbb{F}^m_{kl} - \mathbb{F}^m_{lk}) E_m, \end{aligned}$$

and so

$$d\omega_i \cdot (E_k, E_l) = -\eta_{im} (\mathbb{F}^m_{kl} - \mathbb{F}^m_{lk}). \quad (66.4)$$

Now we calculate

$$-\omega_{ij} \omega^j \cdot (E_j, E_l) = -(\omega_{ij} \cdot E_k)(\omega^j \cdot E_l) + (\omega_{ij} \cdot E_l)(\omega^j \cdot E_k),$$

and this collapses to the right-hand side of equation 66.4, proving the result.

We can solve systematically for the  $\mathbb{F}^i_{jk}$  coefficients by expanding the  $d\omega_i$  in our basis

$$d\omega_i = c^i_{|jk|} \omega^j \omega^k,$$

and then a computation just like that leading from equation 58.1 to equation 58.2 gives us

$$\mathbb{F}^i_{qp} = \frac{1}{2} \eta^{il} (c_{lpq} + c_{pql} - c_{qpl}),$$

with

$$c_{ijk} = \eta_{il} c^l_{jk}.$$

Finally, we need to show that the solution of equation 66.3 is unique; otherwise it will not be correct merely to guess a solution. Since the equations are linear in the connection 1-forms, we look at the equations for the difference between two solutions

$$\alpha_{ij} \omega^j = 0 \quad \alpha_{ij} + \alpha_{ji} = 0$$

and show that this difference must vanish. Expanding in our basis we have

$$\alpha_{ij} = a_{ijk} \omega^k$$

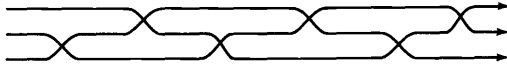
and our two equations tell us that

$$a_{ijk} = a_{ikj}$$

and

$$a_{ijk} = -a_{jik}. \quad (66.5)$$

If we perform a full braid on the indices as diagrammed in Figure 66.1, then we will have returned the indices to their original locations. But we will have used the symmetry operation in equation 66.5 three times, and



**Figure 66.1.** The pattern for permuting the indices on the  $a_{ijk}$ .

this changes the sign. Thus the  $a_{ijk}$  must vanish, and so our solution must be unique.

### *Curvature*

In a similar fashion we represent the curvature tensor by a set of 2-forms  $\Omega_{ij}$  such that

$$\mathcal{G}(E_i, R(A, B)E_j) = \Omega_{ij}(A, B). \quad (66.6)$$

The operator  $R(\cdot, \cdot)$  was defined in equation 56.4. These curvature 2-forms are easily computed from the connection 1-forms

$$\Omega_{ij} = d\omega_{ij} + \omega_{il}\omega^l{}_j. \quad (66.7)$$

These curvature 2-forms determine the frame components of the curvature tensor

$$\Omega_{ij} = R_{ij|kl|} \omega^k \wedge \omega^l$$

(ordered summation), and so they incorporate the efficiencies of an explicit notation. It is a routine (but page-length) calculation to show that the 2-forms defined by equation 66.7 satisfy equation 66.6.

The scheme in which we pick an orthonormal frame  $\omega^i$ , guess solutions for the connection 1-forms, and then use equation 66.7 to compute the curvature is usually the most efficient way to calculate curvature tensors. Also, working in terms of frame components is useful in many other problems, since they have the correct physical dimensions and sizes.





# Bibliography

- Abraham, Ralph, and Jerrold E. Marsden (1978). *Foundations of Mechanics*, 2nd ed. San Mateo, Calif.: Benjamin/Cummings.
- Akhiezer, Naum I. (1962). *The Calculus of Variations*. New York: Blaisdell.
- Arnold, V. I. (1978). *Mathematical Methods of Classical Mechanics*. New York: Springer.
- Bilby, B. A., R. Bullough, and E. Smith (1965). *Proc. Roy. Soc. London A231*, 263–73.
- Bott, Raoul, and Loring W. Tu (1982). *Differential Forms in Algebraic Topology*. New York: Springer.
- Burke, William L. (1970). *Phys. Rev. A2*, 1501–5.
- Burke, William L. (1980). *Spacetime, Geometry, Cosmology*. Mill Valley, Calif.: University Science Books.
- Choquet-Bruhat, Yvonne, Cecile DeWitt-Morette, and Margaret Dillard-Bleick (1982). *Analysis, Manifolds and Physics*. Amsterdam: North-Holland.
- Courant, R., and D. Hilbert (1953). *Methods of Mathematical Physics, Vol. I*. New York: Interscience.
- Courant, R., and D. Hilbert (1962). *Methods of Mathematical Physics, Vol. II*. New York: Interscience.
- de Rham, G. (1960). *Variétés différentiables*. Paris: Hermann.
- Faulkner, John, and Brian P. Flannery (1978). *Astrophys. J.* 220, 1125–32.
- Gelfand, I. M., and S. V. Fomin (1963). *Calculus of Variations*. Englewood Cliffs, N.J.: Prentice-Hall.
- Guillemin, Victor, and Alan Pollack (1974). *Differential Topology*. Englewood Cliffs, N.J.: Prentice-Hall.
- Goldstein, Herbert (1959). *Classical Mechanics*. Reading, Mass.: Addison-Wesley.
- Grossman, N. (1974). *J. Geophys. Res.* 79, 689–94.
- Harrison, B. K., and F. B. Estabrook (1971). *J. Math. Phys.* 12, 653–66.
- Hermann, Robert (1968). *Differential Geometry and the Calculus of Variations*. New York: Academic Press.
- Hermann, Robert (1973). *Geometry, Physics, and Systems*. New York: Marcel Dekker.

- Hilbert, D., and S. Cohn-Vossen (1952). *Geometry and the Imagination*. New York: Chelsea.
- Hirsch, Morris W., and Stephen Smale (1974). *Differential Equations, Dynamical Systems, and Linear Algebra*. New York: Academic Press.
- Israel, R. B. (1979). *Convexity in the Theory of Lattice Gases*. Princeton, N.J.: Princeton Univ. Press.
- Jackson, J. D. (1962). *Classical Electrodynamics*. New York: Wiley.
- Loomis, L. H., and S. Sternberg (1968). *Advanced Calculus*. Reading, Mass.: Addison-Wesley.
- Martin, R., and R. Reeves (1977). *The 99 Critical Shots in Pool*. New York: Quadrangle Books.
- Misner, Charles W., Kip S. Thorne, and John Archibald Wheeler (1973). *Gravitation*. San Francisco: Freeman.
- Nelson, Edward (1967). *Tensor Analysis*. Princeton, N.J.: Princeton Univ. Press
- Newman, William M., and Robert F. Sproull (1973). *Principles of Interactive Computer Graphics*. New York: McGraw-Hill.
- Panofsky, Wolfgang K. H., and Melba Phillips (1955). *Classical Electricity and Magnetism*. Reading, Mass.: Addison-Wesley.
- Pars, L. A. (1962). *An Introduction to the Calculus of Variations*. New York: Wiley.
- Pars, L. A. (1965). *A Treatise on Analytical Dynamics*. New York: Wiley.
- Penrose, R. (1968). In Cecile M. DeWitt and John A. Wheeler, *Batelle Rencontres*. New York: Benjamin.
- Robertson, H. P., and T. Noonan (1968). *Relativity and Cosmology*. Philadelphia: Saunders.
- Saletan, E. J., and A. H. Cromer (1971). *Theoretical Mechanics*. New York: Wiley.
- Santilli, Ruggero Maria (1978). *Foundations of Theoretical Mechanics, Vol. I*. New York: Springer.
- Schouten, J. A. (1951). *Tensor Analysis for Physicists*. Oxford, Eng.: Clarendon Press, Oxford.
- Schouten, J. A. (1954). *Ricci-Calculus*. New York: Springer.
- Slebdozinski, Wladyslaw (1970). *Exterior Forms and Their Applications*. Warsaw: Polish Scientific Publishers.
- Sommerfeld, A. (1952). *Mechanics*. New York: Academic Press.
- Soper, Davison Eugene (1976). *Classical Field Theory*. New York: Wiley-Interscience.
- Spivak, M. (1965). *Calculus on Manifolds*. New York: Benjamin.
- Sternberg, S. (1964). *Lectures on Differential Geometry*. Englewood Cliffs, N.J.: Prentice-Hall.
- Syngé, J. L., and A. Schild (1949). *Tensor Calculus*. New York: Dover.
- Warner, Frank W. (1971). *Foundations of Differentiable Manifolds and Lie Groups*. Glenview, Ill.: Scott, Foresman.
- Weinberg, S. (1972). *Gravitation and Cosmology*. New York: Wiley.
- Weyl, H. (1922). *Space-Time-Matter*. New York: Dover.
- Whittaker, E. T. (1959). *Analytical Dynamics of Particles and Rigid Bodies*. Cambridge: Cambridge University Press.

# Index

- aberration, 45
- action, 240, 324
  - Lie group, 116
- adiabatic demagnetization, 114
- adjoint, 200
- affine space, 8, 11, 13
- affine structure of spacetime, 39
- alternating product, 31, 154
- angle operator, 153
- anholonomic constraints in Lagrangian mechanics, 312
- atlas, 55
  - defective, 57
- automorphism, 81
  
- balance law, electrodynamics, 287
- base space, 85
- basis,
  - dual, 29
  - cotangent space, 66
  - tangent space, 65
- billiard balls, 310
- bivector, 32
  - simple, 36
  - twisted, 187
- block index, 177
- bound charge, 294
- boundary, orientation, 199
- brightness, 334
- bundle, contact, 102
  
- calculus of variations, 225
- canonical 1-form, 89
- Cartan form, 265
- Cartan identity, 158
- Cartesian product, 6
- Cauchy characteristic, 261
- chain rule, 24
- chains, 205
  
- characteristics, 98
- chart, 53
- Clebsch form, 176
- clock, 39
- cohomology, 202
- components, tensor, 30
- conductor, 293
  - fluid, 126
  - moving, 298
- conformal structure of spacetime, 39
- conformal symmetry, 170
- connection, 349
  - crystal, 358
  - geodesic Lagrangian for calculation, 376
  - geodesy, 356
  - Lie group, 356
  - metric, 355, 367
  - symmetric, 366
- connection 1-forms, 352, 404
- conservation laws, 219
  - calculus of variations, 233, 239, 244
  - curved spacetime, 382
  - exterior differential systems, 264
- conservative systems, 306
- constraints, 233, 234
- contact bundle, 99
- contact element, 103
- contact structure, 103
- contact symmetry, 144
- contact transformations, 136
- contraction, 30
  - angle operator for, 153
- convexity, 111
- coordinate transformations, 81
- coordinates, spacetime, 37
- cosmology, 398
- cotangent bundle, 84
  - natural structure of, 88
- covariance, 11, 44

- covariant derivatives, 365, 366
- covectors, 18, 27
  - free, 19
- Curie law, 113
- curl, 157
- curvature,
  - connection, 359
  - moving frames, 407
- curvature tensor, 364
- cycles, 205
  
- deformations of integrals, 201
- degenerate metrics, 164
- densities, 149, 183
- de Rham operator, 168, 200
- derivation operator, 67
- determinants, xii, 180
- differential, 22
- differential calculus, 21
  - second differential, 24
- differential forms, 147, 272
- diffusion, 207, 211
- dihedral, 46
- dispersion relation, 339
- distribution, 134
- divergence, 158
  - metric induced, 369
  - vector field, 328
- divergence theorem, 200
- Doppler shift, 44
- dragging, 125
- dual space, 19
- dynamical system, 94
  - singular, 96
  
- elastic beam, 317, 342
- elastic string, 317
- electrodynamics as exterior differential system, 250
- energy momentum tensor for scalar field, 269
- enthalpy, 140
- equation of state, 111
- equivalence class, 8
- equivalence relation, 7
- etendue*, 336
- Euler–Lagrange equations, 244
- event, 37
- extension, 148
- exterior calculus, 153
- exterior derivative, 156
- exterior differential system, 248
- exterior product, 31, 154
- extrema, 25
  
- fiber bundle, 85
- field theory, 85, 314
  
- flow, of vector field, 95
- fluid flow, isentropic, 252
- force, electrostatic, 285
- force, as 1-form, 271
- free charge in conductor, 299
- free particles, 38
  - as exterior differential system, 266
- Frobenius Theorem, 71, 134
- frozen magnetic flux, 301
- function
  - composition of, 6
  - generalized, 20
  - inverse, 6
  - partial evaluation, 6
- fundamental equation, 109
  
- Gaussian coordinates, 360
- Gaussian optics, 336
- generalized force, 308
- geodesic deviation, 377–381
- geodesics, 374
  - as Hamiltonian system, 377
- geodesy, 356
- Gibbs–Duhem relation, 111
- gravitational lens, 395
- Green’s Theorem, 200
- group velocity, 342
  
- Hamiltonian dynamical system, 96, 144, 320
  - genesis, 327
- Hamiltonian optics, 333
- Hamilton’s equations, 322
- Hamilton’s principle, 225, 307, 315, 324
- heat capacity, 247
- heat equation, 207, 248, 264
  - isovectors for, 258, 259
  - similarity solutions, 261
  - symmetries, 210
- Henon map, 81
- Hessian, 75
- Hodge star operator, 159
  - twisted, 193
- holonomy, 132
  - in thermodynamics, 245
- homogeneous coordinates, 8
- horizontal contact elements, 349
- hypersurface element, 101
  
- ideal, 6
  - contact structure represented by, 105
  - differential, 213
  - partial differential equation represented by, 213
- ideal gas, 111
- index notation, 6, 7, 31, 176
- initial value problem, electrodynamic, 277

- integrability conditions, 135, 278
- integrable vector fields, 134
- integral curve, 94
- integral invariants, 323, 329
- integral submanifold, 134
  - for partial differential equation, 215
- integration, 194
  - informal definition, 79
- isometry, 141, 169
- isovectors, 258
  
- Jacobi identity, 70
- jet bundle, 107
- junction conditions, in electrodynamics, 279
  
- Killing vector, 142, 368
- Killing's equation, 143
- kinetic theory, 223
- Kronecker deltas, 177
  
- Lagrange multiplier, 25, 27
  - in calculus of variations, 241
- Lagrange's equations, 311
- Lagrangian, 225
- Lagrangian field theory, 314
- Lagrangian submanifold, 326
- Laplace-de Rham operator, 168, 200
- Laplace's equation, 166
  - regular integral submanifolds for, 254
- Legendre transformation, 136, 138
- Lie algebra, 117
- Lie bracket, 68
- Lie derivative, 121, 158
- Lie group, 115
- Lie transport, 125
- lift, 88, 92, 354
- light deflection, gravitational, 393
- light propagation, cosmological, 400
- light second, 273
- light signals, 39
- line element, 99
- linear operator, notation for, 6
- linear space, 12
- linear transformations, covariance under, 12
- Liouville's Theorem, 323
- Lorentz gauge condition, 284
- Lorentz group action, 117
- Lorentz metric, signature, 36
- Lorentz transformation, 41
  
- magnetization, 296
- manifold, 51, 52
  - implicit argument for, 83
- map, 5
  - calculus of, 77
  
- Maxwell relations, 245
- Maxwell's equations, 273, 275
  - conservation law for, 224
  - exterior differential system for, 250
- metric tensor, 75
- moving frames, 402
  
- normal forms for tensors, 173
- null cone, 119
  
- Ohm's law, 293
- 1-form, 27, 61
- operator, bilinear, 27
- ordered summation convention, 177
- orientation, 8, 150, 184
  - of a boundary, 199
- orthogonal matrices, as a manifold, 84
  
- parallel transport, 348, 354
- parameter, special affine, 374
- parking a car, 132
- particles, conservation of, 219
- partition of unity, 58, 105
- perfect lens, impossibility of, 338
- permutation symbol, 156, 180
- Poincaré's Lemma, 202
- point charge, 285
- Poisson's equation, 278
- polarizable materials, 294
- potentials, 202
  - electrodynamic, 283
  - for exterior differential system, 267
- Poynting-Heaviside Theorem, 291
- projective plane, 8, 53
- pseudopotentials, 268
- pseudosphere, 49, 56
- pseudospherical geometry, 49
- pullback, 79
  - of twisted forms, 192
- pushforward, 78
  
- reference frame, 40
- regular integral submanifold, 254
- Reynold's number, magnetic, 301
- Riemann tensor, 364
  - symmetries, 368
- rotating disc, 76
- runaway charges, 289
  
- scalars, 30
- section, of a bundle, 86
- sharp operator, 160
- shock waves, 253
- similarity solutions, 258, 261
- slow manifold, 97
- smoothing, in electrodynamics, 297
- special relativity, 37
  - metric structure, 43

- sphere,  $n$ -dimensional, 55
- star operator, 159
- statistical mechanics, 106
- Steiner's surface, 82
- stereographic projection, 55, 119
- Stokes' Theorem, 199
- stress, on conductor, 288
- submanifold, 82, 102
- summation convention, 6
- surjective map, 83
- symmetries, 141
  - calculus of variations, 233, 239, 244
  - mechanical systems, 310
  - similarity solutions, 258
- symplectic geometry, 325
- tangency, of maps, 59
- tangent bundle, 90
  - natural structure of, 91
- tangent map, 93
- tangent space, 52
- tangent vector, 8, 11, 60
- telegraph equation, 212
- tensor, 30
  - algebra, 27
  - odd kind, 183
  - symmetric, 73
- tensor product, 28, 30
- thermodynamics, 108, 245
- tides, 379
- time, 43
- topology, and cohomology, 205
- torsion, 361
- traffic flow, 223
- transformation, infinitesimal, 96
- transformation group, 116
- translation, in Lie group, 116
- transverse orientation, 184
- trivial bundle, 87
- twisted differential forms, 151, 183
  - integration of, 197
- twisted vector, 13
- 2-form, geometric representation of, 32
- uniqueness, electrodynamic, 292
- units, electrodynamic, 273
- variational principles for exterior systems, 264
- vector, 16, 17
- vector field, 93
- velocity addition, 140
- wave diagram, 341
- wave equation, 167
  - as exterior differential system, 249
- wave packets, 338
- wedge product, 31, 154
- worldline, 37