THE LARGE SCALE STRUCTURE OF SPACE-TIME

50th Anniversary Edition

STEPHEN W. HAWKING
GEORGE F. R. ELLIS

Foreword by Abhay Ashtekar

CAMBRIDGE MONOGRAPHS ON MATHEMATICAL PHYSICS

THE LARGE SCALE STRUCTURE OF SPACE-TIME

First published in 1973, this influential work discusses Einstein's General Theory of Relativity to show how two of its predictions arise: first, that the ultimate fate of many massive stars is to undergo gravitational collapse to form 'black holes'; and second, that there was a singularity in the past at the beginning of the universe. Starting with a precise formulation of the theory, including the necessary differential geometry, the authors discuss the significance of spacetime curvature and examine the properties of a number of exact solutions of Einstein's field equations. They develop the theory of the causal structure of a general spacetime, and use it to prove a number of theorems establishing the inevitability of singularities under certain conditions. A foreword contributed by Abhay Ashtekar and a new preface from George Ellis help put the volume into context of the developments in the field over the past 50 years.

STEPHEN W. HAWKING (1942–2018) was an English theoretical physicist, cosmologist, and author who was director of research at the Centre for Theoretical Cosmology at the University of Cambridge. He was the Lucasian Professor of Mathematics at Cambridge from 1979 to 2009 and is the author of numerous books, including the international best-seller *A Brief History of Time*.

GEORGE F. R. ELLIS is the emeritus distinguished professor of complex systems in the Department of Mathematics and Applied Mathematics at the University of Cape Town, South Africa. He is considered one of the world's leading theorists in cosmology and, in recent years, he has been prolific in areas relating to the philosophy of science. He is author or co-author of more than a dozen books, including *Relativistic Cosmology* (with Roy Maartens and Malcolm MacCallum).

CAMBRIDGE MONOGRAPHS ON MATHEMATICAL PHYSICS

S. J. Aarseth *Gravitational N-Body Simulations: Tools and Algorithms*†
D. Ahluwalia *Mass Dimension One Fermions*
J. Ambjørn, B. Durhuus and T. Jonsson *Quantum Geometry: A Statistical Field Theory Approach*†
A. M. Anile *Relativistic Fluids and Magneto-fluids: With Applications in Astrophysics and Plasma Physics*
J. A. de Azcárraga and J. M. Izquierdo *Lie Groups, Lie Algebras, Cohomology and Some Applications in Physics*†
O. Babelon, D. Bernard and M. Talon *Introduction to Classical Integrable Systems*†
F. Bajardi and S. Capozziello *Noether Symmetries in Theories of Gravity*
F. Bastianelli and P. van Nieuwenhuizen *Path Integrals and Anomalies in Curved Space*†
D. Baumann and L. McAllister *Inflation and String Theory* †
V. Belinski and M. Henneaux *The Cosmological Singularity*†
V. Belinski and E. Verdaguer *Gravitational Solitons*†
J. Bernstein *Kinetic Theory in the Expanding Universe*†
G. F. Bertsch and R. A. Broglia *Oscillations in Finite Quantum Systems*†
N. D. Birrell and P. C. W. Davies *Quantum Fields in Curved Space*†
K. Bolejko, A. Krasiński, C. Hellaby and M-N. Célérier *Structures in the Universe by Exact Methods: Formation, Evolution, Interactions*
D. M. Brink *Semi-Classical Methods for Nucleus-Nucleus Scattering*†
M. Burgess *Classical Covariant Fields*†
E. A. Calzetta and B.-L. B. Hu *Nonequilibrium Quantum Field Theory*
S. Carlip *Quantum Gravity in 2+1 Dimensions*†
P. Cartier and C. DeWitt-Morette *Functional Integration: Action and Symmetries*†
J. C. Collins *Renormalization: An Introduction to Renormalization, the Renormalization Group and the Operator-Product Expansion*†
P. D. B. Collins *An Introduction to Regge Theory and High Energy Physics*†
M. Creutz *Quarks, Gluons and Lattices*†
P. D. D'Eath *Supersymmetric Quantum Cosmology*†
J. Dereziński and C. Gérard *Mathematics of Quantization and Quantum Fields*
F. de Felice and D. Bini *Classical Measurements in Curved Space-Times*
F. de Felice and C. J. S Clarke *Relativity on Curved Manifolds*†
B. DeWitt *Supermanifolds, 2nd edition*†
P. G. O. Freund *Introduction to Supersymmetry*†
F. G. Friedlander *The Wave Equation on a Curved Space-Time*†
J. L. Friedman and N. Stergioulas *Rotating Relativistic Stars*
Y. Frishman and J. Sonnenschein *Non-Perturbative Field Theory: From Two Dimensional Conformal Field Theory to QCD in Four Dimensions*
J. A. Fuchs *Affine Lie Algebras and Quantum Groups: An Introduction, with Applications in Conformal Field Theory*†
J. Fuchs and C. Schweigert *Symmetries, Lie Algebras and Representations: A Graduate Course for Physicists*†
Y. Fujii and K. Maeda *The Scalar-Tensor Theory of Gravitation*†
J. A. H. Futterman, F. A. Handler, R. A. Matzner *Scattering from Black Holes*†
A. S. Galperin, E. A. Ivanov, V. I. Ogievetsky and E. S. Sokatchev *Harmonic Superspace*†
R. Gambini and J. Pullin *Loops, Knots, Gauge Theories and Quantum Gravity*†
T. Gannon *Moonshine beyond the Monster: The Bridge Connecting Algebra, Modular Forms and Physics*†
A. García-Díaz *Exact Solutions in Three-Dimensional Gravity*
M. Göckeler and T. Schücker *Differential Geometry, Gauge Theories, and Gravity*†
C. Gómez, M. Ruiz-Altaba and G. Sierra *Quantum Groups in Two-Dimensional Physics*†
M. B. Green, J. H. Schwarz and E. Witten *Superstring Theory Volume 1: Introduction*
M. B. Green, J. H. Schwarz and E. Witten *Superstring Theory Volume 2: Loop Amplitudes, Anomalies and Phenomenology*
V. N. Gribov *The Theory of Complex Angular Momenta: Gribov Lectures on Theoretical Physics*†
J. B. Griffiths and J. Podolský *Exact Space-Times in Einstein's General Relativity*†
T. Harko and F. Lobo *Extensions of f(R) Gravity: Curvature-Matter Couplings and Hybrid Metric-Palatini Gravity*
S. W. Hawking and G. F. R. Ellis *The Large Scale Structure of Space-Time*†
J. Harnad and F. Balogh *Tau Functions and their Applications*
B. B. Hu and E. Verdaguer *Semiclassical and Stochastic Gravity*
F. Iachello and A. Arima *The Interacting Boson Model*†
F. Iachello and P. van Isacker *The Interacting Boson-Fermion Model*†
C. Itzykson and J. M. Drouffe *Statistical Field Theory Volume 1: From Brownian Motion to Renormalization and Lattice Gauge Theory*†

C. Itzykson and J. M. Drouffe *Statistical Field Theory Volume 2: Strong Coupling, Monte Carlo Methods, Conformal Field Theory and Random Systems*†
G. Jaroszkiewicz *Principles of Discrete Time Mechanics*
G. Jaroszkiewicz *Quantized Detector Networks*
C. V. Johnson *D-Branes*†
P. S. Joshi *Gravitational Collapse and Spacetime Singularities*†
J. I. Kapusta and C. Gale *Finite-Temperature Field Theory: Principles and Applications, 2nd edition*†
V. E. Korepin, N. M. Bogoliubov and A. G. Izergin *Quantum Inverse Scattering Method and Correlation Functions*†
K. Krasnov *Formulations of General Relativity*
J. Kroon *Conformal Methods in General Relativity*
M. Le Bellac *Thermal Field Theory*†
L. Lusanna *Non-Inertial Frames and Dirac Observables in Relativity*
Y. Makeenko *Methods of Contemporary Gauge Theory*†
S. Mallik and S. Sarkar *Hadrons at Finite Temperature*
A. Malyarenko and M. Ostoja-Starzewski *Tensor-Valued Random Fields for Continuum Physics*
N. Manton and P. Sutcliffe *Topological Solitons*†
N. H. March *Liquid Metals: Concepts and Theory*†
I. Montvay and G. Münster *Quantum Fields on a Lattice*†
P. Nath *Supersymmetry, Supergravity, and Unification*
L. O'Raifeartaigh *Group Structure of Gauge Theories*†
T. Ortín *Gravity and Strings, 2nd edition*
A. M. Ozorio de Almeida *Hamiltonian Systems: Chaos and Quantization*†
M. Paranjape *The Theory and Applications of Instanton Calculations*
L. Parker and D. Toms *Quantum Field Theory in Curved Spacetime: Quantized Fields and Gravity*
R. Penrose and W. Rindler *Spinors and Space-Time Volume 1: Two-Spinor Calculus and Relativistic Fields*†
R. Penrose and W. Rindler *Spinors and Space-Time Volume 2: Spinor and Twistor Methods in Space-Time Geometry*†
S. Pokorski *Gauge Field Theories, 2nd edition*†
J. Polchinski *String Theory Volume 1: An Introduction to the Bosonic String*†
J. Polchinski *String Theory Volume 2: Superstring Theory and Beyond*†
J. C. Polkinghorne *Models of High Energy Processes*†
V. N. Popov *Functional Integrals and Collective Excitations*†
L. V. Prokhorov and S. V. Shabanov *Hamiltonian Mechanics of Gauge Systems*
S. Raychaudhuri and K. Sridhar *Particle Physics of Brane Worlds and Extra Dimensions*
A. Recknagel and V. Schiomerus *Boundary Conformal Field Theory and the Worldsheet Approach to D-Branes*
M. Reuter and F. Saueressig *Quantum Gravity and the Functional Renormalization Group*
R. J. Rivers *Path Integral Methods in Quantum Field Theory*†
R. G. Roberts *The Structure of the Proton: Deep Inelastic Scattering*†
P. Romatschke and U. Romatschke *Relativistic Fluid Dynamics In and Out of Equilibrium: And Applications to Relativistic Nuclear Collisions*
C. Rovelli *Quantum Gravity*†
W. C. Saslaw *Gravitational Physics of Stellar and Galactic Systems*†
R. N. Sen *Causality, Measurement Theory and the Differentiable Structure of Space-Time*
M. Shifman and A. Yung *Supersymmetric Solitons*
Y. M. Shnir *Topological and Non-Topological Solitons in Scalar Field Theories*
H. Stephani, D. Kramer, M. MacCallum, C. Hoenselaers and E. Herlt *Exact Solutions of Einstein's Field Equations, 2nd edition*†
J. Stewart *Advanced General Relativity*†
J. C. Taylor *Gauge Theories of Weak Interactions*†
T. Thiemann *Modern Canonical Quantum General Relativity*†
D. J. Toms *The Schwinger Action Principle and Effective Action*†
A. Vilenkin and E. P. S. Shellard *Cosmic Strings and Other Topological Defects*†
R. S. Ward and R. O. Wells, Jr *Twistor Geometry and Field Theory*†
E. J. Weinberg *Classical Solutions in Quantum Field Theory: Solitons and Instantons in High Energy Physics*†
J. R. Wilson and G. J. Mathews *Relativistic Numerical Hydrodynamics*†

†Available in paperback

THE LARGE SCALE STRUCTURE OF SPACE-TIME

50TH ANNIVERSARY EDITION

STEPHEN W. HAWKING
University of Cambridge

GEORGE F. R. ELLIS
University of Cape Town

FOREWORD BY
ABHAY ASHTEKAR
Pennsylvania State University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009253154

DOI: 10.1017/9781009253161

First published 1973
First paperback edition 1974
20th printing 2006
50th anniversary edition 2023

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

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

ISBN 978-1-009-25315-4 Hardback

To
D. W. SCIAMA

Contents

Foreword
to the 50$^{\text{th}}$ Anniversary Edition

In 1921, Cambridge University Press published Arthur Eddington's monograph, *The Mathematical Theory of Relativity,* arguably the first systematic and comprehensive textbook on the theory. It embodies Eddington's view that 'The investigation of the external world is a quest for structure rather than substance'. It had a deep influence on how researchers thought of general relativity in subsequent decades.

Five decades later, the Press published another monograph, *The Large Scale Structure of Space-Time* by Stephen Hawking and George Ellis in 1973. Hailed immediately as 'a masterpiece, written by sure hands' it too focuses on 'structure' – but now on *global aspects* of spacetime structure, which had been almost entirely ignored in earlier books. The monograph solidified the new approach to understand gravitational phenomena, introduced by Roger Penrose through his use of global methods and causal structures, which transformed the way the community thought of strong gravity. It has had even greater impact on the development of relativistic gravity than Eddington's monograph because it helped shape the 'golden age' of general relativity during the 1970s.

Before the appearance of this monograph, contributions to general relativity were by and large dominated by tensor calculus and partial differential equations in local coordinates. The monograph served as a powerful catalyst that changed our way of understanding the physics of general relativity. Thanks in large part to its influence, a sizable fraction of researchers started thinking invariantly, in geometrical terms, using spacetime diagrams and light cones. The emphasis shifted to global issues. In subsequent years, this shift led to numerous novel directions that created new frontiers of research: black hole uniqueness theorems; detailed investigations of the cosmic censorship hypothesis; introduction of quasi-local horizons that now play a key role in numerical relativity; and unforeseen connections between relativistic gravitation, quantum physics, and statistical mechanics, through black holes.

The transformative impact of the monograph is not confined to physics and astrophysics. Even in the mathematical community that provides us with rigorous proofs, the emphasis has shifted from local results based on partial differential equations to 'geometric analysis' that focuses on global existence and uniqueness results for solutions to Einstein's equations, obtained using geometric structures that emphasize causality.

Hallmarks of this monograph are its conceptual clarity, mathematical rigor, and concise and precise statements that capture the essential underlying structures. The authors reverse the Machian view that the local laws are determined by large scale structure, and instead 'take the local physical laws to be experimentally determined' and explore 'what these laws imply about the large scale structure of the universe'. This insightful switch guides their discussion throughout the monograph.

The organization of the monograph was also novel at the time. It used invariantly defined structures in differential geometry to present general relativity through a systematic set of postulates. Five decades have passed and yet this approach continues to be contemporary! Similarly, almost nothing new can be added to the presentation of the physical effects of curvature on test particles, the detailed mathematical discussion of energy conditions and the masterful treatment of the global structure of space-times – such as de Sitter, anti-de Sitter, Schwarzschild, and Kerr – that continue to feature prominently in the contemporary literature. The discussion of singularity theorems and strong field dynamics associated with gravitational collapse and binary black hole mergers are the crowning achievements of the monograph. A series of influential works from the then Soviet school led by Khalatnikov and Lifshitz suggested that the formation of singularities in gravitational collapse is an artifact of the high degree of symmetry assumed in the analysis, and generic solutions would be singularity-free. The comprehensive treatment of singularity theorems in the monograph was instrumental in causing a decisive shift in the community, away from this paradigm. Similarly, at the time, many astronomers and physicists did not believe that black holes were physical entities. Inclusion of a detailed discussion of black hole dynamics in a monograph shows incredible foresight and confidence. It has been handsomely rewarded through discoveries of binary black hole mergers by the LIGO–Virgo collaboration. Discussions of these events routinely include not only the technical statements from the monograph, but even some of the diagrams!

In his preface to this golden jubilee edition, George Ellis has included a list of topics that are not covered by the book. Almost all of them refer to discoveries that were made since publication. However, the omission of gravitational waves is somewhat puzzling, given that Bondi, Sachs, Penrose, Newman, and others had developed the subject in detail during the preceding decade, and the subject matter is intimately related to the large scale structure of spacetime. Its inclusion would have made the work even more prescient! Perhaps it was left out because the volume is already close to 400 pages. Indeed, even as it stands, the monograph is peerless in the way it served to guide the subsequent developments in the field.

When it first appeared, I was a graduate student. I distinctly remember the excitement we all felt as we slowly absorbed the grandeur of the new vistas that the monograph opened before us. When I moved to Oxford as a postdoctoral researcher, I eagerly went to Blackwell's to buy my own paperback copy, which had just appeared. At £3.95, it is the best book purchase I have ever made! I still refer to it.

Abhay Ashtekar
University Park, PA, USA

Preface to the 50^{th} Anniversary Edition

This book, written by Stephen Hawking (see Carr *et al.* 2019) and myself between 1971 and 1973, presents a systematic overview of Einstein's General Theory of Relativity as a theory of gravity. We wrote it in the middle of what has come to be called the 'golden age' of general relativity: a time when a largely ignored theory, regarded by many as being at a dead end, transitioned to being truly dynamic, with the foundations being laid for developments in many directions in later years.

The book is dedicated to our research supervisor Dennis Sciama, FRS (Ellis and Penrose 2010), who was an outstanding physicist and supervisor. I arrived in Cambridge from South Africa in 1961, and started as his first research student in the University Department of Applied Mathematics and Theoretical Physics (DAMTP) in January 1962. Stephen arrived from Oxford in 1962, and the third student in the group who would focus on related issues in general relativity and cosmology was Brandon Carter, who came from Australia in 1962. The convivial way the research group was run is recalled in Ellis and Penrose (2010) and Ellis (2014).

The key issue we were involved in at the time was whether the universe had a beginning or not. Dennis was debating with Fred Hoyle, Hermann Bondi, and Tommy Gold whether their Steady State theory of the universe, which had no start, was a better model than the Standard Model, which did have a beginning. However, the Steady State model did not obey the field equations of general relativity: would models obeying those equations necessarily have a beginning? This would represent the earliest time the universe existed. Using the data of the time, this seemed to be the case: the universe would start at a *singularity*, an edge to spacetime where physical quantities such as the density would diverge. The universe – and physics – did not exist before that time.

However, the standard cosmological models had a highly simplified geometry: they were spherically symmetric about every point as well as being spatially homogeneous, hence there was no rotation or acceleration that could avoid a singularity. We wanted to know if more general geometries could allow a non-singular start. Our method was to look at specific anisotropic but spatially homogeneous models; but we could not prove it either way.

A related issue, driven primarily by John Wheeler at Princeton, was whether a spacetime singularity would occur at the centre of gravitational collapse when a star had used up all its nuclear energy. The same issue arose: simple models said this would happen when they were over a certain mass, and collapse to a singularity was unavoidable if they were massive enough. But they were spherically symmetric models. Could rotation of a collapsing star avoid a singularity?

The whole topic was blown wide open in 1965 by a truly innovative paper by Roger Penrose, who was then at Birkbeck College, London, showing singularities would indeed occur at the endpoint of gravitational collapse (Penrose 1965); he would much later receive the Nobel Prize in Physics for this work (Nobel Prize 2020). The paper involved innovative examination of global properties and causal structures of spacetimes, energy inequalities rather than exact equations, the crucial concept of a closed trapped surface, and a characterization of existence of singularities via geodesic incompleteness.

The Cambridge group (mainly Hawking, Carter, and myself) went into overdrive to learn the details of these new methods, jointly with colleagues Felix Pirani and others from King's College, London, and DAMTP visitors John Wheeler and Charles Misner from Princeton and Maryland, respectively. Stephen and I wrote a paper showing that these methods would indeed work in the restricted case of spatially homogeneous models (Hawking and Ellis 1965), and he then rapidly produced a series of existence proofs for generic cases, based on the idea of a time-reversed closed trapped surface together with suitable causal conditions. An initial such theorem was given in both Hawking (1965) and the last chapter of his PhD thesis (Hawking 1966a); a further series of singularity theorems with different details were presented in Hawking (1966b, 1966c, 1966d, 1967).

The Adams Prize is awarded jointly by the Cambridge University Faculty of Mathematics and St John's College for an essay in a stipulated topic in mathematics. In 1966 the topic was 'Geometric Problems of Relativity, with special reference to the foundations of general relativity and cosmology'. The adjudicators were H. Bondi, W. V. D. Hodge, and A. G. Walker. The prize was awarded to Roger Penrose for his essay entitled 'An analysis of the structure of spacetime', presenting the methods he had used in his 1965 paper, while Stephen was awarded an auxiliary prize for his essay 'Singularities and the geometry of spacetime' (Hawking 1966e), reprinted with commentary in Ellis (2014). This essay summarized global properties of general relativity theory, and on this basis developed a series of cosmological singularity theorems he had proved. Neither Adams Prize essay was published as a book, although preprint versions of both were circulated in the relativity community.

Further important work developing causal relations and global properties was carried out *inter alia* by Penrose, Robert Geroch (who was at Birkbeck with Penrose), Carter, Hawking, Werner Israel, Misner, and others; see for example Hawking (1968, 1970, 1971). A major summary theorem was developed by Hawking and Penrose (1970).

On the observational side, crucial new data became available about the nature of the expanding universe via the discovery in 1965 of the Cosmic Microwave Background (CMB) radiation, giving evidence of the nature of the evolution of the early universe and the existence of a Hot Big Bang epoch. Its implications were rapidly explored by Sciama, his students John Stewart and Martin Rees, and many others. Stephen and I wrote a paper (Hawking and Ellis 1968) showing how the very existence of that radiation showed a time-reversed closed trapped universe must exist in the early universe, and so provide evidence of the existence of an initial singularity.

To follow these developments in detail required pulling together a variety of mathematical topics that were not well known to the relativity community at that time, so a summary book was discussed between Stephen and myself in 1966, encouraged by Dennis. A contract for such a book with Cambridge University Press was accepted by them on 24 April 1967 and signed on 18 May 1967 under the title *Singularities, Causality and Cosmology,* to be published in the *Cambridge Series on General Relativity*, edited by W. H. McCrea and D. W. Sciama. By the time of publication in 1973 this had become the *Cambridge Monographs on Mathematical Physics*, with J. C. Polkinghorne added as third editor of the series.

The real writing of the book only started in 1970, with the focus being in 1971–1972, because we were both doing other things. This was the pre-LaTeX era. Stephen was having trouble coordinating his muscles so I typed the text myself, inserting handwritten equations in the text sent to the Press, who typeset it and sent proofs back. Then several rounds of corrections to proofs followed. The diagrams were drawn by a draftsman in the geography department under my guidance. The writing was completed in January 1973. The title had changed to *The Large Scale Structure of Space-Time.*

The book is a book of its time, and does not include major developments that have come later. In particular, it was written at a key time in the development of black hole theory; while it made a contribution to that development, later papers and books developed that theory much further. The same is true for cosmology and gravitational waves. Nobel prizes have been awarded in each of these areas since those times.

We were asked later on if we wished to do an updated version to take some of these developments into account, but declined. Given Stephen's physical condition, this would not have been practical. In any case, the book had a terse style because giving all details in depth would have made it much longer, and no more readable. We did not want to change this.

What this book does not cover:

- **Alternative theories of gravity**: Scalar–tensor theories, higher-order gravity theories
- **Experimental tests of general relativity theory**: Solar systems tests, tests via cosmology and astrophysics
- **Inflationary cosmology and structure formation**: Structure formation in cosmology, CMB anisotropies, dark energy/dark matter existence and nature
- **Black hole thermodynamics**: The four laws of black hole thermodynamics**,** Hawking radiation, astrophysical black holes: formation, accretion, and associated radiation
- **Gravitational radiation**: Carrying off energy and momentum, emission and detection of gravitational radiation
- **Quantum gravity**: Supergravity, string theory, loop quantum gravity, etc.; and the wave function of the universe

Major advances have been made in all these areas since the book was written.

What has changed?

A key point to notice is the following: the status of the energy conditions has completely altered due to the advent of inflationary cosmology theory through the pioneering work of Alan Guth (1981), followed by many others (Guth 2007). This is now widely accepted as a correct model of the universe (Mukhanov 2005, Peter and Uzan 2009), with a slow rolling scalar field dominating early universe dynamics so that the energy conditions required for the singularity theorems no longer hold (Ellis 2014).

This possible breakdown of the energy conditions is essentially recognized in our book on page 96, but it is suggested there that this will be on such a small scale as to not alter the conclusions as regards the singularity theorems. But now that dominance of scalar field dynamics in the early universe is generally accepted, that conclusion is called into question. Singularity-free universes are in principle possible (Ellis and Maartens 2003).

However, in the end, the issue of whether the universe has a start or not depends on a resolution of the issue of the nature of quantum gravity: and we simply do not know what that answer is. The key question that led to the book is unsolved. But we acknowledge this in the conclusion on page 364.

George Ellis
Cape Town
June 2022

References

Carr, B. J., G. F. R. Ellis, G. W. Gibbons, J. B. Hartle, T. Hertog, R. Penrose, M. J. Perry, and K. S. Thorne (2019). 'Stephen William Hawking CH CBE. 8 January 1942–14 March 2018', *Biogr. Mems Fell. R. Soc.* **66**, 267–308.

Ellis, G. F. R. (2014) 'Stephen Hawking's 1966 Adams Prize essay', *Eur. Phys. J.* H **39**, 403–11.

Ellis, G. F. R. and Maartens, R. (2003) 'The emergent universe: inflationary cosmology with no singularity', *Class. Quantum Gravity* **21**, 223.

Ellis, G. F. R. and Penrose, R. (2010) 'Dennis William Sciama', *Biogr. Mems Fell. R. Soc*. **56**, 401–22.

Guth, A. H. (1981) 'Inflationary universe: a possible solution to the horizon and flatness problems', *Phys. Rev.* D **23**, 347.

Guth, A. H. (2007) 'Eternal inflation and its implications', *J. Phys.* A **40**, 6811.

Hawking, S. W. (1965) 'Occurrence of singularities in open universes', *Phys. Rev. Lett.* **15**, 689–90.

Hawking, S. W. (1966a) *Properties of expanding universes*. PhD thesis, University of Cambridge. Available at: www.repository.cam.ac.uk/handle/1810/251038.

Hawking, S. W. (1966b) 'Singularities in the universe', *Phys. Rev. Lett.* **17**, 444–5.

Hawking, S. W. (1966c) 'The occurrence of singularities in cosmology', *Proc. Roy. Soc. Lond.* A **294**, 511–21.

Hawking, S. W. (1966d) 'The occurrence of singularities in cosmology. II', *Proc. Roy. Soc. Lond.* A **295**, 490–3.

Hawking, S. W. (1966e) 'Singularities and the geometry of spacetime', University of Cambridge Adams Prize Essay. Republished with commentary in Ellis (2014).

Hawking, S. W. (1967) 'The occurrence of singularities in cosmology. III. Causality and singularities', *Proc. Roy. Soc. Lond.* A **300**, 187–201.

Hawking, S. W. (1968) 'The existence of cosmic time functions', *Proc. Roy. Soc. Lond.* A **308**, 433–5.

Hawking, S. W. (1970) 'The conservation of matter in general relativity', *Commun. Math. Phys.* **18**, 301–6.

Hawking, S. W. (1971) 'Stable and generic properties in general relativity', *Gen. Rel. Grav.* **1**, 393–400.

Hawking, S. W. and Ellis, G. F. R. (1965) 'Singularities in homogeneous world models', *Phys. Lett*. **17,** 246–7. (doi:10.1016/0031-9163(65)90510-X)

Hawking S. W. and Ellis, G. F. R. (1968) 'The cosmic black body radiation and the existence of singularities in our universe', *Astrophys. J.* **152**, 25–36.

Hawking, S. W. and Penrose, R. (1970) 'The singularities of gravitational collapse and cosmology', *Proc. Roy. Soc. Lond*. A **314**, 529–48.

Mukhanov, V. (2005) *Physical Foundations of Cosmology* (Cambridge University Press).

Nobel Prize (2020) Available at: www.nobelprize.org/prizes/physics/2020/penrose/169729-penrose-interview-march-2021/.

Penrose, R. (1965) 'Gravitational collapse and space–time singularities', *Phys. Rev. Lett.* **14**, 57–9. https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.14.57

Peter, P. and Uzan, J.-P. (2009) *Primordial Cosmology* (Oxford University Press).

Preface

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

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

Cambridge S. W. Hawking
January 1973 G. F. R. Ellis

1
The role of gravity

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

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

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

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

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

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

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

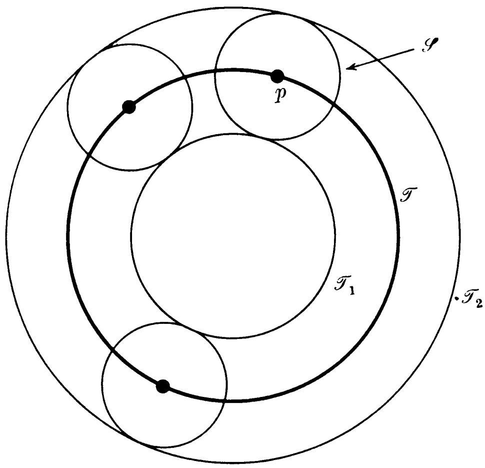

FIGURE 1. At some instant, the sphere $\mathscr{T}$ emits a flash of light. At a later time, the light from a point p forms a sphere $\mathscr{S}$ around p, and the envelopes $\mathscr{T}_1$ and $\mathscr{T}_2$ form the ingoing and outgoing wavefronts respectively. If the areas of *both* $\mathscr{T}_1$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

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

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

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

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

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

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

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

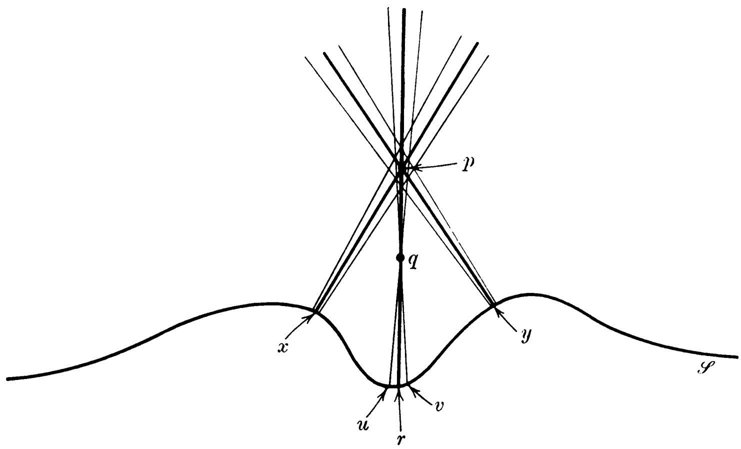

FIGURE 2. The line pr cannot be the shortest line from p to $\mathscr{S}$, because there is a focal point q between p and r. In fact either px or py will be the shortest line from p to $\mathscr{S}$.

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

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

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

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

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

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

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

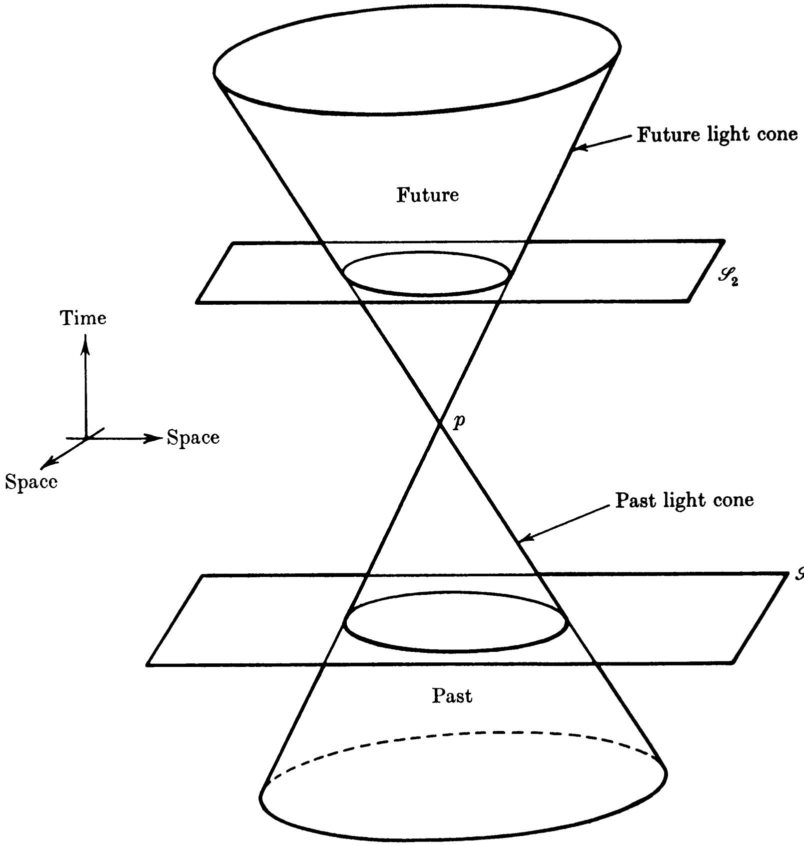

FIGURE 3. In Special Relativity, the light cone of an event p is the set of all light rays through p. The past of p is the interior of the past light cone, and the future of p is the interior of the future light cone.

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

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

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

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

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

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

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

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

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

2
Differential geometry

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

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

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

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

2.1 Manifolds

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

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

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

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

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

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

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

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

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

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

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

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

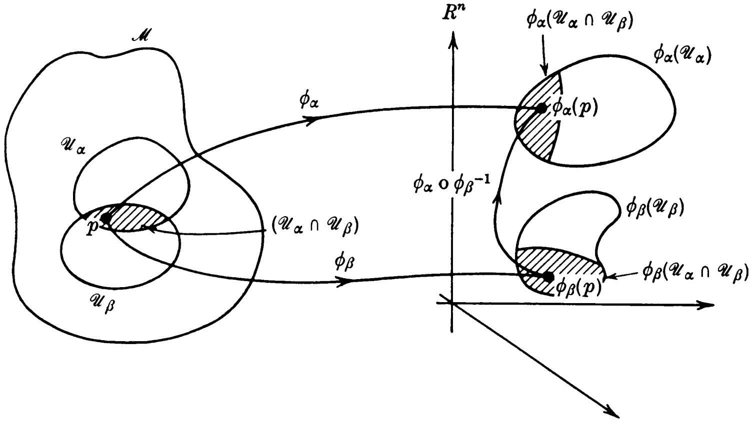

FIGURE 4. In the overlap of coordinate neighbourhoods $\mathscr{U}_\alpha$ and $\mathscr{U}_\beta$, coordinates are related by a C^r map $\phi_\alpha \circ \phi_\beta^{-1}$.

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

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

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

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

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

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

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

in R^{n+1}.

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

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

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

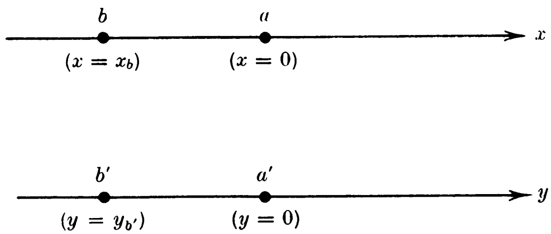

FIGURE 5. An example of a non-Hausdorff manifold. The two lines above are identical for $x = y < 0$. However the two points a $(x = 0)$ and a' $(y = 0)$ are not identified.

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

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

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

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

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

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

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

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

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

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

2.2 Vectors and tensors

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Then the linearity implies that, for example,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$S(\mathbf{T})(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) = \frac{1}{2!}\{T(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)+T(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1)\}$$

for all $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2\in T^*{}_p$. We shall denote the components $S(\mathbf{T})^{ab}$ of $S(\mathbf{T})$ by $T^{(ab)}$; then

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

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

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

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

$$T_{(a_1\ldots a_r)}{}^{b\ldots f}$$

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

and

$$T_{[a_1\ldots a_r]}{}^{b\ldots f}$$

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

For example,

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