

**SERIES ON KNOTS AND EVERYTHING**

*Editor-in-charge:* Louis H. Kauffman

---

*Published:*

Vol. 1: **Knots and Physics**


*L. H. Kauffman*

Vol. 2: **How Surfaces Intersect in Space**

*J. S. Carter*

Vol. 3: **Quantum Topology**

*edited by L. H. Kauffman & R. A. Baadhio*

 Series on Knots and Everything — Vol. 4

---

# **GAUGE FIELDS, KNOTS AND GRAVITY**

---

**John Baez**

Department of Mathematics  
University of California  
Riverside

**Javier P. Muniain**

Department of Physics  
University of California  
Riverside

Published by

World Scientific

P O Box 128, Fairview, NJ 07641

USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661

UK office: 73 Lynton Mead, Totteridge, London N20 8DH

## *To Our Parents*

### Library of Congress Cataloging-in-Publication Data

Baez, John C., 1961–

Gauge fields, knots, and gravity / John C. Baez and Javier P.

Muniain.

p. cm. -- (Series on knots and everything ; vol. 4)

Includes index.

ISBN 9810217293 -- ISBN 9810220340 (pbk)

1. Gauge fields (Physics) 2. Quantum gravity. 3. Knot theory.  
4. General relativity (Physics) 5. Electromagnetism. I. Muniain,  
Javier P. II. Title. III. Series: K & E series on knots and  
everything ; vol. 4.

QC793 .3.F5B33 1994

530.1'4--dc20

94-3438

CIP

Copyright © 1994 by World Scientific Publishing Co. Pte. Ltd.

*All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.*

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 27 Congress Street, Salem, MA 01970, USA.

Printed in Singapore by Uto-Print

# Preface

Two of the most exciting developments of 20th century physics were general relativity and quantum theory, the latter culminating in the ‘standard model’ of particle interactions. General relativity treats gravity, while the standard model treats the rest of the forces of nature. Unfortunately, the two theories have not yet been assembled into a single coherent picture of the world. In particular, we do not have a working theory of gravity that takes quantum theory into account. Attempting to ‘quantize gravity’ has led to many fascinating developments in mathematics and physics, but it remains a challenge for the 21st century.

The early 1980s were a time of tremendous optimism concerning string theory. This theory was very ambitious, taking as its guiding philosophy the idea that gravity could be quantized only by unifying it with all the other forces. As the theory became immersed in ever more complicated technical issues without any sign of an immediate payoff in testable experimental predictions, some of this enthusiasm diminished among physicists. Ironically, at the same time, mathematicians found string theory an ever more fertile source of new ideas. A particularly appealing development to mathematicians was the discovery by Edward Witten in the late 1980s that Chern-Simons theory — a quantum field theory in 3 dimensions that arose as a spin-off of string theory — was intimately related to the invariants of knots and links that had recently been discovered by Vaughan Jones and others. Quantum field theory and 3-dimensional topology have become firmly bound together ever since, although there is much that remains mysterious about the relationship.

While less popular than string theory, a seemingly very different ap-

proach to quantum gravity also made dramatic progress in the 1980s. Abhay Ashtekar, Carlo Rovelli, Lee Smolin and others discovered how to rewrite general relativity in terms of 'new variables' so that it more closely resembled the other forces of nature, allowing them to apply a new set of techniques to the problem of quantizing gravity. The philosophy of these researchers was far more conservative than that of the string theorists. Instead of attempting a 'theory of everything' describing all forces and all particles, they attempted to understand quantum gravity *on its own*, following as closely as possible the traditional guiding principles of both general relativity and quantum theory. Interestingly, they too were led to the study of knots and links. Indeed, their approach is often known as the 'loop representation' of quantum gravity. Furthermore, quantum gravity in 4 dimensions turned out to be closely related to Chern-Simons theory in 3 dimensions. Again, there is much that remains mysterious about this. For example, one wonders why Chern-Simons theory shows up so prominently both in string theory and the loop representation of quantum gravity. Perhaps these two approaches are not as different as they seem!

It is the goal of this text to provide an *elementary* introduction to some of these developments. We hope that both physicists who wish to learn more differential geometry and topology, and mathematicians who wish to learn more gauge theory and general relativity, will find this book a useful place to begin. The main prerequisites are some familiarity with electromagnetism, special relativity, linear algebra, and vector calculus, together with some of that undefinable commodity known as 'mathematical sophistication'.

The book is divided into three parts that treat electromagnetism, gauge theory, and general relativity, respectively. Part I of this book introduces the language of modern differential geometry, and shows how Maxwell's equations can be drastically simplified using this language. We stress the coordinate-free approach and the relevance of global topological considerations in understanding such things as the Bohm-Aharonov effect, wormholes, and magnetic monopoles. Part II introduces the mathematics of gauge theory — fiber bundles, connections and curvature — and then introduces the Yang-Mills equation, Chern classes, and Chern-Simons classes. It also includes a brief introduction to knot theory and its relation to Chern-Simons theory. Part

III introduces the basic concepts of Riemannian and semi-Riemannian geometry and then concentrates on topics in general relativity of special importance to quantum gravity: the Einstein-Hilbert and Palatini formulations of the action principle for gravity, the ADM formalism, and canonical quantization. Here we emphasize tensor computations written in the notation used in general relativity. We conclude this part with a sketch of Ashtekar's 'new variables' and the way Chern-Simons theory provides a solution to the Wheeler-DeWitt equation (the basic equation of canonical quantum gravity).

While we attempt to explain everything 'from scratch' in a self-contained manner, we really hope to lure the reader into further study of differential geometry, topology, gauge theory, general relativity and quantum gravity. For this reason, we provide copious notes at the end of each part, listing our favorite reading material on all these subjects. Indeed, the reader who wishes to understand any of these subjects in depth may find it useful to read some of these references in parallel with our book. This is especially true because we have left out many relevant topics in order to keep the book coherent, elementary, and reasonable in size. For example, we have not discussed fermions (or mathematically speaking, spinors) in any detail. Nor have we treated principal bundles. Also, we have not done justice to the experimental aspects of particle physics and general relativity, focusing instead upon their common conceptual foundation in gauge theory. The reader will thus have to turn to other texts to learn about such matters.

One really cannot learn physics or mathematics except by doing it. For this reason, this text contains over 300 exercises. Of course, far more exercises are assigned in texts than are actually done by the readers. At the very least, we urge the reader to read and ponder the exercises, the results of which are often used later on. The text also includes 130 illustrations, since we wish to emphasize the geometrical and topological aspects of modern physics. Terms appear in boldface when they are defined, and all such definitions are referred to in the index.

This book is based on the notes of a seminar on knot theory and quantum gravity taught by J.B. at U. C. Riverside during the school year 1992-1993. The seminar concluded with a conference on the subject, the proceedings of which will appear in a volume entitled *Knots*

and Quantum Gravity.

We would like to thank Louis Kauffman for inviting us to write this book, and also Chris Lee and Ms. H. M. Ho of World Scientific for helping us at every stage of the writing and publication process. We also wish to express our thanks to Edward Heflin and Dardo D. Píriz for reading parts of the manuscript and to Carl Yao for helping us with some  $\text{\LaTeX}$  complications. Scott Singer of the Academic Computing Graphics and Visual Imaging Lab of the U. C. Riverside deserves special thanks and recognition for helping us to create the book cover. Some of the graphics used for the design of the cover were generated with *Mathematica*, by Wolfram Research, Inc.; these were kindly given to us by Joe Grohens of WRI.

J.B. is indebted to many mathematicians and physicists for useful discussions and correspondence, but he would particularly like to thank Abhay Ashtekar, whose work has done so much to unify the study of gauge fields, knots and gravity. He would also like to thank the readers of the USENET physics and mathematics newsgroups, who helped in many ways with the preparation of this book. He dedicates this book to his parents, Peter and Phyllis Baez, with profound thanks for their love. He also gives thanks and love to his mathematical muse, Lisa Raphals.

J.M. dedicates this book to his parents, Luis and Crescencia Pérez de Muniáin y Mohedano, for their many years of continued love and support. He is grateful to Eleanor Anderson for being a patient and inspiring companion during the long and hard hours taken to complete this book. He also acknowledges José Wudka for many discussions on quantum field theory.

# Contents

Preface	vii
<b>I Electromagnetism</b>	<b>1</b>
1 Maxwell's Equations	3
2 Manifolds	15
3 Vector Fields	23
4 Differential Forms	39
5 Rewriting Maxwell's Equations	69
6 DeRham Theory in Electromagnetism	103
Notes to Part I	153
<b>II Gauge Fields</b>	<b>159</b>
1 Symmetry	161
2 Bundles and Connections	199
3 Curvature and the Yang-Mills Equation	243
4 Chern-Simons Theory	267

<b>5</b>	<b>Link Invariants from Gauge Theory</b>	<b>291</b>
	<b>Notes to Part II</b>	<b>353</b>
<b>III</b>	<b>Gravity</b>	<b>363</b>
<b>1</b>	<b>Semi-Riemannian Geometry</b>	<b>365</b>
<b>2</b>	<b>Einstein's Equation</b>	<b>387</b>
<b>3</b>	<b>Lagrangians for General Relativity</b>	<b>397</b>
<b>4</b>	<b>The ADM Formalism</b>	<b>413</b>
<b>5</b>	<b>The New Variables</b>	<b>437</b>
	<b>Notes to Part III</b>	<b>451</b>
	<b>Index</b>	<b>457</b>

## Gauge Fields, Knots, and Gravity

**Part I**

**Electromagnetism**

# Chapter 1

## Maxwell's Equations

*Our whole progress up to this point may be described as a gradual development of the doctrine of relativity of all physical phenomena. Position we must evidently acknowledge to be relative, for we cannot describe the position of a body in any terms which do not express relation. The ordinary language about motion and rest does not so completely exclude the notion of their being measured absolutely, but the reason of this is, that in our ordinary language we tacitly assume that the earth is at rest.... There are no landmarks in space; one portion of space is exactly like every other portion, so that we cannot tell where we are. We are, as it were, on an unruffled sea, without stars, compass, sounding, wind or tide, and we cannot tell in what direction we are going. We have no log which we can case out to take a dead reckoning by; we may compute our rate of motion with respect to the neighboring bodies, but we do not know how these bodies may be moving in space. – James Clerk Maxwell, 1876.*

Starting with Maxwell's beautiful theory of electromagnetism, and inspired by it, physicists have made tremendous progress in understanding the basic forces and particles constituting the physical world. Maxwell showed that two seemingly very different forces, the electric and magnetic forces, were simply two aspects of the 'electromagnetic field'. In so doing, he was also able to explain *light* as a phenomenon in which ripples in the electric field create ripples in the magnetic field, which in turn create new ripples in the electric field, and so on. Shockingly, however, Maxwell's theory also predicted that light emitted by a moving body would travel no faster than light from a stationary body.



Eventually this led Lorentz, Poincaré and especially Einstein to realize that our ideas about space and time had to be radically revised. That the motion of a body can only be measured relative to another body had been understood to some extent since Galileo. Taken in conjunction with Maxwell's theory, however, this principle forced the recognition that in addition to the rotational symmetries of space there must be symmetries that mingle the space and time coordinates. These new symmetries also mix the electric and magnetic fields, charge and current, energy and momentum, and so on, revealing the world to be much more coherent and tightly-knit than had previously been suspected.

There are, of course, forces in nature besides electromagnetism, the most obvious of which is gravity. Indeed, it was the simplicity of gravity that gave rise the first conquests of modern physics: Kepler's laws of planetary motion, and then Newton's laws unifying celestial mechanics with the mechanics of falling bodies. However, reconciling the simplicity of gravity with relativity theory was no easy task! In seeking equations for gravity consistent with his theory of special relativity, Einstein naturally sought to copy the model of Maxwell's equations. However, the result was not merely a theory in which ripples of some field propagate through spacetime, but a theory in which the geometry of spacetime itself ripples and bends. Einstein's equations say, roughly, that energy and momentum affect the *metric* of spacetime (whereby we measure time and distance) much as charges and currents affect the electromagnetic field. This served to heighten hopes that much or perhaps even all of physics is fundamentally *geometrical* in character.

There were, however, severe challenges to these hopes. Attempts by Einstein, Weyl, Kaluza and Klein to further unify our description of the forces of nature using ideas from geometry were largely unsuccessful. The reason is that the careful study of atoms, nuclei and subatomic particles revealed a wealth of phenomena that do not fit easily into any simple scheme. Each time technology permitted the study of smaller distance scales (or equivalently, higher energies), new puzzles arose. In part, the reason is that physics at small distance scales is completely dominated by the principles of *quantum theory*. The naive notion that a particle is a point tracing out a path in spacetime, or that a field assigns a number or vector to each point of spacetime, proved to be wholly inadequate, for one cannot measure the position and velocity

of a particle simultaneously with arbitrary accuracy, nor the value of a field and its time derivative. Indeed, it turned out that the distinction between a particle and field was somewhat arbitrary. Much of 20th century physics has centered around the task of making sense of microworld and developing a framework with which one can understand subatomic particles and the forces between them in the light of quantum theory.

Our current picture, called the standard model, involves three forces: electromagnetism and the weak and strong nuclear forces. These are all 'gauge fields', meaning that they are described by equations closely modelled after Maxwell's equations. These equations describe *quantum* fields, so the forces can be regarded as carried by particles: the electromagnetic force is carried by the photon, the weak force is carried by the  $W$  and  $Z$  particles, and the strong force is carried by gluons. There are also charged particles that interact with these force-carrying particles. By 'charge' here we mean not only the electric charge but also its analogs for the other forces. There are two main kinds of charged particles, quarks (which feel the strong force) and leptons (which do not). All of these charged particles have corresponding antiparticles of the same mass and opposite charge.

Somewhat mysteriously, the charged particles come in three families or 'generations'. The first generation consists of two leptons, the electron  $e$  and the electron neutrino  $\nu_e$ , and two quarks, the up and down, or  $u$  and  $d$ . Most of the matter we see everyday is made out of these first-generation particles. For example, according to the standard model the proton is a composite of two up quarks and one down, while the neutron is two downs and an up. There is a second generation of quarks and leptons, the muon  $\mu$  and muon neutrino  $\nu_\mu$ , and the charmed and strange quarks  $c$ ,  $s$ . For the most part these are heavier than the corresponding particles in the first generation, although all the neutrinos appear to be massless or nearly so. For example, the muon is about 207 times as massive as the electron, but almost identical in every other respect. Then there is a third, still more massive generation, containing the tau  $\tau$  and tau neutrino  $\nu_\tau$ , and the top and bottom quarks  $t$  and  $b$ . For many years the top quark was merely conjectured to exist, but just as this book went to press, experimentalists announced that it may finally have been found.

Finally, there is a very odd charged particle in the standard model,

the Higgs particle, which is neither a quark nor a lepton. This has not been observed either, and is hypothesized to exist primarily to explain the relation between the electromagnetic and weak forces.

Even more puzzling than all the complexities of the standard model, however, is the question of where *gravity* fits into the picture! Einstein's equations describing gravity do *not* take quantum theory into account, and it has proved very difficult to 'quantize' them. We thus have not one picture of the world, but two: the standard model, in which all forces except gravity are described in accordance with quantum theory, and general relativity, in which gravity alone is described, not in accordance with quantum theory. Unfortunately it seems difficult to obtain guidance from experiment; simple considerations of dimensional analysis suggest that quantum gravity effects may become significant at distance scales comparable to the **Planck length**,

$$\ell_p = (\hbar\kappa/c^3)^{1/2},$$

where  $\hbar$  is Planck's constant,  $\kappa$  is Newton's gravitational constant, and  $c$  is the speed of light. The Planck length is about  $1.616 \cdot 10^{-35}$  meters, far below the length scales we can probe with particle accelerators.

Recent developments, however, hint that gravity may be closer to the gauge theories of the standard model than had been thought. Fascinatingly, the relationship also involves the study of *knots* in 3-dimensional space. While this work is in its early stages, and may not succeed as a theory of physics, the new mathematics involved is so beautiful that it is difficult to resist becoming excited. Unfortunately, understanding these new ideas depends on a thorough mastery of quantum field theory, general relativity, geometry, topology, and algebra. Indeed, it is almost certain that *nobody* is sufficiently prepared to understand these ideas fully! The reader should therefore not expect to understand them when done with this book. Our goal in this book is simply to start fairly near the beginning of the story and bring the reader far enough along to see the frontiers of current research in dim outline.

We must begin by reviewing some geometry. These days, when mathematicians speak of geometry they are usually referring not to Euclidean geometry but to the many modern generalizations that fall

under the heading of 'differential geometry'. The first theory of physics to *explicitly* use differential geometry was Einstein's general relativity, in which gravity is explained as the curvature of spacetime. The gauge theories of the standard model are of a very similar geometrical character (although quantized). But there is also a lot of differential geometry lurking in Maxwell's equations, which after all were the inspiration for both general relativity and gauge theory. So, just as a good way to master auto repair is to take apart an old car and put in a new engine so that it runs better, we will begin by taking apart Maxwell's equations and putting them back together using modern differential geometry.

In their classic form, Maxwell's equations describe the behavior of two vector fields, the **electric field**  $\vec{E}$  and the **magnetic field**  $\vec{B}$ . These fields are defined throughout space, which is taken to be  $\mathbb{R}^3$ . However, they are also functions of time, a real-valued parameter  $t$ . The electric and magnetic fields depend on the electric **charge density**  $\rho$ , which is a time-dependent function on space, and also on the electric **current density**  $\vec{j}$ , which is time-dependent vector field on space. (For the mathematicians, let us note that unless otherwise specified, functions are assumed to be real-valued, and functions and vector fields on  $\mathbb{R}^n$  are assumed to be **smooth**, that is, infinitely differentiable.)

In units where the speed of light is equal to 1, **Maxwell's equations** are:

$$\begin{aligned}\nabla \cdot \vec{B} &= 0 \\ \nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} &= 0 \\ \nabla \cdot \vec{E} &= \rho \\ \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} &= \vec{j}.\end{aligned}$$

There are a number of interesting things about these equations that are worth understanding. First, there is the little fact that we can only determine the direction of the magnetic field experimentally if we know the difference between right and left. This is easiest to see from the **Lorentz force law**, which says that the force on a charged particle with charge  $q$  and velocity  $\vec{v}$  is

$$\vec{F} = q (\vec{E} + \vec{v} \times \vec{B}).$$

To measure  $\vec{E}$ , we need only measure the force  $\vec{F}$  on a static particle and divide by  $q$ . To figure out  $\vec{B}$ , we can measure the force on charged particles with a variety of velocities. However, recall that the definition of the cross product involves a completely arbitrary right-hand rule! We typically define

$$\vec{v} \times \vec{B} = (v_y B_z - v_z B_y, v_z B_x - v_x B_z, v_x B_y - v_y B_x).$$

However, this is just a convention; we could have set

$$\vec{v} \times \vec{B} = (v_z B_y - v_y B_z, v_x B_z - v_z B_x, v_y B_x - v_x B_y),$$

and all the mathematics of cross products would work just as well. If we used this 'left-handed cross product' when figuring out  $\vec{B}$  from measurements of  $\vec{F}$  for various velocities  $\vec{v}$ , we would get an answer for  $\vec{B}$  with the opposite of the usual sign! It may seem odd that  $\vec{B}$  depends on an arbitrary convention this way. In fact, this turns out to be an important clue as to the mathematical structure of Maxwell's equations.

Secondly, Maxwell's equations naturally come in two pairs. The pair that does not involve the electric charge and current densities

$$\nabla \cdot \vec{B} = 0 \quad \nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = 0,$$

looks very much like the pair that *does*:

$$\nabla \cdot \vec{E} = \rho \quad \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = \vec{j}.$$

Note the funny minus sign in the second pair. The symmetry is clearest in the **vacuum** Maxwell equations, where the charge and current densities vanish:

$$\nabla \cdot \vec{B} = 0 \quad \nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = 0,$$

$$\nabla \cdot \vec{E} = 0 \quad \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = 0.$$

Then the transformation

$$\vec{B} \mapsto \vec{E}, \quad \vec{E} \mapsto -\vec{B}$$

takes the first pair of equations to the second and vice versa! This symmetry is called **duality** and is a clue that the electric and magnetic fields are part of a unified whole, the electromagnetic field. Indeed, if we introduce a complex-valued vector field

$$\vec{\mathcal{E}} = \vec{E} + i\vec{B},$$

duality amounts to the transformation

$$\vec{\mathcal{E}} \mapsto -i\vec{\mathcal{E}},$$

and the vacuum Maxwell equations boil down to two equations for  $\vec{\mathcal{E}}$ :

$$\nabla \cdot \vec{\mathcal{E}} = 0, \quad \nabla \times \vec{\mathcal{E}} = i \frac{\partial \vec{\mathcal{E}}}{\partial t}$$

This trick has very practical applications. For example, one can use it to find solutions that correspond to plane waves moving along at the speed of light, which in the units we are using equals 1.

**Exercise 1.** Let  $\vec{k}$  be a vector in  $\mathbb{R}^3$  and let  $\omega = |\vec{k}|$ . Fix  $\vec{E} \in \mathbb{C}^3$  with  $\vec{k} \cdot \vec{E} = 0$  and  $\vec{k} \times \vec{E} = i\omega \vec{E}$ . Show that

$$\vec{\mathcal{E}}(t, \vec{x}) = \vec{E} e^{-i(\omega t - \vec{k} \cdot \vec{x})}$$

satisfies the vacuum Maxwell equations.

The symmetry between  $\vec{E}$  and  $\vec{B}$  does not, however, extend to the non-vacuum Maxwell equations. We can consider making  $\rho$  and  $\vec{j}$  complex, and writing down:

$$\nabla \cdot \vec{\mathcal{E}} = \rho, \quad \nabla \times \vec{\mathcal{E}} = i \left( \frac{\partial \vec{\mathcal{E}}}{\partial t} + \vec{j} \right).$$

However, this amounts to introducing magnetic charge and current density, since if we split  $\rho$  and  $\vec{j}$  into real and imaginary parts, we see that

the imaginary parts play the role of magnetic charge and current densities:

$$\begin{aligned}\rho &= \rho_e + i\rho_m, \\ \vec{j} &= \vec{j}_e + i\vec{j}_m.\end{aligned}$$

We get

$$\begin{aligned}\nabla \cdot \vec{B} &= \rho_m & \nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} &= \vec{j}_m, \\ \nabla \cdot \vec{E} &= \rho_e & \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} &= \vec{j}_e.\end{aligned}$$

These equations are quite charming, but unfortunately no magnetic charges — so called **magnetic monopoles** — have been observed! (We will have a bit more to say about this in Chapter 6.) We could simply keep these equations and say that  $\rho$  and  $\vec{j}$  are real-valued on the basis of experimental evidence. But it is a mathematical as well as a physical challenge to find a better way of understanding this phenomenon. It turns out that the formalism of gauge theory makes it seem quite natural.

Finally, there is the connection between Maxwell's equations and special relativity. The main idea of special relativity is that in addition to the symmetries of space (translations and rotations) and time (translations) there are equally important symmetries mixing space and time, the Lorentz transformations. The idea is that if you and I are both unaccelerated, so that my velocity with respect to you is constant, the coordinates I will naturally use, in which I am at rest, will differ from yours, in which you are at rest. If your coordinate system is  $(t, x, y, z)$  and I am moving with velocity  $v$  in the  $x$  direction with respect to you, for example, the coordinates in which I am at rest are given by

$$\begin{aligned}t' &= (\cosh \phi)t - (\sinh \phi)x \\ x' &= -(\sinh \phi)t + (\cosh \phi)x \\ y' &= y \\ z' &= z,\end{aligned}$$

where  $\phi$  is a convenient quantity called the **rapidity**, defined so that  $\tanh \phi = v$ . Note the close resemblance to the formula for rotations in

space. The idea is that just as the  $x$ ,  $y$ , and  $z$  components of position are all just aspects of something more important, the position itself, space and time are just aspects of a unitary whole, *spacetime*.

Maxwell's equations are invariant under these Lorentz transformations — indeed, this was the main fact that led Einstein to special relativity! He realized that Maxwell's equations predict that *any* unaccelerated observer will measure light moving in *any* direction in the vacuum to have the *same* speed. Mathematically speaking, the point is that if we have a solution of Maxwell's equations and we do a Lorentz transformation on the coordinates together with a certain transformation of  $\vec{E}$ ,  $\vec{B}$ ,  $\rho$  and  $\vec{j}$ , we again have a solution.

For example, suppose that we do a Lorentz transformation of velocity  $v$  in the  $x$  direction, as above. The precise recipe for transforming the charge and current densities is

$$\begin{aligned}\rho' &= (\cosh \phi)\rho - (\sinh \phi)j_x \\ j'_x &= -(\sinh \phi)\rho + (\cosh \phi)j_x \\ j'_y &= j_y \\ j'_z &= j_z.\end{aligned}$$

Note that  $\rho$  and  $\vec{j}$  get mixed up together. In fact, we shall see that they are really just two aspects of a single thing called the 'current', which has  $\rho$  as its component in the time direction and  $j_x, j_y, j_z$  as its components in the space directions.

The formula for transforming the electric and magnetic fields under the same Lorentz transformation is somewhat more complicated:

$$\begin{aligned}E'_x &= E_x \\ E'_y &= (\cosh \phi)E_y - (\sinh \phi)B_z \\ E'_z &= (\sinh \phi)B_y + (\cosh \phi)E_z, \\ B'_x &= B_x \\ B'_y &= (\cosh \phi)B_y + (\sinh \phi)E_z \\ B'_z &= -(\sinh \phi)E_y + (\cosh \phi)B_z.\end{aligned}$$

The most important message here is that the electric and magnetic fields are two aspects of a unified 'electromagnetic field'. Also, we see

that the electromagnetic field is more complicated in character than the current, since it has six independent components that transform in a more subtle manner. It turns out to be a '2-form'.

When we have rewritten Maxwell's equations using the language of differential geometry, all the things we have just discussed will be much clearer — at least if we succeed in explaining things well. The key step, which is somewhat shocking to the uninitiated, is to work as much as possible in a manner that does not require a choice of coordinates. After all, as far as we can tell, the world was *not* drawn on graph paper. Coordinates are merely something *we* introduce for our own convenience, and the laws of physics should not care which coordinates we happen to use. If we postpone introducing coordinates until it is actually necessary, we will not have to do anything to show that Maxwell's equations are invariant under Lorentz transformations; it will be *manifest*.

Just for fun, let us write down the new version of Maxwell's equations right away. We will explain what they *mean* quite a bit later, so do not worry if they are fairly cryptic. They are:

$$\begin{aligned} dF &= 0 \\ \star d \star F &= J. \end{aligned}$$

Here  $F$  is the 'electromagnetic field' and  $J$  is the 'current', while the  $d$  and  $\star$  operators are slick ways of summarizing all the curls, divergences and time derivatives that appear in the old-fashioned version. The equation  $dF = 0$  is equivalent to the first pair of Maxwell's equations, while the equation  $\star d \star F = J$  is equivalent to the second pair. The 'funny minus sign' in the second pair will turn out to be a natural consequence of how the  $\star$  operator works.

If the reader is too pragmatic to get excited by the terse beauty of this new-fangled version of Maxwell's equations, let us emphasize that this way of writing them is a warm-up for understanding gauge theory, and allows us to study Maxwell's equations and gauge theory on curved spacetimes, as one needs to in general relativity. Indeed, we will start by developing enough differential geometry to do a fair amount of physics on general spacetimes. Then we will come back to Maxwell's equations. We warn the reader that the next few sections are not really

a solid course in differential geometry. Whenever something is at all tricky to prove we will skip it! The easygoing reader can take some facts on faith; the careful reader may want to get ahold of a good book on differential geometry to help fill in these details. Some suggestions on books appear in the notes at the end of Part I.

## Chapter 2

### Manifolds

*We therefore reach this result: In the general theory of relativity, space and time cannot be defined in such a way that differences of the spatial co-ordinates can be directly measured by the unit measuring-rod, or differences in the time co-ordinate by a standard clock.*

*The method hitherto employed for laying co-ordinates into the space-time continuum in a definite manner thus breaks down, and there seems to be no other way which would allow us to adapt systems of co-ordinates to the four-dimensional universe so that we might expect from their application a particularly simple formulation of the laws of nature. So there is nothing for it but to regard all imaginable systems of co-ordinates, on principle, as equally suitable for the description of nature. This comes to requiring that:*

The general laws of nature are to be expressed by equations which hold good for all systems of co-ordinates, that is, are co-variant with respect to any substitutions whatever (generally covariant). — *Albert Einstein*

In order to do modern physics we need to be able to handle spaces and spacetimes that are more general than good old  $\mathbb{R}^n$ . The kinds of spaces we will be concerned with are those that look *locally* like  $\mathbb{R}^n$ , but perhaps not *globally*. Such a space is called an  $n$ -dimensional ‘manifold’. For example, the sphere

$$x^2 + y^2 + z^2 = 1,$$

looks locally like the plane  $\mathbb{R}^2$ , which is why some people thought the Earth was flat. These days we call this sphere  $S^2$  — the **2-sphere** — to indicate that it is a 2-dimensional manifold. Similarly, while the space

we live in looks locally like  $\mathbb{R}^3$ , we have no way yet of ruling out the possibility that it is really  $S^3$ , the **3-sphere**:

$$w^2 + x^2 + y^2 + z^2 = 1,$$

and indeed, in many models of cosmology space is a 3-sphere. In such a universe one could, if one had time, sail around the cosmos in a spaceship just as Magellan circumnavigated the globe. More generally, it is even possible that spacetime has more than 4 dimensions, as is assumed in so-called 'Kaluza-Klein theories'. For a while, string theorists seemed quite sure that the universe must either be 10 or 26-dimensional! More pragmatically, there is a lot of interest in low-dimensional physics, such as the behavior of electrons on thin films and wires. Also, classical mechanics uses 'phase spaces' that may have very many dimensions.

These are some of the physical reasons why it is good to generalize vector calculus so that it works nicely on any manifold. On the other hand, mathematicians have many reasons of their own for dealing with manifolds. For example, the set of solutions of an equation is often a manifold (see the equation for the 3-sphere above).

We now head towards a precise definition of a manifold. First of all, we remind the reader that a **topological space** is a set  $X$  together with a family of subsets of  $X$ , called the **open sets**, required to satisfy the conditions:

- 1) The empty set and  $X$  itself are open,
- 2) If  $U, V \subseteq X$  are open, so is  $U \cap V$ ,
- 3) If the sets  $U_\alpha \subseteq X$  are open, so is the union  $\bigcup U_\alpha$ .

The collection of sets taken to be open is called the **topology** of  $X$ . An open set containing a point  $x \in X$  is called a **neighborhood** of  $x$ . The complement of an open set is called **closed**.

A basic example is  $\mathbb{R}^n$ , where a set  $U$  is taken to be open if for every  $x \in U$ , all points sufficiently close to  $x$  are also in  $U$ :

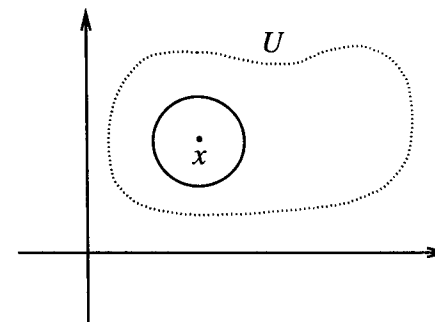


Fig. 1. An open set in  $\mathbb{R}^2$

The use of a topology is that it allows us to define continuous functions. Roughly speaking, a function is continuous if it sends nearby points to nearby points. The trick is making the notion of 'nearby' precise using open sets. A function  $f: X \rightarrow Y$  from one topological space to another is defined to be **continuous** if, given any open set  $U \subseteq Y$ , the inverse image  $f^{-1}U \subseteq X$  is open.

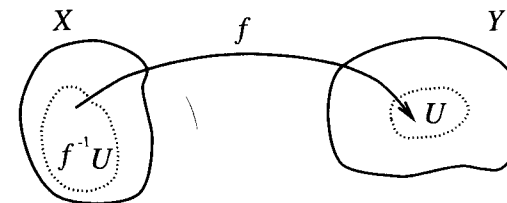


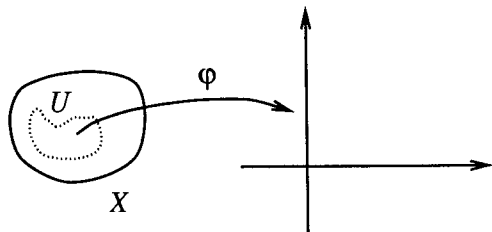
Fig. 2. A continuous function from  $X$  to  $Y$

If one has not yet, one should do the following exercise.

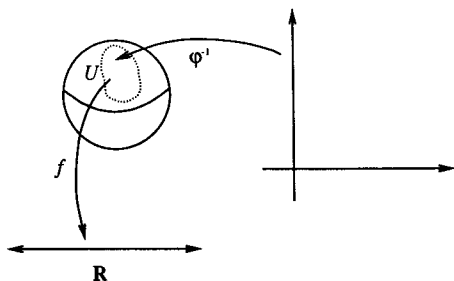
**Exercise 2.** Show that a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is continuous according to the above definition if and only if it is according to the epsilon-delta definition: for all  $x \in \mathbb{R}^n$  and all  $\epsilon > 0$ , there exists  $\delta > 0$  such that  $\|y - x\| < \delta$  implies  $\|f(y) - f(x)\| < \epsilon$ .

The idea of a manifold is that, like the globe, we can cover it with patches that look just like  $\mathbb{R}^n$ . More precisely, we say that a collection

$U_\alpha$  of open sets **covers** a topological space  $X$  if their union is all of  $X$ . Given a topological space  $X$  and an open set  $U \subseteq X$ , we define a **chart** to be a continuous function  $\varphi: U \rightarrow \mathbb{R}^n$  with a continuous inverse (the inverse being defined on the set  $\varphi(U)$ ).

Fig. 3. A chart on  $X$ 

As long as we work 'in the chart  $\varphi$ ' we can pretend we are working in  $\mathbb{R}^n$ , just as the Europeans could pretend they lived on  $\mathbb{R}^2$  as long as they did not go too far from home. For example, if we have a function  $f: U \rightarrow \mathbb{R}$ , we can turn it into a function on  $\mathbb{R}^n$  by using  $f \circ \varphi^{-1}: \mathbb{R}^n \rightarrow \mathbb{R}$ .

Fig. 4. Turning a function on  $U$  into a function on  $\mathbb{R}^n$ 

Finally, we say that an  $n$ -dimensional **manifold**, or  $n$ -**manifold**, is a topological space  $M$  equipped with charts  $\varphi_\alpha: U_\alpha \rightarrow \mathbb{R}^n$ , where  $U_\alpha$  are open sets covering  $M$ , such that the **transition function**  $\varphi_\alpha \circ \varphi_\beta^{-1}$  is smooth where it is defined. Such a collection of charts is called an

**atlas**.

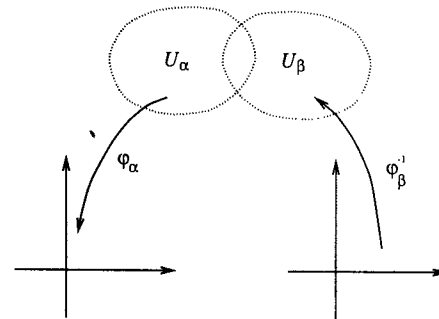


Fig. 5. Two charts and the transition function

What does this definition mean? First, every point of  $M$  lives in some open subset  $U_\alpha$  that looks like  $\mathbb{R}^n$ , or in other words, we can 'patch together' the whole manifold out of bits that look like  $\mathbb{R}^n$ . Second, it means that we can tell using charts if a function on  $M$  is smooth, without any ambiguity, because the transition functions between charts are smooth. To be precise, we say a function  $f: M \rightarrow \mathbb{R}$  is **smooth** if for all  $\alpha$ ,  $f \circ \varphi_\alpha^{-1}: \mathbb{R}^n \rightarrow \mathbb{R}$  is smooth. Suppose you are using the chart  $\varphi_\alpha: U_\alpha \rightarrow \mathbb{R}^n$  and I am using the chart  $\varphi_\beta: U_\beta \rightarrow \mathbb{R}^n$ , and let  $V = U_\alpha \cap U_\beta$  be the overlap of our two charts. Suppose that you think the function  $f$  is smooth on  $V$ , that is, suppose  $f \circ \varphi_\alpha^{-1}$  is smooth on  $\varphi_\alpha V$ , as below:

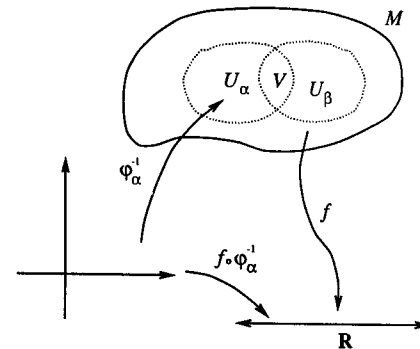


Fig. 6. Your picture

Then I will agree that  $f$  is smooth on  $V$ , that is,  $f \circ \varphi_\beta^{-1}$  will be



smooth on  $\varphi_\beta V$  too:

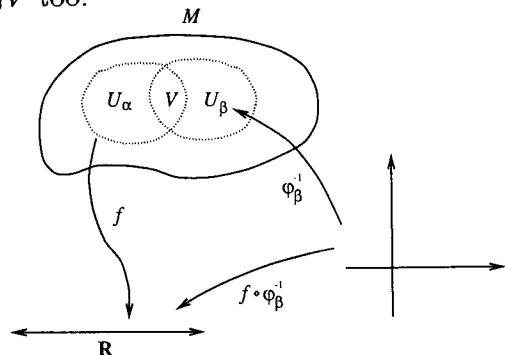


Fig. 7. My picture

Why? Because we can express my function in terms of your function and the transition function:

$$f \circ \varphi_\beta^{-1} = (f \circ \varphi_\alpha^{-1}) \circ (\varphi_\alpha \circ \varphi_\beta^{-1}).$$

Strictly speaking, the sort of manifold we have defined here is called a **smooth manifold**. There are also, for example, **topological manifolds**, where the transition functions are only required to be continuous. For us, 'manifold' will always mean 'smooth manifold'. Also, we will always assume our manifolds are 'Hausdorff' and 'paracompact'. These are topological properties that we prefer to avoid explaining here, which are satisfied by all but the most bizarre and useless examples.

In the following exercises we describe some examples of manifolds, leaving the reader to check that they really are manifolds.

**Exercise 3.** Given a topological space  $X$  and a subset  $S \subseteq X$ , define the **induced topology** on  $S$  to be the topology in which the open sets are of the form  $U \cap S$ , where  $U$  is open in  $X$ . Let  $S^n$ , the  $n$ -sphere, be the unit sphere in  $\mathbb{R}^{n+1}$ :

$$S^n = \{\vec{x} \in \mathbb{R}^{n+1} \mid \sum_{i=1}^{n+1} (x^i)^2 = 1\}.$$

Show that  $S^n \subset \mathbb{R}^{n+1}$  with its induced topology is a manifold.

**Exercise 4.** Show that if  $M$  is a manifold and  $U$  is an open subset of  $M$ , then  $U$  with its induced topology is a manifold.

**Exercise 5.** Given topological spaces  $X$  and  $Y$ , we give  $X \times Y$  the **product topology** in which a set is open if and only if it is a union of sets of the form  $U \times V$ , where  $U$  is open in  $X$  and  $V$  is open in  $Y$ . Show that if  $M$  is an  $m$ -dimensional manifold and  $N$  is an  $n$ -dimensional manifold,  $M \times N$  is an  $(m + n)$ -dimensional manifold.

**Exercise 6.** Given topological spaces  $X$  and  $Y$ , we give  $X \cup Y$  the **disjoint union topology** in which a set is open if and only if it is the union of an open subset of  $X$  and an open subset of  $Y$ . Show that if  $M$  and  $N$  are  $n$ -dimensional manifolds the disjoint union  $M \cup N$  is an  $n$ -dimensional manifold.

There are many different questions one can ask about a manifold, but one of the most basic is whether it extends indefinitely in all directions like  $\mathbb{R}^3$  or is 'compact' like  $S^3$ . There is a way to make this precise which proves to be very important in mathematics. Namely, a topological space  $X$  is said to be **compact** if for every cover of  $X$  by open sets  $U_\alpha$  there is a finite collection  $U_{\alpha_1}, \dots, U_{\alpha_n}$  that covers  $X$ . For manifolds, there is an equivalent definition: a manifold  $M$  is compact if and only if every sequence in  $M$  has a convergent subsequence. A basic theorem says that a subset of  $\mathbb{R}^n$  is compact if and only if it is closed and fits inside a ball of sufficiently large radius.

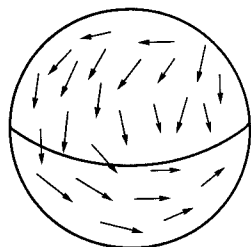
The study of manifolds is a fascinating business in its own right. However, since our goal is to do physics on manifolds, let us turn to the basic types of fields that live on manifolds: vector fields and differential forms.

## Chapter 3

# Vector Fields

*And it is a noteworthy fact that ignorant men have long been in advance of the learned about vectors. Ignorant people, like Faraday, naturally think in vectors. They may know nothing of their formal manipulation, but if they think about vectors, they think of them as vectors, that is, directed magnitudes. No ignorant man could or would think about the three components of a vector separately, and disconnected from one another. That is a device of learned mathematicians, to enable them to evade vectors. The device is often useful, especially for calculating purposes, but for general purposes of reasoning the manipulation of the scalar components instead of the vector itself is entirely wrong. — Oliver Heaviside*

Heaviside was one of the first advocates of modern vector analysis, as well as a very sarcastic fellow. In the quote above, he was making the point that the great physicist Faraday did not need to worry about coordinates, because Faraday had a direct physical understanding of vectors. Pictorially, a vector field on a manifold can be visualized as a field of arrows. For example, a vector field on  $S^2$  is basically just a field of arrows tangent to the sphere:

Fig. 1. Vector field on  $S^2$ 

To do calculations with vector fields, however, it is nice to define them in an algebraic sort of way. The key to defining vector fields on manifolds is to note that given a field of arrows, one can differentiate a function in the direction of the arrows. In particular, given a function  $f$  and a vector field  $v$  on  $\mathbb{R}^n$ , we can form the directional derivative of  $f$  in the direction  $v$ , which we will write simply as  $vf$ .

Let us write a formula for  $vf$  in this case. The formula for a directional derivative should not be news to the readers of this book, but we will rewrite it using some slick physics notation. We will write  $x^1, \dots, x^n$  for the coordinates on  $\mathbb{R}^n$ , and write just  $\partial_\mu$  for the partial derivative  $\partial/\partial x^\mu$ . (When we are dealing with three or fewer dimensions we will sometimes write  $x, y, z$  instead of  $x^1, x^2, x^3$ , and write  $\partial_x, \partial_y, \partial_z$  for  $\partial_1, \partial_2, \partial_3$ .) Also, we will use the **Einstein summation convention** and always sum over repeated indices that appear once as a subscript and once as a superscript. Then if  $v$  has components  $(v^1, \dots, v^n)$ , we have the formula

$$vf = v^\mu \partial_\mu f.$$

If this seems enigmatic, remember that it is just short for

$$vf = v^1 \frac{\partial f}{\partial x^1} + \dots + v^n \frac{\partial f}{\partial x^n}.$$

In fact, since the formula  $vf = v^\mu \partial_\mu f$  holds for all  $f$ , we can be even more slick and write

$$v = v^\mu \partial_\mu.$$

What does this mean, though? The sight of the partial derivatives  $\partial_\mu$  sitting there with nothing to differentiate is only slightly unnerving; we can always put a function  $f$  to the right of them whenever we want. Much odder is that we are saying the vector field  $v$  is the linear combination of these partial derivatives. What we are doing might be regarded as rather sloppy, since we are identifying two different, although related, things: the vector field  $v$ , and the operator  $v^\mu \partial_\mu$  that takes a directional derivative in the direction of  $v$ . In fact, this 'sloppy' attitude turns out to be extremely convenient, and next we will go even further and use it to *define* vector fields on manifolds. It is important to realize that in mathematics it is often crucial to think about familiar objects in a new way in order to generalize them to a new situation.

Now let us define vector fields on a manifold  $M$ . Following the philosophy outlined above, these will be entities whose sole ambition in life is to differentiate functions. First a bit of jargon. The set of smooth (real-valued) functions on a manifold  $M$  is written  $C^\infty(M)$ , where the  $C^\infty$  is short for 'having infinitely many continuous derivatives'. Note that  $C^\infty(M)$  is an **algebra** over the real numbers, meaning that it is closed under (pointwise) addition and multiplication, as well as multiplication by real numbers, and the following batch of rules holds:

$$\begin{aligned} f + g &= g + f \\ f + (g + h) &= (f + g) + h \\ f(gh) &= (fg)h \\ f(g + h) &= fg + fh \\ (f + g)h &= fh + gh \\ 1f &= f \\ \alpha(\beta f) &= (\alpha\beta)f \\ \alpha(f + g) &= \alpha f + \alpha g \\ (\alpha + \beta)f &= \alpha f + \beta f, \end{aligned}$$

where  $f, g, h \in C^\infty(M)$  and  $\alpha, \beta \in \mathbb{R}$ . Of course it is a **commutative algebra**, that is,  $fg = gf$ .

Now, a **vector field**  $v$  on  $M$  is defined to be a function from  $C^\infty(M)$  to  $C^\infty(M)$  satisfying the following properties:

$$\begin{aligned}
v(f+g) &= v(f) + v(g) \\
v(\alpha f) &= \alpha v(f) \\
v(fg) &= v(f)g + fv(g),
\end{aligned}$$

for all  $f, g \in C^\infty(M)$  and  $\alpha \in \mathbb{R}$ . Here we have isolated all the basic rules a directional derivative operator should satisfy. The first two simply amount to linearity, and it is the third one, the product rule or **Leibniz law**, that really captures the essence of differentiation.

This definition may seem painfully abstract. We will see in a bit that it really is just a way of talking about a field of arrows on  $M$ . For now, note the main good feature of this definition: *it does not rely on any choice of coordinates on  $M$ !* A basic philosophy of modern physics is that the universe does not come equipped with a coordinate system. While coordinate systems are necessary for doing specific concrete calculations, the choice of the coordinate system to use is a matter of convenience, and there is often no ‘best’ coordinate system. One should strive to write the laws of physics in a *manifestly* coordinate-independent manner, so one can see what they are really saying and not get distracted by things that might depend on the coordinates.

Let  $\text{Vect}(M)$  denote the set of all vector fields on  $M$ . We leave it to the reader to check that one can add vector fields and multiply them by functions on  $M$  as follows. Given  $v, w \in \text{Vect}(M)$ , we define  $v + w$  by

$$(v + w)(f) = v(f) + w(f),$$

and given  $v \in \text{Vect}(M)$  and  $g \in C^\infty(M)$ , we define  $gv$  by

$$(gv)(f) = gv(f).$$

**Exercise 7.** Show that  $v + w$  and  $gv \in \text{Vect}(M)$ .

**Exercise 8.** Show that the following rules for all  $v, w \in \text{Vect}(M)$  and  $f, g \in C^\infty(M)$ :

$$\begin{aligned}
f(v + w) &= fv + fw \\
(f + g)v &= fv + gv \\
(fg)v &= f(gv) \\
1v &= v.
\end{aligned}$$

(Here ‘1’ denotes the constant function equal to 1 on all of  $M$ .) Mathematically, we summarize these rules by saying that  $\text{Vect}(M)$  is a module over  $C^\infty(M)$ .

It turns out that the vector fields  $\{\partial_\mu\}$  on  $\mathbb{R}^n$  **span**  $\text{Vect}(\mathbb{R}^n)$  as a module over  $C^\infty(M)$ . In other words, every vector field on  $\mathbb{R}^n$  is a linear combination of the form

$$v^\mu \partial_\mu = v^1 \partial_1 + \cdots + v^n \partial_n,$$

for some functions  $v^\mu \in C^\infty(\mathbb{R}^n)$ . It is also true that the vector fields  $\{\partial_\mu\}$  on  $\mathbb{R}^n$  are **linearly independent**.

**Exercise 9.** Show that if  $v^\mu \partial_\mu = 0$ , that is,  $v^\mu \partial_\mu f = 0$  for all  $f \in C^\infty(\mathbb{R}^n)$ , we must have  $v^\mu = 0$  for all  $\mu$ .

This implies that every vector field  $v$  on  $\mathbb{R}^n$  has a unique representation as a linear combination  $v^\mu \partial_\mu$ ; we say that the vector fields  $\{\partial_\mu\}$  form a **basis** of  $\text{Vect}(\mathbb{R}^n)$ . The functions  $v^\mu$  are called the **components** of the vector field  $v$ .

## Tangent Vectors

Often it is nice to think of a vector field on  $M$  as really assigning an ‘arrow’ to each point of  $M$ . This kind of arrow is called a tangent vector. For example, we may think of a tangent vector at a point  $p \in S^2$  as a vector in the plane tangent to  $p$ :

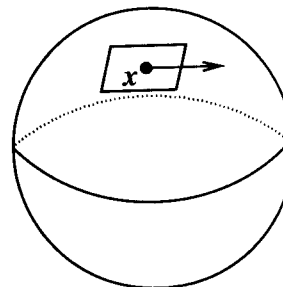


Fig. 2. Tangent vector

To get a precise definition of a tangent vector at  $p \in M$ , note that a tangent vector should let us take directional derivatives at the point  $p$ . For example, given a vector field  $v$  on  $M$ , we can take the derivative  $v(f)$  of any function  $f \in C^\infty(M)$ , and then evaluate the function  $v(f)$  at  $p$ . We can think of the result,  $v(f)(p)$ , as being the result of differentiating  $f$  in the direction ' $v_p$ ' at the point  $p$ . In other words, we can *define*

$$v_p: C^\infty(M) \rightarrow \mathbb{R}$$

by

$$v_p(f) = v(f)(p),$$

and think of  $v_p$  as a tangent vector at  $p$ . We call  $v_p$  the value of  $v$  at  $p$ .

Note that  $v_p$  has three basic properties, which follow from the definition of a vector field:

$$\begin{aligned} v_p(f + g) &= v_p(f) + v_p(g) \\ v_p(\alpha f) &= \alpha v_p(f) \\ v_p(fg) &= v_p(f)g(p) + f(p)v_p(g). \end{aligned}$$

Henceforth, we will simply *define* a **tangent vector** at  $p \in M$  to be a function from  $C^\infty(M)$  to  $\mathbb{R}$  satisfying these three properties. Let  $T_p M$ , the **tangent space** at  $p$ , denote the set of all tangent vectors at  $p \in M$ .

It now follows rigorously from our definitions that for each  $p \in M$ , a vector field  $v \in \text{Vect}(M)$  determines a tangent vector  $v_p \in T_p M$ . One can also show, though it takes a bit of work, that *every* tangent vector at  $p$  is of the form  $v_p$  for some vector field or other. A related fact, which is much easier to show, is the following:

**Exercise 10.** Let  $v, w \in \text{Vect}(M)$ . Show that  $v = w$  if and only if  $v_p = w_p$  for all  $p \in M$ .

Why do tangent vectors as we have defined them 'look like arrows'? First of all, we can add two tangent vectors  $v, w \in T_p M$  by

$$(v + w)(f) = v(f) + w(f),$$

and multiply tangent vectors by real numbers:

$$(\alpha v)(f) = \alpha v(f).$$

(Now we are using the letters  $v, w$  to denote tangent vectors, not vector fields!) With addition and multiplication defined this way, the tangent space is really a vector space. For example, in Figure 2 we have drawn a tangent space to look like a little plane. The tangent vectors can be thought of as arrows living in this vector space.

**Exercise 11.** Show that  $T_p M$  is a vector space over the real numbers.

Another reason why tangent vectors really look like arrows is that curves have tangent vectors:

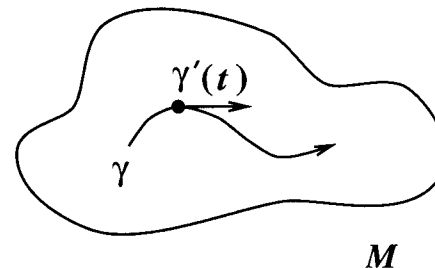


Fig. 3. The tangent vector to a curve in  $M$

By a **curve** we will always mean a function from  $\mathbb{R}$  or some interval to  $M$  that is smooth, i.e., such that for any  $f \in C^\infty(M)$ ,  $f(\gamma(t))$  depends smoothly on  $t$ . Given a curve  $\gamma: \mathbb{R} \rightarrow M$  and any  $t \in \mathbb{R}$ , the tangent vector  $\gamma'(t)$  should be a vector in the tangent space  $T_{\gamma(t)} M$ . We define  $\gamma'(t)$  in the only sensible way possible: it is the function from  $C^\infty(M)$  to  $\mathbb{R}$  that sends any function  $f \in C^\infty(M)$  to the derivative

$$\frac{d}{dt} f(\gamma(t)).$$

In other words, the tangent vector  $\gamma'(t)$  differentiates functions in the direction that  $\gamma$  is moving in at time  $t$ .

**Exercise 12.** Check that  $\gamma'(t) \in T_{\gamma(t)} M$  using the definitions.

If the curve  $\gamma$  describes the motion of a particle through space, the tangent vector  $\gamma'(t)$  represents its velocity. For this reason, we will sometimes write

$$\frac{d\gamma}{dt}$$

for  $\gamma'(t)$ , especially when we are not particularly concerned with which value of  $t$  we are talking about.

Note that for manifolds it generally makes no sense to say that a tangent vector  $v \in T_p M$  is 'the same' as another one,  $w \in T_q M$ , unless the points  $p$  and  $q$  are the same. For example, there is no 'best' way to compare tangent vectors at the north pole of  $S^2$  to tangent vectors at the equator. It also makes no sense to add tangent vectors at different points!

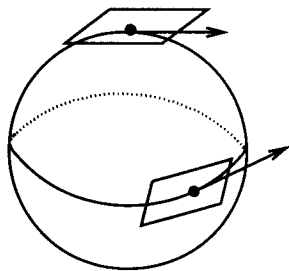


Fig. 4. Tangent vectors at different points of  $S^2$

We mention this because the reader may be used to  $\mathbb{R}^n$ , where one often says the following two vectors are 'the same', even though they are at different points in  $\mathbb{R}^n$ :

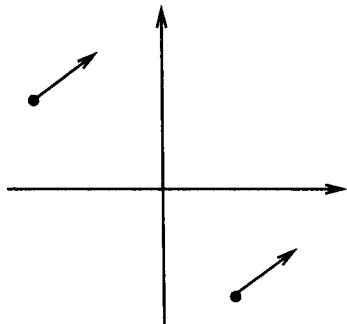


Fig. 5. Tangent vectors at different points of  $\mathbb{R}^n$

The reason why one can get away with this is that for any point  $p$  in

$\mathbb{R}^n$ , the tangent vectors

$$(\partial_\mu)_p \in T_p \mathbb{R}^n,$$

form a basis. This allows one to relate tangent vectors at different points of  $\mathbb{R}^n$  — one can sloppily say that the vector

$$v = v^\mu (\partial_\mu)_p \in T_p \mathbb{R}^n$$

and the vector

$$w = w^\mu (\partial_\mu)_q \in T_q \mathbb{R}^n$$

are 'the same' if  $v^\mu = w^\mu$ , even though  $v$  and  $w$  are not literally equal. Later we will get a deeper understanding of this issue, which requires a theory of 'parallel transport', the process of dragging a vector at one point of a manifold over to another point. This turns out to be a crucial idea in physics, and in fact the root of gauge theory!

## Covariant Versus Contravariant

A lot of modern mathematics and physics requires keeping track of which things in life are covariant, and which things are contravariant. Let us begin to explain these ideas by comparing functions and tangent vectors. Say we have a function  $\phi: M \rightarrow N$  from one manifold to another. If we have a real-valued function on  $N$ , say  $f: N \rightarrow \mathbb{R}$ , we can get a real-valued function on  $M$  by composing it with  $f$ .

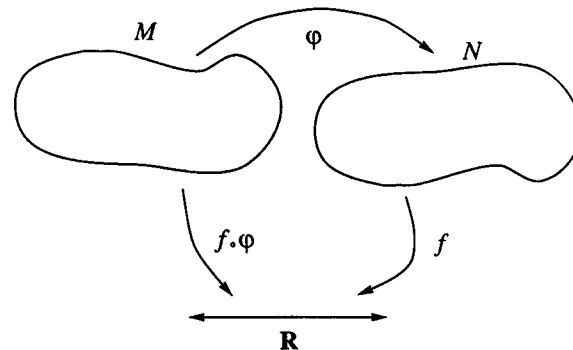


Fig. 6. Pulling back  $f$  from  $N$  to  $M$

We call this process **pulling back**  $f$  from  $N$  to  $M$  by  $\phi$ . We define

$$\phi^* f = f \circ \phi,$$

and call  $\phi^* f$  the **pullback** of  $f$  by  $\phi$ . The point is that while  $\phi$  goes 'forwards' from  $M$  to  $N$ , the pullback operation  $\phi^*$  goes 'backwards', taking functions on  $N$  to functions on  $M$ . We say that real-valued functions on a manifold are **contravariant** because of this perverse backwards behavior.

**Exercise 13.** Let  $\phi: \mathbb{R} \rightarrow \mathbb{R}$  be given by  $\phi(t) = e^t$ . Let  $x$  be the usual coordinate function on  $\mathbb{R}$ . Show that  $\phi^* x = e^x$ .

**Exercise 14.** Let  $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  be rotation counterclockwise by an angle  $\theta$ . Let  $x, y$  be the usual coordinate functions on  $\mathbb{R}^2$ . Show that

$$\begin{aligned}\phi^* x &= (\cos \theta)x - (\sin \theta)y \\ \phi^* y &= (\sin \theta)x + (\cos \theta)y.\end{aligned}$$

By the way, we say that  $\phi: M \rightarrow N$  is **smooth** if  $f \in C^\infty(N)$  implies that  $\phi^* f \in C^\infty(M)$ . Henceforth we will assume functions from manifolds to manifolds are smooth unless otherwise stated, and we will often call such functions **maps**.

**Exercise 15.** Show that this definition of smoothness is consistent with the previous definitions of smooth functions  $f: M \rightarrow \mathbb{R}$  and smooth curves  $\gamma: \mathbb{R} \rightarrow M$ .

Using our new jargon, we have: given any map

$$\phi: N \rightarrow M,$$

pulling back by  $\phi$  is an operation

$$\phi^*: C^\infty(M) \rightarrow C^\infty(N).$$

Tangent vectors, on the other hand, are **covariant**: a tangent vector  $v \in T_p M$  and a smooth function  $\phi: M \rightarrow N$  gives a tangent vector  $\phi_* v \in T_{\phi(p)} N$ , called the **pushforward** of  $v$  by  $\phi$ . This is defined by

$$(\phi_* v)(f) = v(\phi^* f).$$

We say we are **pushing forward**  $v$  by  $\phi$ . Note that we use a subscript asterisk for pushforwards and a superscript for pullbacks! One way to think of the pushforward is that if  $\gamma$  is a curve in  $M$  with tangent vector  $\gamma'(t) \in T_p(M)$ , the curve  $\phi \circ \gamma$  is a curve with tangent vector

$$(\phi \circ \gamma)'(t) = \phi_*(\gamma'(t)) \in T_{\phi(p)}(N).$$

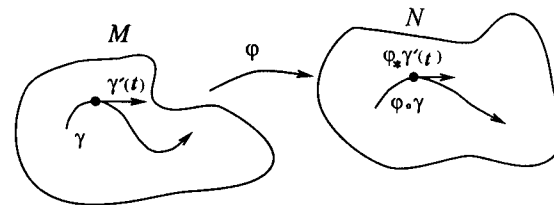


Fig. 7. Pushing forward the tangent vector of a curve from  $M$  to  $N$

**Exercise 16.** Prove that  $(\phi \circ \gamma)'(t) = \phi_*(\gamma'(t))$ .

**Exercise 17.** Show that the pushforward operation

$$\phi_*: T_p M \rightarrow T_{\phi(p)} N$$

is linear.

**Exercise 18.** Show that if  $\phi: M \rightarrow N$  we can push forward a vector field  $v$  on  $M$  to obtain a vector field  $\phi_*$  on  $N$  satisfying

$$(\phi_* v)_q = \phi_*(v_p)$$

whenever  $\phi(p) = q$ .

**Exercise 19.** Let  $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  be rotation counterclockwise by an angle  $\theta$ . Let  $\partial_x, \partial_y$  be the coordinate vector fields on  $\mathbb{R}^2$ . Show that at any point of  $\mathbb{R}^2$

$$\begin{aligned}\phi_* \partial_x &= (\cos \theta) \partial_x - (\sin \theta) \partial_y \\ \phi_* \partial_y &= (\sin \theta) \partial_x + (\cos \theta) \partial_y.\end{aligned}$$

It is traditional in mathematics, by the way, to write pushforwards and other covariant things with lowered asterisks, and to write pull-backs and other contravariant things with raised asterisks. It might help as a mnemonic to remember that the tangent vectors  $\partial_\mu$  are written with the  $\mu$  downstairs, and are covariant. In the next chapter we will discuss things similar to tangent vectors, but which are contravariant! These things will have their indices upstairs. We warn the reader, however, that while the vector field  $\partial_\mu$  is covariant and has its indices downstairs, physicists often think of a vector field  $v$  as *being* its components  $v^\mu$ . These have their indices upstairs, so physicists say that the  $v^\mu$  are contravariant! This is one of those little differences that makes communication between the two subjects a bit more difficult.

## Flows and the Lie Bracket

One sort of vector field that comes up in physics is the velocity vector field of a fluid, such as water. Imagine that the velocity vector field  $v$  is constant as a function of time, so that each molecule of water traces out a curve  $\gamma(t)$  as time passes, with the tangent vector of  $\gamma$  equal to the value of  $v$  at the point  $\gamma(t)$ :

$$\gamma'(t) = v_{\gamma(t)}$$

for all  $t$ . If the curve starts at some point  $p \in M$ , that is  $\gamma(0) = p$ , we call  $\gamma$  the **integral curve through  $p$**  of the vector field  $v$ :

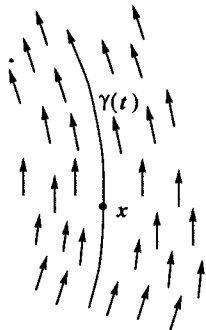


Fig. 8. Integral curve through  $p$  of the vector field  $v$

Calculating the integral curves of a vector field amounts to solving a first-order differential equation. One has to be careful, because the solution might ‘shoot off to infinity’ in a finite amount of time:

**Exercise 20.** Let  $v$  be the vector field  $x^2\partial_x + y\partial_y$  on  $\mathbb{R}^2$ . Calculate the integral curves  $\gamma(t)$  and see which ones are defined for all  $t$ .

We say that the vector field  $v$  is **integrable** if all the integral curves are defined for all  $t$ .

Suppose  $v$  is an integrable vector field on  $M$ , which we think of as the velocity vector field of some water. If we keep track of how *all* the molecules of water are moving along, we have something called a ‘flow’. Let  $\phi_t(p)$  be the integral curve of  $v$  through the point  $p \in M$ . For each time  $t$ , the map

$$\phi_t: M \rightarrow M$$

turns out to be smooth, by a result on the smooth dependence of solutions of differential equations on the initial conditions. Water that was at  $p$  at time zero will be at  $\phi_t(p)$  by time  $t$ , so we call the family of maps  $\{\phi_t\}$  the **flow generated by  $v$** . The defining equation for the flow is (rewriting our equation for  $\gamma$ ):

$$\frac{d}{dt}\phi_t(p) = v_{\phi_t(p)}.$$

**Exercise 21.** Show that  $\phi_0$  is the identity map  $\text{id}: X \rightarrow X$ , and that for all  $s, t \in \mathbb{R}$  we have  $\phi_t \circ \phi_s = \phi_{t+s}$ .

There is an important way to get new vector fields from old ones that is related to the concept of flows. This is called the **Lie bracket** or **commutator** of vector fields. Given  $v, w \in \text{Vect}(M)$ , the Lie bracket  $[v, w]$  is defined by

$$[v, w](f) = v(w(f)) - w(v(f)),$$

for all  $f \in C^\infty(M)$ , or, for short,

$$[v, w] = vw - wv.$$



Let us show that the Lie bracket defined in this way actually is a vector field on the manifold  $M$ . It is easy to prove linearity, so the crucial thing is the Leibniz rule: if  $u = [v, w]$ , we have

$$\begin{aligned} u(fg) &= (vw - wv)(fg) \\ &= v[w(f)g + fw(g)] - w[v(f)g + fv(g)] \\ &= vw(f)g + fvw(g) - wv(f)g - f wv(g) \\ &= u(f)g + fu(g). \end{aligned}$$

Here we used the Leibniz law twice and then used the definition of the Lie brackets.

The Lie bracket measures the failure of 'mixed directional derivatives' to commute. Of course, ordinary mixed partial derivatives *do* commute:

$$[\partial_\mu, \partial_\nu] = 0.$$

We can think of this pictorially, as follows: flowing a little bit first in the  $\partial_\mu$  direction and then in the  $\partial_\nu$  direction gets us to the same place as if we had done it in the other order:

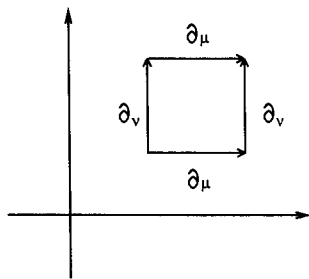


Fig. 9.  $[\partial_\mu, \partial_\nu] = 0$

However, if we take some other vector fields, this does not usually work:

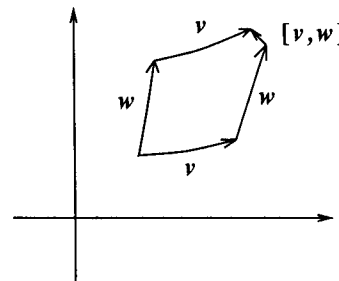


Fig. 10.  $[v, w] \neq 0$

We say in this case that the vector fields do not **commute**.

**Exercise 22.** Consider the normalized vector fields in the  $r$  and  $\theta$  directions on the plane in polar coordinates (not defined at the origin):

$$\begin{aligned} v &= \frac{x\partial_x + y\partial_y}{\sqrt{x^2 + y^2}} \\ w &= \frac{x\partial_y - y\partial_x}{\sqrt{x^2 + y^2}}. \end{aligned}$$

Calculate  $[v, w]$ .

To make the relationship with flows precise, suppose that  $v$  generates the flow  $\phi_t$ , and  $w$  generates the flow  $\psi_s$ . Then for any  $f \in C^\infty(M)$

$$(vf)(p) = \left. \frac{d}{dt} f(\phi_t(p)) \right|_{t=0},$$

and similarly

$$(wf)(p) = \left. \frac{d}{ds} f(\psi_s(p)) \right|_{s=0},$$

so one can check that

$$[v, w](f)(p) = \left. \frac{d^2}{dt ds} f(\phi_t(\psi_s(p))) - f(\psi_s(\phi_t(p))) \right|_{s=t=0}.$$

If you think about it, this is related to what we said above. In  $f(\phi_t(\psi_s(p)))$  we are starting at  $p$ , flowing along  $w$  a little bit, then along  $v$  a little bit, and then evaluating  $f$ , while in  $f(\psi_s(\phi_t(p)))$  we are flowing first along  $v$  and then  $w$ . The Lie bracket measures (infinitesimally, as it were) how these flows fail to commute!

**Exercise 23.** Check the equation above.

The Lie bracket of vector fields satisfies some identities which we will come back to in Part II. For now, we simply let the reader prove them:

**Exercise 24.** Show that for all vector fields  $u, v, w$  on a manifold, and all real numbers  $\alpha$  and  $\beta$ , we have:

$$1) [v, w] = -[w, v].$$

$$2) [u, \alpha v + \beta w] = \alpha[u, v] + \beta[u, w].$$

$$3) \text{ The Jacobi identity: } [u, [v, w]] + [v, [w, u]] + [w, [u, v]] = 0.$$

## Chapter 4

# Differential Forms

*As a herald it's my duty  
to explain those forms of beauty. — Goethe, Faust.*

## 1-forms

The electric field, the magnetic field, the electromagnetic field on space-time, the current — all these are examples of differential forms. The gradient, the curl, and the divergence can all be thought of as different aspects of single operator  $d$  that acts on differential forms. The fundamental theorem of calculus, Stokes' theorem, and Gauss' theorem are all special cases of a single theorem about differential forms. So while they are somewhat abstract, differential forms are a powerful unifying notion.

We begin with 1-forms. Our goal is to generalize the concept of the gradient of a function to functions on arbitrary manifolds. What we will do is to make up, for each smooth function  $f$  on  $M$ , an object called  $df$  that is supposed to be like the usual gradient  $\nabla f$  defined on  $\mathbb{R}^n$ . Remember that the directional derivative of a function  $f$  in the on  $\mathbb{R}^n$  in the direction  $v$  is just the dot product of  $\nabla f$  with  $v$ :

$$\nabla f \cdot v = v f.$$

In other words, the gradient of  $f$  is a thing that keeps track of the directional derivatives of  $f$  in all directions. We want our ' $df$ ' to do the same job on any manifold  $M$ .

The gradient of a function on  $\mathbb{R}^n$  is a vector field, so one might want to say that  $df$  should be a vector field. The problem is the dot product in the formula above. On  $\mathbb{R}^n$  there is a well-established way to take the dot product of tangent vectors, but manifolds do not come pre-equipped with a way to do this. Geometers call a way of taking dot products of tangent vectors a ‘metric’. In fact, we will see that in general relativity the gravitational field is described by the metric on spacetime. Far from there being a single ‘best’ metric on a manifold, there are typically *lots* that satisfy Einstein’s equations of general relativity. This makes it nice to avoid using a particular metric unless we actually *need* to. Therefore we will not think of  $df$  as a vector field, but as something else, a ‘1-form’.

The trick is to realize what  $\nabla f$  is doing in the formula  $\nabla f \cdot v = vf$ . For each vector field  $v$  that we choose, this formula spits out a function  $vf$ , the directional derivative of  $f$  in the direction  $v$ . In other words, what really matters is the *operator*

$$v \mapsto \nabla f \cdot v,$$

or, what is the same thing,

$$v \mapsto vf.$$

Let us isolate the essential properties of this map. There is really only one: linearity! This means that

$$\nabla f \cdot (v + w) = \nabla f \cdot v + \nabla f \cdot w$$

for any vector fields  $v$  and  $w$ , and

$$\nabla f \cdot (gv) = g(\nabla f \cdot v)$$

where  $g$  is any smooth function on  $\mathbb{R}^n$ . Since we can pull out any function  $g \in C^\infty(\mathbb{R}^n)$  in the above formula, not just constants, mathematicians say that

$$v \mapsto \nabla f \cdot v$$

is **linear over**  $C^\infty(\mathbb{R}^n)$  — not just linear over the real numbers.

So, abstracting a bit, we define a **1-form** on any manifold  $M$  to be a map from  $\text{Vect}(M)$  to  $C^\infty(M)$  that is linear over  $C^\infty(M)$ . In other

words, if we feed a vector field  $v$  to a 1-form  $\omega$ , it spits out a function  $\omega(v)$  in a way satisfying

$$\omega(v + w) = \omega(v) + \omega(w),$$

$$\omega(gv) = g\omega(v).$$

We use  $\Omega^1(M)$  to denote the space of all 1-forms on a manifold  $M$ . Later on we will talk about 2-forms, 3-forms, and so on.

The basic example of a 1-form is this: for any smooth function  $f$  on  $M$  there is a 1-form  $df$  defined by

$$df(v) = vf.$$

(Think of this as a slick way to write  $\nabla f \cdot v = vf$ .) To show that  $df$  is really a 1-form, we just need to check linearity:

$$df(v + w) = (v + w)f = vf + wf = df(v) + df(w),$$

and

$$df(gv) = (gv)(f) = gv(f) = gdf(v).$$

We call the 1-form  $df$  the **differential** of  $f$ , or the **exterior derivative** of  $f$ .

Just as we can add vector fields or multiply them by functions, we can do the same for 1-forms. We can add two 1-forms  $\omega$  and  $\mu$  and get a 1-form  $\omega + \mu$  by defining

$$(\omega + \mu)(v) = \omega(v) + \mu(v),$$

and we can multiply a 1-form  $\omega$  by a smooth function  $f$  and get a 1-form  $f\omega$  by defining

$$(f\omega)(v) = f\omega(v).$$

**Exercise 25.** Show that  $\omega + \mu$  and  $f\omega$  are really 1-forms, i.e., show linearity over  $C^\infty(M)$ .

**Exercise 26.** Show that  $\Omega^1(M)$  is a module over  $C^\infty(M)$  (see the definition in Exercise 8.)

The map  $d: C^\infty(M) \rightarrow \Omega^1(M)$  that sends each function  $f$  to its differential  $df$  is also called the differential, or exterior derivative. It is interesting in its own right, and has the following nice properties:

**Exercise 27.** *Show that*

$$\begin{aligned} d(f+g) &= df + dg \\ d(\alpha f) &= \alpha df \\ (f+g)dh &= f dh + g dh \\ d(fg) &= f dg + g df, \end{aligned}$$

for any  $f, g, h \in C^\infty(M)$  and any  $\alpha \in \mathbb{R}$ .

The first three properties in the exercise above are just forms of linearity, but the last one is a version of the product rule, or Leibniz law:

$$d(fg) = f dg + g df.$$

It is the Leibniz law that makes the exterior derivative really act like a derivative, so if you only want to do *part* of Exercise 27 check that the Leibniz law holds! It is worth mentioning, by the way, that when Leibniz was inventing calculus he first guessed that  $d(fg) = df dg$ , and only got it right the next day.

In fact, the reader has seen differentials before, in calculus. They start out as part of the expressions for differentiation

$$\frac{dy}{dx}$$

and integration

$$\int f(x) dx$$

but soon take on a mysterious life of their own, as in

$$d \sin x = \cos x dx!$$

We bet you remember wondering what the heck these differentials *really are!* In physics one thinks of  $dx$  as an ‘infinitesimal change in position’, and so on — but this is mystifying in its own right. Early in the history of calculus, the philosopher Berkeley complained about these

infinitesimals, writing “They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them ghosts of departed quantities?” More recently, people have worked out an alternative approach to the real numbers, called ‘nonstandard analysis’, that includes a logically satisfactory theory of infinitesimals — puny numbers that are greater than zero but less than any ‘standard’ real number. Most people these days, however, prefer to think of differentials as 1-forms.

Let us show that  $d \sin x = \cos x dx$  is really true as an equation concerning 1-forms on the real line. We need to show that no matter what vector field we feed these two 1-forms, they spit out the same thing. This is not hard. Any vector field  $v$  on  $\mathbb{R}$  is of the form  $v = f(x)\partial_x$ , so on one hand we have

$$(d \sin x)(v) = v \sin x = f(x)\partial_x \sin x = f(x) \cos x,$$

and on the other hand:

$$(\cos x dx)(v) = (\cos x) v(x) = f(x) \cos x \partial_x x = f(x) \cos x.$$

This is in fact just a special case of the following:

**Exercise 28.** *Suppose  $f(x^1, \dots, x^n)$  is a function on  $\mathbb{R}^n$ . Show that*

$$df = \partial_\mu f dx^\mu.$$

This means that on  $\mathbb{R}^n$  the exterior derivative of a function is really just a different way of thinking about its gradient, since in old-fashioned language we had

$$\nabla f = (\partial_1 f, \dots, \partial_n f).$$

To do the exercise above one needs to use the fact that the vector fields  $\{\partial_\mu\}$  form a basis of vector fields on  $\mathbb{R}^n$ . In fact, this implies that the 1-forms  $\{dx^\mu\}$  form a basis of 1-forms on  $\mathbb{R}^n$ . The key is that

$$dx^\mu(\partial_\nu) = \partial_\nu x^\mu = \delta_\nu^\mu$$

where the **Kronecker delta**  $\delta_\nu^\mu$  equals 1 if  $\mu = \nu$  and 0 otherwise. Now suppose we have a 1-form  $\omega$  on  $\mathbb{R}^n$ . Then we can define some functions

$$\omega_\mu = \omega(\partial_\mu),$$

and we claim that

$$\omega = \omega_\mu dx^\mu.$$

This will imply that the 1-forms  $\{dx^\mu\}$  span the 1-forms on  $\mathbb{R}^n$ . To show that  $\omega$  equals  $\omega_\mu dx^\mu$ , we just need to feed both of them a vector field and show that they spit out the same function! Feed them  $v = v^\nu \partial_\nu$ , for example. Then on the one hand

$$\omega(v) = \omega(v^\nu \partial_\nu) = v^\nu \omega(\partial_\nu) = v^\nu \omega_\nu,$$

while on the other hand,

$$(\omega_\mu dx^\mu)(v) = (\omega_\mu dx^\mu)(v^\nu \partial_\nu) = \omega_\mu v^\nu dx^\mu(\partial_\nu) = \omega_\nu v^\nu$$

using the fact that  $dx^\mu(\partial_\nu) = \delta_\nu^\mu$ .

We leave it to the reader to finish the proof that the 1-forms  $\{dx^\mu\}$  form a basis of  $\Omega^1(\mathbb{R}^n)$ :

**Exercise 29.** Show that the 1-forms  $\{dx^\mu\}$  are linearly independent, i.e., if

$$\omega = \omega_\mu dx^\mu = 0$$

then all the functions  $\omega_\mu$  are zero.

## Cotangent Vectors

Just as a vector field on  $M$  gives a tangent vector at each point of  $M$ , a 1-form on  $M$  gives a kind of vector at each point of  $M$  called a 'cotangent vector'. Given a manifold  $M$  and a point  $p \in M$ , a **cotangent vector**  $\omega$  at  $p$  is defined to be a linear map from the tangent space  $T_p M$  to  $\mathbb{R}$ . Let  $T_p^* M$  denote the space of all cotangent vectors at  $p$ .

For example, if we have a 1-form  $\omega$  on  $M$ , we can define a cotangent vector  $\omega_p \in T_p^* M$  by saying that for any vector field  $v$  on  $M$ ,

$$\omega_p(v_p) = \omega(v)(p).$$

Here the right-hand side stands for the function  $\omega(v)$  evaluated at the point  $p$ .

**Exercise 30.** For the mathematically inclined: show that the  $\omega_p$  is really well-defined by the formula above. That is, show that  $\omega(v)(p)$  really depends only on  $v_p$ , not on the values of  $v$  at other points. Also, show that a 1-form is determined by its values at points. In other words, if  $\omega, \nu$  are two 1-forms on  $M$  with  $\omega_p = \nu_p$  for every point  $p \in M$ , then  $\omega = \nu$ .

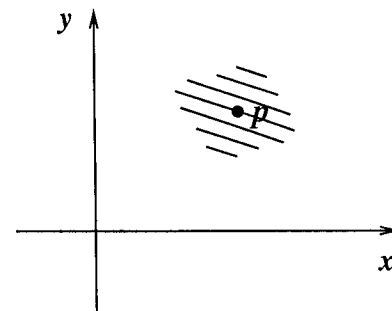
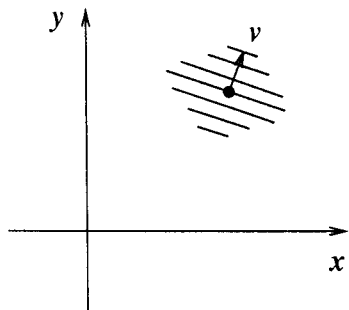


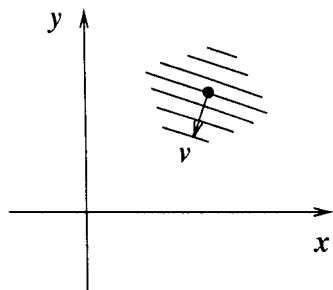
Fig. 1. A picture of the cotangent vector  $(df)_p$

How can we visualize a cotangent vector? A tangent vector is like a little arrow; it points somewhere. A cotangent vector does not. A nice heuristic way to visualize a cotangent vector is as a little stack of parallel hyperplanes. For example, if we have a function  $f$  on a manifold  $M$ , we can visualize  $df$  at a point  $p \in \mathbb{R}^2$  by drawing the level curves of  $f$  right near  $p$ , which look like a little stack of parallel lines. The picture in Figure 1 is two-dimensional, so level surfaces are just contour lines, and hyperplanes are just lines.

The bigger  $df$  is, the more tightly packed the hyperplanes are. When we take a tangent vector  $v \in T_p M$ , the number  $df(v)$  basically just counts how many little hyperplanes in the stack  $df$  the vector  $v$  crosses. In Figure 2 we show a situation where  $df(v) = 3$ . By definition, of course, the number  $df(v)$  is just the directional derivative  $v(f)$ !

Fig. 2.  $df(v) = 3$ 

Actually we must be a bit careful about thinking about  $df(v)$  in terms of pictures, because it could be negative! If we think of the little stack of hyperplanes as ‘contour lines’, we should really count the number of them  $v$  crosses with a plus sign if  $v$  is pointing ‘uphill’ and a minus sign if it is pointing ‘downhill’.

Fig. 3.  $df(-v) = -3$ 

If this way of thinking of 1-forms is confusing, feel free to ignore it — but people with a strong taste for visualization may find it very handy.

Now let us explain precisely what we mean by 1-forms being dual to vector fields. First of all, given any vector space  $V$ , the **dual** vector space  $V^*$  is defined to be the space of all linear functionals  $\omega: V \rightarrow \mathbb{R}$ . In particular, the cotangent space  $T_p^*M$  is the dual of the tangent space  $T_pM$ . More generally, if we have a linear map from one vector space to

another,

$$f: V \rightarrow W,$$

we automatically get a map from  $W^*$  to  $V^*$ , the **dual** of  $f$ , written

$$f^*: W^* \rightarrow V^*$$

and defined by

$$(f^*\omega)(v) = \omega(f(v)).$$

Thus the dual of a vector space is a contravariant sort of beast: linear maps between vector spaces give rise to maps between their duals that go ‘backwards’.

**Exercise 31.** Show that the dual of the identity map on a vector space  $V$  is the identity map on  $V^*$ . Suppose that we have linear maps  $f: V \rightarrow W$  and  $g: W \rightarrow X$ . Show that  $(gf)^* = f^*g^*$ .

This means that cotangent vectors are contravariant. In other words, suppose we have a map  $\phi: M \rightarrow N$  from one manifold to another with  $\phi(p) = q$ . We saw in the last section that there is a linear map

$$\phi_*: T_pM \rightarrow T_qN.$$

This gives a dual map, which we write as  $\phi^*$ , going the other way:

$$\phi^*: T_q^*N \rightarrow T_p^*M.$$

If  $\omega$  is a cotangent vector at  $\phi(x)$ , we call  $\phi^*\omega$  the **pullback** of  $\omega$  by  $\phi$ . Explicitly, if  $v \in T_pM$  and  $\omega \in T_q^*N$ , we have

$$(\phi^*\omega)(v) = \omega(\phi_*v).$$

We can also do this ‘pulling back’ globally. That is, given a 1-form  $\omega$  on  $N$ , we get a 1-form  $\phi^*\omega$  on  $M$  defined by

$$(\phi^*\omega)_p = \phi^*(\omega_q)$$

where  $\phi(p) = q$ .

**Exercise 32.** Show that the pullback of 1-forms defined by the formula above really exists and is unique.

Recall from the previous section that we can also pull back functions on  $N$  to functions on  $M$  when we have a map  $\phi: M \rightarrow N$ . There is a marvelous formula saying that the exterior derivative is compatible with pullbacks. Namely, given a function  $f$  on  $N$  and a map  $\phi: M \rightarrow N$ , we have

$$\phi^*(df) = d(\phi^*f).$$

Mathematicians summarize this by saying that the exterior derivative is **natural**. For example, if  $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a diffeomorphism representing some change of coordinates, the above formula implies that we can compute  $d$  of a function on  $\mathbb{R}^n$  either before or after changing coordinates, and get the same answer. (We discuss this a bit more in the next section.) So naturality can be regarded as a grand generalization of coordinate-independence.

To prove the above equation we just need to show that both sides, which are 1-forms on  $M$ , give the same cotangent vector at every point  $p$  in  $M$ :

$$(\phi^*(df))_p = (d(\phi^*f))_p.$$

This, in turn, means that

$$(\phi^*(df))_p(v) = (d(\phi^*f))_p(v)$$

for all  $v \in T_p M$ . To prove this, work out the left hand side using all the definitions and show it equals the right hand side:

$$\begin{aligned} (\phi^*(df))_p(v) &= (df)_q(\phi_*v) \\ &= ((\phi_*v)f)(p) \\ &= v(\phi^*f)(p) \\ &= (d(\phi^*f))_p(v) \end{aligned}$$

To make this more concrete it might be good to work out some examples:

**Exercise 33.** Let  $\phi: \mathbb{R} \rightarrow \mathbb{R}$  be given by  $\phi(t) = \sin t$ . Let  $dx$  be the usual 1-form on  $\mathbb{R}$ . Show that  $\phi_*dx = \cos t dt$ .

**Exercise 34.** Let  $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  denote rotation counterclockwise by the angle  $\theta$ . Let  $dx, dy$  be the usual basis of 1-forms on  $\mathbb{R}^2$ . Show that

$$\begin{aligned} \phi^*dx &= \cos \theta dx - \sin \theta dy \\ \phi^*dy &= \sin \theta dx + \cos \theta dy. \end{aligned}$$

The formula

$$\phi^*(df) = d(\phi^*f)$$

is a very good reason why the differential of a function has to be a 1-form instead of a vector field. Both functions and 1-forms are contravariant, so if  $\phi: M \rightarrow N$  and  $f \in C^\infty(N)$ , both sides above are 1-forms on  $N$ . If one tried to make the differential of a function be a vector field, there would be no way to write down a sensible formula like this, since vector fields are covariant. (Try it!)

## Change of Coordinates

*Indeed, from childhood we have become familiar with the appearance of physical equations in non-Cartesian systems, such as polar coordinates, and in non-inertial systems, such as rotating coordinates. — Steven Weinberg*

*The introduction of numbers as coordinates [...] is an act of violence whose only practical vindication is the special calculatory manageability of the ordinary number continuum with its four basic operations. — Hermann Weyl*

So far we have been avoiding coordinates as much as possible. The reason, of course, is that the world does not come equipped with coordinates! As far as we can tell, coordinates are something *we* impose upon the world when we want to talk about where things are. They are extremely useful, and in many applications quite essential. Unfortunately, different people might pick different coordinates! So it is good to know how the components of a vector field or 1-form depend on the coordinates used.

First let us describe how one can use coordinates locally on any manifold to work with vector fields and differential forms. We described the basic idea back in Chapter 2: given an  $n$ -dimensional manifold  $M$ , a chart is a diffeomorphism  $\varphi$  from an open set  $U$  in  $M$  to  $\mathbb{R}^n$ . This allows us to do turn calculations on  $U$  into calculations on  $\mathbb{R}^n$ .

For example, we can use  $\varphi$  to pull back the coordinate functions  $x^\mu$  from  $\mathbb{R}^n$  to  $U$ . Instead of calling these functions  $\varphi^*x^\mu$  as one really should, we usually call them simply  $x^\mu$ . This is not too confusing as long as we know we are ‘working in the chart’  $\varphi: U \rightarrow \mathbb{R}^n$ . The functions

$x^\mu$  on  $U$  are known as **local coordinates** on  $U$ . Any function on  $U$  can be written as a function  $f(x^1, \dots, x^n)$  of these local coordinates.

Similarly, the coordinate vector fields  $\partial_\mu$  are a basis of vector fields on  $\mathbb{R}^n$ , and we may push these forwards by  $\varphi^{-1}$  to a basis of vector fields on  $U$ . As with the local coordinates, people usually denote these vector fields simply as  $\partial_\mu$ . These are called the **coordinate vector fields** associated to the local coordinates  $x^\mu$  on  $U$ . One thus writes any vector field  $v$  on  $U$  as

$$v = v^\mu \partial_\mu.$$

In the same way, the coordinate 1-forms  $dx^\mu$  are a basis of 1-forms on  $\mathbb{R}^n$ , which we may pull back to  $U$  by  $\varphi$ , obtaining a basis of 1-forms on  $U$ . These are called the **coordinate 1-forms** associated to the local coordinates  $x^\mu$ . These are written simply as  $dx^\mu$ . Note that our use of  $x^\mu$  and  $dx^\mu$  to denote functions and 1-forms on  $U$ , while sloppy, is consistent:

**Exercise 35.** Show that the coordinate 1-forms  $dx^\mu$  really are the differentials of the local coordinates  $x^\mu$  on  $U$ .

We can write any 1-form  $\omega$  on  $U$  as

$$\omega = \omega_\mu dx^\mu.$$

We should emphasize that it is bad to think of vector fields  $v$  or 1-forms  $\omega$  as *being* their components  $v^\mu$  or  $\omega_\mu$ . Instead, we should think of them as *having* components, which depend on the basis used. For example, the usual coordinate functions  $x^1, \dots, x^n$  on  $\mathbb{R}^n$  give a basis  $\{\partial_\mu\}$  for  $\text{Vect}(\mathbb{R}^n)$ . Given any vector field  $v$  on  $\mathbb{R}^n$ , I can write it uniquely as

$$v = v^\mu \partial_\mu,$$

where the  $v^\mu$  are functions on  $\mathbb{R}^n$ . But suppose you chose some other coordinates on  $\mathbb{R}^n$  — that is, some functions  $x'^1, \dots, x'^n$  on  $\mathbb{R}^n$  such that  $\{\partial'_\nu\}$  was another basis for  $\text{Vect}(\mathbb{R}^n)$ . Then you would write

$$v = v'^\nu \partial'_\nu.$$

The vector field  $v$  is the same in both cases — it is blissfully unaware of which coordinates we mere mortals are using. But its components

depend on a coordinate system, and for us to talk to each other, we need to know how your components,  $v'^\nu$ , are expressed in terms of mine,  $v^\mu$ .

First, since your vector fields form a basis, we can express mine as linear combinations of yours:

$$\partial_\mu = T_\mu^\nu \partial'_\nu,$$

where the  $T_\mu^\nu$  are a matrix of functions on  $\mathbb{R}^n$ . It is not too hard to figure out these functions. Just apply both sides of the equation, which are vector fields, to the coordinate function  $x'^\lambda$ :

$$\partial_\mu x'^\lambda = T_\mu^\nu \partial'_\nu x'^\lambda.$$

The partial derivative  $\partial'_\nu x'^\lambda$  is just the Kronecker delta  $\delta_\nu^\lambda$ , so actually we just have

$$\partial_\mu x'^\lambda = T_\mu^\lambda.$$

We can write this out somewhat more impressively as follows:

$$T_\mu^\lambda = \frac{\partial x'^\lambda}{\partial x^\mu}.$$

This implies that

$$\partial_\mu = \frac{\partial x'^\nu}{\partial x^\mu} \partial'_\nu.$$

Then, to express the components  $v'^\mu$  in terms of the components  $v^\mu$ , start with the fact that  $v'^\mu \partial'_\mu = v^\mu \partial_\mu$ , and use the equation above to get

$$v'^\nu \partial'_\nu = v^\mu \frac{\partial x'^\nu}{\partial x^\mu} \partial'_\nu.$$

Equating coefficients, we get

$$v'^\nu = \frac{\partial x'^\nu}{\partial x^\mu} v^\mu.$$

Now we can talk to each other! In short, to translate from my components to yours, I simply multiply by a matrix of partial derivatives corresponding to the change of coordinates.

1-forms work the same way, and we leave them as an important exercise for the reader:



**Exercise 36.** In the situation above, show that

$$dx'^\nu = \frac{\partial x'^\nu}{\partial x^\mu} dx^\mu.$$

Show that for any 1-form  $\omega$  on  $\mathbb{R}^n$ , writing

$$\omega = \omega_\mu dx^\mu = \omega'_\nu dx'^\nu,$$

your components  $\omega'_\nu$  are related to my components  $\omega_\mu$  by

$$\omega'_\nu = \frac{\partial x^\mu}{\partial x'^\nu} \omega_\mu.$$

There is an interesting distinction between ‘active’ or ‘passive’ coordinate transformations. A **passive** coordinate transformation is a change of coordinate functions (on  $\mathbb{R}^n$ , or on a chart), which is what we have just been considering. We are not moving points of our space around, just changing the functions we use to describe them. An **active** coordinate transformation is just another name for a diffeomorphism

$$\phi: M \rightarrow M;$$

it moves the points of  $M$  around. We can push vector fields forwards by a diffeomorphism, and pull functions and 1-forms back. It is nice to know how these look in the special case of  $\mathbb{R}^n$  (or a chart). Not surprisingly, the formulas look similar to the formulas for passive coordinate transformations that we have just derived!

There is, however, something a bit tricky about this business. The simplest example of this trickiness occurs when people in certain places switch from standard time to daylight saving time in the spring. The mnemonic formula is ‘spring forward, fall back’. This is supposed to remind you to set your clock forward in the spring and back in the fall. The hard part is remembering what setting a clock ‘forward’ means! Is one supposed to move the hour hand to a *later* time, so one has to wake up *earlier* than one otherwise would? Or is one supposed to move the hour hand to an *earlier* time, so one can stay in bed *later*? Note that it takes a clock and a point in time to give a number that we call the ‘time’  $t$ . More generally, it takes a coordinate system *together* with a point in spacetime to give a number. Changing the coordinate system

one way has a similar effect to moving points of spacetime around the opposite way.

Let us now consider the effect a map  $\phi: \mathbb{R}^m \rightarrow \mathbb{R}^n$  has on coordinate vector fields and 1-forms. If  $n = m$  and  $\phi$  is a diffeomorphism, this is an ‘active coordinate transformation’, but it is actually easier to keep things straight if we work in the general case. Write  $x^1, \dots, x^m$  for the coordinates on  $\mathbb{R}^m$ , and  $x'^1, \dots, x'^n$  for the coordinates on  $\mathbb{R}^n$ . First note that we can pull *back* the coordinate functions  $x'^\nu$  on  $\mathbb{R}^n$  to functions  $\phi^* x'^\nu$  on  $\mathbb{R}^m$  using  $\phi$ . The definition is that

$$(\phi^* x'^\nu)(p) = x'^\nu(\phi(p))$$

for any point  $p$  in  $\mathbb{R}^m$ . In what follows, we will be sloppy and write

$$\frac{\partial x'^\nu}{\partial x^\mu}$$

when we really mean

$$\frac{\partial}{\partial x^\mu} \phi^* x'^\nu.$$

The reason we do this is simply that everyone does it, and the reader will have to get used to it.

Now consider the coordinate vector field  $\partial_\mu$  on  $\mathbb{R}^m$ . We can push  $\partial_\mu$  *forward* by  $\phi$ , and we claim that

$$\phi_* \partial_\mu = \frac{\partial x'^\nu}{\partial x^\mu} \partial'_\nu$$

To see this, just apply both sides to any coordinate function  $x'^\lambda$  on  $\mathbb{R}^n$  and show that we get the same answer. The left hand side gives

$$\begin{aligned} (\phi_* \partial_\mu)(x'^\lambda) &= \partial_\mu(\phi^* x'^\lambda) \\ &= \frac{\partial x'^\lambda}{\partial x^\mu}, \end{aligned}$$

where in the last step we are being sloppy in the way described above. The right hand side gives

$$\begin{aligned} \frac{\partial x'^\nu}{\partial x^\mu} \partial'_\nu x'^\lambda &= \frac{\partial x'^\nu}{\partial x^\mu} \delta_\nu^\lambda \\ &= \frac{\partial x'^\lambda}{\partial x^\mu}, \end{aligned}$$

which is the same.

Finally, consider a coordinate 1-form  $dx^\nu$ . We can pull this *back* by  $\phi$ . We claim that

$$\phi^*(dx^\nu) = \frac{\partial x^\nu}{\partial x^\mu} dx^\mu.$$

**Exercise 37.** Show this.

With these basic formulas in hand, you should be able to transform between coordinates both actively and passively!

To conclude, we should note that sometimes it is nice to be more general and work with a basis  $e_\mu$  of vector fields on a chart that are not the coordinate vector fields. These are easy to come by:

**Exercise 38.** Let

$$e_\mu = T_\mu^\nu \partial_\nu,$$

where  $\partial_\nu$  are the coordinate vector fields associated to local coordinates on an open set  $U$ , and  $T_\mu^\nu$  are functions on  $U$ . Show that the vector fields  $e_\mu$  are a basis of vector fields on  $U$  if and only if for each  $p \in U$  the matrix  $T_\mu^\nu(p)$  is invertible.

If we have such a basis, we automatically get a **dual basis** of 1-forms  $f^\mu$  on  $U$  such that

$$f^\mu(e_\nu) = \delta_\nu^\mu,$$

the Kronecker delta.

**Exercise 39.** Use the previous exercise to show that the dual basis exists and is unique.

We can write any vector field  $v$  on  $U$  as a linear combination

$$v = v^\mu e_\mu$$

where  $v^1, \dots, v^n$  are functions on  $U$ , called the **components** of  $v$  in the basis  $e_\mu$ . Similarly, we can write any 1-form  $\omega$  on  $U$  as a linear combination

$$\omega = \omega_\mu f^\mu.$$

We will use these more general bases quite a bit in the next chapter, when we discuss the notion of a 'metric'. This is like an inner product,

and it will be handy to work with 'orthonormal' bases of vector fields and 1-forms on a chart. We leave it to the reader to work out how the components of a vector field or 1-form change when we perform an arbitrary change of basis:

**Exercise 40.** Let  $e_\mu$  be a basis of vector fields on  $U$  and let  $f^\mu$  be the dual basis of 1-forms. Let

$$e'_\mu = T_\mu^\nu e_\nu$$

be another basis of vector fields, and let  $f'^\mu$  be the corresponding dual basis of 1-forms. Show that

$$f'^\mu = (T^{-1})^\mu_\nu f^\nu.$$

Show that if  $v = v^\mu e_\mu = v'^\mu e'_\mu$ , then

$$v'^\mu = (T^{-1})^\mu_\nu v^\nu,$$

and that if  $\omega = \omega_\mu f^\mu = \omega'_\mu f'^\mu$  then

$$\omega'_\mu = T_\mu^\nu \omega_\nu.$$

## p-forms

By the geometrical product of two vectors, we mean the surface content of the parallelogram determined by these vectors; we however fix the position of the plane in which the parallelogram lies. We refer to two surface areas as geometrically equal only when they are equal in content and lie in parallel planes. By the geometrical product of three vectors we mean the solid (a parallelepiped) formed from them. — Hermann Grassman

If you ever seriously wondered how to take cross products in 4 dimensions, you were well on your way to reinventing differential forms. In fact, if you ever wondered why the definition of cross products requires a 'right-hand rule', you were getting close. (This rule is especially irksome to those who happen to be left-handed.) Differential forms allow one to generalize cross products to any number of dimensions, and it turns out that if one does things correctly, no right-hand rule is necessary! Interestingly, though, it turns out to be better to define the cross product not for tangent vectors (or vector fields) but for cotangent vectors (or 1-forms). If we do this, we get an extra bonus. Namely, we can

show that the gradient, curl, and divergence are all different versions of the same thing, and see how to define them on arbitrary manifolds.

Let us plunge right in. Let  $V$  be a vector space. We want to be able to multiply two vectors in  $V$  somehow, and we want the basic property of the cross product, the antisymmetry,

$$\vec{v} \times \vec{w} = -\vec{w} \times \vec{v},$$

to hold. But we will call this generalized sort of cross product the ‘wedge product’ (or ‘exterior product’) and write it with a  $\wedge$ . We proceed in an abstract, algebraic sort of way. Namely, we will define a bigger vector space  $\Lambda V$ , in fact an algebra, so that the wedge product of any number of vectors in  $V$  will lie in this algebra. First we will give the definition as a mathematician would: the **exterior algebra** over  $V$ , denoted  $\Lambda V$ , is the algebra generated by  $V$  with the relations

$$v \wedge w = -w \wedge v$$

for all  $v, w \in V$ . What does this mean? Roughly, it means that we start with the vectors in  $V$  together with an element 1, and then form an algebra by taking all linear combinations of formal products of the form  $v_1 \wedge \cdots \wedge v_p$ , where  $v_i \in V$ ; the only relations we impose upon these linear combinations are those in the definition of an algebra (as defined above in Chapter 3) together with the ‘anticommutative’ rule  $v \wedge w = -w \wedge v$ .

For example, say  $V$  is 3-dimensional. Then everything in  $\Lambda V$  is a linear combination of wedge products of elements of  $V$ . Suppose  $V$  has a basis  $dx, dy, dz$ . (We write the basis this way because in a bit we will want  $V$  to be a space of cotangent vectors.) Then for starters we have

$$1 \in \Lambda V$$

and

$$dx, dy, dz \in \Lambda V,$$

along with all linear combinations of these. But we can also take the wedge product of any two elements  $v, w \in V$  and get an element of  $\Lambda V$ .

If

$$\begin{aligned} v &= v_x dx + v_y dy + v_z dz \\ w &= w_x dx + w_y dy + w_z dz \end{aligned}$$

then we have

$$v \wedge w = (v_x dx + v_y dy + v_z dz) \wedge (w_x dx + w_y dy + w_z dz)$$

$$(v_x w_y - v_y w_x) dx \wedge dy + (v_y w_z - v_z w_y) dy \wedge dz + (v_z w_x - v_x w_z) dz \wedge dx,$$

where all we did is use the definition of an algebra together with the ‘anticommutative’ rule. Notice that this looks a whole lot like the formula for the cross product! If we have a third element of  $V$ , say

$$u = u_x dx + u_y dy + u_z dz,$$

we can get another element of  $\Lambda V$ , namely  $u \wedge v \wedge w$ . This triple wedge product is closely related to the ‘triple product’ of three vectors in  $\mathbb{R}^3$ ,  $\vec{u} \cdot (\vec{v} \times \vec{w})$ . We can also take wedge products of four or more vectors, but if  $V$  is 3-dimensional, this is always zero:

**Exercise 41.** Show that

$$u \wedge v \wedge w = \det \begin{pmatrix} u_x & u_y & u_z \\ v_x & v_y & v_z \\ w_x & w_y & w_z \end{pmatrix} dx \wedge dy \wedge dz.$$

Compare this to  $\vec{u} \cdot (\vec{v} \times \vec{w})$ .

**Exercise 42.** Show that if  $a, b, c, d$  are four vectors in a 3-dimensional space then  $a \wedge b \wedge c \wedge d = 0$ .

**Exercise 43.** Describe  $\Lambda V$  if  $V$  is 1-dimensional, 2-dimensional, or 4-dimensional.

In general, for any vector space  $V$ , we define  $\Lambda^p V$  to be the subspace of  $\Lambda V$  consisting of linear combinations of  $p$ -fold products of vectors in  $V$ , e.g.

$$v_1 \wedge \cdots \wedge v_p.$$

Elements of  $\Lambda V$  that lie in  $\Lambda^p V$  are said to have **degree**  $p$ . For example,  $\Lambda^1 V$  is just  $V$  itself, while  $\Lambda^0 V$  is by convention defined to be  $\mathbb{R}$ , since numbers can be regarded as wedge products of *no* vectors. Copying the example above, one can show the following:

**Exercise 44.** Let  $V$  be an  $n$ -dimensional vector space. Show that  $\Lambda^p V$  is empty for  $p > n$ , and that for  $0 \leq p \leq n$  the dimension of  $\Lambda^p V$  is  $n!/p!(n-p)!$

Recall that a vector space  $V$  is a **direct sum** of subspaces  $L_1, \dots, L_n$  if every vector  $v \in V$  can be uniquely expressed as  $v_1 + \dots + v_n$ , where  $v_i \in L_i$ . In this situation, we may think of vectors in  $V$  as  $n$ -tuples  $(v_1, \dots, v_n)$  where  $v_i \in L_i$ . Alternatively, given vector spaces  $V_1, \dots, V_n$ , the **direct sum**  $V_1 \oplus \dots \oplus V_n$ , sometimes written

$$\bigoplus_{i=1}^n V_i,$$

is defined as the vector space of all  $n$ -tuples  $(v_1, \dots, v_n)$  with  $v_i \in V_i$ , where addition and scalar multiplication are defined componentwise. The exterior algebra is an example of such a direct sum:

**Exercise 45.** Show that  $\Lambda V$  is the direct sum of the subspaces  $\Lambda^p V$ :

$$\Lambda V = \bigoplus \Lambda^p V,$$

and that the dimension of  $\Lambda V$  is  $2^n$  if  $V$  is  $n$ -dimensional.

There is something very special about the exterior algebra in 3 dimensions! The wedge product of two vectors in  $V$  lies in  $\Lambda^2 V$ . Only in dimension 3 is the dimension of  $\Lambda^2 V$  equal to that of  $V$  itself. So only in 3 dimensions can we pretend, if we so desire, that the wedge product of two vectors is again a vector! The way to do this (as we will see in Chapter 5) is to define a linear map called the ‘star operator’ that turns elements of  $\Lambda^2(V)$  into  $\Lambda V$ . When  $V$  has the basis  $dx, dy, dz$ , the star operator is given by

$$\begin{aligned} \star: dx \wedge dy &\mapsto dz \\ \star: dy \wedge dz &\mapsto dx \\ \star: dz \wedge dx &\mapsto dy. \end{aligned}$$

The cross product really amounts to taking the wedge product and then applying the star operator. Note, however, that our definition of

the star operator incorporates a right-hand rule. We could just as well have defined  $\star: \Lambda^2 V \rightarrow V$  by

$$\begin{aligned} \star: dy \wedge dx &\mapsto dz \\ \star: dz \wedge dy &\mapsto dx \\ \star: dx \wedge dz &\mapsto dy \end{aligned}$$

which would amount to a left-hand rule. In short, the ‘right-hand rule’ nonsense enters when we unnaturally try to make the product of two elements of  $V$  to come out to an element of  $V$ , instead of  $\Lambda^2 V$ . This is noted in some physics books, where they say that the cross product of two vectors is a ‘pseudovector’ or ‘axial vector’, rather than a true vector. We prefer to say that the wedge product of 2 vectors lies in  $\Lambda^2 V$  — this is true in all dimensions.

Exterior algebra is an interesting subject in itself, but we do not just want to generalize the cross product of vectors; we want to generalize the cross product of vector fields. Actually, as already mentioned, it is much better to take products of 1-forms! We will do this by copying our construction of  $\Lambda V$ , with the smooth functions  $C^\infty(M)$  on some manifold  $M$  taking the place of the real numbers, and the 1-forms  $\Omega^1(M)$  taking the place of the vector space  $V$ . Namely, we define the **differential forms** on  $M$ , denoted  $\Omega(M)$ , to be the algebra generated by  $\Omega^1(M)$  with the relations

$$\omega \wedge \mu = -\mu \wedge \omega$$

for all  $\omega, \mu \in \Omega^1(M)$ . To be precise, we should emphasize that we form  $\Omega(M)$  as an algebra ‘over  $C^\infty(M)$ ’. This means, first of all, that  $\Omega(M)$  consists of linear combinations of wedge products of 1-forms with *functions* as coefficients. We allow all **locally finite** linear combinations, that is, those for which every point  $p$  in  $M$  has a neighborhood where only finitely many terms are nonzero. Secondly, it means that  $\Omega(M)$  satisfies the rules of an algebra with *functions* taking the place of numbers. Maybe we should say again what all these rules are. We have, for all  $\omega, \mu, \nu \in \Omega(M)$  and  $f, g \in C^\infty(M)$ ,

$$\omega + \mu = \mu + \omega, \quad \omega + (\mu + \nu) = (\omega + \mu) + \nu, \quad \omega \wedge (\mu \wedge \nu) = (\omega \wedge \mu) \wedge \nu,$$

$$\omega \wedge (\mu + \nu) = \omega \wedge \mu + \omega \wedge \nu, \quad (\omega + \mu) \wedge \nu = \omega \wedge \nu + \mu \wedge \nu,$$

$$1\omega = \omega, \quad f(g\omega) = (fg)\omega, \quad f(\mu + \nu) = f\mu + f\nu, \quad (f + g)\omega = f\omega + g\omega.$$

We define the 0-forms,  $\Omega^0(M)$ , to be the functions themselves, and define the wedge product of a function with a differential form to be the ordinary product:  $f \wedge \omega = f\omega$ . We define the product of a number  $c$  and a differential form  $\omega$  to be the product of the constant function  $c \in \Omega^0(M)$  and  $\omega$ . Elements that are linear combinations of products of  $p$  1-forms are called  **$p$ -forms**, and we write the space of  $p$ -forms on  $M$  as  $\Omega^p(M)$ . We have

$$\Omega(M) = \bigoplus_p \Omega^p(M).$$

For example, suppose  $M = \mathbb{R}^n$ . The 0-forms on  $\mathbb{R}^n$  are just functions, like

$$f.$$

The 1-forms all look like

$$\omega_\mu dx^\mu$$

where the coefficients  $\omega_\mu$  are functions. It is easy to check that the 2-forms all look like

$$\frac{1}{2} \omega_{\mu\nu} dx^\mu \wedge dx^\nu$$

where we have put in a factor of  $\frac{1}{2}$  because  $dx^\mu \wedge dx^\nu = -dx^\nu \wedge dx^\mu$ . Also for this reason, we may as well assume that  $\omega_{\mu\nu} = -\omega_{\nu\mu}$ . Then on  $\mathbb{R}^3$ , for example, we have

$$\omega = \omega_{12} dx^1 \wedge dx^2 + \omega_{23} dx^2 \wedge dx^3 + \omega_{31} dx^3 \wedge dx^1.$$

Similarly, the 3-forms look like

$$\frac{1}{3!} \omega_{\mu\nu\lambda} dx^\mu \wedge dx^\nu \wedge dx^\lambda,$$

and we may as well assume that  $\omega_{\mu\nu\lambda}$  is totally antisymmetric (that is, switches sign when we switch any two indices). On  $\mathbb{R}^3$  we get

$$\omega = \omega_{123} dx^1 \wedge dx^2 \wedge dx^3.$$

There are no nonzero 4-forms, 5-forms, etc., on  $\mathbb{R}^3$ . In general, there are no nonzero  $p$ -forms on an  $n$ -dimensional manifold if  $p > n$ .

We leave it for the reader to show some important facts about differential forms in the following exercises.

**Exercise 46.** Given a vector space  $V$ , show that  $\Lambda V$  is a **graded commutative or supercommutative algebra**, that is, if  $\omega \in \Lambda^p V$  and  $\mu \in \Lambda^q V$ , then

$$\omega \wedge \mu = (-1)^{pq} \mu \wedge \omega.$$

Show that for any manifold  $M$ ,  $\Omega(M)$  is graded commutative.

**Exercise 47.** Show that differential forms are contravariant. That is, show that if  $\phi: M \rightarrow N$  is a map from the manifold  $M$  to the manifold  $N$ , there is a unique **pullback map**

$$\phi^*: \Omega(N) \rightarrow \Omega(M)$$

agreeing with the usual pullback on 0-forms (functions) and 1-forms, and satisfying

$$\begin{aligned} \phi^*(\alpha\omega) &= \alpha\phi^*\omega \\ \phi^*(\omega + \mu) &= \phi^*\omega + \phi^*\mu \\ \phi^*(\omega \wedge \mu) &= \phi^*\omega \wedge \phi^*\mu \end{aligned}$$

for all  $\omega, \mu \in \Omega(N)$  and  $\alpha \in \mathbb{R}$ .

**Exercise 48.** Compare how 1-forms and 2-forms on  $\mathbb{R}^3$  transform under parity. That is, let  $P: \mathbb{R}^3 \rightarrow \mathbb{R}^3$  be the map

$$P(x, y, z) = (-x, -y, -z),$$

known as the ‘parity transformation’. Note that  $P$  maps right-handed bases to left-handed bases and vice versa. Compute  $\phi^*(\omega)$  when  $\omega$  is the 1-form  $\omega_\mu dx^\mu$ , and when it is the 2-form  $\frac{1}{2} \omega_{\mu\nu} dx^\mu \wedge dx^\nu$ .

In physics, the electric field  $\vec{E}$  is called a vector, while the magnetic field  $\vec{B}$  is called an axial vector, because  $\vec{E}$  changes sign under parity transformation, while  $\vec{B}$  does not. In Chapter 5 we will see that it is best to think of the electric field as a 1-form on space, and the magnetic field as a 2-form. In other words, while we may be used to thinking of

$\vec{E} = (E_x, E_y, E_z)$  and  $\vec{B} = (B_x, B_y, B_z)$  as vector fields, it is better to use

$$E = E_x dx + E_y dy + E_z dz$$

and

$$B = B_x dy \wedge dz + B_y dz \wedge dx + B_z dx \wedge dy.$$

By the above exercise, this means that they transform differently under parity.

If the reader is frustrated because exterior algebras and differential forms seem difficult to *visualize*, we suggest taking a peek ahead to Figures 3 and 4 of Chapter 5. Grassman, the inventor of the exterior algebra, visualized a wedge product  $v_1 \wedge \cdots \wedge v_p$  as an oriented parallelepiped with sides given by the vectors  $v_1, \dots, v_p$ . One must be careful, however, because the wedge product of 1-forms corresponds to a parallelepiped in the *cotangent* space.

## The Exterior Derivative

We know from the first section of this chapter that the differential is a nice way to generalize the good old 'gradient' to manifolds. As we saw, the differential of a function, or 0-form, is a 1-form. Now we will show how to take the differential of a  $p$ -form and get a  $(p+1)$ -form:

$$d: \Omega^p(M) \rightarrow \Omega^{p+1}(M).$$

This will let us generalize the gradient, the curl and the divergence in one fell swoop, and see that they are secretly all the same thing. The big clue is that the curl of a gradient is zero:

$$\nabla \times (\nabla f) = 0$$

This suggests that we make  $d$  satisfy  $d(df) = 0$  for any function  $f$ . Another clue is that the various product rules

$$\begin{aligned} \nabla(fg) &= (\nabla f)g + f\nabla g \\ \nabla \times (fv) &= \nabla f \times v + f\nabla \times v \\ \nabla \cdot (fv) &= \nabla f \cdot v + f\nabla \cdot v \\ \nabla \cdot (v \times w) &= (\nabla \cdot v)w - v\nabla \cdot w \end{aligned}$$

should all be special cases of some sort of Leibniz law for differential forms. Since the differential forms are graded commutative, it turns out that we need a graded version of the Leibniz law.

After scratching our head for a while, we define the **exterior derivative**, or **differential**, to be the unique set of maps

$$d: \Omega^p(M) \rightarrow \Omega^{p+1}(M)$$

such that the following properties hold:

- 1)  $d: \Omega^0(M) \rightarrow \Omega^1(M)$  agrees with our previous definition.
- 2)  $d(\omega + \mu) = d\omega + d\mu$  and  $d(c\omega) = cd\omega$  for all  $\omega, \mu \in \Omega(M)$  and  $c \in \mathbb{R}$ .
- 3)  $d(\omega \wedge \mu) = d\omega \wedge \mu + (-1)^p \omega \wedge d\mu$  for all  $\omega \in \Omega^p(M)$  and  $\mu \in \Omega(M)$ .
- 4)  $d(d\omega) = 0$  for all  $\omega \in \Omega(M)$ .

To show that these properties uniquely determine the exterior derivative, one just needs the fact that any 1-form is a locally finite linear combination of those of the form  $df$  (with functions as coefficients). This fact is easy to see on  $\mathbb{R}^n$ , and can be shown in general using charts. Then to calculate  $d$  of any differential form, say

$$fdg \wedge dh,$$

we just use rules 1) - 4):

$$\begin{aligned} d(fdg \wedge dh) &= df \wedge (dg \wedge dh) + f \wedge d(dg \wedge dh) \\ &= df \wedge dg \wedge dh + fd(dg) \wedge dh - fdg \wedge d(dh) \\ &= df \wedge dg \wedge dh. \end{aligned}$$

To show that  $d$  with these properties is actually well-defined, it suffices (by the black magic of algebra) to show that this way of calculating  $d$  is compatible with the relations in the definition of differential forms. The most important one of these is the anticommutative law

$$\omega \wedge \mu = -\mu \wedge \omega$$

for 1-forms. For  $d$  to be well-defined, it had better be true that calculating  $d(\omega \wedge \mu)$  gives the same answer as  $d(-\mu \wedge \omega)$ . This is where the

graded Leibniz law is necessary: when  $\omega$  and  $\mu$  are 1-forms, we have

$$\begin{aligned} d(-\mu \wedge \omega) &= -d(\mu \wedge \omega) \\ &= -d\mu \wedge \omega + \mu \wedge d\omega, \\ &= -\omega \wedge d\mu + d\omega \wedge \mu \\ &= d(\omega \wedge \mu). \end{aligned}$$

Let us calculate the exterior derivative of 1-forms and 2-forms on  $\mathbb{R}^3$ . Taking any 1-form

$$\omega = \omega_x dx + \omega_y dy + \omega_z dz,$$

we get

$$d\omega = d\omega_x \wedge dx + d\omega_y \wedge dy + d\omega_z \wedge dz,$$

hence by the rule for  $d$  of a function and a little extra work

$$d\omega = (\partial_y \omega_z - \partial_z \omega_y) dy \wedge dz + (\partial_z \omega_x - \partial_x \omega_z) dz \wedge dx + (\partial_x \omega_y - \partial_y \omega_x) dx \wedge dy.$$

In other words, the exterior derivative of a 1-form on  $\mathbb{R}^3$  is essentially just the curl! We need right-hand to define the curl, however, while the exterior derivative involves no right-hand rule. This is because  $d$  of a 1-form is a 2-form; the right-hand rule only comes in when one tries to pretend that this 2-form is a 1-form, using the star operator as follows:

$$\star d\omega = (\partial_y \omega_z - \partial_z \omega_y) dx + (\partial_z \omega_x - \partial_x \omega_z) dy + (\partial_x \omega_y - \partial_y \omega_x) dz.$$

And, as noted, this pretense is only possible in 3 dimensions, while we can take  $d$  of a 1-form in any dimension:

**Exercise 49.** Show that on  $\mathbb{R}^n$  the exterior derivative of any 1-form is given by

$$d(\omega_\mu dx^\mu) = \partial_\nu \omega_\mu dx^\nu \wedge dx^\mu.$$

Next, taking a 2-form on  $\mathbb{R}^3$ :

$$\omega = \omega_{xy} dx \wedge dy + \omega_{yz} dy \wedge dz + \omega_{zx} dz \wedge dx$$

we get

$$\begin{aligned} d\omega &= d\omega_{xy} \wedge dx \wedge dy + d\omega_{yz} \wedge dy \wedge dz + d\omega_{zx} \wedge dz \wedge dx \\ &= \partial_z \omega_{xy} dz \wedge dx \wedge dy + \partial_x \omega_{yz} dx \wedge dy \wedge dz + \partial_y \omega_{zx} dy \wedge dz \wedge dx \\ &= (\partial_z \omega_{xy} + \partial_x \omega_{yz} + \partial_y \omega_{zx}) dx \wedge dy \wedge dz. \end{aligned}$$

Thus the exterior derivative of a 2-form on  $\mathbb{R}^3$  is just the divergence in disguise. In short, the exterior derivative has as special cases the following familiar operators:

- ◇ *Gradient*  $d: \Omega^0(\mathbb{R}^3) \rightarrow \Omega^1(\mathbb{R}^3)$
- ◇ *Curl*  $d: \Omega^1(\mathbb{R}^3) \rightarrow \Omega^2(\mathbb{R}^3)$
- ◇ *Divergence*  $d: \Omega^2(\mathbb{R}^3) \rightarrow \Omega^3(\mathbb{R}^3)$

In fact, there is a simple formula for the exterior derivative of any differential form on  $\mathbb{R}^n$ . Let  $I$  stand for a **multi-index**, that is, a  $p$ -tuple  $(i_1, \dots, i_p)$  of distinct integers between 1 and  $n$ . Let  $dx^I$  stand for the  $p$ -form

$$dx^{i_1} \wedge \dots \wedge dx^{i_p}$$

on  $\mathbb{R}^n$ . Then any  $p$ -form on  $\mathbb{R}^n$  can be expressed as

$$\omega = \omega_I dx^I$$

where following the Einstein summation convention we sum over all multi-indices  $I$ . We have

$$d\omega = d\omega_I \wedge dx^I$$

by the Leibniz law, since  $d(dx^I) = 0$  (as can easily be checked). More concretely, using the formula for  $d$  of a function, we have

$$d\omega = (\partial_\mu \omega_I) dx^\mu \wedge dx^I.$$

Using this formula it is easy to derive an amazing identity:

$$d(d\omega) = 0$$

for any differential form on  $\mathbb{R}^n$ . Just compute:

$$\begin{aligned} d(d\omega) &= d(\partial_\mu \omega_I dx^\mu \wedge dx^I) \\ &= \partial_\nu \partial_\mu \omega_I dx^\nu \wedge dx^\mu \wedge dx^I \end{aligned}$$

and note that on the one hand

$$\partial_\nu \partial_\mu \omega_I = \partial_\mu \partial_\nu \omega_I$$

by the equality of mixed partials, but on the other hand

$$dx^\nu \wedge dx^\mu = -dx^\mu \wedge dx^\nu$$

by the anticommutative law. With a little thought one can see this means that  $d(d\omega)$  is equal to the negative of itself, so it is zero. This rule is so important that people often write it as

$$d^2\omega = 0$$

or even just

$$d^2 = 0.$$

On  $\mathbb{R}^3$ ,  $d$  acts like the gradient on 0-forms, the curl on 1-forms and the divergence on 2-forms, so the identity  $d^2 = 0$  contains within it the identities

$$\nabla \times (\nabla f) = 0$$

and

$$\nabla \cdot (\nabla \times v) = 0.$$

But this identity is better, since it applies to differential forms in any dimension. In fact it applies to any manifold! Here is an easy proof that does not use coordinates. By definition, any  $p$ -form on a manifold is a linear combination — with *constant* coefficients — of  $p$ -forms like

$$\omega = f_0 df_1 \wedge \cdots \wedge df_p.$$

So it suffices to prove the identity for  $p$ -forms of this sort. We have

$$d\omega = df_0 \wedge df_1 \wedge \cdots \wedge df_p$$

by the Leibniz law and the fact that  $d(df) = 0$  for any function. Using the Leibniz law and  $d(df) = 0$  again, we obtain

$$d(d\omega) = 0.$$

It turns out that the identity  $d^2 = 0$  and its generalizations have profound consequences for physics, starting with Maxwell's equations. It is also the basis of a very important connection between geometry and topology, called deRham theory. We will explore these in Chapter 6. When we do, it is important to remember that this identity is just a way of saying that partial derivatives commute! As so often the case, the simplest facts in mathematics lie at the root of some of the most sophisticated developments.

We will wrap up this section by showing that the exterior derivative is **natural**. We already discussed this for functions in Section 4; it simply meant that  $d$  commutes with pullbacks. In fact, this is true for differential forms of any degree. In other words, for any map  $\phi: M \rightarrow N$  between manifolds, and any differential form  $\omega \in \Omega^p(M)$ , we have

$$\phi^*(d\omega) = d(\phi^*\omega).$$

The proof is easy. By Exercise 47,  $\phi^*$  is real-linear, so it suffices to treat the case where

$$\omega = f_0 df_1 \wedge \cdots \wedge df_p.$$

We then have, using Exercise 47 again together with the naturality of  $d$  on functions,

$$\begin{aligned} \phi^*(d\omega) &= \phi^*(df_0 \wedge df_1 \wedge \cdots \wedge df_p) \\ &= \phi^*df_0 \wedge \cdots \wedge \phi^*df_p \\ &= d\phi^*f_0 \wedge \cdots \wedge d\phi^*f_p \\ &= d(\phi^*f_0 \wedge d\phi^*f_1 \wedge \cdots \wedge d\phi^*f_p) \\ &= d(\phi^*f_0 \wedge \phi^*df_1 \wedge \cdots \wedge \phi^*df_p) \\ &= d(\phi^*(f_0 \wedge df_1 \wedge \cdots \wedge df_p)) \\ &= d(\phi^*\omega) \end{aligned}$$

as desired.



for any differential form on  $\mathbb{R}^n$ . Just compute:

$$\begin{aligned} d(d\omega) &= d(\partial_\mu \omega_I dx^\mu \wedge dx^I) \\ &= \partial_\nu \partial_\mu \omega_I dx^\nu \wedge dx^\mu \wedge dx^I \end{aligned}$$

and note that on the one hand

$$\partial_\nu \partial_\mu \omega_I = \partial_\mu \partial_\nu \omega_I$$

by the equality of mixed partials, but on the other hand

$$dx^\nu \wedge dx^\mu = -dx^\mu \wedge dx^\nu$$

by the anticommutative law. With a little thought one can see this means that  $d(d\omega)$  is equal to the negative of itself, so it is zero. This rule is so important that people often write it as

$$d^2\omega = 0$$

or even just

$$d^2 = 0.$$

On  $\mathbb{R}^3$ ,  $d$  acts like the gradient on 0-forms, the curl on 1-forms and the divergence on 2-forms, so the identity  $d^2 = 0$  contains within it the identities

$$\nabla \times (\nabla f) = 0$$

and

$$\nabla \cdot (\nabla \times v) = 0.$$

But this identity is better, since it applies to differential forms in any dimension. In fact it applies to any manifold! Here is an easy proof that does not use coordinates. By definition, any  $p$ -form on a manifold is a linear combination — with *constant* coefficients — of  $p$ -forms like

$$\omega = f_0 df_1 \wedge \cdots \wedge df_p.$$

So it suffices to prove the identity for  $p$ -forms of this sort. We have

$$d\omega = df_0 \wedge df_1 \wedge \cdots \wedge df_p$$

by the Leibniz law and the fact that  $d(df) = 0$  for any function. Using the Leibniz law and  $d(df) = 0$  again, we obtain

$$d(d\omega) = 0.$$

It turns out that the identity  $d^2 = 0$  and its generalizations have profound consequences for physics, starting with Maxwell's equations. It is also the basis of a very important connection between geometry and topology, called deRham theory. We will explore these in Chapter 6. When we do, it is important to remember that this identity is just a way of saying that partial derivatives commute! As so often the case, the simplest facts in mathematics lie at the root of some of the most sophisticated developments.

We will wrap up this section by showing that the exterior derivative is **natural**. We already discussed this for functions in Section 4; it simply meant that  $d$  commutes with pullbacks. In fact, this is true for differential forms of any degree. In other words, for any map  $\phi: M \rightarrow N$  between manifolds, and any differential form  $\omega \in \Omega^p(M)$ , we have

$$\phi^*(d\omega) = d(\phi^*\omega).$$

The proof is easy. By Exercise 47,  $\phi^*$  is real-linear, so it suffices to treat the case where

$$\omega = f_0 df_1 \wedge \cdots \wedge df_p.$$

We then have, using Exercise 47 again together with the naturality of  $d$  on functions,

$$\begin{aligned} \phi^*(d\omega) &= \phi^*(df_0 \wedge df_1 \wedge \cdots \wedge df_p) \\ &= \phi^*df_0 \wedge \cdots \wedge \phi^*df_p \\ &= d\phi^*f_0 \wedge \cdots \wedge d\phi^*f_p \\ &= d(\phi^*f_0 \wedge \phi^*f_1 \wedge \cdots \wedge \phi^*f_p) \\ &= d(\phi^*f_0 \wedge \phi^*df_1 \wedge \cdots \wedge \phi^*df_p) \\ &= d(\phi^*(f_0 \wedge df_1 \wedge \cdots \wedge df_p)) \\ &= d(\phi^*\omega) \end{aligned}$$

as desired.



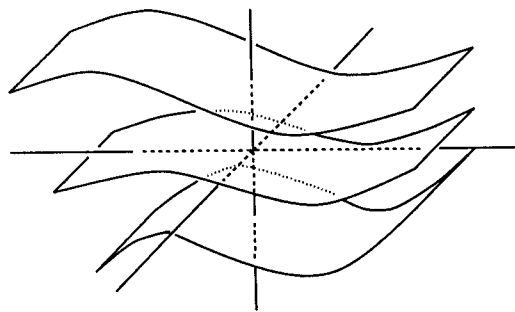


Fig. 1. Splitting spacetime into space and time

One advantage of the differential form language is its generality. We can take our spacetime to be any manifold  $M$ , of any dimension, and define the electromagnetic field to be a 2-form  $F$  on  $M$ . The first pair of Maxwell equations says just that

$$dF = 0.$$

Sometimes — but not always — we can split spacetime up into space and time, that is, write  $M$  as  $\mathbb{R} \times S$  for some manifold  $S$  we call ‘space’. If so, we can write  $t$  for the usual coordinate on  $\mathbb{R}$ , and split  $F$  into an electric and magnetic field:

**Exercise 50.** Show that any 2-form  $F$  on  $\mathbb{R} \times S$  can be uniquely expressed as  $B + E \wedge dt$  in such a way that for any local coordinates  $x^i$  on  $S$  we have  $E = E_i dx^i$  and  $B = \frac{1}{2} B_{ij} dx^i \wedge dx^j$

We can also split the exterior derivative into spacelike and timelike parts as before:

**Exercise 51.** Show that for any form  $\omega$  on  $\mathbb{R} \times S$  there is a unique way to write  $d\omega = dt \wedge \partial_t \omega + d_S \omega$  such that for any local coordinates  $x^i$  on  $S$ , writing  $t = x^0$ , we have

$$\begin{aligned} d_S \omega &= \partial_i \omega_I dx^i \wedge dx^I, \\ dt \wedge \partial_t \omega &= \partial_0 \omega_I dx^0 \wedge dx^I. \end{aligned}$$

When we split spacetime up into space and time,  $dF = 0$  becomes equivalent to the pair of equations

$$d_S B = 0, \quad \partial_t B + d_S E = 0.$$

In the static case, when  $\partial_t E = \partial_t B = 0$ , we can forget about the  $t$  coordinate entirely and treat  $E$  and  $B$  as forms on space satisfying the static equations

$$d_S B = 0, \quad d_S E = 0.$$

Note that the electric and magnetic fields are only defined after we choose a way of splitting spacetime into space and time! If someone hands us a manifold  $M$ , it may be diffeomorphic to  $\mathbb{R} \times S$  in many different ways, or in no way at all. In special relativity one learns that different inertial frames (corresponding to observers moving at constant velocity) will give different splittings of spacetime into  $\mathbb{R} \times \mathbb{R}^3$ , which are related by Lorentz transformations. This means that the electric and magnetic fields will get mixed up when we do a Lorentz transformation, as described in Chapter 1. More drastically, we could split spacetime into space and time in a wiggly way as in Figure 1 above. This may seem perverse, but there is usually no ‘best’ way to split spacetime into space and time, particularly in the context of general relativity.

## The Metric

*In the Space and Time marriage we have the greatest Boy meets Girl story of the age. To our great-grandchildren this will be as poetical as the ancient Greek marriage of Cupid and Psyche seems to us. — Lawrence Durrell, Balthazar*

The first pair of Maxwell equations does not involve measuring distances in spacetime. That is why they are ‘generally covariant’, i.e., one can pull back a solution by any diffeomorphism, no matter how much it stretches or distorts spacetime, and get another solution. This is not the case for the second pair, which require for their formulation a way of measuring distances and times. The key idea of relativity is that distances and time intervals are two aspects of a single concept, the ‘spacetime interval’. Mathematically, spacetime intervals are calculated using a ‘metric’ on spacetime.

In ordinary Euclidean  $\mathbb{R}^3$  we measure distances and angles using

$$v \cdot w = v^1 w^1 + v^2 w^2 + v^3 w^3,$$

$$\|v\|^2 = v \cdot v.$$

$$v \cdot w = -v^0 w^0 + v^1 w^1 + v^2 w^2 + v^3 w^3.$$

$$g: V \times V \rightarrow \mathbb{R},$$

$$\begin{aligned} g(cv + v', w) &= cg(v, w) + g(v', w) \\ g(v, cw + w') &= g(v, w) + cg(v, w'), \end{aligned}$$

$$g(v, w) = g(w, v),$$

and **nondegenerate**: if  $g(v, w) = 0$  for all  $w \in V$ , then  $v = 0$ . We say that  $v \in V$  is spacelike, timelike or null depending on whether  $g(v, v)$  is positive, negative or zero. If  $g(v, w) = 0$ , we say that  $v$  and  $w$  are orthogonal. Note that null vectors are orthogonal to themselves!

Given a metric on  $V$ , we can always find an **orthonormal basis** for  $V$ , that is, a basis  $\{e_\mu\}$  such that  $g(e_\mu, e_\nu)$  is 0 if  $\mu \neq \nu$ , and  $\pm 1$  if  $\mu = \nu$ . The number of  $+1$ 's and  $-1$ 's is independent of the orthonormal basis, and if the number of  $+1$ 's is  $p$  and the number of  $-1$ 's is  $q$ , we say the metric has **signature**  $(p, q)$ . For example, Minkowski spacetime has signature  $(3, 1)$ , with the **Minkowski metric** given by

$$\eta(v, w) = -v^0 w^0 + v^1 w^1 + v^2 w^2 + v^3 w^3.$$

So far we have been talking about spacetimes that are vector spaces. Now let  $M$  be a manifold and consider a situation where the metric depends on where one is. A **metric**  $g$  on  $M$  assigns to each point  $p \in M$  a metric  $g_p$  on the tangent space  $T_p M$ , in a smoothly varying way. By 'smoothly varying' we mean that if  $v$  and  $w$  are smooth vector fields on  $M$ , the inner product  $g_p(v_p, w_p)$  is a smooth function on  $M$ . By the way, we usually write this function simply as  $g(v, w)$ .

One can show that the smoothness condition implies that the signature of  $g_p$  is constant on any connected component of  $M$ . We are really only interested in cases where the signature is constant on all of  $M$ . If the signature of  $g$  is  $(n, 0)$ , where  $\dim M = n$ , we say that  $g$  is a **Riemannian** metric, while if the signature is  $(n - 1, 1)$ , we say that  $g$  is **Lorentzian**. By a **semi-Riemannian manifold** we mean a manifold equipped with a metric, and similarly for a **Riemannian manifold** and a **Lorentzian manifold**.

In relativity, spacetime is a Lorentzian manifold, which in the real world appears to be 4-dimensional, although other cases are certainly interesting. The easiest way to get ahold of a 4-dimensional Lorentzian manifold is to take a 3-dimensional manifold  $S$ , 'space', with a Riemannian metric  ${}^3g$ , and let  $M$ , 'spacetime', be given by  $\mathbb{R} \times S$ . Then we can define a Lorentzian metric

$$g = -dt^2 + {}^3g$$

on  $M$  as follows. Let  $x^i$  ( $i = 1, 2, 3$ ) be local coordinates on an open subset  $U \subseteq S$ , and let  $t$  or  $x^0$  denote the coordinate on  $\mathbb{R}$ , that is,

'time'. Then  $x^\mu$  ( $\mu = 0, 1, 2, 3$ ) are local coordinates on  $\mathbb{R} \times U \subseteq M$ , and we can define the metric  $g$  to be that with components

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & & & \\ 0 & & {}^3g_{ij} & \\ 0 & & & \end{pmatrix}$$

This represents a special sort of **static** spacetime, in which space has a metric that is independent of time.

The most basic use of a Lorentzian metric is to measure distances and times. For example, if a path  $\gamma: [0, 1] \rightarrow M$  is spacelike, that is, if its tangent vector is everywhere spacelike, we define its **arclength** to be

$$\int_0^1 \sqrt{g(\gamma'(t), \gamma'(t))} dt.$$

If  $\gamma$  is timelike, we define the **proper time** along  $\gamma$  — that is, the time ticked off by a clock moving along  $\gamma$  — to be

$$\int_0^1 \sqrt{-g(\gamma'(t), \gamma'(t))} dt.$$

We will mainly be interested in some more sophisticated applications of the metric, however. The most fundamental of these is 'raising and lowering indices', that is, converting between tangent and cotangent vectors. If  $V$  is a vector space equipped with a metric  $g$ , there is a natural way to turn an element  $v \in V$  into an element of  $V^*$ , namely the linear functional  $g(v, \cdot)$  which eats another element of  $V$  and spits out a number.

**Exercise 52.** Use the nondegeneracy of the metric to show that the map from  $V$  to  $V^*$  given by

$$v \mapsto g(v, \cdot)$$

is an isomorphism, that is, one-to-one and onto.

It follows that if  $M$  is a semi-Riemannian manifold the metric defines an isomorphism between each tangent space  $T_p M$  and the corresponding cotangent space  $T_p^* M$ . We can picture this as follows: if the tangent vector  $v$  is a little arrow, the cotangent vector  $\omega = g(v, \cdot)$  is

a stack of hyperplanes perpendicular to  $v$ , as in Figure 2. The reason for this is that  $\omega$  vanishes on vectors orthogonal to  $v$ . The key point is that one needs the metric to know what 'orthogonal' means!

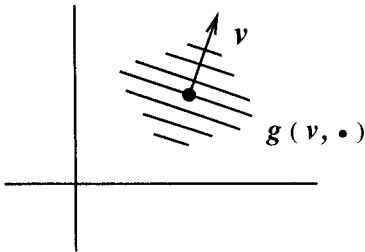


Fig. 2. Tangent vector  $v$  and cotangent vector  $g(v, \cdot)$

Similarly, we can convert between vector fields and 1-forms on  $M$ . By using the metric on space, for example, we can think of the electric field as a vector field instead of a 1-form. We need to do this in order to think of the electric field as 'pointing' in some direction.

Suppose  $M$  is a semi-Riemannian manifold. Now that we can visualize 1-forms on  $M$  as fields of little arrows, there is a nice way for us to visualize  $p$ -forms for higher  $p$  as well. We can draw a wedge product  $\omega \wedge \mu$  of two cotangent vectors at  $p$  as a little parallelogram, as in Figure 3. So we can visualize a 2-form on  $M$  as field of such 'area elements'. Similarly, we can draw a wedge product  $\omega \wedge \mu \wedge \nu$  of three cotangent vectors at  $p$  as a little parallelepiped, as in Figure 4, and visualize a 3-form as a field of these 'volume elements' — and so on for higher  $p$ -forms.

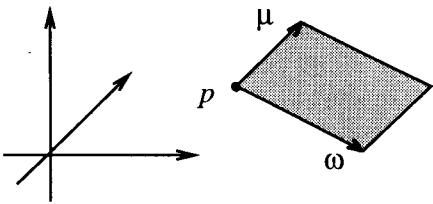


Fig. 3. Picture of  $\omega \wedge \mu \in \Lambda^2 T_p^* M$

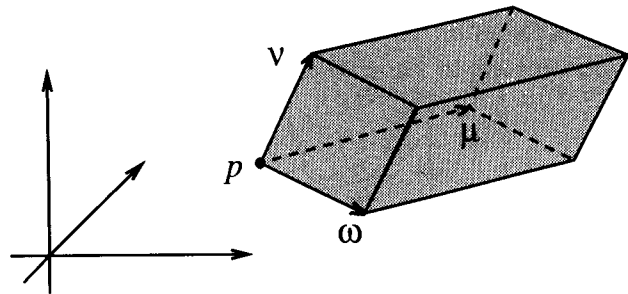


Fig. 4. Picture of  $\omega \wedge \mu \wedge \nu \in \Lambda^3 T_p^* M$

If  $M$  is  $n$ -dimensional,  $g_{\mu\nu}$  is an  $n \times n$  matrix. The nondegeneracy condition implies this matrix is invertible, so let  $g^{\mu\nu}$  denote the inverse matrix. Then we have the following handy formulas, which explain why the process of converting between vector fields and 1-forms using the metric is called **raising and lowering indices**:

**Exercise 53.** Let  $v = v^\mu e_\mu$  be a vector field on a chart. Show that the corresponding 1-form  $g(v, \cdot)$  is equal to  $v_\nu f^\nu$ , where  $f^\nu$  is the dual basis of 1-forms and

$$v_\nu = g_{\mu\nu} v^\mu.$$

**Exercise 54.** Let  $\omega = \omega_\mu f^\mu$  be a 1-form on a chart. Show that the corresponding vector field is equal to  $\omega^\nu e_\nu$ , where

$$\omega^\nu = g^{\mu\nu} \omega_\mu.$$

**Exercise 55.** Let  $\eta$  be the Minkowski metric on  $\mathbb{R}^4$  as defined above. Show that its components in the standard basis are

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

In general, if we have any quantity with some indices, such as

$$A^{\alpha\beta\cdots\gamma}{}_{\delta\epsilon\cdots\zeta},$$

we can lower or raise any index with the metric and its inverse, using the Einstein summation convention. E.g., we can lower  $\alpha$  and get

$$A_\alpha{}^{\beta\cdots\gamma}{}_{\delta\epsilon\cdots\zeta} = g_{\alpha\mu} A^{\mu\beta\cdots\gamma}{}_{\delta\epsilon\cdots\zeta},$$

or raise  $\delta$  and get

$$A^{\alpha\beta\cdots\gamma\delta}{}_{\epsilon\cdots\zeta} = g^{\delta\mu} A^{\alpha\beta\cdots\gamma}{}_{\mu\epsilon\cdots\zeta}.$$

If we have a lot indices floating around it is important to keep track of their order when we raise and lower them; otherwise things get confusing. Note that we can even raise and lower indices on the metric itself:

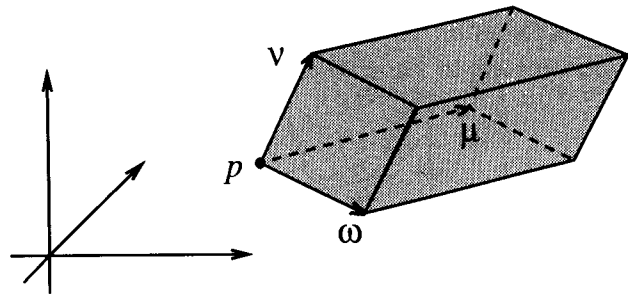


Fig. 4. Picture of  $\omega \wedge \mu \wedge \nu \in \Lambda^3 T_p^* M$

If  $M$  is  $n$ -dimensional,  $g_{\mu\nu}$  is an  $n \times n$  matrix. The nondegeneracy condition implies this matrix is invertible, so let  $g^{\mu\nu}$  denote the inverse matrix. Then we have the following handy formulas, which explain why the process of converting between vector fields and 1-forms using the metric is called **raising and lowering indices**:

**Exercise 53.** Let  $v = v^\mu e_\mu$  be a vector field on a chart. Show that the corresponding 1-form  $g(v, \cdot)$  is equal to  $v_\nu f^\nu$ , where  $f^\nu$  is the dual basis of 1-forms and

$$v_\nu = g_{\mu\nu} v^\mu.$$

**Exercise 54.** Let  $\omega = \omega_\mu f^\mu$  be a 1-form on a chart. Show that the corresponding vector field is equal to  $\omega^\nu e_\nu$ , where

$$\omega^\nu = g^{\mu\nu} \omega_\mu.$$

**Exercise 55.** Let  $\eta$  be the Minkowski metric on  $\mathbb{R}^4$  as defined above. Show that its components in the standard basis are

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

In general, if we have any quantity with some indices, such as

$$A^{\alpha\beta\cdots\gamma}{}_{\delta\epsilon\cdots\zeta},$$

we can lower or raise any index with the metric and its inverse, using the Einstein summation convention. E.g., we can lower  $\alpha$  and get

$$A_\alpha{}^{\beta\cdots\gamma}{}_{\delta\epsilon\cdots\zeta} = g_{\alpha\mu} A^{\mu\beta\cdots\gamma}{}_{\delta\epsilon\cdots\zeta},$$

or raise  $\delta$  and get

$$A^{\alpha\beta\cdots\gamma\delta}{}_{\epsilon\cdots\zeta} = g^{\delta\mu} A^{\alpha\beta\cdots\gamma}{}_{\mu\epsilon\cdots\zeta}.$$

If we have a lot indices floating around it is important to keep track of their order when we raise and lower them; otherwise things get confusing. Note that we can even raise and lower indices on the metric itself:

$$g_{\mu\nu} = g(e_\mu, e_\nu).$$



## The Volume Form

Since a metric allows us to measure distances on a manifold, it should allow us to measure volumes as well, and thus allow us to do integrals. This is in fact the case. We will postpone the study of integration on manifolds to Chapter 6, but we will define a basic ingredient of it here, the 'volume form'. This concept is needed to write down Maxwell's equations in differential form language. It turns out that a closely related concept is that of an 'orientation', that is, a globally well-defined way to tell the difference between left and right.

Given an  $n$ -dimensional vector space  $V$  with two bases  $\{e_\mu\}$ ,  $\{f_\mu\}$ , there is always a unique linear transformation  $T: V \rightarrow V$  taking one basis to the other:

$$Te_\mu = f_\mu.$$

This is necessarily invertible, so its determinant is nonzero. Let us say that  $\{e_\mu\}$  and  $\{f_\mu\}$  have the same orientation if  $\det T > 0$ , and the opposite orientation if  $\det T < 0$ . For example, any right-handed basis in  $\mathbb{R}^3$  has the same orientation as the usual right-handed basis  $(e_1, e_2, e_3)$ :

$$e_1 = (1, 0, 0), \quad e_2 = (0, 1, 0), \quad e_3 = (0, 0, 1),$$

while any left-handed basis, like  $(-e_1, -e_2, -e_3)$ , has the opposite orientation.

**Exercise 60.** *Show that any even permutation of a given basis has the same orientation, while any odd permutation has the opposite orientation.*

Let us define an **orientation** on  $V$  to be a choice of an equivalence class of bases of  $V$ , where two bases are deemed equivalent if they have the same orientation. E.g., on  $\mathbb{R}^3$  there is the right-handed orientation, which contains the basis  $(e_1, e_2, e_3)$  and all other bases with the same orientation, and the left-handed orientation. There are always only two orientations on  $V$ .

There is another way to think about orientations. Suppose  $V$  is an  $n$ -dimensional vector space with basis  $\{e_\mu\}$ . Then

$$e_1 \wedge \cdots \wedge e_n$$

is a nonzero element of  $\Lambda^n V$  which we call the **volume element** associated to the basis  $\{e_\mu\}$ . We can picture it as a little parallelepiped in  $n$  dimensions.

Let us see how the volume element depends on a change of basis. Note that any element  $\omega \in \Lambda^n V$  can be written as

$$ce_1 \wedge \cdots \wedge e_n,$$

for some constant  $c$ , since a wedge product that contains any  $e_\mu$  twice automatically vanishes. Suppose  $\{f_\nu\}$  is another basis of  $V$  and let  $T_\mu^\nu$  be the matrix with

$$f_\nu = T_\nu^\mu e_\mu.$$

Then

$$\begin{aligned} f_1 \wedge \cdots \wedge f_n &= (T_1^1 e_1 + \cdots + T_1^n e_n) \wedge \cdots \wedge (T_n^1 e_1 + \cdots + T_n^n e_n) \\ &= (\det T) e_1 \wedge \cdots \wedge e_n \end{aligned}$$

since in the first line one is really summing over all expressions of the form

$$\text{sign}(\sigma) T_1^{\sigma(1)} \cdots T_n^{\sigma(n)} e_1 \wedge \cdots \wedge e_n$$

where  $\sigma$  is a permutation and  $\text{sign}(\sigma)$  is its sign, which comes in from the anticommutativity of the wedge product. Thus two bases have the same orientation if the corresponding volume elements differ by a *positive* scalar multiple. Or, if we like, we can think of an orientation as being a choice of a volume form modulo positive scalar multiples.

Now let us turn from vector spaces to manifolds in general. As usual, let  $M$  be an  $n$ -dimensional manifold. We define a **volume form**  $\omega$  on  $M$  to be a nowhere vanishing  $n$ -form. Thus for each point  $p \in M$ ,  $\omega_p$  is a volume element on  $T_p^* M$ . The standard volume form on  $\mathbb{R}^n$  is

$$\omega = dx^1 \wedge \cdots \wedge dx^n.$$

As we will see, when we do a multiple integral like

$$\int_{\mathbb{R}^3} f \, dx \, dy \, dz$$

we are really integrating the 3-form  $f \, dx \wedge dy \wedge dz$ .



We say  $M$  is **orientable** if there exists a volume form on  $M$ . By an **orientation** on  $M$  we mean a choice of an equivalence class of volume forms on  $M$ , where two volume forms  $\omega$  and  $\omega'$  are equivalent if  $\omega' = f\omega$  for some positive function  $f$ . Any volume form that is in the chosen equivalence class is said to be **positively oriented**; otherwise it is said to be **negatively oriented**. In particular, the **standard orientation** on  $\mathbb{R}^n$  is the equivalence class containing the volume form  $dx^1 \wedge \cdots \wedge dx^n$ .

If we have an orientation on  $M$ , we can decide unambiguously whether any basis  $e^\mu$  of a cotangent space  $T_p^*M$  is **right-handed** or **left-handed**, as follows. Just pick a volume form  $\omega$  in the equivalence class, write  $e^1 \wedge \cdots \wedge e^n$  as a constant times  $\omega$ , and check to see whether the constant is positive or negative. This is the precise sense in which an orientation gives a global definition of right vs. left. Since a basis of the tangent space gives a dual basis of the cotangent space, we can also define right-handed and left-handed bases of the tangent space.

The classic example of a nonorientable manifold is the Möbius strip:

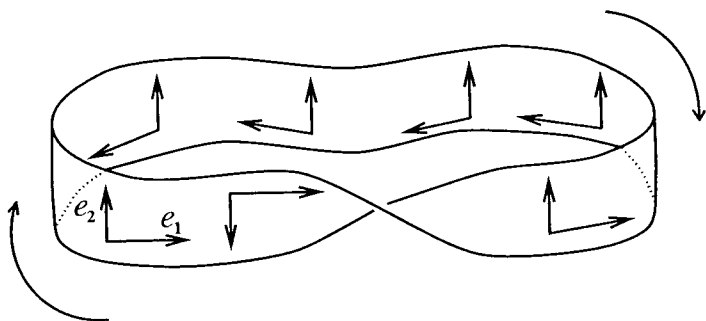


Fig. 5. The Möbius strip is nonorientable

As the figure indicates, there is no way to define the notion of a right-handed basis of  $T_p^*M$  for the Möbius strip in a smoothly varying way. Using a Riemannian metric we can identify  $T_p^*M$  with the tangent space  $T_pM$ . We have drawn a 'right-handed' basis of  $T_pM$  at one point, and show how if one drags it smoothly around a noncontractible loop it become 'left-handed'. If space was nonorientable, we might take a long journey in a spaceship around a noncontractible loop and come back home as a mirror-image version of ourselves. (However, we would not feel reflected; we would think everything *else* had been reflected.)

A manifold equipped with an orientation is said to be **oriented**. One can also think of an oriented manifold as one having 'oriented charts' as follows:

**Exercise 61.** Let  $M$  be an oriented manifold. Show that we can cover  $M$  with oriented charts  $\varphi_\alpha: U_\alpha \rightarrow \mathbb{R}^n$ , that is, charts such that the basis  $dx^\mu$  of cotangent vectors on  $\mathbb{R}^n$ , pulled back to  $U_\alpha$  by  $\varphi_\alpha$ , is positively oriented.

**Exercise 62.** Given a diffeomorphism  $\phi: M \rightarrow N$  from one oriented manifold to another, we say that  $\phi$  is **orientation-preserving** if the pullback of any right-handed basis of a cotangent space in  $N$  is a right-handed basis of a cotangent space in  $M$ . Show that if we can cover  $M$  with charts such that the transition functions  $\varphi_\alpha \circ \varphi_\beta^{-1}$  are orientation-preserving, we can make  $M$  into an oriented manifold by using the charts to transfer the standard orientation on  $\mathbb{R}^n$  to an orientation on  $M$ .

Now suppose that  $M$  is an oriented  $n$ -dimensional manifold with metric  $g$ . There is a canonical volume form on  $M$  which we can construct as follows. First, cover  $M$  with oriented charts  $\varphi_\alpha: U_\alpha \rightarrow \mathbb{R}^n$ . In any chart set

$$g_{\mu\nu} = g(\partial_\mu, \partial_\nu),$$

and define

$$\text{vol} = \sqrt{|\det g_{\mu\nu}|} dx^1 \wedge \cdots \wedge dx^n.$$

Clearly this is a volume form on  $U$ . What we need to show is that given any overlapping chart  $\varphi': U' \rightarrow \mathbb{R}^n$ , and defining

$$g'_{\mu\nu} = g(\partial'_\mu, \partial'_\nu),$$

then the volume form

$$\text{vol}' = \sqrt{|\det g'_{\mu\nu}|} dx'^1 \wedge \cdots \wedge dx'^n,$$

agrees with  $\text{vol}$  on the overlap  $U \cap U'$ . This will imply the existence of a volume form on all of  $M$ , defined by this sort of formula, and independent of choice of chart.

On the overlap we have

$$dx'^\nu = T^\nu_\mu dx^\mu$$

where the matrix-valued function  $T$  is given by

$$T_{\mu}^{\nu} = \frac{\partial x'^{\nu}}{\partial x^{\mu}}.$$

Thus we have

$$dx'^1 \wedge \cdots \wedge dx'^n = (\det T) dx^1 \wedge \cdots \wedge dx^n,$$

so to show  $\text{vol} = \text{vol}'$  we need to show

$$\sqrt{|\det g'_{\mu\nu}|} = (\det T)^{-1} \sqrt{|\det g_{\mu\nu}|}.$$

To see this, note that

$$\begin{aligned} g'_{\mu\nu} &= g(\partial'_{\mu}, \partial'_{\nu}) \\ &= g\left(\frac{\partial x^{\alpha}}{\partial x'^{\mu}} \partial_{\alpha}, \frac{\partial x^{\beta}}{\partial x'^{\nu}} \partial_{\beta}\right) \\ &= (T^{-1})_{\mu}^{\alpha} (T^{-1})_{\nu}^{\beta} g_{\alpha\beta} \end{aligned}$$

or, taking determinants,

$$\det g'_{\mu\nu} = (\det T)^{-2} \det g_{\mu\nu}.$$

Since both charts are oriented,  $\det T > 0$ , so

$$\sqrt{|\det g'_{\mu\nu}|} = (\det T)^{-1} \sqrt{|\det g_{\mu\nu}|}$$

as desired.

We call  $\text{vol}$  the volume form on  $M$  associated to the metric  $g$ . People often write the volume form as  $\sqrt{|\det g|} d^n x$ . In the Lorentzian case, this is just

$$\text{vol} = \sqrt{-\det g} d^n x,$$

since the determinant of  $g_{\mu\nu}$  is negative. In general relativity, people often write the volume form as simply  $\sqrt{-g} d^n x$ , using  $g$  to stand for the determinant of  $g_{\mu\nu}$ .

In Chapter 6 we will describe integration theory on an oriented manifold, and show how to integrate functions on an oriented semi-Riemannian manifold  $M$ . The basic idea is that when we integrate a function  $f$  over  $M$ , we are really doing the integral

$$\int_M f \text{ vol},$$

that is, integrating the  $n$ -form  $f \text{ vol}$ . Right now our main goal is to describe the second pair of Maxwell equations in differential form language. For this, we need the volume form to define something called the Hodge star operator. The following fact will come in handy:

**Exercise 63.** Let  $M$  be an oriented  $n$ -dimensional semi-Riemannian manifold and let  $\{e_{\mu}\}$  be an oriented orthonormal basis of cotangent vectors at some point  $p \in M$ . Show that

$$e_1 \wedge \cdots \wedge e_n = \text{vol}_p,$$

where  $\text{vol}$  is the volume form associated to the metric on  $M$ , and  $\text{vol}_p$  is its value at  $p$ .

## The Hodge Star Operator

The Hodge star operator is the key to understanding the 'duality' symmetry of the vacuum Maxwell equations, as described in Chapter 1. This symmetry is the reason why the second pair of Maxwell equations look similar (but not quite the same) as the first pair. Think about these equations in ordinary Minkowski space. In old-fashioned notation, they are:

$$\begin{aligned} \nabla \cdot \vec{B} &= 0 \\ \nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} &= 0 \\ \nabla \cdot \vec{E} &= \rho \\ \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} &= \vec{j}. \end{aligned}$$

In differential form notation, the first pair becomes:

$$\begin{aligned} d_S B &= 0 \\ \partial_t B + d_S E &= 0, \end{aligned}$$

where  $B$  is a 2-form on space and  $E$  is a 1-form on space (both functions of time). The funny thing is that the second pair seems to have the roles of  $E$  and  $B$  reversed (modulo the minus sign). This would amount

to treating  $E$  as a 2-form and  $B$  as a 1-form! The Hodge star operator saves the day, since it converts 1-forms on 3-dimensional space into 2-forms, and vice versa. However, it does so at a price: it requires a choice of metric and also a choice of orientation.

How does the Hodge star operator do this? Here is where our way of drawing differential forms comes in handy. At any point  $p$  in a 3-dimensional Riemannian manifold  $M$ , the Hodge star operator maps a 1-form  $\nu$ , which we draw as a little arrow, into a 2-form  $\omega \wedge \mu$  that corresponds to an area element that is orthogonal to  $\nu$ , as follows:

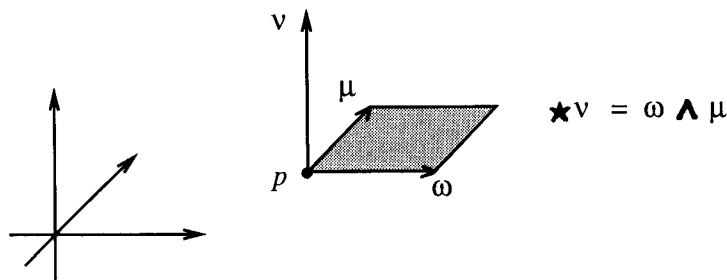


Fig. 6. The Hodge star of  $\nu$  is  $\omega \wedge \mu$

Conversely, it maps  $\omega \wedge \mu$  to  $\nu$ . In general, in  $n$  dimensions the Hodge star operator maps  $p$ -forms to  $(n-p)$ -forms in a very similar way, taking each little ' $p$ -dimensional area element' to an orthogonal ' $(n-p)$ -dimensional area element'.

The precise definition of the Hodge star operator uses the inner product of differential forms. Let  $M$  be an  $n$ -dimensional oriented semi-Riemannian manifold. Then the inner product of two  $p$  forms  $\omega$  and  $\mu$  on  $M$  is a function  $\langle \omega, \mu \rangle$  on  $M$ . We define the **Hodge star operator**

$$\star: \Omega^p(M) \rightarrow \Omega^{n-p}(M)$$

to be the unique linear map from  $p$ -forms to  $(n-p)$ -forms such that for all  $\omega, \mu \in \Omega^p(M)$ ,

$$\omega \wedge \star \mu = \langle \omega, \mu \rangle \text{vol}$$

Note that both sides of the equation are  $n$ -forms. We often call  $\star \mu$  the **dual** of  $\mu$ .

It might not be obvious from this definition that the Hodge star operator really exists, or how to compute it! For this, it is nice to have a formula for it. Suppose that  $e^1, \dots, e^n$  are a positively oriented orthonormal basis of 1-forms on some chart. Thus

$$\langle e^\mu, e^\nu \rangle = 0$$

if  $\mu \neq \nu$ , and

$$\langle e^\mu, e^\mu \rangle = \epsilon(\mu)$$

where  $\epsilon(\mu) = \pm 1$ . Then we claim that for any distinct  $1 \leq i_1, \dots, i_p \leq n$ ,

$$\star(e^{i_1} \wedge \dots \wedge e^{i_p}) = \pm e^{i_{p+1}} \wedge \dots \wedge e^{i_n}$$

where  $\{i_{p+1}, \dots, i_n\}$  consists of the integers from 1 to  $n$  not included in  $\{i_1, \dots, i_p\}$ :

$$\{i_{p+1}, \dots, i_n\} = \{1, \dots, n\} - \{i_1, \dots, i_p\}.$$

The sign  $\pm$  is given by

$$\text{sign}(i_1, \dots, i_n) \epsilon(i_1) \dots \epsilon(i_p),$$

where  $\text{sign}(i_1, \dots, i_n)$  denotes the sign of the permutation taking  $(1, \dots, n)$  to  $(i_1, \dots, i_n)$ .

**Exercise 64.** Show that if we define the Hodge star operator in a chart using this formula, it satisfies the property  $\omega \wedge \star \mu = \langle \omega, \mu \rangle \text{vol}$ . Use the result from Exercise 63.

The formula for the Hodge star operator might seem complicated, so consider an example. Take  $dx, dy, dz$  as a basis of 1-forms on  $\mathbb{R}^3$  with its usual Euclidean metric and orientation. Then we have

$$\star dx = dy \wedge dz, \quad \star dy = dz \wedge dx, \quad \star dz = dx \wedge dy,$$

and conversely

$$\star dx \wedge dy = dz, \quad \star dy \wedge dz = dx, \quad \star dz \wedge dx = dy.$$

If one interprets the definition correctly, one also can work out what the Hodge star operator does to the 0-form (or function) 1 and the volume form  $dx \wedge dy \wedge dz$ :

$$\star 1 = dx \wedge dy \wedge dz, \quad \star dx \wedge dy \wedge dz = 1.$$

Since Hodge star operator on  $\mathbb{R}^3$  lets us turn 1-forms into 2-forms and vice versa, it sheds some new light on familiar operations like the cross product, curl and divergence. Given two 1-forms  $\omega$  and  $\mu$  on  $\mathbb{R}^3$ , their wedge product is a 2-form, and is perfectly well-defined without reference to a metric and orientation. But if we allow ourselves to use a metric and orientation, we can take the Hodge star of  $\omega \wedge \mu$  and obtain a 1-form! If

$$\omega = \omega_i dx^i, \quad \mu = \mu_i dx^i,$$

then using the standard metric and orientation we get

$$\star(\omega \wedge \mu) = (\omega_y \mu_z - \omega_z \mu_y) dx + (\omega_z \mu_x - \omega_x \mu_z) dy + (\omega_x \mu_y - \omega_y \mu_x) dz.$$

This is basically just the cross product! The reader may wonder why we have done all this work to get back to concepts that everyone knows from basic vector calculus. Part of the point is that we now can work in spacetimes of arbitrary dimension, with arbitrary metrics and orientations. But it is also nice to see just where the metric and orientation are needed in the definition of the cross product in  $\mathbb{R}^3$ : only when we want to take a 2-form and convert it into a 1-form are they necessary.

Moreover, if  $\omega$  is a 1-form on  $\mathbb{R}^3$ ,  $d\omega$  is a 2-form, but  $\star d\omega$  is a 1-form again, and if we use the standard metric and orientation this is basically just the curl of  $\omega$ :

**Exercise 65.** Calculate  $\star d\omega$  when  $\omega$  is a 1-form on  $\mathbb{R}^3$ .

Similarly, if  $\omega$  is a 1-form on  $\mathbb{R}^3$ ,  $d\star\omega$  is a 3-form, but  $\star d\star\omega$  is a 0-form, or function, and this basically amounts to taking the divergence of  $\omega$ :

**Exercise 66.** Calculate  $\star d\star\omega$  when  $\omega$  is a 1-form on  $\mathbb{R}^3$ .

We encourage the reader to do the following exercises, too:

**Exercise 67.** Give  $\mathbb{R}^4$  the Minkowski metric and the orientation in which  $(dt, dx, dy, dz)$  is positively oriented. Calculate the Hodge star operator on all wedge products of  $dx^\mu$ 's. Show that on  $p$ -forms

$$\star^2 = (-1)^{p(4-p)+1}.$$

**Exercise 68.** Let  $M$  be an oriented semi-Riemannian manifold of dimension  $n$  and signature  $(s, n-s)$ . Show that on  $p$ -forms

$$\star^2 = (-1)^{p(n-p)+s}.$$

**Exercise 69.** Let  $M$  be an oriented semi-Riemannian manifold of dimension  $n$  and signature  $(s, n-s)$ . Let  $e^\mu$  be an orthonormal basis of 1-forms on some chart. Define the Levi-Civita symbol for  $1 \leq i_j \leq n$  by

$$\epsilon_{i_1 \dots i_n} = \begin{cases} \text{sign}(i_1, \dots, i_n) & \text{all } i_j \text{ distinct} \\ 0 & \text{otherwise} \end{cases}$$

Show that for any  $p$ -form

$$\omega = \frac{1}{p!} \omega_{i_1 \dots i_p} e^{i_1} \wedge \dots \wedge e^{i_p}$$

we have

$$(\star\omega)_{j_1 \dots j_{n-p}} = \frac{1}{p!} \epsilon^{i_1 \dots i_p}_{j_1 \dots j_{n-p}} \omega_{i_1 \dots i_p}.$$

## The Second Pair of Equations

We now use the Hodge star operator to write the second pair of Maxwell equations in terms of differential forms. The key thing to understand is the effect of taking the dual  $\star F$  of the electromagnetic field  $F$ .

First consider the case where  $M$  is Minkowski spacetime with its usual coordinates  $x^\mu$ . We will sometimes write  $t$  for the time coordinate  $x^0$ . Then we can split  $F$  into electric and magnetic fields,

$$F = B + E \wedge dt,$$

where  $B$  is a time-dependent 2-form on space and  $E$  is a time-dependent 1-form on space. If one likes components, we have  $F = \frac{1}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu$

where

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix}$$

Now introduce the Minkowski metric on spacetime:

$$\eta(v, w) = -v^0 w^0 + v^1 w^1 + v^2 w^2 + v^3 w^3.$$

This allows us to define the Hodge star operator. A little calculation using Exercise 67 shows that

$$(\star F)_{\mu\nu} = \begin{pmatrix} 0 & B_x & B_y & B_z \\ -B_x & 0 & E_z & -E_y \\ -B_y & -E_z & 0 & E_x \\ -B_z & E_y & -E_x & 0 \end{pmatrix}.$$

In other words, taking the dual of  $F$  amounts to doing the replacements

$$E_i \mapsto -B_i, \quad B_i \mapsto E_i.$$

This is the main difference between the first pair of Maxwell equations — which in old-fashioned form are

$$\nabla \cdot \vec{B} = 0 \quad \nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = 0,$$

and the second pair:

$$\nabla \cdot \vec{E} = \rho \quad \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = \vec{j}.$$

The other difference between the first and second pairs is that the latter contain  $\rho$  and  $\vec{j}$ . To speak of these in the language of differential forms, we use the fact that the metric allows us to turn vector fields into 1-forms. Thus we can turn the good old current density

$$\vec{j} = j^1 \partial_1 + j^2 \partial_2 + j^3 \partial_3$$

into the 1-form

$$j = j_1 dx^1 + j_2 dx^2 + j_3 dx^3.$$

Similarly, we can combine the current density and the electric charge density  $\rho$  in a single vector field on Minkowski spacetime:

$$\vec{J} = \rho \partial_0 + j^1 \partial_1 + j^2 \partial_2 + j^3 \partial_3,$$

and by using the Minkowski metric, we can turn this vector field into a 1-form

$$J = j - \rho dt$$

which we call the **current**.

Now we claim that just as the first pair of Maxwell equations are really

$$dF = 0,$$

the second pair are really

$$\star d \star F = J.$$

This is not so surprising, because at least on Minkowski space, the second pair of Maxwell equations

$$\nabla \cdot \vec{E} = \rho, \quad \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = \vec{j}$$

can be rewritten as

$$\begin{aligned} \star_S d_S \star_S E &= \rho, \\ -\partial_t E + \star_S d_S \star_S B &= j, \end{aligned}$$

where  $\star_S$  denotes the Hodge star operator on 'space', that is,  $\mathbb{R}^3$  with its usual Euclidean metric.

**Exercise 70.** Check this result.

These look very similar to the version of the first pair of Maxwell equations in which we have split spacetime into space and time:

$$\begin{aligned} d_S B &= 0, \\ \partial_t B + d_S E &= 0. \end{aligned}$$

The difference really amounts to using the Hodge star operator twice.

More generally, start by assuming that spacetime  $M$  is any manifold. Then the electromagnetic field  $F$  is a 2-form on  $M$  the current  $J$  is a 1-form on  $M$ , and the first Maxwell equation is  $dF = 0$ . We must assume  $M$  is semi-Riemannian and oriented to write down the second pair of Maxwell's equations, that is,  $\star d \star F = J$ . To introduce electric and magnetic fields we must assume  $M = \mathbb{R} \times S$ , where  $S$  is space, and write  $F = B + E \wedge dt$ . Similarly we write  $J = j - \rho dt$ . Then the first Maxwell equation splits into

$$d_S B = 0, \quad \partial_t B + d_S E = 0.$$

Suppose also that space is 3-dimensional and that the metric on  $M$  is a static one of the form  $g = -dt^2 + {}^3g$  where  ${}^3g$  is a Riemannian metric on space,  $S$ . Let  $\star_S$  denote the Hodge star operator on (time-dependent) differential forms on  $S$ . Then

$$\star F = \star_S E - \star_S B \wedge dt$$

so

$$d \star F = \star_S \partial_t E \wedge dt + d_S \star_S E - d_S \star_S B \wedge dt$$

and

$$\star d \star F = -\partial_t E - \star_S d_S \star_S E \wedge dt + \star_S d_S \star_S B.$$

Setting  $\star d \star F = J$  and equating like terms, we obtain

$$\star_S d_S \star_S E = \rho, \quad -\partial_t E + \star_S d_S \star_S B = j,$$

as desired.

**Exercise 71.** Check the calculations above.

It is interesting to note that in the **static Maxwell equations**, where  $E$  and  $B$  are independent of  $t$ , there is a pair involving only  $E$ :

$$dE = 0, \quad \star_S d_S \star_S E = \rho,$$

and a pair involving only  $B$ :

$$dB = 0, \quad \star_S d_S \star_S B = 0.$$

This makes it clear that only when the electric and magnetic fields are time-dependent do they affect each other. Historically, it was Faraday who first discovered in 1831 that a changing magnetic field causes a nonzero curl in the electric field. He is responsible for the

$$\frac{\partial \vec{B}}{\partial t}$$

term in the equations of electromagnetism. Maxwell's brilliant contribution to the equations came when he hypothesized in 1861 that a changing electric field causes a nonzero curl in the magnetic field. In other words, he guessed there should be a

$$\frac{\partial \vec{E}}{\partial t}$$

term, too. It is only when both of these effects are taken into account that we get electromagnetic radiation, in which ripples in  $E$  cause ripples in  $B$  and vice versa, causing waves that move through space.

Interestingly enough, the reason Maxwell made his hypothesis was not an experiment, but a problem with the equations of electromagnetism as they stood at the time. This was the problem of charge conservation. Not only is the total electric charge of the world constant, the only way charge can get from one place to another is by moving through the intervening regions. This is called a 'local conservation law'. Mathematically, one can formulate it in Minkowski spacetime by saying that any increase or decrease in the charge density at any point is solely due to the divergence of the current density. In old-fashioned language one expresses this by the **continuity equation**

$$\frac{d\rho}{dt} = -\nabla \cdot \vec{j}.$$

Maxwell realized that the  $\partial \vec{E} / \partial t$  term would make the continuity equation an automatic consequence of the laws of electromagnetism! This can be seen by starting with

$$\nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = \vec{j},$$

taking the divergence of both sides to obtain

$$-\nabla \cdot \frac{\partial \vec{E}}{\partial t} = \nabla \cdot \vec{j},$$

and then interchanging the order of the derivatives on the left hand side and using the fact that

$$\nabla \cdot \vec{E} = \rho.$$

In fact, the continuity equation can be expressed more elegantly in differential form language as

$$d \star J = 0,$$

and this law is a simple consequence of Maxwell's equations in their most general modern form. Starting with  $\star d \star F = J$  and taking the dual of both sides we obtain  $d \star F = \pm \star J$ , where the sign depends on the value of  $\star^2$  on 1-forms (see Exercise 68). Taking the exterior derivative of both sides and using  $d^2 = 0$ , we get  $d \star J = 0$ . In terms of components, this equation is written  $\partial^\mu J_\mu = 0$ .

This is a good example of how the identity  $d^2 = 0$  has powerful physical consequences. When we get to gauge theories we will see that Maxwell's equations are a special case of the Yang-Mills equations, which describe not only electromagnetism but also the strong and weak nuclear forces. A generalization of the identity  $d^2 = 0$ , the Bianchi identity, implies conservation of 'charge' in all of these theories — although these theories have different kinds of 'charge'. Similarly, we will see when we get to general relativity that due to the Bianchi identity, Einstein's equations for gravity automatically imply local conservation of energy and momentum! So what we are seeing here is only the tip of the iceberg.

It is also interesting to consider the **vacuum Maxwell equations**, that is, the case  $J = 0$ :

$$dF = 0, \quad d \star F = 0.$$

These are preserved by duality:

$$F \mapsto \star F.$$

Recall that when spacetime  $M$  is of the form  $\mathbb{R} \times S$ , so that  $F = B + E \wedge dt$ , we have  $\star F = \star_S E - \star_S B \wedge dt$ , so duality amounts to:

$$B \mapsto \star_S E, \quad E \mapsto -\star_S B,$$

or when  $S = \mathbb{R}^3$ ,

$$\vec{B} \mapsto \vec{E}, \quad \vec{E} \mapsto -\vec{B}$$

in old-fashioned language.

In 4 dimensions something very interesting happens, since then the dual of a 2-form is a 2-form. Note from Exercise 67 that if  $M$  is a Lorentzian 4-dimensional manifold, the operator

$$\star: \Omega^2(M) \rightarrow \Omega^2(M)$$

has

$$\star^2 = -1,$$

while if  $M$  is Riemannian, we have

$$\star^2 = 1.$$

In the Riemannian case things are very nice: we say  $F \in \Omega^2(M)$  is **self-dual** if  $\star F = F$ , and **anti-self-dual** if  $\star F = -F$ . Since  $\star^2 = 1$ , it is not surprising that the Hodge star operator has eigenvalues  $\pm 1$ . That is, we can write any  $F \in \Omega^2(M)$  as a sum of self-dual and anti-self-dual parts:

$$F = F_+ + F_-, \quad \star F_\pm = \pm F_\pm.$$

**Exercise 72.** Show this is true if we take

$$F_\pm = \frac{1}{2}(F \pm \star F).$$

In the Lorentzian case things are not quite as nice, since  $\star^2 = -1$  implies its eigenvalues are  $\pm i$ . This means that we should really consider complex-valued differential forms on  $M$ . If we do that, we can write any  $F \in \Omega^2(M)$  as

$$F = F_+ + F_-$$

where

$$\star F_\pm = \pm i F_\pm.$$

**Exercise 73.** Show that this result is true.

Let us bend words a bit and say in this case too that  $F_+$  is self-dual and  $F_-$  is anti-self-dual.

In either the Riemannian or Lorentzian case, if we have a self-dual (or anti-self-dual) 2-form  $F$  satisfying the first pair of vacuum Maxwell equations:

$$dF = 0,$$

it automatically satisfies the second pair:

$$d \star F = 0.$$

Of course, in the Lorentzian case  $F$  will need to be complex-valued, which is not very sensible physically. However, since Maxwell's equations are linear, we can always take the real part (or imaginary part) of a solution and get a real-valued solution.

The trick of turning two pairs of vacuum Maxwell equations into one turns out to be the tip of another iceberg. First, the Hodge star operator and the exterior derivative interact with each other in a very nice way that has a lot to do with topology. This leads to a subject called Hodge theory. Self-duality is also important in the Yang-Mills equations. These are a lot harder to solve than Maxwell's equations, because they are nonlinear, but using self-duality one can find some solutions in the Riemannian case. These self-dual (or anti-self-dual) solutions are called 'instantons', because they start out small near  $t = -\infty$ , get big for a little while, and then get small again near  $t = +\infty$ . Instantons are of importance both in the physics of the strong force and in studying the topology of 4-dimensional manifolds.

Self-duality also turns out to be important for the Einstein equations. This was emphasized by Penrose, who used a method called 'twistors' to find self-dual solutions to the Einstein equations. Self-duality of a somewhat different sort is also crucial in Ashtekar's reformulation of general relativity, which we discuss in Chapter 5 of Part III.

We can get a bit of the flavor of this business by using self-duality to find some solutions of the vacuum Maxwell's equations on Minkowski space. These solutions represent light moving around through empty

space! If we write

$$F = B + E \wedge dt$$

we have

$$\star F = \star_S E - \star_S B \wedge dt,$$

so  $F$  will be self-dual if

$$\star_S E = iB, \quad \star_S B = -iE.$$

**Exercise 74.** Show that these equations are equivalent, and both hold if at every time  $t$  we have

$$E = E_1 dx^1 + E_2 dx^2 + E_3 dx^3,$$

$$B = -i(E_1 dx^2 \wedge dx^3 + \text{cyclic permutations}).$$

Let us assume  $F$  is self-dual and that  $E$  is a **plane wave**, that is, of the form

$$E(x) = \mathbf{E} e^{ik_\mu x^\mu}$$

where  $\mathbf{E} = E_j dx^j$  is a constant complex-valued 1-form on  $\mathbb{R}^3$  and  $k \in (\mathbb{R}^4)^*$  is a fixed covector, called the **energy-momentum**. Recall that the covector  $k$  eats the vector  $x \in \mathbb{R}^4$  corresponding to a point in Minkowski space and spits out a number  $k(x)$  in a linear way: in coordinates this is just

$$k(x) = k_\mu x^\mu.$$

By self-duality, we have

$$B(x) = \mathbf{B} e^{ik_\mu x^\mu}$$

where  $\mathbf{B} = -i \star_S \mathbf{E}$ . Thus the first Maxwell equation,  $d_S B = 0$ , implies that

$$\mathbf{B} \wedge d_S e^{ik_\mu x^\mu} = 0$$

at all points  $x$ . Let us write  ${}^3k$  for  $k_j dx^j$ , the spatial part of the energy-momentum, called the **momentum** of the plane wave. Then

$$d_S e^{ik_\mu x^\mu} = e^{ik_\mu x^\mu} {}^3k,$$

so the first Maxwell equation holds precisely when

$$\mathbf{B} \wedge {}^3k = 0.$$



Expressing  $\mathbf{B}$  in terms of  $\mathbf{E}$ , this equation is equivalent to

$$\star_S \mathbf{E} \wedge {}^3k = 0,$$

or, by the definition of the Hodge star operator,

$$\langle \mathbf{E}, {}^3k \rangle = 0.$$

This says that the electric field must be orthogonal to the momentum of the plane wave.

Similarly, the second Maxwell equation,  $\partial_t B + d_S E = 0$ , says that

$${}^3k \wedge \mathbf{E} = k_0 \mathbf{B}.$$

**Exercise 75.** *Check the above result.*

This equation is really just a fancy way of saying that the cross product of the electric field and the momentum is proportional to the magnetic field. The number  $k_0$  is called the **frequency** of the plane wave. Writing  $\mathbf{B}$  in terms of  $\mathbf{E}$ , we obtain an equation  $\mathbf{E}$  must satisfy:

$${}^3k \wedge \mathbf{E} = -ik_0 \star_S \mathbf{E}.$$

**Exercise 76.** *Show this equation implies  $k_\mu k^\mu = 0$ . Thus the energy-momentum of light is light-like!*

If we solve the first pair of vacuum Maxwell's equations this way, duality automatically implies we have solved the second pair. A simple example of a solution is

$$k = dt - dx, \quad \mathbf{E} = dy - idz.$$

Note that  ${}^3k$  and  $\mathbf{E}$  are really orthogonal, and also

$${}^3k \wedge \mathbf{E} = -dx \wedge dy + idx \wedge dz = -ik_0 \star_S \mathbf{E},$$

as required.

It is enlightening to express this solution in old-fashioned language. It gives:

$$\vec{E} = (0, e^{i(t-x)}, -ie^{i(t-x)}), \quad \vec{B} = (0, -ie^{i(t-x)}, -e^{i(t-x)}).$$

**Exercise 77.** *Check the above result.*

Of course, to get an honest, *real* solution of Maxwell's equations we can take the real part:

$$\vec{E} = (0, \cos(t-x), \sin(t-x)), \quad \vec{B} = (0, \sin(t-x), -\cos(t-x)).$$

In other words, the plane wave moves in the  $x$  direction at the speed of light, with the electric and magnetic fields orthogonal to each other rotating counterclockwise in the  $yz$  plane. A plane wave in which  $\vec{E}$  and  $\vec{B}$  rotate counterclockwise when viewed as the wave moves towards one is said to be **left circularly polarized**. As it turns out, all the self-dual plane wave solutions of Maxwell's equations are left circularly polarized. To get right circularly polarized plane waves, we need the anti-self-dual plane wave solutions. General plane wave solutions will be linear combinations of self-dual and anti-self-dual ones.

One thing we see here is a close connection between the Hodge star operator and chirality, or handedness. In a more sophisticated quantum-field theoretic picture of light, we may think of it as made of photons that spin either clockwise or counterclockwise about their axis of motion. Light has no preferred chirality. However, a different sort of massless particle, the neutrino, *does* have a preferred chirality — one of the puzzles of nature.

**Exercise 78.** *Prove that all self-dual and anti-self-dual plane wave solutions are left and right circularly polarized, respectively.*

**Exercise 79.** *Let  $P: \mathbb{R}^4 \rightarrow \mathbb{R}^4$  be parity transformation, that is,*

$$P(t, x, y, z) = (t, -x, -y, -z).$$

*Show that if  $F$  is a self-dual solution of Maxwell's equations, the pullback  $P^*F$  is an anti-self-dual solution, and vice versa.*

## Chapter 6

# DeRham Theory in Electromagnetism

*I received your paper, and thank you very much for it. I do not say I venture to thank you for what you have said about "Lines of Force", because I know you have done it for the interests of philosophical truth; but you must suppose it is work grateful to me, and gives me much encouragement to think on. I was at first almost frightened when I saw such mathematical force made to bear upon the subject, and then wondered to see that the subject stood it so well. — Michael Faraday, to James Clerk Maxwell*

### Closed and Exact 1-forms

As we have seen, the first pair of Maxwell equations simply say that electromagnetic field  $F$  has  $dF = 0$ . In the static case, they say that the electric field has  $dE = 0$  and the magnetic field  $B$  has  $dB = 0$ . Equations of this sort are especially charming because they are 'generally covariant', that is, independent of any fixed choice of metric or other geometrical structure on spacetime. This implies that they are preserved by any diffeomorphism. In other words, if  $\omega$  is a form on a manifold  $M$  satisfying the equation  $d\omega = 0$ , the pullback of  $\omega$  under any diffeomorphism of  $M$  again satisfies this equation. Since a diffeomorphism is a kind of change of coordinates, this means that the first pair of Maxwell equations is invariant, not just under Lorentz transformations, rotations, and translations, but under *all* coordinate transformations.

Now let us try to solve these equations. It is easy to come up with lots of solutions, because  $d^2 = 0$ . If  $F$  is  $d$  of something, it automatically satisfies  $dF = 0$ , and similarly for  $E$  and  $B$  in the static case. This simple observation is the basis of a surprisingly large amount of mathematics and physics. It leads to a very interesting question: can one get *all* the solutions of the first pair of Maxwell equations this way? The branch of mathematics that answers this sort of question is called deRham cohomology.

Let us first introduce some standard terminology. In general, if the exterior derivative of a differential form is zero, we say the differential form is **closed**. On the other hand, a differential form that is the exterior derivative of some other differential form is called **exact**. The equation  $d^2 = 0$  may thus be expressed in words by saying 'all exact forms are closed'. For example, if the electric field  $E$  is  $d$  of some function on space, we will automatically have  $dE = 0$ . In physics one calls a function (or 0-form)  $\phi$  with

$$E = -d\phi$$

a **scalar potential** for  $E$ ; the minus sign is just a convention. Similarly, if the magnetic field  $B$  is  $d$  of some 1-form on space, we will automatically have  $dB = 0$ . One calls a 1-form  $A$  with

$$B = dA$$

a **vector potential** for  $B$ . Also, if the electromagnetic field  $F$  satisfies

$$F = dA$$

for some 1-form  $A$  on spacetime, we automatically have  $dF = 0$ , and we call  $A$  a vector potential for  $F$ .

Now let us study when a closed 1-form is exact. Say we have a manifold  $S$ , with a 1-form  $E$  on it satisfying  $dE = 0$ . Can we cook up a function  $\phi$  on  $S$  with  $E = -d\phi$ ? Let us try and see what, if anything, prevents us. We will attempt to find such a function  $\phi$  by integrating the 1-form  $E$  along paths in  $S$ . Technically, a **path**  $\gamma$  in  $S$  is a piecewise smooth map from  $\gamma: [0, T] \rightarrow S$ , but in this section we will be lazy and only work with smooth paths. If  $\gamma$  is a path,  $\gamma'(t)$  is a tangent vector

at the point  $\gamma(t)$ , and applying the cotangent vector  $E_{\gamma(t)}$  at the same point we get a number; then we integrate this from 0 to  $T$ . We write this as

$$\int_{\gamma} E = \int_0^T E_{\gamma(t)}(\gamma'(t)) dt.$$

Our plan will be to define  $\phi$  as follows: fix any point  $p \in S$  and for any  $q \in S$  let

$$\phi(q) = - \int_{\gamma} E$$

where  $\gamma$  is some path from  $p$  to  $q$ . The reader may be familiar with this strategy in the special case when  $S = \mathbb{R}^3$ ; this is how one writes a curl-free vector field as the gradient of a function.

There are a number of potential problems with this plan. First, there might not be any path from  $p$  to  $q$ ! It is rather odd to imagine in terms of physics, but mathematically there is nothing to stop  $S$  from being made of several pieces, or 'components', with no paths from one to another. For example,  $S$  might be the disjoint union of two copies of  $\mathbb{R}^3$  — two separate universes, as it were — and there would be no path from one to the other. We will have to rule out this case. If there is a path between any two points in  $S$ , we say that  $S$  is **connected** (or more precisely, **arc-connected**). If not, a maximal connected subset of  $S$  is called a **connected component**. Henceforth in our quest to solve  $dE = 0$  we will assume  $S$  is connected. (If not, it would be easy to apply our technique to each connected component separately.)

The next problem, which is more serious, is that the integral  $\int_{\gamma} E$  will in general depend on the details of the path  $\gamma$ , not just its endpoints  $\gamma(0) = p$  and  $\gamma(T) = q$ . We want to see what conditions are necessary to rule out this problem. First, let us see how the integral changes when we smoothly vary the path  $\gamma$ . In other words, suppose that we have a smoothly varying family of paths from  $p$  to  $q$  labelled by some parameter  $s \in [0, T]$ . We can describe all these by a function  $\gamma(s, t)$ . For each  $s$ ,  $\gamma(s, \cdot)$  is a smooth path with  $\gamma(s, 0) = p$  and  $\gamma(s, T) = q$ , and  $\gamma(s, t)$  should depend smoothly on  $s$  as well as  $t$ .

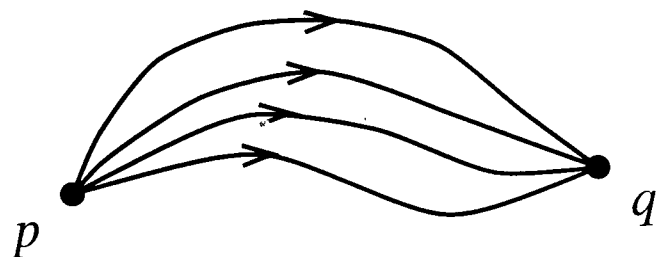


Fig. 1. A smoothly varying family of paths from  $p$  to  $q$

To see how

$$I_s = \int_0^T E_{\gamma(s,t)}(\gamma'(s,t)) dt$$

depends on  $s$ , let us differentiate it with respect to  $s$ . To do computations we can assume we are working in a coordinate chart on  $S$  – if not, break up the integral into pieces that each fit in a chart. Using coordinates to describe the pairing of the 1-form  $E$  and the tangent vector  $\gamma'$ , we have

$$I_s = \int_0^T E_\mu(\gamma(s,t)) \partial_t \gamma^\mu(s,t) dt,$$

Thus

$$\begin{aligned} \partial_s I_s &= \int \partial_s [E_\mu(\gamma(s,t)) \partial_t \gamma^\mu(s,t)] dt \\ &= \int [\partial_s E_\mu(\gamma(s,t)) \partial_t \gamma^\mu(s,t) + E_\mu(\gamma(s,t)) \partial_s \partial_t \gamma^\mu(s,t)] dt \\ &= \int [\partial_s E_\mu(\gamma(s,t)) \partial_t \gamma^\mu(s,t) - \partial_t E_\mu(\gamma(s,t)) \partial_s \gamma^\mu(s,t)] dt \\ &= \int \partial_\nu E_\mu(\gamma(s,t)) [\partial_s \gamma^\nu \partial_t \gamma^\mu - \partial_t \gamma^\nu \partial_s \gamma^\mu] dt \end{aligned}$$

using the product rule, then integration by parts, and then the chain rule. Recalling that

$$dE = (\partial_\mu E_\nu - \partial_\nu E_\mu) dx^\mu dx^\nu,$$

we obtain

$$\partial_s I_s = \int (dE)_{\mu\nu} \partial_s \gamma^\mu \partial_t \gamma^\nu dt.$$

Thus  $I_s$  is independent of  $s$  when  $dE = 0$ . This shows that  $I_s$  will be the same for two different paths as long as we can find a smoothly varying family of paths interpolating between them.

In math jargon, we say two paths  $\gamma_0, \gamma_1: [0, T] \rightarrow S$  from  $p$  to  $q$  are **homotopic** if there exists a smooth function  $\gamma: [0, 1] \times [0, T] \rightarrow S$  such that  $\gamma(s, \cdot)$  is a path from  $p$  to  $q$  for each  $s$ , and

$$\gamma(0, t) = \gamma_0(t), \quad \gamma(1, t) = \gamma_1(t).$$

We call the function  $\gamma$  a **homotopy** between  $\gamma_0$  and  $\gamma_1$ . In this terminology, what we have shown is that a closed 1-form has the same integral along any two homotopic paths.

There still may be a problem with defining

$$\phi(q) = - \int_\gamma E$$

where  $\gamma$  is any path from  $p$  to  $q$ . Perhaps not all paths from  $p$  to  $q$  are homotopic! A nice example is the plane with the origin removed: this is a manifold, and the two paths from  $(-1, 0)$  to  $(1, 0)$  shown below are not homotopic:

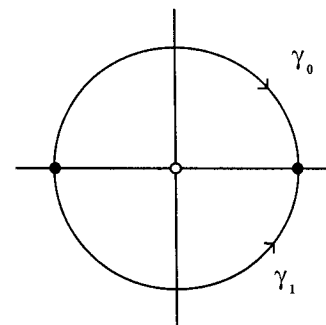


Fig. 2. Two paths that are not homotopic in  $\mathbb{R}^2 - \{0\}$

It is pretty obvious that there is no way to smoothly deform the path  $\gamma_0$  to the path  $\gamma_1$  without getting snagged on the hole at the origin. Of course, being 'obvious' does not count as a proof! However, we can really *prove* this fact by finding a closed 1-form that has different integrals along the two paths. It is not hard: try

$$E = \frac{xdy - ydx}{x^2 + y^2}.$$

This 1-form ‘wraps around the hole’, so it has different integrals along  $\gamma_0$  and  $\gamma_1$ :

**Exercise 80.** Show that this 1-form  $E$  is closed. Show that  $\int_{\gamma_0} E = -\pi$  and  $\int_{\gamma_1} E = \pi$ .

This means that we cannot use  $\phi(q) = -\int_{\gamma} E$  to define  $\phi$  in a path-independent manner. We can visualize how  $E$  wraps around the hole if we draw  $E$  in the manner described in Chapter 4:

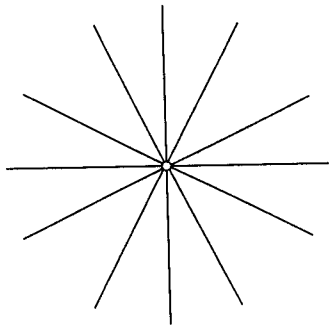


Fig. 3. Picture of  $E = (xdy - ydx)/(x^2 + y^2)$

The fact that  $E$  is not exact simply means that there is no function whose level curves are the lines in the figure. If there were such a function, say  $-\phi$ , we would have  $E = -d\phi$ .

Given a connected manifold  $S$ , we say that  $S$  is **simply connected** if any two paths between two points  $p, q$  are homotopic. If  $S$  is simply connected, we can carry out our plan and define  $\phi$  unambiguously when  $E$  is closed. In particular, things are fine on  $\mathbb{R}^n$ :

**Exercise 81.** Show that  $\mathbb{R}^n$  is simply connected by exhibiting an explicit formula for a homotopy between any two paths between arbitrary points  $p, q \in \mathbb{R}^n$ .

Now let us show that when  $S$  is simply connected our plan really succeeds! Namely, suppose that  $S$  is simply connected and  $E$  is a closed 1-form on  $S$ . Pick any point  $p \in S$  and define a function  $\phi$  on  $S$  by

$$\phi(q) = - \int_{\gamma} E$$

where  $\gamma$  is any path from  $p$  to  $q \in S$ . Let us show that

$$E = -d\phi.$$

To show that these 1-forms agree at some point  $q$ , it suffices to show that they agree when applied to any tangent vector  $v \in T_q S$ . By the definition of  $d\phi$ , this means we need to show

$$E(v) = -v(\phi).$$

To do this, pick a path  $\gamma: [0, 2] \rightarrow S$  with  $\gamma(0) = p$  and  $\gamma(1) = q$ , and such that  $\gamma'(1) = v$ , as shown below. Then we have

$$\begin{aligned} E(v) &= E(\gamma'(1)) \\ &= \frac{d}{ds} \int_0^s E(\gamma(t)) dt \Big|_{s=1} \\ &= - \frac{d}{ds} \phi(\gamma(s)) \Big|_{s=1} \\ &= -v(\phi) \end{aligned}$$

using the fact that the derivative of  $\phi(\gamma(s))$  with respect to  $s$  is the same as the derivative of  $\phi$  in the direction  $\gamma'(s) = v$ .



Fig. 4. Proof that  $E = -d\phi$

To summarize, we have shown that on a simply connected manifold, every closed 1-form is exact. In this case, we can always find a scalar potential for the electric field. Later, we will show how to generalize this result to  $p$ -forms for higher  $p$ . For 2-forms, this will let us understand when we can find a vector potential for the magnetic field or electromagnetic field.

Let us finish this section with a few words about loops! A path  $\gamma: [0, T] \rightarrow S$  is a **loop** if it ends where it starts, that is, if  $\gamma(0) = \gamma(T) = p$  for some point  $p \in S$ . We also say then that  $\gamma$  is a loop **based at**  $p$ , or that  $p$  is the **basepoint** of  $\gamma$ . Loops play a special role in electromagnetism, gauge theory and in the new approach to quantum gravity known as the 'loop representation', for which this book is intended as preparation. The basic idea is that we can understand fields in a very natural way by imagining a particle that goes around a loop and is altered somehow in the process. For example, we will explain later in this chapter how when we move a charged particle around a loop in space, its wavefunction is multiplied by a number  $e^{i\theta}$ , where  $\theta$  is proportional to the integral of the vector potential around the loop! A similar fact holds for loops in spacetime, with the electromagnetic field  $F$  taking the place of the magnetic field. And a grand generalization of this fact holds for all the forces in the standard model — this is why we say they are all 'gauge fields'. Gravity is similar but in a sense even simpler: gravity is just a manifestation of the curvature of spacetime, where by 'curvature' we refer to the fact that if we take an object and move it around a loop, trying our best to 'parallel transport' it, nonetheless it comes back *rotated*.

We conclude this section by describing the role loops play in electrostatics. Let us suppose, as above, that space is some manifold  $S$  and the electric field on  $S$  is a 1-form  $E$  on  $S$ . Consider the integral of  $E$  around a loop  $\gamma$ , i.e.  $\int_{\gamma} E$ . If we wish to emphasize that  $\gamma$  is a loop we can write this as

$$\oint_{\gamma} E.$$

In certain important cases this will be zero! We say that a loop  $\gamma: [0, T] \rightarrow S$  based at  $p$  is **contractible** if it is homotopic to a constant loop  $\eta$  that just stays at  $p$ :

$$\eta(t) = p$$

for all  $t \in [0, T]$ . Below we show a contractible loop  $\gamma$  and a noncontractible loop  $\delta$  in  $\mathbb{R}^2 - \{0\}$ .

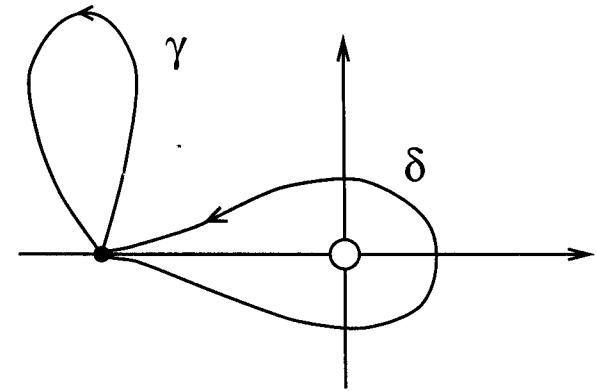


Fig. 5. A contractible loop  $\gamma$  and a noncontractible loop  $\delta$

By the result established earlier, if  $dE = 0$  then we must have  $\int_{\gamma} E = 0$  if  $\gamma$  is contractible, since the integral of  $E$  around a constant loop is zero. In particular, if  $S$  is simply connected,  $\int_{\gamma} E = 0$  for *all* loops if  $dE = 0$ . This is definitely *not* true when  $S$  is not simply connected; for example, our friend the 1-form

$$E = \frac{xdy - ydx}{x^2 + y^2}$$

on  $\mathbb{R}^2 - \{0\}$  gives an integral of  $2\pi$  around the loop  $\delta$  shown above. More generally, it gives  $2\pi$  times the **winding number** of the loop, that is, the number of times the loop goes around the origin, counted with a plus sign when it goes around counterclockwise, and with a minus sign when it goes around clockwise.

There is a converse, too, that allows us to rephrase the electrostatic equation  $dE = 0$  purely in terms of integrals around loops. This converse is a consequence of Stokes' theorem relating the curl of a vector field to its integral around a loop bounding a surface. Let us pick a chart giving coordinates  $x^{\mu}$  about some point  $p \in S$ , and consider the integral of  $E$  around a square loop  $\gamma$  in the  $x^{\mu}$ - $x^{\nu}$  plane:

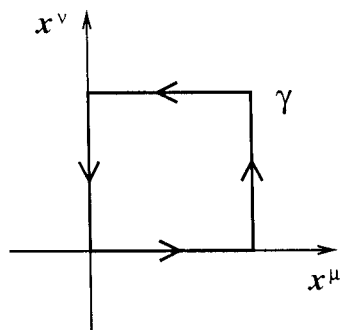


Fig. 6. The integral of  $E$  around a small square

Suppose this square is given by

$$\{0 \leq x^\mu \leq \epsilon, 0 \leq x^\nu \leq \epsilon\}.$$

Then by Green's theorem,

$$\int_\gamma E = \int_0^\epsilon \int_0^\epsilon (\partial_\mu E_\nu - \partial_\nu E_\mu) dx^\mu dx^\nu$$

and in the limit as  $\epsilon \rightarrow 0$  this is equal to

$$\epsilon^2 (\partial_\mu E_\nu - \partial_\nu E_\mu) = \epsilon^2 (dE)_{\mu\nu}$$

evaluated at  $p$ , plus terms of order  $\epsilon^3$ . So if  $\int_\gamma E$  vanishes for all contractible loops in  $S$ , then  $dE = 0$ .

In short, a 1-form  $E$  is closed if and only if  $\int_\gamma E = 0$  for all contractible loops  $\gamma$ . Similarly, it follows from things we have already shown that  $E$  is exact if and only if  $\int_\gamma E = 0$  for *all* loops. In the next sections we will generalize this result to  $p$ -forms. For this, we will need to generalize Stokes' theorem.

**Exercise 82.** Show that a 1-form  $E$  is exact if and only if  $\int_\gamma E = 0$  for all loops  $\gamma$ . (Hint: if  $\omega$  is not exact, show that there are two smooth paths  $\gamma, \gamma'$  from some point  $x \in M$  to some point  $y \in M$  such that  $\int_\gamma \omega \neq \int_{\gamma'} \omega$ . Use these paths to form a loop, perhaps only piecewise smooth.)

**Exercise 83.** For any manifold  $M$ , show the manifold  $S^1 \times M$  is not simply connected by finding a 1-form on it that is closed but not exact.

## Stokes' Theorem

*The objects which we shall study are called exterior differential forms. These are the things which occur under integral signs. — Harley Flanders*

We have been so busy showing what differential forms have to do with Maxwell's equations that we have neglected to properly emphasize that differential forms are just things that one integrates! This is a terrible omission, which we now correct. We will see that  $n$ -forms can be integrated over  $n$ -manifolds, or more generally  $n$ -manifolds with a 'boundary', and that the concepts of exterior derivative and boundary are tied together by the modern version of Stokes' theorem.

The modern version of Stokes' theorem is beautiful because it shows that a number of important theorems of calculus are really all aspects of the same thing. Let us give rough statements of these to point out how similar they are. First, there is the fundamental theorem of calculus. This says that if one has a function  $f: [a, b] \rightarrow \mathbb{R}$ , then

$$\int_a^b f'(x) dx = f(b) - f(a).$$

It relates the integral of the derivative of  $f$  over the closed interval  $[a, b]$  to the values of  $f$  on the 'boundary', that is, the endpoints. Second, there is the good old version of Stokes' theorem. This says that if one has a 2-dimensional surface  $S$  in  $\mathbb{R}^3$  whose boundary  $\partial S$  is traced out by a loop  $\gamma: [0, T] \rightarrow \mathbb{R}^3$ , and  $\vec{A}$  is a vector field on  $\mathbb{R}^3$ , then

$$\int_S (\nabla \times \vec{A}) \cdot \vec{n} = \int_\gamma \vec{A},$$

where  $\vec{n}$  is the unit normal to  $S$ . Again, this relates the integral of the derivative of  $\vec{A}$  over  $S$  to the integral of  $\vec{A}$  over the boundary  $\partial S$ . Third, there is Gauss' theorem. This says that if one has a 3-dimensional region  $R \subset \mathbb{R}^3$  with smooth boundary  $\partial R$ , and  $\vec{A}$  is a vector field defined on  $R$ , then

$$\int_R \nabla \cdot \vec{A} = \int_{\partial R} \vec{A} \cdot \vec{n}$$

where  $\vec{n}$  is the outwards-pointing unit normal to  $\partial R$ . This too relates the integral of the derivative of  $\vec{A}$  over  $R$  to the integral of  $\vec{A}$  over the

boundary  $\partial R$ . In physics, we call  $\int_{\partial R} \vec{A} \cdot \vec{n}$  the **flux** of  $\vec{A}$  through the surface  $\partial R$ .

Now that we know about differential forms, it is clear that in the fundamental theorem of calculus we are starting with a function, or 0-form,  $f$ , forming the 1-form  $df = f'(x)dx$ , and integrating it over a closed interval. A closed interval is not quite a manifold, since the two endpoints do not have neighborhood that looks like  $\mathbb{R}$ , but we will see that it is a 1-dimensional 'manifold with boundary'. We have also seen that the curl really amounts to  $d$  of a 1-form. Thus in Stokes' theorem we are really taking  $d$  of a 1-form, obtaining a 2-form, and integrating it over a 2-dimensional manifold with boundary,  $S$ . We have also seen that the divergence in  $\mathbb{R}^3$  is really  $d$  of a 2-form. So in Gauss' theorem we are really taking  $d$  of a 2-form, obtaining a 3-form, and integrating it over a 3-dimensional manifold with boundary,  $R$ .

Roughly speaking, the general Stokes' theorem says that under certain conditions, if  $M$  is a  $n + 1$ -dimensional manifold with boundary and  $\omega$  is an  $n$ -form on  $M$ , then

$$\int_M d\omega = \int_{\partial M} \omega.$$

We will not prove this theorem, but we will make sense of all the pieces involved. To do this, we will define a manifold with boundary, and then explain how to integrate differential forms over manifolds with boundary. We refer the reader to the notes at the end of Part I for books that prove the theorem — it is not really all that hard!

The concept of a manifold with boundary is a simple generalization of that of an ordinary manifold. A simple example would be the annulus

$$\{(x, y) \in \mathbb{R}^2: 1 \leq x^2 + y^2 \leq 2\}.$$

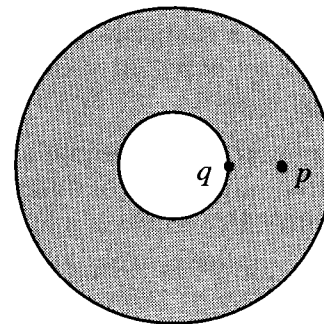


Fig. 7. A manifold with boundary: the annulus

The point  $p = (3/2, 0)$  has a neighborhood that looks just like  $\mathbb{R}^2$ , but the point  $q = (1, 0)$ , which is on the boundary, does not. It does, however, have a neighborhood that looks like the closed half-plane

$$\mathbb{H}^2 = \{(x, y): y \geq 0\}.$$

Thus in a manifold with boundary we want to allow charts that look like the closed half-space

$$\mathbb{H}^n = \{(x^1, \dots, x^n): x^n \geq 0\}.$$

We have to worry a bit about the fact that we have not yet defined what it means for a function on  $\mathbb{H}^n$  to be smooth! We want such functions to be smooth 'up to and including the boundary'. Perhaps the simplest way to say this is that a function on  $\mathbb{H}^n$  is **smooth** if it extends to a smooth function on the manifold

$$\{(x^1, \dots, x^n): x^n > -\epsilon\}$$

for some  $\epsilon > 0$ .

So: we define a  $n$ -dimensional manifold **with boundary** to be a topological space  $M$  equipped with charts of the form  $\varphi_\alpha: U_\alpha \rightarrow \mathbb{R}^n$  or  $\varphi_\alpha: U_\alpha \rightarrow \mathbb{H}^n$ , where  $U_\alpha$  are open sets covering  $M$ , such that the transition function  $\varphi_\alpha \circ \varphi_\beta^{-1}$  is smooth where it is defined. (We also assume some technical conditions, namely that  $M$  is Hausdorff and paracompact. We will have a bit more to say about these in a bit.) Note that a plain old manifold is automatically a manifold with boundary,