UNIVERSAL
LIBRARY

OU_156602

UNIVERSAL
LIBRARY

Call No. Accession No.

Author

Title

This book should be returned on or before the date last marked below.

THE NATURE OF PHYSICAL THEORY

THE NATURE OF PHYSICAL THEORY

A Study in Theory of Knowledge

BY

VICTOR F. LENZEN, Ph.D.

Associate Professor of Physics in the University of California

NEW YORK

JOHN WILEY & SONS, INC.

LONDON: CHAPMAN & HALL, LIMITED

1931

COPYRIGHT, 1931,
BY
VICTOR F. LENZEN

All Rights Reserved
This book or any part thereof must not be reproduced in any form without the written permission of the publisher.

PRINTED IN U. S. A.

PRESS OF
BRAUNWORTH & CO., INC.
BOOK MANUFACTURERS
BROOKLYN, NEW YORK

PREFACE

In view of the disposition of contemporary thinkers to self-revelation, it may not be amiss to sketch the somewhat diverse activities which have resulted in the present philosophical study of physical theory.

In the first place, the book is an expression of ideas developed in teaching elementary experimental physics, primarily dynamics, and the mathematical theory of higher dynamics, at the University of California during the last ten years. The most characteristic methods of the book were developed in reflecting upon the fact that I found it desirable to present dynamics in one way for elementary students and in another for graduate students of physics. The outcome of my investigations is the exposition of dynamics by a method of successive definition, in which dynamical concepts are successively expounded from the qualitative, operational and postulational points of view. Furthermore, an analysis of the theory of relativity and quantum theory revealed that physical theory progresses by successive approximation. In the present book I have attempted to expound systematically the whole of physics by showing how it can be constructed by successive definition and approximation. The development of my ideas has been aided by the critical comments of my colleague, Professor W. H. Williams.

The character of this critical study of physical theory owes much, however, to an earlier philosophical training at California and Harvard, during which period my interest in the logic of mathematics and natural science was stimulated by Professors B. A. Bernstein, C. I. Lewis, Josiah Royce, E. V. Huntington and Bertrand Russell. When Mr. Russell lectured at Harvard during the spring of 1914 he had completed his work on *Principia Mathematica* and was turning his atten-

tion to the logical analysis of physics. In a conference he suggested that I write a paper on causality, which he declared should be investigated by an analysis of the differential equations of physics. I was unable to write the paper at that time, but hope that the task has been partially, even if belatedly, performed in this book. For methods of analysis I am therefore particularly indebted to Mr. Russell. But the book is also reminiscent of Professor R. B. Perry's seminary in theory of knowledge, of Professor M. P. Mason's careful lectures on Kant, of Professor W. E. Hocking's seminary in post-Kantian idealism, and of Professor R. F. A. Hoernlé's lectures on Bosanquet. I especially place high value upon my intensive study of Fichte's *Wissenschaftslehre*, Hegel's *Phenomenologie*, and Bosanquet's *Logic*. For in my attempt to construct a theory of physics the most important and difficult step was to free myself from the exclusive use of the methods of the logic of pure mathematics. I doubt whether I could have arrived at the characteristic methods of this book, if I had not studied the idealistic theories of the development of thought. It is more than ten years since I have read Bosanquet, yet I am aware that my thinking has been molded by such principles as " truth is the whole," " thought is the nisus to totality," etc.

But the activity most immediately responsible for the present book has been the preparation of papers for the Philosophical Union in Berkeley, since the spring of 1923. I was encouraged to write these papers through the friendly interest of philosophers in Berkeley and elsewhere, and through the encouragement of the former chairman of the physics department, the late Professor E. P. Lewis, and the present chairman, Professor E. E. Hall. These papers have been published in the University of California Publications in Philosophy, and I am indebted to the editors of the series and the University of California Press for permission to use material from them, especially from the last one, " The Metaphysical Import of Scientific Facts," which is being published in a volume on " The Nature of Facts."

The interest accorded my Union papers encouraged me to plan a more systematic and detailed study of physical theory. In the carrying out of my project I have been greatly aided by

the opportunity to spend the year 1927–28 in Göttingen as John Simon Guggenheim Memorial Fellow. The plan of the book was largely worked out in a graduate course in the department of philosophy during the spring of 1930. I have to thank Professor J. Loewenberg for his encouragement and initiative in inviting me to give this course.

Thus the stimulus to writing the book has come primarily from philosophers. Since it was partly my purpose to provide philosophical students of science with a rigorous systematic exposition of physical theory I have used a minimum of mathematics. But I have developed the subject with sufficient detail so that the book will be useful also to physicists who are interested in methodology and systematic construction. The book thus stands between the popular expositions of physics and the detailed treatises for experts. Again, since the principal problem of the book is philosophical analysis, I have confined myself to those theories which appear to have been definitely incorporated, at least for the time being, in the system of physics.

To some of the chapters I have appended references. These do not constitute a bibliography of the subject, but are acknowledgments. In particular, I am indebted to the writings of Max Born; also to those of A. Einstein, A. S. Eddington, A. Sommerfeld, M. Planck, H. Weyl, N. Bohr, P. Jordan, W. Heisenberg, P. and T. Ehrenfest, H. Thirring, R. C. Tolman, P. W. Bridgman, P. Hertz, W. F. G. Swann. For the more philosophical phases of the subject I am especially indebted to the writings of R. Carnap; also to those of H. Reichenbach, M. Schlick, H. Dingler.

While writing the final draft of the book it was my great privilege to discuss critical problems of physics with Professor W. F. G. Swann, and I have borrowed much from him. The final form of my exposition has also been greatly influenced by C. I. Lewis's *Mind and the World Order*. My method of building up physical theory may be viewed as an illustration of the function of a variable *a priori* in knowledge. But, whereas Professor Lewis has arrived at the theory by reflection upon the structure of pure mathematics and the problem of its applicability to experience, I have arrived at the theory

from an analysis of the process of transformation of empirical generalizations into definitions, which process was first noted by H. Poincaré.

I am especially indebted to Professors G. P. Adams and S. C. Pepper for their continued interest and encouragement, and to my colleague, Professor L. B. Loeb, for his interest and advice. Finally, I wish to extend my expression of appreciation to Professor J. H. Woods, whose guidance made possible the successful conclusion of my philosophical studies at Harvard, and to Professor R. B. Abbott, who arranged my transition from philosophy to physics and guided my first and uncertain steps as an assistant in the physical laboratory.

V. F. L.

BERKELEY, CALIFORNIA
July, 1931

CONTENTS

PAGE

INTRODUCTION v

THE GENERAL CHARACTER OF PHYSICAL THEORY 1
1. The Criterion of Physical Reality 1
2. *A priori* Elements in Physical Theory 10

PART I.—SPECIAL CLASSICAL PHYSICAL THEORIES 13

CHAPTER

I. FUNDAMENTAL CONCEPTS AND METHODS 15
1. Fundamental Concepts 15
2. Quantity and Magnitude 19
3. Theory of Extensive Quantity 23
4. The Meaning of Physical Concepts 29
5. Scalars, Vectors and Tensors 30
6. The Elements of Analysis 33
7. Vector Analysis 35
8. Stages of Physical Science 39
9. The Origin of Physical Concepts 42
10. The Inductive Exposition of Physical Theory 45

II. EUCLIDEAN GEOMETRY 50
1. Space and Geometry 50
2. Metrical Geometry 53
3. Synthetic and Analytic Geometry 58
4. Geometrical Transformations 59
5. The Concept of Group 62
6. The Structure of Space 63
7. Ideal Geometry 64
8. Abstract Geometry 65
9. Conclusions 69
10. Geometrical Optics 72

III. THE THEORY OF TIME 75
1. The Concept of Time 75
2. The Measurement of Local Time 77
3. Physical Theory and Abstract Theory 80
4. Extended Time 82

IV. KINEMATICS 85
1. Kinematics of a Point 85
2. Applications of Kinematics 90
3. Transformation of System of Reference 91
4. Wave Motion 93
5. Conclusion 95

CHAPTER PAGE

V. DYNAMICS 96
1. Introduction 96
2. The Statical Concept of Force 97
3. The Kinetical Concept of Force 104
4. Classical Dynamics 110
5. The Concept of Energy 112
6. Applications of Classical Dynamics 113
7. The Principle of Relativity 113
8. The Fundamental Principle of Dynamics 115
9. Reversibility 117
10. Conclusions 118

VI. ELECTRODYNAMICS 119
1. Electrostatics 119
2. Permanent Magnetism 124
3. The Electric Current 125
4. The First Equation of the Electromagnetic Field 127
5. The Second Equation of the Electromagnetic Field 129
6. Conclusions 131

VII. THERMODYNAMICS. 133
1. Temperature 133
2. Quantity of Heat 136
3. The Mechanical Theory of Heat 140
4. Thermodynamics 142
5. The First Law of Thermodynamics 144
6. The Second Law of Thermodynamics 146

PART II.—UNITARY SYSTEMS OF PHYSICAL THEORY 152

VIII. THE MECHANICAL CONCEPTION OF NATURE 153
1. Introduction 153
2. Gravitation 154
3. Elasticity 156
4. Sound 157
5. Light 158
6. The Kinetic Theory of Matter 161
7. The Kinetic Theory of Gases 162
8. The Method of Statistical Mechanics 169
9. The Electromagnetic Field 175

IX. THE ELECTRODYNAMIC CONCEPTION OF NATURE. *A. Microscopic Electrodynamics* 178
1. Introduction 178
2. The Atomistic Theory of Electricity 179
3. The Electron Theory 182
4. The Electromagnetic Theory of Mass 188
5. The Electrodynamic Theory of Mechanical Force 190

CHAPTER PAGE

X. THE ELECTRODYNAMIC CONCEPTION OF NATURE (*Continued*). *B. The Special Theory of Relativity* 192
1. The Contradiction in Classical Electrodynamics 192
2. The Theory of Time 195
3. The Postulates of the Theory of Relativity 197
4. The Lorentz Transformation 199
5. Space-Time 203

XI. THE GEOMETRICAL CONCEPTION OF NATURE. *A. Riemannian Geometry* 207
1. Introduction 207
2. Euclidean Space 208
3. Gauss's Theory of Surfaces 210
4. Riemannian Geometry 218
5. The Parallel Displacement of a Vector 220

XII. THE GEOMETRICAL CONCEPTION OF NATURE (*Continued*). *B. The Theory of Gravitation* 223
1. Accelerated Frames of Reference 223
2. The Principle of Equivalence 224
3. Variable Space-Time 226
4. The Law of Gravitation 229
5. Geometrical Physics 230

PART III.—THE QUANTUM THEORY

XIII. THE CLASSICAL QUANTUM THEORY 235
1. Introduction 235
2. The Quantum Hypothesis 237
3. The Hypothesis of Light Quanta 241
4. The Bohr Theory of the Atom 242
5. The Quantum Conditions for Multiply Periodic Systems . . 246
6. The Correspondence Principle 248

XIV. SYMBOLIC QUANTUM MECHANICS 252
1. Introduction 252
2. Matrix Mechanics 253
3. Wave Mechanics 258
4. The Interpretation of Quantum Mechanics 264

PART IV.—METHODOLOGICAL PRINCIPLES 269

XV. THE CONCEPT OF SUBSTANCE 271
1. The Criterion of Truth 271
2. Methodological Principles 273
3. Substance as Substratum 274
4. Substance as the Permanent 275
5. Constants 277
6. Invariants 281
7. The Character of Physical Reality 282

CHAPTER		PAGE
XVI.	THE CONCEPT OF CAUSALITY	285
	1. The Primitive Concept of Causality	285
	2. The Functional Concept of Causality	289
	3. Classical Dynamics	290
	4. The Physics of Fields	294
	5. Quantum Physics	295
INDEX		297

THE NATURE OF PHYSICAL THEORY

INTRODUCTION

THE GENERAL CHARACTER OF PHYSICAL THEORY

1. THE CRITERION OF PHYSICAL REALITY

THE problem of physical science is the construction of concepts of things, characters and processes which will serve to describe, interpret and order reality from a selected point of view. Things and phenomena considered as instances of physical characters and processes constitute the physical order. The most general characteristic of physical reality is that it is ordered in space and time. According to a more specific classical definition the physical order is the realm of matter and energy. However, in order to give a detailed exposition of the concept of physical reality we must ultimately have recourse to examples of physical properties and processes which one may assume to be known. Thus the motion of material bodies, the elastic properties of bodies, the propagation of light, the phenomena correlated with changes in the hotness and coldness of bodies, electrical and magnetic actions, the structure of matter—these exemplify the content of the physical order.

The physical order is assumed to be objective, that is, it may be known by several minds. In the first instance the elements of the physical order are given in sensation. In physical science one constructs the fundamental concepts of the things, characters and processes revealed by sensation. A physical theory is the systematic representation of certain characters of the physical order. The goal of physical science is a unitary system of physical theory.

It is the purpose of this book to analyze the concepts, principles and systems of physical theory. The method of procedure will be to build up physical theory from the aspects of the physical order given in sensation. Now the process of con-

structing physical theory presupposes a criterion of physical reality which should be made explicit at the outset.

Probably the first stage in science is the almost unreflective assumption or recognition of the existence of a world of things. Preceding the conscious development of science primitive men recognized things, invented names for them, described them, etc. We who live in the historical era have inherited concepts for the representation and interpretation of our experience; the application of these concepts is learned in earliest childhood so that their use appears to be instinctive. The fashioning of such a conceptual system for the description of common things in space and time is the problem of a primitive qualitative physics. Now starting from this level, which is that of ordinary experience, one may proceed in two directions.

On the one hand one may develop science systematically. Certain qualities of the world known in unreflective experience are selected for special observation, classification and correlation. One seeks an explanation of phenomena and an analysis of things. The physical properties of the world are especially characterized by the fact that numbers may be assigned to them by the process of measurement. The attempt to extend physical methods to other sciences means that the common goal of science is a numerical representation of reality given in experience.

It will serve to illuminate the nature of physical science if we examine in a preliminary way the principal historical examples of comprehensive physical theories. During the modern period the development of physical science has been molded primarily by the mechanical conception of nature, according to which all physical phenomena are to be interpreted as manifestations of mechanical force and the motions of bodies. The mechanical theory was based upon the dynamics of the seventeenth century, and it was furthered by the development of hydrodynamics, generalized dynamics, and celestial mechanics during the eighteenth century; but the mechanical theory achieved its greatest recognition during the nineteenth century upon the successes of the kinetic theory of gases and upon the attempts at a mechanical theory of the ether and electromagnetism. However, the development of microscopic elec-

trodynamics during the closing decades of the nineteenth century, and the subsequent construction of the special theory of relativity, have brought about the abandonment of mechanical physics for an electrodynamic conception of nature. The contemporary emphasis upon the quantum theory of atomic processes and radiation is further evidence that electrodynamics has become the legislative branch of theoretical physics. Now although contemporary physics is guided primarily by an electrodynamic theory, the possibility of still another point of view, a geometrical conception of nature, has been indicated by the interpretation of gravitation as a manifestation of the metrical structure of space-time. The mechanical, electrodynamic and geometrical conceptions of nature are the most significant unitary systems of physical theory.

Instead of building science upon its common-sense foundation, one may, on the other hand, subject the concept of the world of experience to a philosophical analysis, in order to determine the ultimate significance of our scientific concepts. Ordinarily the physicist does not undertake such a philosophical analysis of the criteria of physical reality. In an experiment he assumes that his laboratory table, meter-rods, standard weights, glass tubes, etc., are real things; he adopts from ordinary experience the criteria by which these assumptions can be justified. But the full significance of physical theory can not be grasped without an understanding of the concept of physical reality. I shall therefore begin my analysis of physical theory with an examination of the criteria of physical reality.

We may begin with the fact that one has knowledge of a world of real things. This world, which is known by acquaintance and description, consists of the stones, tables, chairs and other familiar things of common experience. It is convenient to initiate a discussion of the nature of reality with an analysis of the concept of thing. Now suppose that I am perceiving a thing such as a billiard ball. On analysis I characterize the ball as something having aspects: thus the ball has a visible aspect which is round and white; the ball has a tangible aspect which is smooth, hard, cool, round, etc. We may say that aspects of a thing are given in experience. However, the concept of thing implies more than the concept of given aspects; there are also

possible aspects. While I am merely seeing the ball some visible aspects are given, whereas the tangible aspects are merely possible. Thus a thing can be characterized as an entity which has given and possible aspects.

Now an important characteristic of reality is that at least some aspects can be given to some mind. Thus an important relation to which aspects may be terms is the relation of object to subject. The discussion of this specific very pervasive relationship raises the problem of the status of aspects with respect to their dependence on, or independence of, being given to some mind. I think that there are three possible solutions to this problem: namely, subjectivism, monistic realism and dualistic realism. Subjectivism is the theory that aspects exist only when they are given. Thus while I am looking at my table the visible aspect exists; I close my eyes and the aspect ceases to be. Monistic realism is the doctrine that aspects are independent realities; they exist when they are not given. According to dualistic realism experienced aspects are subjective representations of an independent reality. The foregoing three theories are possible metaphysical theories of the ultimate status of the aspects of physical reality.

Dr. Carnap has advanced the view that the metaphysical problem, as I have defined it, is an illusory one. I do not wish, however, to discuss this question in the present book; it appears to me that it is possible to construct science without answering, or even raising, the metaphysical problem. I have endeavored to use the term aspect without giving it a subjectivistic or realistic interpretation; my concept of aspect coincides with Mach's concept of element. We can describe given and possible aspects, or elements, without considering the problem of their subjectivity or independent reality. The systematic description and correlation of aspects of reality is the problem of science.

Now the reference to the metaphysical problem has provided us with additional concepts for the expression of the criterion of physical reality. The fundamental principle is that in the construction of a physical theory one employs a subjectivistic criterion of reality and assumes a realistic attitude. A thing is real because it is an object of possible experience; yet one describes

it as if it were independent of experience, although it may be dependent upon experience for its reality. On account of these apparently contradictory factors it is difficult to expound the theory of aspects without making it appear that one is either a subjectivist or a realist. I do not wish to base physical science upon subjectivism or realism because in my view a theory of aspects is merely preliminary to the metaphysical problem; a metaphysical subjectivism or realism can be superposed upon physical science.

As I have remarked, science, for example physical science, is the detailed study of certain characters of given and possible aspects. The characters of the aspects can be described independently of a theory of the ultimate status of the aspects. In the history of physics, however, physical theories have been expressed in terms of a special metaphysical theory, in particular, dualistic realism.

Thus at the beginning of the modern era Galileo made the distinction between primary qualities exemplified by extension and figure, and secondary qualities exemplified by hotness and sound. The primary qualities were assumed to be independently real and the secondary qualities were assumed to be subjective. Then Newton assumed absolute space and absolute time as independent realities; absolute time, for example, was presupposed to exist even if it eluded observation. Thus in Newton we find the explicit rejection of a subjectivistic criterion of physical reality. During the nineteenth century a dualistic metaphysics was asserted by Helmholtz and has recently been reiterated by Planck.

However, the development of dualism was paralleled by the subjectivistic criticisms of Berkeley and Hume. In the latter part of the nineteenth century the subjectivistic criterion of reality became effective in Mach's discussion of physical theory. The program of Mach's positivism was to eliminate the speculative metaphysical elements from physics; the realistic criterion of reality was replaced by a subjectivistic criterion. The data of physics consist of elements given in sensation. In my interpretation positivism is not a new metaphysics, but a pre-metaphysical theory. The assumptions of my theory of aspects are not a substitute for dualism, but the basis of a non-specu-

lative methodology of science. I quite admit the possibility of giving a dualistic interpretation of science, but such a dualistic interpretation is entirely apart from the solution of the scientific problem. The so-called elimination of metaphysics from physics means that the latter is to be restricted to the theory of aspects. The problem of physics is the description of the characters and relations of aspects which may be represented by numbers. Thus the spatial properties of the aspects of a body furnish the basis for certain numbers; similarly, the hotness and color of aspects can be represented by other numbers. The distinction between primary and secondary qualities is replaced by that between measurable and non-measurable characters of aspects.

In agreement with the empiricistic point of view my theory of science assumes a subjectivistic criterion of reality. A physical body is a class of aspects which are or can be given to some mind. In general one may say that the criterion of the reality of aspects is that they be possibilities of experience; that an aspect is a possibility of experience may be verified by experience or inferred from a general law connecting given aspects with possible aspects. The objectivity of the physical order is grounded in correlations between aspects given to different minds.

The subjectivistic criterion of physical reality is readily exemplified by an analysis of concepts which are applicable to observable bodies. Physical science, however, assumes the reality of molecules, atoms and electrons; if these entities are real they must be bodies with experienceable aspects. The principles of the theory of aspects may thus be illustrated by a discussion of the problem of the reality of molecules. electrons and other microscopic entities.

At the outset I remark that microscopic entities are endowed with properties the concepts of which are derived from directly experienceable aspects; for example, position, shape, mass, energy, have been ascribed to molecules. The physical quantity mass is assigned to a macroscopic body by certain operations which re[illegible]the experience of specific aspects; if mass is assigned to a molecule it implies that the molecule must have aspects which permit, at least theoretically, the same operations. Since molecules are described by the same concepts as are

macroscopic bodies it would seem that the meaning of their reality should be the same. However, it requires a special discussion in order to formulate a criterion of the reality of microscopic entities.

Molecules and other particles were first introduced into quantitative physics hypothetically during the nineteenth century. For example, in the kinetic theory of gases it was found possible to make assumptions about the properties of molecules and about the relations between microscopic and macroscopic quantities, and then by a process of deduction derive the laws which are confirmed by experience. As hypotheses the microscopic concepts proved to be successful. It was not, however, generally conceded that the derivation of experimental laws from microscopic hypotheses was a confirmation of the reality of molecules and atoms. Mach and Ostwald rejected molecules and atoms as fictions; it is evident that these scientists assumed that the possibility of experience is the criterion of reality. It is not entirely clear what criteria were employed by the proponents of microscopic concepts, but I think that the principal ground for belief in the reality of microscopic entities was the classical realistic criterion.

Since the advent of the twentieth century physicists have generally accepted the view that molecules, atoms and electrons exist. This general agreement is based on the fact that it has been possible to devise experiments in which one can observe individual effects of microscopic entities. I think that this indicates that in the last analysis the physicist employs a subjectivistic criterion of reality.

Now in the macroscopic order we test the reality of aspects by experience. But we also discover correlations between types of aspect, so that if one type of aspect is experienced we may infer the possibility of experience of a second type. For example, there are correlations between visible and tangible aspects, and hence the experience of the visible aspect is the basis for the inference of the possible tangible aspect. In this way we infer that the moon has a possible tangible aspect. An experience of an aspect and the employment of a law of correlation is the method of determining the existence of microscopic entities.

The existence of molecules can be inferred from observation

of the Brownian movement. If small particles are suspended in a liquid, observation through a powerful microscope reveals the existence of aspects which move with an irregular motion. The aspects are attributed to a particle which is bombarded by the surrounding molecules; the motion of the particle is that which would be expected from random impacts with molecules in the neighborhood. Thus the observation of an aspect together with the statistical derivation of the motion from the hypothesis of irregular molecular bombardments is evidence for the existence of molecules.

The discovery of radio-activity gave rise to the concept of alpha particle. Now if an alpha particle strikes against a zinc sulphide screen there is a flash of light; that is, the observer experiences an aspect which is the effect of an alpha particle. The directly observed aspect may be said to belong to the alpha particle just as the visible aspect of a star belongs to the star. Again, if an alpha particle or electron passes through saturated water vapor a condensation track is produced. An individual condensation track, which is an entity with observable aspects, is the effect of an individual particle.

We thus see that the criterion of possibility of experience is the test of the reality of atomic entities. There is no need for the interpretation of an atom as an independent reality which produces subjective effects. A microscopic entity is something to which physical quantities can be assigned and to which aspects belong, such as flashes of light on a zinc sulphide screen, condensation tracks, etc. Thus all entities of physical science can be characterized in terms of experienceable aspects. If, as a matter of fact, some aspects are not experienceable, for example, the tangible aspects of the moon, then this is merely a technical limitation, and not one of principle. In my opinion, the discovery of experimental methods which enable one to observe the aspects of individual microscopic entities has definitely rendered possible the general application of a subjectivistic criterion of physical reality; it has freed physical science from the necessity of operating with fictions or with a realistic criterion. The concepts of microscopic entities do not merely work—they are directly applicable to experience.

The observation of aspects of individual microscopic bodies

made possible the general application of a subjectivistic criterion of reality. In the special theory of relativity the realistic criterion is definitely repudiated. The characteristic result of the special theory is that absolute simultaneity at a distance can not be perceived and therefore is not a physical reality; an observable time order is relative to a frame of reference. The assumption of a subjectivistic criterion of reality as the basis for relativism has been set forth with admirable clarity by Professor Lovejoy. In conformity with his dualistic realism he rejects the relativistic logic. I admit that from the standpoint of metaphysical dualism it is possible to assert the existence of a non-observable absolute time system, but it is irrelevant to physics. Time as an order of given and possible aspects is relative.

The subjectivistic criterion of physical reality has achieved definitive expression in the quantum theory, which is the outcome of the recognition of an essential discontinuity in microscopic processes. It has been found possible to formulate only statistical laws for the determination of elementary processes; exact determinism is thereby abandoned in physics. In the classical era one would have called upon a dualistic metaphysics to rescue strict determinism for a real world which exists independently of our observations. The new quantum theory, however, has been constructed in the light of a subjectivistic criterion of reality.

The principles of the quantum theory describe the regularities of microscopic processes. Now the process of observing an atomic system disturbs it; this renders impossible the simultaneous precise determination of all physical quantities belonging to the system; hence there is an essential indeterminacy in our description of microscopic processes. As a consequence, only statistical laws represent what is observable in phenomena. The concept of a precisely determined microscopic order is a speculation which was able to arise because in the study of macroscopic processes one can neglect the interactions with measuring instruments. In the study of microscopic processes the agency of measurement can not be neglected, and this yields a fundamental indeterminacy and statistical regularity. A statistical theory of physical processes is thus a consequence of the strict

application of the subjectivistic criterion of reality to microscopic systems. This employment of the subjectivistic criterion is clearly set forth by Heisenberg in the paper in which he announced the principle of indeterminacy. He said, "Since the statistical character of the quantum theory is so clearly bound to the indeterminacy of all perception one might erroneously conjecture that behind the perceived statistical world there is a real world in which causality holds. But we emphatically assert that such speculations seem to us unfruitful and meaningless. It is the problem of physics to describe only the formal connection of perceptions." Again, Dirac remarks, "A fraction of a photon is never observed, so that we may safely assume it cannot exist."

To summarize: The methodological basis of science requires only the theory of aspects. However, it is an historical fact that classical physical theory was usually expressed in terms of a dualistic realism. The criticism of Mach was directed to the elimination of dualistic metaphysics from physics.

The dualistic interpretation of physics was based upon a realistic criterion of reality; the positivistic criticism is that a subjectivistic criterion should be employed. Thus metaphysical dualism is replaced by methodological subjectivism. The realistic criterion of classical physics was the basis of dualism, absolutism, determinism. The subjectivistic criterion of contemporary physics leads to monism, relativism, and indeterminism.

The program of the present book is to build the systems of physical theory, classical as well as contemporary, with a subjectivistic criterion of reality.

2. *A PRIORI* ELEMENTS IN PHYSICAL THEORY

My purpose is to analyze critically the concepts, principles, and systems of physical theory. Through such a study one may hope to obtain light on the nature of knowledge, more specifically of exact science, and to discover the respective contributions of experience and reason in the construction of science. To the student of theory of knowledge the science of theoretical physics is indeed a fruitful subject for analysis, for this science is both empirical and rational, inductive and deductive. On the

one hand, physics is an empirical science, it takes its rise in observation and experiment. On the other hand, in its systematic theoretical form, physics partakes of the nature of pure mathematics; a physical theory can be expressed by a deductive system. The philosopher Kant remarked that the degree of development of a science depends on the extent to which it employs mathematics. In this respect the most advanced of the natural sciences is theoretical physics; it is therefore the appropriate basis of a theory of knowledge.

Now it is an historical fact that fundamental periods in the history of physics have been correlated with the creation of new philosophical systems. Thus the speculations of the early Greek physicists concerning the substance of things provided the stimulus to the initial period in European philosophy. The experimental and theoretical physics of the seventeenth century was reflected in fundamental problems of modern philosophy. Indeed, the critical philosophy was partly the outcome of Kant's analysis of Newtonian dynamics. Since the eighteenth century, however, there has been a remarkable development of physics. Especially since the dawn of the twentieth century have new experimental results given rise to theories based upon new concepts and principles. In the light of history one may expect that the new physical theories will furnish the basis for a new theory of knowledge.

According to the Kantian theory, knowledge requires the union of two factors: empirical and rational, material and formal, *a posteriori* and *a priori*. In the present book physical theory will be expounded in a manner which exhibits the respective contributions of experience and reason to our knowledge of the physical order.

In Kant's epoch the only existing system of theoretical physics was Newtonian mechanics; this furnished the basis for static *a priori* elements in knowledge, invariable forms which are conditions of the possibility of experience. We shall see, however, that the new constructions reveal a fluid *a priori* in knowledge.

A general theory of a variable *a priori* has been expounded by C. I. Lewis in *Mind and the World Order*. In the present study we shall deal with the more specific problem of the

occurrence of *a priori* elements in our knowledge of the physical order. The *a priori* element in physical theory is grounded in the definitions of the fundamental concepts. Now it is a thesis of this book that some concepts are abstractions from experience. The definitions of fundamental physical concepts likewise are based on experience, but they also involve an element that is not entirely subject to the control of experience. Thus we define a standard rod to have unit length and postulate that it remains constant upon a displacement; we define a standard body to have unit mass and postulate that it remains constant upon a displacement. Such definitions are posited by a free choice which is guided by the methodological principle that the system have the maximum synthetic unity consistent with experience.

The *a priori*, or conventional, element in physical theory, however, is subject to transformation, both in the light of theory and experience. Thus empirical generalizations based upon prior definition of the concepts are transformed into implicit definitions of concepts. The original *a priori* elements are then replaced by empirical elements. The transformation of propositions which express empirical laws into implicit definitions will be an important phase of our critical study of the concepts, principles and systems of physical theory.

REFERENCES

CARNAP, R. Der logische Aufbau der Welt. Berlin, 1928.

DIRAC, P. A. M. Principles of Quantum Mechanics. Oxford, 1930.

HEISENBERG, W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 1927.

LEWIS, C. I. Mind and the World Order. New York, 1929.

LOVEJOY, A. O. The Dialectical Argument against Absolute Simultaneity. Journal of Philosophy, 1931.

PLANCK, M. Positivismus und reale Aussenwelt. Leipzig, 1931.

PART I

SPECIAL CLASSICAL PHYSICAL THEORIES

CHAPTER I

FUNDAMENTAL CONCEPTS AND METHODS

1. FUNDAMENTAL CONCEPTS

In this opening chapter we shall analyze certain basic concepts which are employed in the construction of physical theory. Now we conceive of the elements of the physical order as bodies which undergo certain physical processes. Thus the concept of **body** is fundamental in physical science. One refers to moving bodies, hot bodies, luminous bodies, electrified bodies, sounding bodies, etc. One distinguishes between macroscopic or large-scale bodies, such as a billiard ball, and microscopic or fine-scale bodies, such as an atom or electron. In ordinary experience we think of a body as a more or less permanent entity in space and time. From the standpoint of empiricism, however, such a concept of body is a construction which presupposes a theory. Indeed, we shall see that the common-sense notion of body is rendered impossible by the theory of relativity, which presupposes a plurality of times and spaces. We must then define the concept of body so that it admits of transformation.

I begin with an analysis of the concept of thing. A fundamental property of a thing is that it has **aspects.** If I look at a book upon the table I experience in sensation a partial aspect of the book. A thing presents different aspects from different points of view. We can describe an aspect by saying that it is colored, that it has a certain shape, that it is smooth, etc. Knowledge of a thing is ultimately based upon acquaintance with aspects of the thing. However, the concept of thing implies more than the concept of given aspects; there are also possible aspects. Thus while I am merely seeing the book some visible aspect is given, whereas the tangible aspects are

merely possible. Hence a thing may be characterized as an entity which has given and possible aspects. Now in the present exposition of physical theory I shall assume as far as possible the standpoint of empiricism. Accordingly we shall assume that the starting point for knowledge of the physical order is acquaintance with aspects of bodies. An aspect is a union of particular qualities, complexities and relationships.

In order to characterize the properties of aspects it is necessary to introduce universals. Let us then suppose that we have given two separate aspects, for example, the aspects of two red patches. We then observe that there is a relation of similarity between the two aspects: the two aspects are similar in certain respects and dissimilar in other respects. In the present instance the two aspects are similar with respect to color. In virtue of the similarity of two aspects from a given point of view we ascribe a common entity, a **universal**, to the two aspects; the two similar aspects are thus viewed as instances of the same universal. The problem of the ultimate status of universals is a metaphysical question which I shall not discuss; it suffices for us that it is convenient to describe aspects as if they were instances of a universal.

Universals are qualities or relations; thus the aspect of a red patch is an instance of the quality redness. The relation between two small aspects A and B such that A is to the left of B, is an instance of the relation "to the left of." Another example of a relation is that of the coincidence of two point aspects; this relation is similarity with respect to position.

Entities may be classified as particulars and universals. Particulars are instances of universals. Aspects are examples of particulars which may be given in sensation. Universals are the objects of concepts. A concept is a subjective entity, whereas a universal is objective. In the following discussion we shall refer to both entities, the one designated depending upon the context.

We have previously seen that aspects may be known by acquaintance. They may also be known by description, in which one represents an aspect by concepts of the universals of which the aspect is an instance. Thus I may describe a given aspect of my desk as brown, hard, smooth, etc.; that is,

I describe the aspect as an instance of the qualities brownness, hardness, smoothness, etc.

The program of the present book is to describe how the concepts of physical theory are derived from experiences of aspects. The preceding analysis of aspects now provides a basis for the exposition of the concepts of time, thing, and space. Let us begin with the concept of **time**. In accordance with the empirical standpoint I describe time as a specific relation between aspects. We have already considered aspects which are given in sensation; in addition we remember aspects which are no longer given. Remembered aspects are said to be in the past; memory is a process of knowing the past. Thus we may distinguish between present aspects which are given and past aspects which are remembered. The past and present are related by the relation of succession. Now we can apprehend particular instances of the succession of aspects. From several similar instances we abstract the concept of time as a universal, an order, a schema of aspects. We may further define the future as that relative to which the present is past; we remember the past and anticipate the future. As a basis for the study of the physical order I assume the validity of some memory, but recognize the hypothetical character of the future.

We are now prepared to understand the concept of **thing**. Examples of things are iron rods, clocks, the earth, billiard balls, etc. In order to introduce the concept of thing we must again start from aspects. We have previously distinguished between present and past aspects; now we find that present aspects are frequently similar to past aspects. The concept of thing represents such a class of similar past and present aspects. We also conceive of the existence of other minds which experience aspects of the same thing, but the evidence for this hypothesis falls outside the scope of our study.

Thus the concept of thing is the concept of a class of aspects which are in the past, present, or future for several minds. The proposition that a thing exists means that certain correlations of aspects may be experienced. We conceive of a thing as an entity which persists in time; objective time is the order of persistence of things. Now a temporal series of aspects may be sufficiently similar so that we consider them to be aspects

of a permanent thing, whereas in other respects the aspects may become dissimilar. If the change in the aspects is correlated with changes in the possible aspects for other minds we say that the thing has undergone a change or process. Objective time is thus the order of change of aspects of things.

The foregoing analysis has yielded the concept of a **process in time.** In the theory of time we refer to processes ot things which are called events. An important concept is that of the simultaneity of **events**; this relation is similarity with respect to temporal position.

The remaining fundamental concept is that of **space.** Now an aspect has a structure or form which is called spatial, so that the particular qualities of the aspect constitute the fields of spatial relations. Similarly there are spatial relations between present aspects. Now some aspects belong to things which persist in time. Thus the concept of objective space is based upon the apprehension of the succession of similar spatial patterns of simultaneous aspects; that is, objective space is a form of things. We conceive of space as existing throughout time and of the motion of things in space. Space may then be described as the totality of possibilities of relative position of things and phenomena.

In the analysis of space the most important concept is that of the coincidence of points, which is based upon a succession of similar coincidences of point aspects. Coincidence is similarity with respect to position.

We have seen that an aspect can be represented by a set of concepts of universals of which the aspect is an instance. An aspect belongs to a thing, and hence the concepts which represent the aspects can be employed to represent the thing. However, in constructing the concept of thing we recognize certain aspects as more essential than others. The real shape of a table is the shape of an aspect from a special point of view, the color of a table is the color of an aspect under certain conditions, so that our concept of a thing depends upon the characters of certain standard aspects. Essentially, our concept of thing is based upon tactual aspects. The spatial and temporal positions of things are fundamentally defined in terms of the relations of tactual aspects.

Thus the concepts of qualities and relations of aspects acquire the status of concepts of qualities and relations of things and phenomena; for example, we characterize a thing as long, hot, etc. The problem of natural science is to describe important particular things and to find correlations between characters. Now it has been found possible to correlate some characters with **numbers.** Numbers can be assigned by

(1) *The operation of measurement.*
(2) *The determination of Gaussian coordinates.*
(3) *Calculation.*
(4) *Hypothetical assignment.*

The assignment of numbers by the operation of measurement is of fundamental importance; accordingly I introduce this topic by a discussion of the nature of quantity.

2. QUANTITY AND MAGNITUDE

A thing has properties such as qualities and relations. Some properties have the character that they are measurable, that is, it is possible to assign to the property a number which is the outcome of the operation of measurement. The physical properties of things are the fundamental measurable properties; indeed, Planck has defined physical reality as that which is measurable. Thus the discussion of physical properties leads to the **theory of measurement.**

A thing considered merely from the point of view of physical properties will be called a body; that is, a body is something which has general measurable properties and relations. As a basis for the discussion I assume that we are able to recognize the existence of bodies and the coincidence of points of bodies. The concept of coincidence is fundamental in the theory of measurement.

The relation of coincidence is such that if point P coincides with point Q, then Q coincides with P. Also, if P coincides with Q and Q coincides with R, then P coincides with R.

These propositions express that coincidence is a symmetrical and transitive relation. Now a symmetrical and transitive relation is to be interpreted as indicating the possession of a common

property. We express this character of coincidence by stating that coincidence is similarity with respect to position. The propositions which assert that coincidence is a symmetrical and transitive relation express the formal characters of the relation; in order to understand the concept of coincidence we must also apprehend the respect in which the points are similar, and this requires the apprehension of instances of coincidences.

Let us have given two fixed points on a solid body; these two points determine a segment, A. We can construct another segment, B, such that the terminal points coincide respectively with the terminal points of A. By observation one can then verify the following propositions.

(1) *If the terminal points of* A *coincide with those of* B, *then the terminal points of* B *coincide with those of* A.

(2) *If the terminal points of* A *coincide with those of* B, *and the terminal points of* B *coincide with those of* C, *then the terminal points of* A *coincide with those of* C.

These propositions assert that the relation of coincidence of terminal points is a symmetrical and transitive relation.

The foregoing propositions are approximately true for all orientations and positions of the segments and for all kinds of solids; we express the results of observation by saying that the segments are equal with respect to extension; that is, the propositions about coincidences are transformed into propositions about equality with respect to extension. Thus the concept of equality of two neighboring segments is an abstraction from observations of coincidences of points. The relation of equality of segments with respect to extension has been defined so that the relation is transitive and symmetrical. We employ the concept as if the equal segments had a common property, i.e., the same extension.

The preceding discussion has assumed that the segments are adjacent. If we now separate the segments the concept of coincidence becomes inapplicable. By definition we determine that the segments are equal when separated, thus introducing the notion of equality at a distance by a pure definition. The concept of the equality of segments is explained, therefore, as an abstraction from experience plus a convention.

I next explain the inequality of segments; for this purpose I assume the concept of point in a segment. Now by definition if a terminal point of segment A coincides with a terminal point of segment B, and the other terminal point of A falls within B, then segment A is less than B with respect to extension. Conversely B is greater than A. I further assume that the relations of greater than and less than hold for segments at a distance. The relations of greater than and less than of segments with respect to extension satisfy the following propositions.

If A *is greater than* B, *and* B *is greater than* C, *then* A *is greater than* C.

If A *is greater than* B, *and* B *is equal to* C, *then* A *is greater than* C.

If A *is equal to* B, *and* B *is greater than* C, *then* A *is greater than* C.

We have thus seen that segments are entities which stand in the relations of equality, greater than and less than, with respect to extension.

All the propositions about the equality and inequality of segments can be derived from a set of eight propositions,[1] some of which have been stated above. These propositions may be transformed into postulates which implicitly define the properties of the relations of equality, greater than and less than, with respect to extension. Similarly the postulates define segments to be entities which satisfy these relations. We can now empty the postulates of their specific reference to extension and thereby obtain a definition of the formal characters of the relations of equality, greater than and less than, with respect to any property whatsoever.

Thus one may say that the laws for the equality and inequality of segments are the basis of a definition which serves as a criterion of similar relations between other entities.

The preceding account gives our first example of the transformation of *a priori* elements in knowledge. We initially defined the concepts of equality and inequality of segments with respect to extension upon the basis of experiences of coincidences; propositions which express facts about the relations

[1] Russell, B., Principles of Mathematics.

were then discovered. The axioms of such a system of propositions could then be transformed into postulates which constitute an implicit definition of the concepts. Finally one may deprive the postulates of all reference to extension and then obtain a set of abstract postulates which define the form of generalized equality and inequality. It then becomes an experimental fact that the equality and inequality of segments with respect to extension approximately satisfy the postulates. The abstract postulates constitute a formal element of knowledge which is *a priori* with respect to future experience.

Let us now introduce the traditional distinction between **quantity** and **magnitude**: a quantity is a particular and a magnitude is a universal. Thus a segment is a quantity, whereas an extension is a magnitude. The proposition that two segments are equal with respect to extension is expressed by the proposition that the two segments have the same extension. In general, the assertion that two particulars (quantities) are equal with respect to some property (magnitude) is equivalent to the assertion that the quantities have the same magnitude. Also, the assertion that a quantity is greater or less than another with respect to some property is equivalent to the assertion that the one quantity has a magnitude which is greater or less than that of the other. Magnitudes are characterized by the fact that they may be greater or less.

A distinction has been made between the relational and absolute theories of magnitude. The postulates of the absolute theory express properties of the relations greater than and less than between magnitudes; the postulates of the relational theory express properties of the relations equality, greater than, and less than between quantities. At first sight it would appear that the relational theory avoids the problem of universals. However, inasmuch as the relations between the quantities hold with respect to a property (kind of magnitude), it appears to me that the theories are equivalent. For the most part mathematicians have employed the language of the absolute theory; I prefer the language of the relational theory.

In general, quantities are characterized as particulars which may stand in the relations of equality, greater than or less than, with respect to a universal. Quantities may be further sub-

divided into **extensive** and **intensive** quantities, or in the language of the absolute theory, quantities with extensive and intensive magnitude. The discussion thus far applies to both kinds of quantity.

3. THEORY OF EXTENSIVE QUANTITY

I shall now consider the characterization of extensive quantities. A segment is an extensive quantity, or in the language of the absolute theory, a segment is an instance of an extension which is an extensive magnitude. The statement that a segment is an extensive quantity may be explained as follows. Suppose that we have given a set of segments A, B, C. . . . The operation of adding a segment B to a segment A consists in placing B so that one of its terminal points coincides with a terminal point of A. The remaining terminal points determine a segment which is called the sum of A and B; we may symbolize this by

$$A + B = S.$$

Suppose we have given a set of equal segments; then the sum of two segments each equal to A is a segment B which is defined to be twice as great as A. The ratio of the segment B to the segment A is 2. We may symbolize B by

$$B = 2A.$$

In general, by adding equal segments together one obtains a segment which is a multiple of a given segment,

$$S = nA.$$

One can construct a segment which is defined to be a fractional multiple of a given segment. If the segment X is such that

$$3X = 2A$$

then the segment X is defined to be $\frac{2}{3}$ as great as the segment A. I shall pass over the problem of incommensurable segments.

Thus we obtain a set of segments, A, B, C, etc., which are multiples of a selected segment. In the language of the absolute theory one would say that the magnitudes of the segments A, B, C, are multiples of a unit magnitude. Our set of segments will now be found to obey certain laws of operation; for example, if

$$A = aU$$

$$B = bU$$

we find that

$$aU + bU = (a + b)U$$

$$aU + bU = bU + aU, \text{ etc.}$$

The laws which are obtained by experiments upon segments can be derived from a selected group of them which we may call axioms. We shall consider these axioms for the theory of segments transformed into a set of postulates of an abstract theory. These postulates then define the general concept of continuous extensive quantity, or continuous extensive magnitude. The postulates furnish a criterion which particulars must satisfy in order to be called quantities.

A satisfactory set of postulates for extensive quantities has been given, for example, by Professor Huntington. It is an experimental fact that segments approximately satisfy the postulates. Again, the real numbers are examples of extensive quantities, that is, the operations upon real numbers satisfy the abstract postulates. This renders possible the assignment of numbers to extensive quantities, in other words one can establish a one-to-one correspondence between magnitudes and numbers.

We may illustrate this by the case of segments. Let A, B, and C be segments; then since segments are extensive quantities we can represent them as

$$A = aU$$

$$B = bU$$

$$C = cU.$$

Then $A + B = C$ can be written

$$aU + bU = cU$$

$$(a + b)U = cU,$$

where $$a + b = c.$$

Thus, just as

$$A + B = C,$$

so

$$a + b = c.$$

In general, if the number a corresponds to the quantity A, and b corresponds to the quantity B, then $a + b = c$ corresponds to C. The correspondence is with respect to U as a unit quantity. If now A, B, C, etc., are expressed as

$$A = aU$$

$$B = bU$$

$$C = cU$$

then a, b, c, are the measures of the quantities A, B, C, relative to the standard U, with respect to extension. Thus the measure is the ratio of a quantity to a standard quantity with respect to a magnitude. Employing the language of the absolute theory, one would say that a, b, c, are the measures of the magnitudes of the quantity with respect to the magnitude of the standard. To specify a measure one must state the ratio and the name of the standard. In the case of segments we express a segment A as

$$A = nU$$

where U is the standard segment. n is the ratio of the segment to the standard segment with respect to extension. Using the language of the absolute theory, one would say that n is the ratio of the extension of the segment to the extension of the standard. For example, a particular rod has an extension to which the number 2 is assigned; 2 is the ratio of the extension of the rod

to the extension of the standard meter. Thus we correlate a number with a magnitude and thereby a number is assigned to the body which has the given magnitude.

The assignment of the number 2 to a rod is expressed by a proposition such as, the length of this rod is 2 meters. The physical concept of length is to be interpreted as the concept of a relation between a pure number and the body to which the number is assigned. Thus, the length of this rod is 2 meters, may be written

2 is the length in meters of this rod
2 (length in meters) this rod.

The preceding proposition asserts a relation of the form xRy, where the relation R is one-many. There are many rods to which the number 2 has the relation " length in meters." Now a one-many relation gives rise to a descriptive function; thus one may write

$$x = R`y,$$

$$x = \text{the } R \text{ of } y.$$

Applying this analysis[1] to the present problem one obtains

2 = the length in meters ` this rod

or

the length in meters ` this rod = 2.

If one does not wish to employ the symbolism of *Principia Mathematica* one may write

2 = the length in meters of this rod.

As remarked previously, there are many rods of which one can say that the length in meters = 2; of any one of these rods one can say that it has the relation of length in meters to the number 2. Hence the relation of length in meters to the number two is a universal; it is a relational property of which any particular rod is an instance. Thus we think of " length in meters = 2 " as a property of a particular rod.

[1] For this analysis I am indebted to R. Carnap.

However, the important part of this analysis is the result that a pure number may be denoted by a descriptive function. Thus

2 = the length in meters of this rod.

Similarly one would say

2 = the coordinate in meters of this particle.

Now, as the position of a particle changes, its coordinates change; consequently there is a variable, say x, which can be described by

x = the coordinate in meters of this particle.

The laws of physics are expressed by functional relations between variables which can be denoted by descriptive functions which arise from certain relations which are expressed by physical concepts. Thus we say, if

s = the distance in meters passed over by a falling body,
t = the time,
0 = the initial time,
0 = the initial distance,
0 = the initial velocity,

then

$$s = \tfrac{1}{2}gt^2.$$

Let us now review the preceding discussion of extensive quantity. A segment was taken as an example of an extensive quantity; segments are additive; for example, $A + B = C$ with respect to extension. In the language of the absolute theory, the extension of A + the extension of B = the extension of C. When n is a number and U is the extension of a standard, a given extension can be expressed as nU. As a consequence, to the addition of extensions corresponds the addition of numbers. In general, the addition of extensive magnitudes can be represented by the addition of numbers; hence numbers can be assigned to extensive magnitudes and thereby to extensive quantities. In the case of a segment the number is described as "the length" of the segment. In the example of the segment there are the following factors: A particular segment is an

extensive quantity; the segment has extension, which is an extensive magnitude. When one assigns to the segment a number which arises from the operation of comparing the extension of the segment with the extension of a standard, the statement of the length of the segment expresses a relation between the number and the segment.

There is great confusion in terminology; for example, the word **length** is used in three senses. One says:

(1) *This segment is a length.*

"Length" is here used to denote the quantity, which is a particular.

(2) *This segment has length.*

In this case " length " may be understood to refer to the magnitude, which is a universal. In the previous discussion I have employed the word extension to designate the magnitude.

(3) *The length in meters of this segment is 2.*

The expression " the length " is here used to denote the number. " Length in meters is 2 " is a universal in that it can be assigned to different particulars. Notice, however, that it is a universal which is quite different from the traditional magnitude.

A similar confusion occurs in the use of the word **mass.** Thus one says:

This body is a mass.	(**quantity**)
This body has mass.	(**magnitude**)
The mass of this body is 2.	(**number**)

Intensive quantities do not obey a law of addition, but they are characterized in terms of the relations of equality, greater than, or less than. In the language of the absolute theory intensive magnitudes are entities which may stand in the relations of greater than or less than.

The magnitudes of intensive quantities can be measured, that is, a number can be assigned. This is rendered possible

by the definition of the measures of intensive magnitudes in terms of the measures of extensive magnitudes. Thus

m = the mass of a body;

v = the volume;

ρ = the density;

$\rho = m/v$.

4. THE MEANING OF PHYSICAL CONCEPTS

In contemporary physics the meaning of physical concepts is undergoing a transformation. Consequently one finds several meanings of the term **physical quantity.**

In traditional physics a physical quantity is a **particular** which has a magnitude to which a number can be assigned; for example, a segment has an extension to which a number can be assigned. Physics contains equations such as

$$v = s/t.$$

In traditional physics these equations express relations between magnitudes; speed is a magnitude which consists in the correlation of extension and duration; the symbol of division symbolizes a rate. The fundamental concepts of physics are of extension, duration, inertia, etc.; the concepts of other magnitudes are defined in a manner exemplified in the concept of speed.

In contemporary physics it is coming to be recognized that the essential thing is the number assigned to the magnitude. In practice we replace the concept of magnitude by that of number; thus when we use $v = s/t$ we substitute numbers. Magnitude enters into our thought as something to which a number is assigned according to a specified rule. The empirical fact is that we assign numbers on the basis of space-time coincidences—the number substitutes for the magnitude. The change in point of view has resulted in a change of terminology.

Professors Eddington and Swann apply the term physical quantity to the **number** instead of to the particular, the measure of magnitude of quantity of traditional physics thus becoming the physical quantity itself of contemporary physics.

In setting forth the conceptual apparatus of the traditional view of quantity I was careful to reserve the word "length" for a relation. This was for the purpose of setting up a relational theory of physical concepts. On this view physical concepts, such as the concepts of length, mass, temperature, electric charge, are the concepts of relations between a pure number and some physical object. These relations give rise to descriptive functions so that a number may be denoted by

the length
the mass
the temperature.

This appears to me to be the logical formulation of Bridgman's theory of the operational nature of physical concepts. A physical concept expresses a relation which is set up by an operation, and the concept is defined by the description of the operations.

I shall employ the term physical quantity to designate the number which is denoted by a descriptive function which arises from the concept of the relation between the number and the object.

According to traditional physics the laws of physics are functional relations between the magnitudes of variable physical quantities (particulars). According to the present formulation the laws of physics are expressed by functional relations between variable physical quantities (numbers).

5. SCALARS, VECTORS, AND TENSORS

We have already seen that physical quantities may be classified as extensive or intensive. We shall now study the classification of physical quantities as scalars, vectors, and tensors.

In order to discuss the nature of the various kinds of quantity I shall assume that we have given a rectangular Cartesian frame of reference, which consists of three mutually perpendicular rigid lines. The position of a point is specified by x, y, z, which are the perpendicular distances from the three coordinate planes, respectively. We shall allow a translation or rotation of the frame of reference.

A **scalar** is usually defined in the terminology of traditional physics as a quantity which has magnitude only. A better definition is that a scalar is a quantity whose magnitude is expressed by a number which is unchanged by a transformation of the frame of reference; thus a scalar is represented by an invariant. In accordance with our modified definition of a physical quantity a scalar is an invariant. The length of a segment, the mass of a body, are examples of scalars in classical physics.

In order to define a **vector** I shall first introduce the directed segment of a straight line. The projections of the segment upon the axes of coordinates are called the components and are denoted by Δx, Δy, Δz. Upon transformation of the frame of reference the components of the directed segment are transformed in accordance with a law which enables one to calculate the components in the new frame.

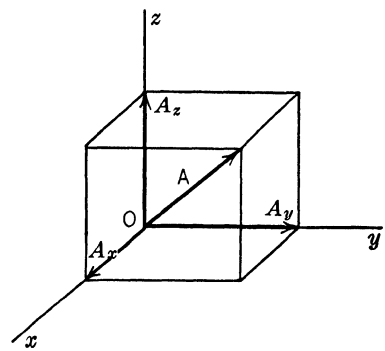


FIG. 1.

A vector is defined in traditional physics as a quantity which has magnitude and direction. A better definition is that a vector is a quantity which can be represented by the directed segment of a straight line, and which has components which are transformed according to the law for the transformation of the components of a directed segment. The vector **A** may thus be specified by the three components, A_x, A_y, A_z. In accordance with our modified definition of a physical quantity a vector is a set of three numbers which transform according to certain laws. I shall use the phrase absolute value, instead of magnitude, of a vector; the absolute value is represented by the length of a directed segment. The physical concepts of force, velocity, acceleration, give rise to descriptive functions which denote vectors. For example, we speak of the velocity of a body.

An invariant representation of physical laws is one which is unchanged upon transformation of the frame of reference. For

example, the expression of a law is invariant in form if it states that two vectors are equal. Thus

$$\mathbf{A} = \mathbf{B}$$

where **A** and **B** denote vectors, is invariant, for this equation stands for three equations which assert the equality of the components of the vectors,

$$A_x = B_x$$

$$A_y = B_y$$

$$A_z = B_z.$$

Upon transformation of the frame of reference the components of both vectors are transformed according to the same law. Thus in the new frame the law is

$$A'_x = B'_x$$

$$A'_y = B'_y$$

$$A'_z = B'_z$$

which is the same in form as before.

A **tensor** is exemplified by the stress at a point in an elastic medium. The stress is defined as the force per unit area acting across the surface between contiguous parts of the body. In order to specify the stress at a point we can consider three mutually perpendicular elementary surfaces, for example, surfaces in the coordinate planes. Then the stress on each surface is resolved into three Cartesian components. X_x, Y_x, Z_x are the components of stress on the surface perpendicular to the x axis. Altogether there are nine components of stress,

$$\begin{matrix} X_x & Y_x & Z_x \\ X_y & Y_y & Z_y \\ X_z & Y_z & Z_z \end{matrix}$$

$$X_y = Y_x, \quad X_z = Z_x, \quad Y_z = Z_y.$$

The stress at a point in an elastic body is a symmetrical tensor of the second rank. In accordance with our modified definition of physical quantity a tensor is a set of numbers, relative to some

frame of reference, which transform according to certain laws. If the components of one tensor are equal to the components of another in one system they are equal in all systems. Thus the expression of a physical law has an invariant form if it asserts the equality of the components of a tensor; likewise if it asserts that the components of a tensor are equal to zero.

The various physical quantities may be described as kinds of tensor. I have given an example of a tensor of the second rank; a vector is a tensor of the first rank, and a scalar a tensor of zero rank.

6. THE ELEMENTS OF ANALYSIS

Since the laws of physics are expressed by functional relations between variable physical quantities (numbers), it is therefore desirable to define some of the fundamental concepts of mathematical analysis.

A variable, x, is a symbol which may denote any one of a given class of elements; the elements of the class are the values of the variable. Examples of variables are variable real numbers. A variable y is said to be a single-valued function of x when if x is determined y is uniquely determined. We write $y = f(x)$. For example, if x and y represent real numbers,

$$y = x^2$$

is a function of x.

x is called the independent and y the dependent variable. A function is said to be continuous if a small change in the independent variable gives rise to a small change in the dependent variable. More precisely, $f(x)$ is continuous at $x = a$ if $\lim_{x \to a} f(x) = f(a)$.

Let x and y be real variables, and suppose that y is a continuous function of x. Then the derivative of y with respect to x is defined by

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

The derivative is the rate of change of y with respect to x. Suppose that z is a function of two independent variables x and y,

$z = f(x, y)$. Then $\frac{\partial z}{\partial x}$ is the partial derivative of z with respect to x. The definition of the partial derivative assumes that y remains constant. The derivative is employed to define the velocity, the acceleration of a body, etc.

The indefinite integral of y with respect to x is a function $U(x) + C$ such that

$$\frac{dU}{dx} = y.$$

$U(x)$ is denoted by $\int ydx$. Hence the indefinite integral is

$$\int ydx + C.$$

Integration is thus the inverse of differentiation.

Suppose that the range of the variable x between the values a and b is subdivided into n small intervals Δx_i. Let $f(x_i)$ be the value of $f(x)$ at any point in the interval Δx_i. Then the definite integral of $f(x)$ with respect to x between the limits a and b is defined by

$$\lim_{\Delta x \to 0} \{f(x_1)\Delta x_1 + f(x_2)\Delta x_2 + \ldots f(x_n)\Delta x_n\}.$$

Thus the definite integral is the limit of a sum. The fundamental theorem of the integral calculus states that the value of the definite integral of $f(x)$ between a and b is equal to the value of the indefinite integral for $x = b$ minus the value of the indefinite integral for $x = a$, i.e.,

$$\left[\int f(x)dx\right]_b - \left[\int f(x)dx\right]_a.$$

The definite integral is denoted by

$$\int_a^b f(x)dx.$$

In physics the work done by a force is a definite integral.

An ordinary differential equation is exemplified by

$$\frac{d^2x}{dt^2} + \omega^2 x = 0.$$

This special equation describes a vibratory motion of a point along a straight line. A solution of the differential equation is a function x of t which satisfies the equation. The general solution of a differential equation contains arbitrary constants; for example, the general solution of the given equation is

$$x = A \cos (\omega t + \delta)$$

where A and δ are arbitrary. The arbitrary constants can be determined by the assignment of initial conditions. For example, the assignment $x = x_0$ and $\frac{dx}{dt} = v_0$ at $t = t_0$, yields a particular solution of the differential equation.

A partial differential equation is one involving partial derivatives. Thus the propagation of a wave is described by

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + \frac{\partial^2 \varphi}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2 \varphi}{\partial t^2} = 0,$$

where φ is a function of x, y, z and t. The general solution of a partial differential equation contains arbitrary functions. A particular solution is determined by the assignment of initial conditions and boundary conditions.

The laws of physics are expressed by differential equations.

7. VECTOR ANALYSIS

A vector has been defined as a set of three numbers which are relative to a coordinate system. A special kind of vector consists of three numbers which are called the components of a directed segment. Some authors restrict the term vector to directed segments; in my terminology the **directed segment** is a particular which is represented by a vector which is a set of three numbers. Conversely, a directed segment may be employed to give a geometrical representation of a set of three numbers, that is, a vector. Thus we may use directed segments to represent the force, the velocity, the electric field, etc., such use expressing the fact that the set of three numbers constituting a vector is a unity. The representation of vectors by directed segments is called the vectorial representation of phys-

ical quantities. In order to express the laws of physics vectorially, a vector algebra and analysis have been developed.

Let the symbols **A** and **B** denote directed segments. A segment **A** has length, direction and sense. The segment $-\mathbf{A}$ is defined to have the same length and direction as **A**, but opposite sense. If m is a real number, the segment $m\mathbf{A}$ is defined to have the same direction and sense as **A** and length m times the length of **A**.

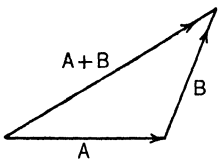


FIG. 2.

The sum of **A** and **B** is the segment which is formed by sliding **B** parallel to itself so that the origin of **B** coincides with the terminus of **A**. The segment which extends from the origin of **A** to the terminus of **B** is defined to be the sum of **A** and **B**. We write

$$\mathbf{A} + \mathbf{B} = \mathbf{C}.$$

The addition of segments obeys the commutative law

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A},$$

and the associative law

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}.$$

It is to be emphasized that the operation applies to **segments** and not to numbers.

The scalar product of two segments is defined to be a scalar equal to the length of **A** $\times$ the length of **B** $\times$ the cosine of the angle between them. The scalar product is denoted by

$$\mathbf{A}\cdot\mathbf{B}.$$

The scalar product satisfies the commutative law

$$\mathbf{A}\cdot\mathbf{B} = \mathbf{B}\cdot\mathbf{A},$$

and the distributive law

$$\mathbf{A}\cdot(\mathbf{B} + \mathbf{C}) = \mathbf{A}\cdot\mathbf{B} + \mathbf{A}\cdot\mathbf{C}.$$

The vector product of two segments is defined to be a segment whose length is equal to the length of **A** $\times$ the length of **B** $\times$ the sine of the angle between them, whose direction is perpendicular

to the plane of **A** and **B** and whose sense is in the direction of the thumb when the fingers of the right hand are curled in the direction in which **A** must be rotated to bring it in the same line as **B**. The vector product is denoted by

$$\mathbf{A} \times \mathbf{B}.$$

We have

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}.$$

We define the derivative of a segment with respect to a scalar variable t by a definition similar to the one for real variables. We write,

$$\frac{d\mathbf{A}}{dt}.$$

Now the operations defined for segments correspond to operations on the vectors which represent the segments. Thus $\mathbf{A} \cdot \mathbf{B}$ corresponds to

$$A_x B_x + A_y B_y + A_z B_z.$$

Conversely, the operations on vectors can be represented by operations on segments.

We have seen that physical vectors can be represented by segments; thus the operations on segments can be employed to represent the operations upon vectors in general; indeed the operations of the theory of segments have been defined in order to represent the important operations of physics.

In physical theory the force is a vector F_x, F_y, F_z, and the differential displacement is a vector dx, dy, dz. The work done by the force is

$$F_x dx + F_y dy + F_z dz.$$

This corresponds to the scalar product of two segments, and therefore the work can be represented by

$$\mathbf{F} \cdot d\mathbf{r},$$

where **F** is a segment which represents the force, and $d\mathbf{r}$ is a segment which represents the displacement. In this way we represent vectors by segments, and combinations of vectors by

operations upon segments. The advantage of this representation is that a single symbol represents the vector. Furthermore, the vectorial representation of physical laws is invariant in form, that is, it is independent of the position or orientation of the frame of reference.

In my interpretation the segment represents the vector (three numbers). Another interpretation is based upon the definition of quantity as the particular reality. We may deprive the laws of the theory of segments of all reference to segments and view them as a set of postulates for an abstract theory. Then we may apply the operations to vectors, not defined as numbers, but as quantities in the traditional sense. This procedure has been expounded by R. C. Tolman.

I shall make considerable use of the vectorial representation; to that end I shall define two quantities which are important for physics. Let **A** be a segment which represents a vector A_x, A_y, A_z; for brevity we usually say that **A** is the vector.

We introduce a scalar called the divergence of a vector. If A_x, A_y, A_z are the Cartesian components of **A** (a set of numbers) then

$$\text{div } \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}.$$

In a fluid the rate at which matter flows out of an element of volume per unit volume is div $(\rho \mathbf{v})$, where ρ is the density and **v** is the velocity. $-\frac{\partial \rho}{\partial t}$ is the rate of decrease of matter per unit volume. The conservation of matter is expressed by

$$\text{div } \rho \mathbf{v} = -\frac{\partial \rho}{\partial t},$$

$$\text{div } \rho \mathbf{v} + \frac{\partial \rho}{\partial t} = 0.$$

In a steady state $\text{div}(\rho \mathbf{v}) = 0$, that is, the rate at which matter flows in through one part of the surface is equal to the rate at which it flows out through another part of the surface.

I now define the curl of a vector **A.** Curl **A** is a vector whose Cartesian components are

$$\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}$$

$$\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}$$

$$\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}.$$

A conservative field of force is such that the curl of the field is zero. Now if an electric charge is carried around a closed path in an electrostatic field the work done is zero. This is expressed by curl $\mathbf{E} = 0$.

8. STAGES OF PHYSICAL SCIENCE

Having examined the basic concepts of physical theory such as space, time, body, quantity and magnitude, and having sketched the elements of mathematical analysis and vector analysis, we are now prepared to consider more specifically the problem of this book: the analysis of the concepts, principles and systems of physical theory.

As I have previously noted, physical science is a union of inductive and deductive factors; in the course of its development physics becomes more and more deductive. Our experience of the physical order begins with the experience of particular aspects of reality. This is accompanied by the observation of relations between aspects, by the construction of concepts of sense qualities and relations, by the discovery of correlations between qualities and relations, and by the construction of concepts of bodies which are characterized by qualities and relations. Quantitative physics begins when we correlate some qualities or relations with numbers; the number substitutes for the quality or relation. The assignment of numbers is based upon the observation of space-time coincidences; the number is denoted by a descriptive function, for example, 2 = the length

in meters of this rod. One then seeks functional relations between variable numbers; these functional relations are constructed into a deductive system by the methods of pure mathematics.

In order to characterize the stages of physics I shall distinguish between

experimental physics
theoretical physics
mathematical physics.

The data of **experimental physics** consist of space-time coincidences. The results are sets of assignments of numbers to descriptive functions.

The data of **theoretical physics** consist of sets of assignments of numbers to descriptive functions. The results are the discovery of functional relations between variables the values of which are the numbers.

The data of **mathematical physics** consist of functional relations between variables. The results are implications between these functional relations and others.

I shall illustrate the preceding discussion by considering an experiment in which a ball is rolled down an inclined groove. In preparation for the experiment a coordinate system is constructed for the purpose of determining the position of the ball at different times. A standard meter rod is placed along the groove so that one end of the rod coincides with the top of the groove, to which the coordinate 0 is assigned, the point of the groove which coincides with the other end of the rod being given the coordinate 1. It is not necessary to describe further the construction of the coordinate system. There is also a clock which ticks seconds.

The experiment is then as follows. We place the ball at the top of the groove, which has the coordinate 0, and release the ball upon hearing a tick of the clock, which we shall label 0. Thus the coincidence of the ball and the point marked 0 is simultaneous with the tick of the clock. This space-time coincidence is a datum. We assign to the ball

$$\begin{aligned} \textit{the coordinate of the ball} &= 0 \\ \textit{the time} &= 0. \end{aligned}$$

The ball rolls down the incline and at the second tick of the clock we observe that the ball coincides with the point marked 1. We write down

$$\textit{the coordinate of the ball} = 1$$
$$\textit{the time} = 1.$$

At the third tick one observes that the ball coincides with the point marked 4, etc. One records the observations in the following table.

t — the time in seconds.....	0	1	2	3	4
s — the distance in meters.....	0	1	4	9	16

This table records the assignment of numbers to descriptive functions on the basis of the observation of space-time coincidences.

The preceding table is the datum for theoretical physics. One discovers that the numbers satisfy the relation

$$s = \tfrac{1}{2}gt^2$$

where g is a constant. The theoretical result is expressed as follows:

If

s = the distance passed over,

t = the time.

then

$$s = \tfrac{1}{2}gt^2.$$

The reference to physical reality is expressed by the conditions: if s = the distance passed over, etc. Using the terminology of Professor Swann, one would say that it is possible to assign numbers s and t such that $s = \frac{1}{2}gt^2$ is true. However, I have an even more sophisticated formulation: It was possible to assign numbers s and t such that $s = \frac{1}{2}gt^2$ is satisfied. This formulation presupposes the validity of memory. I can not find a more skeptical interpretation, however, without risking the dissolution of consciousness. In any case, my mode of expression avoids the rash predictions which are usually exemplified by that favorite of logicians, " The sun will rise tomorrow."

The relation $s = \frac{1}{2}gt^2$ is a datum for mathematical physics. A result of mathematical physics is: If

$$s = \tfrac{1}{2}gt^2,$$

then

$$\frac{d^2s}{dt^2} = g.$$

Nothing is said about the assignment of numbers.

9. THE ORIGIN OF PHYSICAL CONCEPTS

The occurrence in physical theory of concepts derived from experiment, and the employment of the functional relations of pure mathematics, suggest the problem of the origin and nature of physical concepts. What are the respective contributions of experience and reason to their content? We are thus prompted to ask the question, are physical concepts abstractions from experience or are they invented? This question may also be phrased, are universals discovered or invented?

All concepts are invented in the sense that if there were no mental activity of some sort there would be no construction of concepts. Using terms more precisely, I should say that some concepts are abstracted from experience and some are invented.

Concepts of sense qualities are directly abstracted from experience. Consider, for example, the statement that a particular aspect is an instance of the universal redness. This means that the aspect is similar to another in a certain respect. Hence the problem becomes, are similarities discovered or invented? If the point of view may be chosen arbitrarily then similarity is invented, but it appears to me that similarity is something found, because a question such as: are two tactual aspects of a table similar in color? is meaningless. The point of view must be appropriate to the particulars compared. Thus I think that fundamental concepts such as those of sense qualities are abstracted from experience.

As previously defined, the concept of thing is an abstraction from experience; it is defined in terms of the observed similarities and correlations of aspects,

The fundamental physical concept is that of the **space-time coincidence** of two points—the similarity of two points with respect to spatial position and temporal position. I assume that the concept of space-time coincidence is an abstraction from experience. It is with this concept that we test the applicability of a physical concept to the bodies given in experience.

The assignment of numbers by the operation of measurement seems to be a process of invention. However, once the standard or coordinate system has been selected the assignment of the number is by the observation of coincidences. The concept of a relation between a number and an object is defined by an experiment and is abstracted from experience in a secondary sense.

Some concepts, although grounded in experience, express a content which involves a conventional element. Thus the concept of length presupposes that the standard is unchanged upon displacement.

Some concepts are definitely inventions. They are combinations of concepts abstracted from experience or modifications of these concepts; for example, the concepts of the kinetic theory of gases and the general theory of relativity are invented. Initially, invented physical concepts are hypotheses to be tested for applicability to experience, applicability being ultimately determined by the observation of space-time coincidences.

Concepts play a double rôle. Those concepts which have been abstracted from experience are descriptive of previous experience; they serve as principles of order for future experience. Such concepts are *a posteriori* in their origin and *a priori* with respect to the interpretation of future experience.

Concepts which are abstracted from experience, or which are definable in terms of concepts abstracted from experience, are approximate, their objects are not strictly invariable. Thus the concept of the space-time coincidence of two points, and all physical concepts of the assignment of numbers by measurement, are approximate. Hence experimentally determinable physical quantities do not exactly satisfy a definite functional relation between variables.

The concepts of pure mathematics are concepts of objects

and not of assignments. Assuming that mathematical concepts are precise, one may then propose the view that mathematical concepts are inventions, or the apprehension of a Platonic realm of pure forms. Either the invention or apprehension is suggested in many cases by physical concepts. In objective terms one may say that the objects of pure mathematics may be viewed as ideal structures, which may be held to be constructed in pure intuition (Kant), or may subsist in a realm of pure forms. The properties of mathematical objects are definable by postulates which constitute an implicit definition of the concepts of the objects.

Some mathematical objects are assignable to experience approximately; thus the real numbers are assignable to physical objects. Despite the approximate character of physical measurements we write the equations between the physical quantities as if the values of the variables were not selected for the purpose of assignment to physical objects. Since the numbers are not assignable precisely they satisfy determinate functional relations approximately.

In view of the preceding discussion it is convenient to distinguish between three types of systematic physics.

1. **Theoretical physics** consists of propositions of the following kind:

If x = the distance of a particle from an origin and t = the time, then

$$\frac{d^2x}{dt^2} + \omega^2 x = 0.$$

This is the equation of motion for a particle moving with simple harmonic motion. On account of the approximate character of the propositions of theoretical physics one conceives of an ideal theoretical physics.

2. **Ideal theoretical physics** consists of propositions like the following:

If x = the distance as ideally measured with an ideal measuring rod, t = the time as ideally determined, then

$$\frac{d^2x}{dt^2} + \omega^2 x = 0,$$

Ideal physics makes assertions about ideal particles, ideal rigid bodies, ideal fluids, etc. Classical theoretical physics may be viewed as ideal. The idealization of physical objects raises the problem of the metaphysical status of these ideal entities.

3. **Mathematical physics** consists of propositions like the following:

If $$x = x_0, \quad \frac{dx}{dt} = v_0 \quad \text{at } t = 0$$

and

$$\frac{d^2x}{dt^2} + \omega^2 x = 0,$$

then

$$x = x_0 \cos \omega t + \frac{v_0}{\omega} \sin \omega t.$$

This proposition does not presuppose that the numbers are assigned.

10. THE INDUCTIVE EXPOSITION OF PHYSICAL THEORY

In order to analyze physical theory I have chosen the inductive method of building up the structure of theoretical physics. We shall see how physical theory arises out of observation and experiment; how experimental physics furnishes the basis for theoretical physics. This method will enable us to study in detail the respective contributions of experience and reason in the constitution of physical science.

Now the conscious development of science presupposes the crude concept of a world of real things in space and time. The initial phase in the construction of science is qualitative; one selects from the background of unreflective experience certain qualities of things which are made the objects of special study and examination. Thus a qualitative study of the physical order would be based upon the recognition that bodies are long or short, in motion or at rest, heavy or light, hot or cold, capable of attracting light bodies, capable of attracting iron. A phase of qualitative physics is the recognition that qualities are constant or variable. The qualitative investigation of physical objects originated in earliest antiquity, but the system-

atic attempt at a qualitative physics originated among the early Greek philosophers, who sought to explain manifold qualities as attributes of one or more substances. The qualitative phase of physics is an essential constituent of even the most highly developed system of physical theory, for the ultimate significance of physical concepts lies in their connection with immediate qualitative experience. Nevertheless, the significant fact about physical theory is that in the study of the physical order it has been possible to supplement qualitative physics by quantitative physics.

Quantitative physics arises when we correlate some qualities or relations with numbers. A particular body may thereby be characterized by physical quantities, that is, the numbers assigned to its properties. Physical quantities are denoted by descriptive functions; for example,

$$l = \textit{the length in meters of this rod.}$$

The concept of length is that of a relation of a number to a property of a body with respect to the same property of a standard. In general, a physical concept is defined by the description of the operation of establishing the relation. The assignment of a number by direct measurement involves the observation of space-time coincidences. The problem of the construction of physical theory is one of expounding

(*a*) *the definitions of fundamental quantities,*
(*b*) *the definitions of derived quantities,*
(*c*) *the laws and principles which are expressed by functional relations between variable physical quantities.*

Now the definition of fundamental quantities and the inductive discovery of functional relations between them can not be carried out wholly in a straightforward manner. All qualities are relative, for example, the redness of a thing is functionally related to other properties. Similarly, the physical quantities which are assigned to physical objects depend upon accompanying conditions. The definition of a fundamental quantity, therefore, must proceed by successive approximation. We first define a quantity, that is, we describe a method of assigning

a number to a property, upon the assumption that conditions as estimated qualitatively are invariable. Then we discover laws correlating the physical quantity and qualities. These laws then enable us to define quantitatively to the first approximation the conditions under which the quantity is defined to the second approximation. For example, we may define the length of a body as the number obtained by an operation with a standard rod whose qualitative temperature is that of ordinary conditions on the earth. One then discovers a correlation between the length of a rod and its hotness or coldness. We may then define a scale of temperature in terms of the variations in length of some standard rod. This enables us to define quantitatively the conditions of temperature under which the standard of length is defined to the second approximation.

Now in the development of a system of physical theory the definition of the fundamental quantities undergoes a change. The fundamental quantities are first defined independently, then laws are discovered which can be transformed into implicit definitions of the quantities. Thus for Archimedes force is defined in terms of the weight of a body. Newton constructed dynamics into a system based upon the three laws of motion; these laws may now be viewed as postulates which implicitly define the fundamental dynamical quantities. Again, the magnetic field was first defined in terms of the force due to a magnetic pole; but after the discovery of the law which correlates an electric current and the field, the field may be defined in terms of this law, which thereby becomes a definition. The temperature of a body was first defined in terms of a thermometer, such as a gas thermometer. Then the second law of thermodynamics was discovered and can be transformed into a definition of temperature. Thus in the construction of a physical theory there is a union of definition and experimental law. In order to improve the synthetic unity of a theory one is led to change the fundamental definitions so that experimental laws are transformed into definitions and definitions give rise to experimental facts. Fundamental and derived quantities may thereby change places. The ultimate goal of physical theory is the implicit definition of all fundamental quantities by a set of postulates.

Again, with the development of a more refined technique laws may be discovered to be only first approximations. Upon the approximate theories as a basis one may then build theories of higher approximation, for example, the theory of relativity and quantum theory. It must be emphasized, however, that the definition of the experimental conditions under which the classical theories are revealed to be inadequate requires the applicability of the classical theories to the first approximation. Thus the verification of the general theory of relativity presupposes the employment of pre-relativistic physics in the construction of the apparatus, in the determination of the observations, and in the calculation of the mass of the sun which is substituted in the formula. Likewise the definition of the experimental conditions for the verification of the quantum theory presupposes the approximate truth of the classical theories.

In this book I propose to build up physical theory inductively by the method of successive definition and successive approximation. Not only does this schematically reproduce the historical order of discovery, but it is also the order of proof. Furthermore, the proposed method reproduces the stages in which the contemporary student learns physics and would enable him to correlate the several methods of defining the fundamental concepts.

We are now prepared to begin the systematic construction of physical theory. The fundamental empirical concept of physics is that of body, and the fundamental datum is the observation of the space-time coincidence of points. Ultimately the applicability of all physical concepts is tested by the observation of space-time coincidences. The problem of physical science might be described as the definition of the concept of the physical order, but this definition must come at the end of an inductive exposition. If I may borrow a phrase from Bosanquet, physical reality lies ahead, not behind.

REFERENCES

Section 1

CARNAP, R. Der logische Aufbau der Welt. Berlin, 1928.
LEWIS, C. I. Mind and the World Order. New York, 1929.
RUSSELL, B. Our Knowledge of the External World. New York, 1929.

Sections 2, 3, 4

BRIDGMAN, P. W. The Logic of Modern Physics. New York, 1927.
CARNAP, R. Abriss der Logistik. Vienna, 1928.
CARNAP, R. Physikalische Begriffsbildung. Karlsruhe, 1926.
COUTURAT, L. Les principes des mathématiques. Paris, 1905.
DUHEM, P. La théorie physique. Paris, 1906.
EDDINGTON, A. S. The Mathematical Theory of Relativity. Cambridge, 1923.
RUSSELL, B. The Principles of Mathematics. Cambridge, 1903.
TOLMAN, R. C. The Measurable Quantities of Physics. *Physical Review*, Vol. 9, 1917.
WHITEHEAD, A. N. The Principle of Relativity. Cambridge, 1922.
WHITEHEAD and RUSSELL. Principia Mathematica. Cambridge, 1910, 1912.

Section 6

BLISS, G. A. The Function Concept, *in* Monographs on Modern Mathematics. Edited by J. W. A. Young, New York, 1915.
OSGOOD, W. F. Calculus. New York, 1916.

CHAPTER II

EUCLIDEAN GEOMETRY

1. SPACE AND GEOMETRY

THE concept of **space** is one of the fundamental concepts employed in the description of the physical order. Bodies and phenomena are conceived to have position in space; thus motion is described as change of position in space. In view of the generality of application of the concept of space its analysis will reveal a general character of an important part of reality.

In the study of space we shall analyze the concepts and propositions of geometry, which is characterized as the science of the properties of space. It is more accurate, however, to say that geometry in the first instance is a study of the properties of geometrical figures, that is, of constructions in space. A theory of space is based upon the relations exhibited by geometrical figures. A fundamental principle of classical physics was the proposition that space is Euclidean. The significance of this principle we shall determine from a study of Euclidean geometry.

The definition of physical concepts proceeds by successive approximation. In order to define the concepts of geometry we must initially assume that we can make qualitative estimates of accompanying conditions. Thus we presuppose the concept of body as something which persists in time; we assume that bodies remain at rest during our geometrical investigation, except when they are displaced in an experiment, that the state of hotness, stress, electrification remains constant. On this basis we proceed to define the geometrical properties of bodies, to discover laws, and to introduce the concept of space. One can then define the conditions of geometrical experiments

quantitatively to the first approximation, and then define the geometrical concepts to the second approximation.

Furthermore, the development of physical concepts is a process of successive definition. In this chapter I define space for bodies at rest with respect to the surface of the earth; later I shall give a more general definition of space.

I shall begin with a statement of some historical views about the nature of space. In Democritus, space is the **void** which renders possible the motion of the atoms; in Descartes, **extension** is the defining attribute of matter; in Leibniz, space is the **order of co-existence** of things; in Kant, space is an *a priori* form of **intuition.** Each of these historical views reveals an essential element of the concept of space. Our theory of space must account for the following characteristics.

(1) *Space is the ground of the possibility of motion of bodies. Motion is change of position in space.*

(2) *Space is the ground of essential characters of bodies in the physical world. An essential character of a body is its extension.*

From this point of view space appears to be a quality.

(3) *Space is an order of bodies.*

Thus bodies in a straight line have an order which may be expressed by the propositions, A is to the left of B, B is to the left of C, C is to the left of D, etc. Space is thus the ground of relations between things in space.

(4) *Bodies in space are known through sensation.*

In the present exposition of space I shall assume that the most fundamental characteristic of space is expressed by the proposition that things have **position** in space. For example, the position of A is to the left of the position of B. I therefore describe space as the manifold of positions. Now we must admit of positions that are not occupied at a given time. A more adequate description of space is therefore given by the following proposition: *Space is the totality of possibilities of relative position of bodies and phenomena.* The notion of possibility leaves open the possibility of empty space.

A theory of space may be an absolute theory or relational theory. An absolute theory conceives of space as a manifold of **absolute** positions; in a relational theory space is a manifold of **relative** positions. The difference between the two theories can be described in general terms as follows. Space may be said to be constituted of complexes of the form aRb; such a complex consists of terms a and b standing in the relation R. Now the relation R presupposes the terms a and b, but the terms may or may not presuppose the relation. In an absolute theory the terms do not imply the relations between them; the concept of position does not presuppose relations to other positions. In a relational theory, however, the terms presuppose relations to other terms; the relations between positions are internal relations, that is, the concept of position in space implies relations between them. I have adopted a relational theory of space and therefore have defined space as the **totality of possibilities of relative position** of things and phenomena.

The preceding description of space ascribes no definite form to it—space thus far is an amorphous manifold. However, we think of space as having a **structure**, for the very problem as to whether the properties of space are described by Euclidean or a non-Euclidean geometry presupposes that we ascribe it a definite relational structure.

The theory of the structure of space is based upon the properties of a class of bodies in space. Specifically, the entity in terms of which the structure of space is defined is the **solid body**. Assuming that we have a crude concept of solid body—examples are an iron rod, a block of stone, a piece of hard wood—we may state that the structure of space is determined by the possibilities of **relative position** of solid bodies.

My analysis of the structure of space assumes, therefore, the concept of solid body. Solid bodies may stand in various relations; configurations of solid bodies have properties which are the subject matter of geometry. The general propositions of geometry express the properties of certain special figures, such as points, lines, surfaces, regular solids, etc. Points are exemplified by a dot or a pin point; lines, by a thin cord, a ray of light, a thin wire, the edge of some solid body. A surface is

exemplified by a thin sheet of paper; a regular solid is exemplified by a die, the earth, etc. Thus in the first instance geometry is a branch of physics. The subject matter of physical geometry is the general laws of the possibilities of relative position of solid bodies.

Geometrical configurations embody different types of relation. We may consider the following preliminary classification of geometrical properties into topological, projective and metrical properties. The **topological** properties do not presuppose the concept of measurement and may be characterized as qualitative. For instance, the distinction that may be drawn between surfaces having one side and surfaces having two sides is topological. The topological properties of a geometrical figure are unchanged by a continuous point-to-point deformation, so that in topological geometry the concepts of straight line and length do not occur. The concept of **projective** properties presupposes that of the straight line, but not of length. The projective properties are unchanged by a projective transformation. **Metrical** properties presuppose the concept of length; Euclidean metrical properties are unchanged by a set of transformations which includes translation, rotation, the similarity transformation and reflection. The different types of property being the subject matter of different geometries, we may have topological geometry, projective geometry, metrical geometry, etc. I begin the discussion with an account of the fundamental concepts of metrical geometry.

2. METRICAL GEOMETRY

The construction of metrical geometry is based upon the properties of a special class of solid bodies, namely, practically rigid bodies, such as iron rods, blocks of stone, sticks of hard wood, etc. The meaning of the concept of practically rigid body will be exhibited in the following discussion.

The formation of metrical concepts in geometry is undoubtedly preceded by qualitative ideas, for one may declare a certain rod to be long, another to be short, without having a metrical concept of length. Again, it is undoubtedly a primitive experience that upon displacement a rod appears to retain the

same spatial extension. It is to be expected that the metrical concepts will agree with our rough and unanalyzed comparisons.

In order to explain the metrical concepts let us suppose that we have given a number of practically rigid bodies such as iron rods. Upon a rod A we mark two points between which there is a certain relation; upon a second rod B we mark two points which coincide respectively with the points on the first rod. The concept of **coincidence** is a fundamental concept whose meaning is to be understood only from the observation of two coincident points. We then observe that if the points on A coincide with the points on B, then the points on B coincide with the points on A. Also, if the points on A coincide with those on B, and the points on B coincide with those on C, then the points on A coincide with those on C; we express these results by the statement that the two pairs of points are **congruent.** Thus the relation of congruence is based upon a symmetrical and transitive relation between two pairs of points, and the principle of abstraction which asserts that such a relation is to be interpreted as the possession of a common property. Congruence is equality with respect to **distance** and is an example of similarity with respect to a given property. The coincidence of two pairs of points is thus interpreted as the possession of a common distance—the distance between the points of each pair is the same. I have chosen the concept of distance as fundamental in accordance with the relational theory of space. However, as previously explained, we may think of a pair of points as determining a segment. A segment has an extension which is definable in terms of the distance between the two points; in this way we define extension as an attribute (Descartes) in terms of a relational theory of space (Leibniz).

The preceding discussion explains distance or extension at a place; we must now extend the concept throughout space. The possibility of this extension is based upon the following experiments, in the description of which it is convenient to think of the pair of points as determining a segment. Suppose that the terminal points of the segments A and B coincide. If I change the orientation of the segments, or if I displace the segments, the coincidence is not disturbed. Next, suppose

that one segment is kept fixed, and the other is displaced in any manner whatsoever and then brought back to its original position; the terminal points are found to coincide again. To this fact of experience we now add a definition. We postulate that *when two congruent segments are separated in space they remain congruent.* Since the motion of a segment may be analyzed into translation and rotation, we have postulated that the concept of congruence is invariant under a translation and rotation. One might ask whether distance or extension really remains unchanged during a translation or rotation. At this fundamental stage of the discussion, this is a meaningless question, for it presupposes that we have previously defined our metrical concepts in terms of which distance may be determined.

In order to measure the extension of any physical segment determined by points on a practically rigid body, it is necessary to choose some segment as a standard of extension, by comparison with which the measure of extension of the segment may be determined. The operation of measuring extension consists in counting the number of times that the standard segment can be laid off on the line. One terminal point of the standard segment is brought into coincidence with one end point of the segment and the position of the other terminal point of the standard is marked on the body. The standard segment is then displaced so that its first point coincides with the point marked on the body, and so on. We thus see that the measurement of an extension requires the observation of coincidences of points. The operation performed in measurement is a method of realizing the conditions under which one can observe congruence of a part of a segment with the standard segment. The **number** which is obtained by the operation of measurement is called a physical quantity, and the outcome of the measurement is expressed by a statement such as: *The measure of extension of this rod is* 2 *meters.* In the preceding chapter I used the term length for the physical concept of extension—thus the length in meters of this body is 2. I have used the concept of extension in order to designate the measurable property, and the concept of length to designate the relation between the number and the property. For other measurable

IL 999

properties, however, this distinction in language does not occur. The word distance, for example, is employed for the measurable relation which in some sense is given in experience, and also for the physical relation between the measure of distance and the distance.

With the exposition of the method of measuring extension we have completed our account of the physical concept of length. The concept of length is a concept of a method of assigning numbers to segments with respect to a standard. The choice of the standard is determined partly by experience and partly by convention; the standard must be chosen from among the bodies which we know to have a definite length as determined by our empirical criterion, but the particular body chosen is a matter of convention. The selection of some body given in experience as the standard of length has been called by Reichenbach a "Zuordnung" or correlation between some physical body and our concept.

To summarize: we have seen that the fundamental concept of metrical geometry, that of length, is constructed by a definition based upon the experiences of the coincidences of points on practically rigid bodies. The concept of length is here described as an abstraction from experience plus the definition that the character remains invariant during a motion.

In the preceding account of the construction of a physical concept one must recognize the use of matching. The qualitative concept of length is based upon an estimate of coincidence. However, the estimate of coincidence is also an estimate of similarity, but in this case the conditions are such that the problem is more restricted and there is better agreement between consecutive estimates. Thus ultimately quantity is definable in terms of quality.

The preceding discussion of length is only a first approximation. It is, however, adequate to the formulation of laws in terms of length. With the aid of the physical measure of length one can define the measure of temperature, the temperature, which is the physical quantity correlated with the hotness or coldness of a body as determined by our temperature sense. Assuming, then, that the standard of length is constant, one can discover the empirical proposition which expresses the

linear expansion of a rod as a function of the temperature. Then, by adding the condition that the standard of length is to be at a given temperature, one obtains the second approximation in the definition of length. The standard of length in physics is the extension, called the meter, between two lines on a certain bar at zero degree Centigrade. One can further discover laws which reveal that the length of a body changes upon the application of force. Since the measure of force is usually defined in terms of the measure of length, the physical concept of force must not be used in the initial definition of length; as a higher approximation, however, we need to state that the standard of length is not to be subjected to external forces.

The discussion of the definition of length illustrates the general procedure which must be employed in order to define a set of physical concepts. In the definition of the first physical concept one starts with a qualitative concept and a setting of conditions designated by hotness, force, etc., which are assumed to be reproducible. The definition of one physical concept enables one to define other physical concepts which serve to define the conditions under which the first concept is defined to a second approximation.

On the basis of the concept of length we are now able to proceed to the building up of physical geometry. Our first task is to define a physical straight line. Physical lines are exemplified by wires, strings, rays of light, etc. I shall define a straight line as the line of shortest length between two points. A hanging cord to which a weight is attached is an example; the length of the cord between two points is shorter than the length of any other cord. The most important example of a straight line is a ray of light in a homogeneous medium, indeed, we might define a straight line in terms of a ray of light. Practically, comparison with a ray of light furnishes the test of a straight line, such as a straight edge. With the aid of the concept of straight line one can construct a plane surface; surfaces are exemplified by the boundaries of solids. A plane may be defined as a surface such that if any two points on a straight line lie in the surface all points lie in the surface. If I place a straight edge on the table top, the straight edge is in

contact with the table top at all points for all positions of the straight edge; hence the table top is a physical plane surface. Similarly we can construct a surface such that all points of the surface are equidistant from a given point. We need also the concept of angle. An angle is enclosed by two straight half lines which originate in a common point. The congruence of two angles is defined in terms of the coincidence of the lines enclosing them. A line which originates in a point dividing a straight line forms supplementary angles with the two half lines. An angle which is congruent to its supplement is defined to be a right angle. One can further define concepts for the measurement of areas and volumes.

Limiting the discussion to the plane, we can construct such plane figures as triangles, circles, etc. If we measure the sides of a right-angled triangle we discover that the physical measures approximately satisfy the Pythagorean theorem: the square of the length of the hypotenuse is equal to the sum of the squares of the lengths of the other two sides. Similarly, we discover that the sum of the angles of a triangle is equal to two right angles. By these methods one arrives at a set of propositions, expressing relations between variable physical quantities (numbers), which constitute physical plane Euclidean geometry.

3. SYNTHETIC AND ANALYTIC GEOMETRY

The fundamental concepts of distance, length, angle, etc., having been set forth, there are two methods of procedure which may be employed for the further development of geometry: the synthetic method and the analytic method. In the **synthetic** method the propositions of geometry directly express properties of geometrial figures: for example, the sum of the angles of a triangle is two right angles; the diagonals of a parallelogram bisect each other. I shall later point out that an ideal or abstract system of geometry in terms of the synthetic method is based upon a set of postulates patterned after Euclid's *Elements*.

A more powerful method in geometry is the **analytic** method, which employs a coordinate system. As an example I shall describe the rectangular Cartesian system of coordinates in

the plane. This system is based upon two straight lines at right angles to each other, called coordinate axes. The position of a point in the plane is described by its coordinates, which are the perpendicular distances of the point from the axes. One may also think of the coordinate system as a lattice of squares constructed out of unit measuring rods. The coordinates of a point are usually denoted by the symbols x and y, and different values of x and y represent corresponding points in the plane. Now a straight line is the locus of points whose coordinates approximately satisfy a certain linear equation. Similarly a circle is the locus of points whose coordinates approximately satisfy a certain equation of the second degree. Thus the properties of geometrical figures can be represented by equations. The analytic method will be largely used in this book.

4. GEOMETRICAL TRANSFORMATIONS

By the employment of a coordinate system the analytic method introduces an extrinsic element into the analysis of geometrical figures. Hence a problem of the analytic method is the determination of those functions of the coordinates which express intrinsic properties of figures. A method for the solution of the problem is offered by the theory of geometrical transformations. By way of introduction to the characterization of geometrical properties by means of transformations I shall consider the figures of Euclidean geometry. Suppose, for example, we displace a triangle so that it remains parallel to itself; in such a displacement, which is called translation, the lengths of the sides and the measures of the angles are not changed. Again, the translation of a parallelogram does not change the lengths of the sides and their parallelism. Another displacement which does not change the lengths of sides or the measures of angles is a rotation about a fixed point. Thus the metrical properties of Euclidean geometrical figures are not changed by translation or rotation. Let us now describe translation and rotation analytically. The points of our figures are described by Cartesian coordinates, x, y, z. Translation or rotation transforms the points of a figure into other points of space; analytically expressed, the statement is that the point

x', y', z' is correlated with the point x, y, z where x', y', z' are functions of x, y, z. The setting up of such a correspondence between points is called a **transformation**: translation and rotation are thus transformations of the points of a figure into the points of another figure. Now the geometrical properties of figures are expressed in terms of functions of the coordinates of the points; for example, the length of a side of a triangle is

$$s = \{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2\}^{\frac{1}{2}},$$

where x_1, y_1, z_1 and x_2, y_2, z_2 are the coordinates of the end points. In general, functions of the coordinates are changed by a transformation. In geometry, however, one is interested in those functions of the coordinates which are **invariant** under transformations. Thus the length of a line is invariant under translation and rotation. The geometrical properties of figures are best classified by finding the transformations under which they are invariant. I shall, therefore, describe some of the principal types of transformation.

In a given set we may include the following four transformations:

(1) *translation*;
(2) *rotation about a fixed point*;
(3) *reflection, in which a point with coordinates* x, y, z *is transformed into a point with coordinates* $-$x, $-$y, $-$z;
(4) *similarity transformation, by which a figure is transformed into a similar figure.*

All these transformations are expressed by linear transformations of x, y, z into x', y', z'. Euclidean metrical geometry may be characterized as the set of propositions which express properties which are unchanged by such linear transformations; it is the invariant theory of these linear transformations. All non-invariant relations between coordinates do not belong to geometry; for example, the proposition that a point has coordinates 4, 5, 7 does not belong to geometry, but to a science which studies the points as individuals. On the other hand, the function of the coordinates which we call the distance

between two points is invariant under transformations 1 to 3, or is multiplied by a factor independent of the position of the point by the similarity transformation. A similar significance is possessed by the proposition that two straight lines form a definite angle. Thus we see that invariance under a certain set of linear transformations characterizes the analytical expressions which represent the geometrical properties studied in Euclidean metrical geometry. By considering other types of transformation we obtain other geometries.

Let us next consider the **affine** transformation, in which x', y', z' are arbitrary integral linear functions of x, y, z. Thus

$$x' = a_1x + b_1y + c_1z + d_1$$

where a_1, b_1, c_1, d_1 are constants. The affine transformation contains as special cases the linear transformations which characterize metrical geometry. The affine transformation transforms a plane into a plane, a straight line into a straight line, a system of parallel straight lines into a system of parallels. In general, the length of a line and the measure of the angle between two straight lines is changed by the affine transformation; thus a sphere is transformed into an ellipsoid. If we now select from the concepts and theorems of metrical geometry those which are unchanged by arbitrary affine transformations we obtain affine geometry, which is thus the invariant theory of the affine transformation. This geometry does not contain the concepts of distance and angle, and in it the distinction between the ellipse and circle vanishes. There remains the distinction between finite and infinite regions of space, and the concept of the parallelism of two straights.

I next mention the **projective** transformation, which includes the affine transformation as a special case. The analytical definition is that x', y', z' are fractional linear functions of x, y, z which have the same denominator. In a projective transformation the parallelism of two straight lines need no longer be maintained. Projective geometry may be now defined as the invariant theory of the projective transformations.

Lastly, we may consider the most general transformation, one which establishes a one-to-one and continuous transfor-

mation of points. Such a transformation may be exemplified by the arbitrary deformation of a solid piece of rubber without tearing. The properties of figures which are unchanged by such general transformations constitute the subject matter of topological geometry, or analysis situs. An example of a topological concept is the distinction between a surface of one side and a surface of two sides. It is easy to construct a surface with only one side, and this property is unchanged by all continuous point-to-point transformations. Topology is the invariant theory of continuous point-to-point transformations: one may say that the subject matter of topology is the qualitative properties of geometrical figures.

5. THE CONCEPT OF GROUP

The foregoing discussion may now be completed by introducing the concept of **group.** Suppose that we have given a set of translations a, b, c. We shall consider an operation upon translations in which a translation a is succeeded by a translation b, symbolized by aob. The identical transformation i is the one which leaves a figure unchanged, and the inverse a' of a translation a is like the direct translation but opposite in sense. Then we can verify the following laws for the operation upon translations:

(1) *If a translation is succeeded by a translation the result of the operation is a translation.*
(2) *The operation upon translations obeys the associative law,* ao(boc) = (aob)oc.
(3) *If a given translation is succeeded or preceded by the identical translation the result is the given translation,* aoi = ioa = a.
(4) *A translation succeeded by its inverse yields the identical transformation,* aoa′ = i.

The foregoing laws which hold for translations may now be transformed into a set of postulates which serve to define the formal properties of a group of elements with respect to an operation. In abstract terms, a set of elements forms a

group with respect to an operation, if the following postulates are satisfied.

(1) *The result of an operation upon two elements is an element of the group.*
(2) *The operation obeys the associative law.*
(3) *The identical element exists.*
(4) *Corresponding to every element there is an inverse element.*

Now the various transformations which we have discussed always form a group. Thus the linear transformations which leave the metrical properties unchanged form a group, as do, respectively, the affine transformations, the projective transformations, and the continuous point-to-point transformations. For example, Euclidean metrical geometry is the invariant theory of the group of translations, rotations, reflections and similarity transformations. Affine geometry is the invariant theory of the affine group. Thus the interpretation of geometry as the theory of the invariants of groups of transformations affords a systematic method of classifying geometrical properties.

6. THE STRUCTURE OF SPACE

In the foregoing discussion we have seen that physical geometry is the study of the properties of geometrical figures, exemplified by physical points, lines, surfaces, and solid bodies. The properties of practically rigid solid bodies are of especial interest; figures constructed of practically rigid bodies and measured by a practically rigid measuring rod are described approximately by the propositions of Euclidean metrical geometry. To the question, what is meant by the statement that space is approximately Euclidean? we are now ready to give the answer that geometrical figures constructed out of practically rigid bodies and measured by a practically rigid rod have approximately the properties expressed by the propositions of Euclidean geometry. The metrical structure of space is the structure of configurations of practically rigid bodies. Physical geometry reveals the fact that to the first approximation space is Euclidean.

7. IDEAL GEOMETRY

Because of errors of measurement and an element of indeterminateness in the concepts, the propositions of physical geometry are approximate. However, geometry is also pursued as a branch of mathematics in which the propositions are assumed to be **exact.** I shall now explain the concept of ideal geometry as a branch of mathematics. Ideal geometry may be said to arise from physical geometry by the process of idealizing geometrical figures; thus the geometrical elements become points which have no parts, lines which have no thickness, surfaces without depth, etc. It is difficult to give a significant description of such entities; the dubious procedure of defining them as the limits of physical elements has usually been employed. In order to distinguish the elements of ideal geometry from the corresponding ones of physical geometry I shall refer to them as ideal points, lines, surfaces, etc. The physical measurements are to be thought of as ideally carried out with the aid of ideal measuring rods. The propositions expressing metrical and other relations of this geometry are exact.

We have already seen that there are synthetic and analytic methods in geometry. The application of the synthetic method to ideal elements leads to the development of a deductive system based upon a set of axioms which express the properties of the ideal elements and the relations between them. A model for such a deductive system was furnished by Euclid's *Elements.* In the analytic method geometry is based upon equations involving the coordinates of ideal points which are assumed to be determinable with ideal accuracy.

However, the employment of the adjective **ideal** raises the question of the kind of reality possessed by these objects. Having initially presupposed the existence of physical bodies there is no problem of the existence of physical figures. We do, however, have the problem of the existence of the entities whose relations are expressed by the exact propositions of ideal geometry.

Two historically important theories of the mode of existence of ideal geometrical elements are rationalism and intuitionism. **Rationalism** is exemplified by the traditional interpretation

of the Platonic theory of ideas as a logical realism. Geometrical entities have being in a realm of pure forms which are objects of reason and not of sense. The structures of physical geometry participate in the pure forms and therefore approximately exemplify the definite relations of ideal geometry. Thus, a theory of pure forms is an explanation of the possibility of ideal geometry.

Intuitionism in geometry is best illustrated by the Kantian philosophy. According to Kant geometrical figures are constructed in pure intuition; the necessity and universality of geometry is grounded in the fact that space is an *a priori* form of intuition. The theory of intuitionism has had many adherents; the language of the theory is in current use among mathematicians.

8. ABSTRACT GEOMETRY

In recent years two methods have been developed which offer solutions of the existence of the ideal elements. One is Mr. Whitehead's construction of points and other entities by the method of extensive abstraction; the second is the elimination of content from the propositions of ideal geometry and the resultant creation of abstract geometry.

The method of **extensive abstraction** is designed for the construction of entities, having the properties of the ideal entities, out of entities that are actually apprehended in experience. Mr. Whitehead applies the method to the construction of the concepts of a four-dimensional space-time. However, the method can be applied to the elements of space alone and in this way has been used by Mr. Russell and Mr. Broad. A point may be constructed as follows. Suppose that we have given a set of concentric spheres which we say have the same point as center. Such a set of spheres is an abstractive set. Another abstractive set consists of a set of cubical boxes one contained inside the other, and all having the same point as center. The properties of these abstractive sets are defined by certain postulates. Now the class of all such abstractive sets is defined to be a point. Mr. Russell gives an alternative definition in terms of partially overlapping spheres. It seems to me, however, that this method does not obviate the necessity of assum-

ing ideal entities, for the abstractive sets which will serve to define a definite point are themselves ideal constructions. The objects of sense perception have indefinite boundaries; the elements of the abstractive sets must have definite boundaries. As a matter of fact the method of extensive abstraction is purely formal and never comes in contact with physical reality. Inasmuch as the abstractive sets are defined by postulates, they are just as abstract as the points of an abstract geometry as expounded, for example, by Hilbert.

The second method of solving the problem of the existence of geometrical entities is the creation of an **abstract geometry.** Consider first the creation of an abstract geometry out of ideal **synthetic** geometry. The propositions of geometry express relations between points, lines, planes, etc. The traditional interpretation of geometry was that the concepts of the geometrical elements had a meaning independent of the propositions described by the postulates. The abstract interpretation of geometry limits the meaning of the concepts to those properties expressed by the postulates. Instead of the concept of point one uses that of element, and instead of the concepts of concrete geometrical relations one employs the concept of relation. The postulates then express the formal properties of the elements and the relations. The postulates are neither true nor false, but are blank forms. The problem of geometry consists in showing that the postulates imply the theorems. The postulates acquire a meaning and become true or false when an interpretation is given to the elements and the relations.

I have already explained the creation of an abstract theory by the example of the theory of groups. However, it is desirable to illustrate the same process for geometry. As an example, consider a proposition about the order of points A, B, and C on a straight line. *If* A *is to the left of* B, *and* B *is to the left of* C, *then* A *is to the left of* C. According to the traditional interpretation of geometry we know what a point is and know the relation " to the left of," and recognize the proposition as evidently true. We obtain a proposition of abstract geometry when we empty our concepts of all content except that implied by the proposition. Thus the points become undefined elements, and the relation " to the left of " becomes some unspeci-

fied relation. Our proposition then becomes: *If the element* A *has the relation* R *to* B, *and* B *has the relation* R *to* C, *then* A *has the relation* R *to* C. This proposition is neither true nor false. It constitutes an implicit definition of the specific content of the terms " element " and " the relation *R*." The relation *R* is a transitive relation and the elements are such that they enter into such a relation.

If the elements are points on a line and the relation is " to the left of," the resulting proposition is true.

If the elements are real numbers and the relation is " less than," the proposition is true.

If the elements are persons and the relation is " loves," the proposition is sometimes false.

A physical geometry is an approximately true interpretation of a corresponding abstract geometry.

The postulates of an abstract geometry may be said to define the form of a system of elements in relation, that is, a relational structure. One may now raise the problem of the mode of existence of this relational structure. It seems clear that the intuitionism of Kant can not be employed in a theory of the reality of the relational structure defined by the postulates. The relational structure is a complex universal which may be exemplified in various instances, and hence the problem of the reality of mathematical objects is that of the reality of universals. I shall not attempt to solve this problem in the present work.

The foregoing exposition of abstract geometry assumed the synthetic method as a basis for discussion. I shall now consider the formation of an abstract **analytic** geometry. In analytic geometry we choose a coordinate system and determine the positions of points by coordinates. If we use a Cartesian system the coordinates are the distances of the point from the coordinate axes. Geometrical properties are expressed by functions of the coordinates. In ideal analytic geometry we think of the coordinates as ideally determined by ideal measuring rods.

The transition to abstract analytic geometry may occur in two steps. The first step consists in considering the point as merely a set of numbers (x, y, z). In ideal analytic geometry

an equation in these variables represents a geometrical figure —thus $x - y = 0$ is the equation of a straight line through the origin and inclined at an angle of 45° with the positive direction of the x axis. If we consider x, y, z as merely a set of numbers then this equation will be neither true nor false but a functional relation between variable numbers which yields a true proposition for some numbers and a false proposition for others. This is the level of the ordinary algebra and analysis of real variables. On this level true propositions are exemplified by: If $x - y = 0$ then $x = y$.

The final stage of abstraction is that in which we consider x, y, z as elements which satisfy certain postulates. Then, $x - y = 0$ implies $x = y$, becomes a blank form. By these methods one is able to obtain an abstract theory of algebra and analysis which is identical with that for geometry.

In this book we shall remain on the first level of abstraction, that of the algebra and analysis of real variables. I shall call this **pure** analysis in order to reserve the term abstract for the final stage of abstraction. Then (x, y, z) is a set of numbers. These sets of numbers constitute a number space.

In the preceding chapter I distinguished between theoretical physics, ideal theoretical physics and mathematical physics. In the present chapter on geometry the general propositions of physical geometry belong to theoretical physics, the propositions of ideal geometry belong to ideal theoretical physics, and the propositions of the algebra and theory of functions of real variables belong to mathematical physics. I have not considered a physics which belongs to ultra abstract mathematics.

The outcome of the previous discussion is that we need to consider only physical synthetic and analytic geometry, and pure analytic geometry in the sense of pure algebra and analysis. Physical synthetic geometry is an approximately true interpretation of some abstract set of postulates. An algebra of real variables furnishes an exactly true interpretation of some abstract set of postulates. Physical analytic geometry furnishes an interpretation of real variables in virtue of which undetermined functional relations between variables become approximately true. In general, assuming the work of experi-

mental physics, we need only theoretical physics and mathematical physics.

9. CONCLUSIONS

Let us now review the results of the preceding discussion.

About space one may say that things in space can be known in sensation, hence space is a character of experience. Space was defined as the totality of possibilities of relative position of things and phenomena. This definition provides a basis for a relational analysis of space. One can define the concept of extension as a quality in terms of the relation of the end points of a segment. In our study of space we utilized the motion of bodies, and thus space is a frame for motion.

One also ascribes a structure to space. My view is that the structure of space is merely the relational structure of configurations of practically rigid bodies. In this way geometry, which is the study of the spatial properties of configurations, describes the structure of space. In order to develop metrical geometry we define the concept of congruence and postulate that a specified rigid rod is to have unit length. The measurement of configurations of practically rigid bodies then shows that their metrical properties are expressed approximately by the propositions of Euclidean geometry. We express this result by the statement that physical space is Euclidean.

It must be emphasized, however, that this experimental result is relative to a definition, namely, that a rigid rod retains the same length (measure number) upon displacement. This is a natural definition in that it agrees with our qualitative estimates of extension. Strictly speaking, however, it is *a priori* for the present stage of the discussion. Later, it will become desirable to revise the definition in order to take account of changes of temperature, etc. Thus our *a priori* is an empirical one; it is subject to revision. Once the definition has been laid down, the experimental results are based upon the experiences of coincidences of points. It is an experimental fact that the properties of practically rigid bodies with respect to congruence are expressed by the propositions of Euclidean geometry.

In order to express this result we may employ the propositions of the synthetic method, or we may use the analytic

method. In analytical terms the proposition that space is Euclidean means that it is possible to construct a physical Cartesian system of coordinates, that is, we can build up a cubical lattice out of unit measuring rods. The distance between two points is then given by

$$s^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2$$

where the values of the variables are numbers obtained from the observation of coincidences of the points with points of the lattice. This function of the coordinates is invariant under a translation and rotation, that is, a certain linear transformation. The Euclidean properties of figures are unchanged by such a transformation—thus a straight line is transformed into a straight line. The proposition that the sum of the angles of a triangle is two right angles is invariant under the transformation. Thus Euclidean geometry includes the theory of the invariants of the group of transformations under which the distance between two points is invariant.

The discussion thus far has been in terms of physical geometry. Now Euclidean geometry can be based upon a set of axioms from which all the other propositions can be deduced as theorems. We obtain abstract synthetic geometry when we transform the approximate axioms into exact postulates which define the formal properties of the elements and relations. Let us suppose this to have been done. Then the postulates of Euclidean geometry define the concept of rigid body.

At the beginning of the analysis of geometry I assumed a concept of practically rigid body. Our concept of rigid body was partially expressed by the concept of congruence which in turn was expressed by the propositions which assert that congruence is a symmetrical and transitive relation. Then we discovered Euclidean geometry. Experimentally discovered Euclidean geometry is now transformed into a set of abstract postulates which define the formal content of the concepts of rigid body and congruence. The course of development is that special propositions define a concept which gives rise to more general propositions which are then employed to define the concept independently of the special subject matter. Thus the postulates of Euclidean geometry define the formal

properties of rigid bodies. This, however, is not enough for physics—we must also have contact with bodies given in experience, and so we point to a concrete body which serves as a standard. Thus our physical science of the positional relations of practically rigid bodies contains two factors: (1) A formal factor, i.e., the postulates of Euclidean geometry. These postulates define the formal characteristics of rigid bodies. (2) A material factor. We point to a concrete thing and declare that it is a rigid body of length l. Then the postulates serve as a criterion for the recognition of other things as rigid bodies. All those things which have the relation to the standard expressed by the postulates are rigid bodies. It is an empirical fact that there is a system of things which approximately have the properties expressed by the postulates.

Let us now consider the analytic formulation. Adopting Euclidean geometry initially as an induction from experience, we can express it by saying that it is possible to construct Cartesian coordinates so that

$$s^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2.$$

If we deprive these variables of physical significance, we get our pure analytic geometry. We then build up a theory of invariants under a group of transformations which leave the foregoing function invariant.

The physical definition of rigid body now contains two factors:

(1) The formal factor is the proposition that

$$s^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2$$

is to be invariant under certain transformations.

(2) The material factor is that the values of the variables are to be determined by this standard rigid body. "Space is Euclidean" means: It is possible to assign numbers x, y, z by operations with a rod and the observation of coincidences such that

$$s^2 = (x_2 - x_1)^2 + \ldots$$

is invariant under certain transformations.

In the chapter on Riemannian geometry we shall need to assume that the continuum is Euclidean to the first order in the immediate neighborhood of a point. For this purpose we express the Euclidean structure of space in differential form by

$$ds^2 = dx^2 + dy^2 + dz^2.$$

ds is the differential of distance along the arc between the points (x, y, z) and $(x + dx, y + dy, z + dz)$.

10. GEOMETRICAL OPTICS

The application of Euclidean geometry to geometrical optics is important. Among the things given in experience are some which we characterize as **visible.** For example, before me on the table is a piece of white paper, which is an example of a visible body. Visible bodies are characterized by the property that they are capable of being seen. Now we observe that the visibility of bodies is largely dependent upon the existence of other visible bodies. Thus I enter a dark room with a lighted candle, and thereupon other bodies in the room become visible; I remove the candle and the bodies in the room can no longer be seen. A lighted candle is an example of a class of bodies which may be called luminous; the sun and the stars are luminous bodies. On the basis of experiments such as that with the candle we call the luminous body the cause of the visibility of surrounding bodies. Inasmuch as luminous bodies illuminate bodies in the vicinity, we conceive of something propagated from the luminous body to neighboring bodies, thereby illuminating them, that is, rendering them visible. Also an illuminated body may in turn illuminate other bodies. In physics the term **light** is applied to the physical condition which is propagated from a luminous body to other bodies.

The concept of a condition of visibility that is propagated from a luminous body to another body expresses empirical laws. The experimental basis is that there are certain observable relations between visible bodies which may be expressed by the statement that a luminous body is the origin of the con-

dition which renders other bodies visible. The concept of the propagation of light is thus based upon verifiable empirical laws. Using only the criterion of condition of visibility these laws may readily be established.

A fundamental law is expressed by the proposition that to the first approximation light travels in straight lines. Observing that if one places an opaque body between a luminous body and a screen, an approximately sharply defined shadow of the opaque body is formed on the screen, one is able to conceive of light as propagated along straight lines in rays.

Another law relates to the **reflection** of light. If a ray of light strikes a smooth surface it is reflected so that the angle of incidence is equal to the angle of reflection.

If a ray passes from one medium into another the direction of the ray is changed; this is called **refraction.** Experiment shows that the ratio of the sine of the angle of incidence to the sine of the angle of refraction is constant for given media. One then defines a physical quantity, called the index of refraction, by this ratio. By convention the angle of incidence is measured in air.

The study of the propagation of light in rays, and of the reflection and refraction of rays is the subject matter of geometrical optics, a science which contains a minimum of hypotheses.

Geometrical optics may be based upon Fermat's principle of **least time**, namely, that the path of a ray of light between two points is determined by the condition that the time of propagation along the natural path is less than along an adjacent path. Initially, the concepts of geometrical optics are based upon the epistemological priority of **tactual** over **visual** aspects, as was clearly recognized by Berkeley in his *New Theory of Vision.* The laws of geometrical optics are formulated in terms of the tactual aspects as standard for the determination of position in space. The visual aspects belonging to a body are related to the tactual aspects by the laws of perspective and other laws of optics. The stick in the water appears bent, but is " really " straight because the straight tactual aspect is unchanged by the water. The distinction between a real image and a virtual image is that a real image can be formed on something tangible, whereas the virtual

image can not. Once the laws of optics have been obtained, optical methods may be employed to make geometrical measurements. Indeed, one may say that the concept of space is redefined in terms of geometrical optics.

REFERENCES

BROAD, C. D. Scientific Thought. New York and London. 1923.
CARNAP, R. Der Raum, Kant-Studien. 1922.
EINSTEIN, A. Geometrie und Erfahrung. Berlin, 1921.
—— Vier Vorlesungen über Relativitäts-theorie. Braunschweig, 1922.
HUNTINGTON, E. V. The Fundamental Propositions of Algebra, *in* Monographs on Modern Mathematics. Edited by J. W. A. Young. New York, 1915.
KLEIN, F. Elementarmathematik II. Berlin, 1925.
POINCARÉ, H. The Foundations of Science. New York, 1913.
REICHENBACH, H. Philosophie der Raum-Zeit Lehre. Berlin und Leipzig, 1928.
YOUNG, J. W. Fundamental Concepts of Algebra and Geometry. New York, 1911.
VEBLEN, O. The Foundations of Geometry, *in* Monographs on Modern Mathematics.

CHAPTER III

THE THEORY OF TIME

1. THE CONCEPT OF TIME

THE concept of **time** is a fundamental concept of physics. Thus physical processes are said to take place in time; in particular, motion may be described as change of position in space with time. Physical bodies have the characteristic that they persist in time.

Among the historical statements concerning time one finds the following.

(1) *Time passes. Time is the quality of passage.*
(2) *Leibniz described time as the order of succession of phenomena.*
(3) *According to Kant time is an* a priori *form of intuition.*

In the present discussion I shall base the theory of time upon the concept of **event**; an event is a brief physical process. Events are distinguished by the fact that they stand in the relations of simultaneity or succession. I describe time as the totality of possibilities of relative temporal position of events. Like space, time is a manifold, but a one-dimensional manifold: this means that in the analytical description of the temporal position of events only one number is necessary to specify the position of the event relative to some origin and unit of time.

With regard to temporal position one may hold either an absolute theory or a relational theory. I shall adopt the **relational** theory and thus our definition agrees with the view of Leibniz that time is the order of succession of phenomena. The boundary events of a process are related as earlier and later, and the process as a whole may be thought of as possessing the quality of passage. Thus a process has duration

which we define in terms of a relation between the beginning and end of the process. Physical events which have position in time are apprehended in sensation, and thus time is a character of experience.

The preceding metaphysical description of time will afford a basis for the discussion of time in physics. The physical problem is one of the method of **measuring** time. In order to discuss the problem of time in physics it is desirable to distinguish between (*a*) time at a place, **local** time, and (*b*) time throughout space, **extended** time. I shall first consider the measurement of time at a given place.

As a starting point we shall assume the concept of **event**. A fundamental relation of events is that of **simultaneity**; two events may be simultaneous, such as the sounding of a whistle and the starting of a light signal. This meaning of simultaneity is based upon a direct experience of the simultaneous occurrence of two events. The estimate of simultaneity is based upon an observation of temporal coincidence which is analogous to the observation of coincidence of points in space. As in the case of space, the quantitative concepts of time are based upon a matching or estimate of similarity of qualities; simultaneity is similarity with respect to temporal position.

In addition to the concept of simultaneity of two events we also have the concept of the **relation of succession** of two events. Thus I assume that one knows the meaning of the proposition, ball *B* is projected vertically upwards after ball *A*. A temporal process is marked out by two events, the beginning and the end, of which the first is earlier and the second is later.

We have experiences of various processes that are going on in time. The distinguishable processes are simultaneous in the sense that events in one process are simultaneous with events in other processes. Thus we may experience simultaneously the process of breathing, the pulse beat, the rumble of a passing train, the passage of an airplane overhead. The concept of interval of time or duration is the result of abstraction from these simultaneously occurring processes. On the qualitative level we also have crude comparisons of the intervals between events or the durations of processes—thus the

interval of time between the rising and setting of the sun is long in comparison with the interval between two pulse beats.

When we seek to construct a **metrical** theory of time we come upon the problem of the **structure** of time. Now time in itself has no definite structure; the structure of time is the relational structure of certain processes. An example of such a process is the swing of a pendulum. The beginning of the swing may be chosen as the time when the pendulum bob is at one end of its path; the end of the swing is the time when it first returns to this position. Suppose that we have two pendulums of equal length; then they will swing together. If A swings together with B then B swings together with A. Furthermore, if A swings together with B, and B with C, then A swings together with C. Thus there is a symmetrical and transitive relation between the pendulums which swing together. We express this by the proposition that the pendulums are **synchronous**; by the principle of abstraction we ascribe to the synchronous processes a common relation, the same interval of time. In this way a swing of a pendulum is characterized by an interval of time which is a relation between the beginning and end of the process. One can also ascribe a duration to the swing as a whole. Hence a given process may be characterized by a *relation*, interval of time, or by a *quality*, duration. Interval of time, and duration, are measurable properties.

2. THE MEASUREMENT OF LOCAL TIME

The synchronous motion of two pendulums may be the basis of the physical concept of time interval. The vibrations of two synchronous pendulums are equal in interval of time, which may be expressed by the proposition that the vibrations are isochronous. Thus we represent the concept of equality with respect to interval as an abstraction from the experience of the synchronous motion of pendulums of equal length. Then we further make the assumption that successive vibrations of a pendulum are isochronous, that is, equal in time. This is a definition concerning which it is meaningless in the present stage of discussion to ask whether or not it is correct. The assumption, however, agrees with a qualitative estimate of

the temporal similarity of successive vibrations of a pendulum. We have the further problem of the equality in time of vibrations of pendulums separated in space, but this will require a special discussion.

Historical accounts of the measurement of time reveal the fact that various physical processes have been employed to define an interval of time. The interval between successive full moons, the interval between seasons, the interval of one rotation of the earth about its axis, the time required for sand to run out of an hour glass, the interval between pulse beats, the time required for a certain amount of water to escape from the opening of a vessel—these are among the different standards of time which have been employed. Galileo, for example, by means of his pulse beat, determined that the swings of a pendulum take the same time. It is a remarkable fact that all these standards of time agree approximately in the definition of isochronous intervals. This agreement corresponds to the agreement of the coincidence of practically rigid bodies in space. We may therefore maintain that we conceive of time to have a structure which is that of certain isochronous processes.

In order to measure any interval of time we need to choose some time interval as a unit. The unit of interval is embodied in the rhythmical process of some mechanism, called a clock, which is subject to the same conditions as estimated qualitatively. Successive processes of the standard clock define equal intervals of time; the measure of time of any process is obtained by counting the number of standard intervals during the process. Thus one notes the position of the indicator of the standard clock which is simultaneous with the beginning of the process, and then counts the number of intervals of the clock until the end of the process. Hence the measurement of the duration of a physical process is based upon the observation of coincidences in space and time. The unit of time employed in physics is defined by the rotation of the earth; it is called the second and is 1/86,400 of the mean solar day.

In the discussion of length I distinguished between extension, a measurable property, and length. The expression "the length" denotes a physical quantity, that is, the measure

number which is assigned to extension. It is more difficult to distinguish between the measurable property and the physical quantity in the case of time. We need to distinguish between interval of time as a *measurable relation* and "the interval of time" which is a *physical quantity*. We may also distinguish between the duration of a process and "the time" of a process.

Upon the adoption of a unit of time interval one can measure the duration of physical processes and thereby discover laws. Thus obtaining the laws of motion enables one to control the conditions to which the clock is subject, and then one is able to define the unit of time to a higher approximation. Now as a first approximation the unit of time is defined by the rotation of the earth. But it follows from the laws of motion that the angular velocity of the earth is decreasing on account of tidal friction. Hereupon we seem to fall into a contradiction, because the concept of a unit implies that it remains constant. The contradiction is resolved by orders of approximation: to the first order the angular velocity of the earth is constant; to the second order the angular velocity is decreasing slowly. This means that the second approximation in the definition of the unit of time is by means of the laws of motion. The other quantities entering the laws of motion, for example, mass, are measured on the assumption that the unit of time defined by the rotation of the earth remains constant.

Having chosen a unit of time interval we can measure the interval of time between two events by counting the number of unit intervals between the earlier and later events. The interval of time is expressed by a proposition such as: *The time in seconds of this process is* 2. In the representation of temporal properties one may use either a synthetic method or an analytic method. Using the synthetic method, one expresses the time interval between two events directly; for example, the time interval between the emission of light from the sun and its arrival at the earth is 8 minutes. Using the analytic method, one expresses the dates of events with respect to some origin. The time interval between two events is the difference of the dates: thus the light was emitted by the sun

at 12:00, and it arrived at the earth at 12:08. In the analytic method one chooses a certain event as the origin from which to measure time. Time is homogeneous and hence the temporal properties of processes are invariant under a transformation of origin. The general character of a physical process is independent of the initial time.

3. PHYSICAL THEORY AND ABSTRACT THEORY

In the theory of space I distinguished between physical geometry, ideal geometry and abstract geometry; I shall make a similar distinction for the theory of time. Thus far I have presented a **physical** theory of time in which the concept of event is fundamental. In the measurement of time intervals the fundamental process is the determination of simultaneity. The observation of simultaneity, which is similarity with respect to temporal position, is usually combined with the observation of coincidences. Like the physical point, the event lacks precision of character. The indefiniteness of the points and events which occur in the physical theory of time renders inexact our physical measures of time. In order to construct an exact theory of time the physical concepts have been replaced by concepts of **ideal** entities. Events thus become instants which have neither parts nor magnitude; instants have been defined as the limits of events. Moreover, the physical measures are thought of as ideally determined by means of ideal clocks, and the propositions expressing metrical and other relations in time are *exact*. However, as in the case of space, the use of the adjective ideal raises the problem of the kind of reality possessed by instants, ideal simultaneity, etc.

One method of solution of the problem of ideal time is that of Mr. Whitehead. He constructs instants out of finitely extended durations by the method of **extensive abstraction.** But this method does not solve the problem of the reality of the ideal entities, because the abstractive sets, which will serve to define a definite instant, are themselves ideal constructions.

A second method of solving the problem of the existence of ideal temporal entities and relations is the creation of an

abstract theory of time. The propositions about time express relations between instants. In the ideal theory the concept of instant has a meaning in addition to the formal properties expressed by the propositions. In the abstract theory of time the meaning of the concepts is limited to formal properties; one replaces the concept of instant by that of element, and instead of the concepts of concrete temporal relations one employs the general concept of relation. The propositions express the formal properties of the elements and relations; they are blank forms which become true or false when an interpretation is given to the elements and the relations. In the physical theory of time we have the proposition: *If event* A *precedes event* B, *and event* B *precedes event* C, *then event* A *precedes event* C.

In the abstract theory this proposition becomes:

$$ARB \cdot BRC \text{ implies } ARC.$$

This abstract proposition defines the relation R to be a transitive relation. If A, B and C are physical events and the relation R is the relation of earlier and later, then the proposition is approximately satisfied. The abstract theory of time is that it is a one-dimensional continuum. In the traditional abstract theory of time one interprets the elements directly as physical events. In Mr. Whitehead's theory, however, the elements with the relations described by the postulates are to be interpreted as extended processes. In either case the sets of postulates are only approximately satisfied by empirical entities. Mr. Whitehead's theory is an abstract theory with a new set of indefinables and postulates.

There is also a pure analytic theory of time, similar to pure analytic geometry, in which an instant of time is represented by a real number, and the one-dimensional manifold constituting time is represented by the continuum of real numbers. The relations between the numbers are satisfied approximately by the physical measures of time. From the pure point of view time is the independent variable in the equations of dynamics.

Thus for physical time we need only a physical theory and a pure analytic theory. The physical theory is based upon

the concept of physical event, and the concepts of the simultaneity and succession of events. The pure theory is the theory of the continuum of real numbers; an instant of time is represented by a real number. In view of the theory of physical quantities as numbers which are assigned by operations, it seems clear that the pure theory is sufficient.

I have distinguished between abstract synthetic theory and a pure analytic theory. The abstract synthetic theory carries the process of abstraction one step beyond the pure analytic theory, for the correlation of instants with real numbers is already an interpretation of an abstract theory. The relational structure of the continuum of real numbers is identical with that of time as abstractly conceived. However, the pure analytic theory is abstract with respect to the interpretation of the numbers as physical measures.

The relational structure of time may be represented graphically. The space of a line is a manifold of one dimension which has the same relational structure as time; in this way time may be represented by a line. The analytic representation of points on the line in terms of distances from an origin corresponds to the representation of time in terms of intervals measured from an origin.

In concluding this section I recall that we have been considering only local time, that is, time at a place. In experimental physics, which furnishes data to theory, a place is a region which is small in comparison with other regions which may be considered. Expressed in terms of an ideal space, local time is time at a point. We have seen that this is to be expressed in purely analytic terms.

4. EXTENDED TIME

I have now to complete the account of time by a theory of **time throughout space**. Prior to the theory of relativity physicists were hardly aware of this problem; it was assumed that we knew, for example, the meaning of simultaneity throughout space. I think that in classical physics both our metaphysical concepts and mathematical theory outran our physics. The metaphysical concept of absolute time seemed to offer a clear conception of simultaneity extending throughout

space; in the mathematical theory the coordinates of separated particles were expressed as functions of an independent variable t, and since the mathematical problem consisted in the solution of differential equations the question of the measurement of t did not arise. In physics, however, the concept of time throughout space must be defined in such a way that we know how to compare the times of events at separated points.

In order to extend time throughout space we may conceive of similar clocks located at the various points at which we may wish to know the time of events. At a given place we know the meaning of simultaneity; the time of an event is indicated by the position of the hand of the local clock simultaneous with the event. By comparing the readings of the clocks for different events one can determine the temporal relations of the various separated events. However, we must first synchronize the clocks.

There appear to be two methods of synchronizing clocks: the method of transport and the method of signaling. Suppose that at a given place we have two similar clocks that run synchronously. One may extend the time at the given place to another place by moving one of the clocks there, thereby postulating that the rate of the clock is not changed by the motion. When the captain of a ship determines the time by a chronometer, which was set at the port of sailing, he is employing the method of extending time throughout space by the transport of a clock. Similarly the locomotive engineer extends a time system along the railroad track by transport of his watch. This procedure is possible in physics.

The second method of synchronizing clocks is that of signaling. Given clocks at points A and B. Suppose that at time t_A a signal is sent from A to B, where it is immediately reflected, and returns to A at time t'_A. One can then synchronize the two clocks by the postulate that

$$t_B = \frac{t_A + t'_A}{2}.$$

This is equivalent to the assumption that the average speed of the signal from A to B is the same as the average speed of

the signal from B to A, which is a natural assumption in view of the isotropic character of space as determined from experiments on practically rigid bodies. If the speed of the signal is infinite then $t_B = t_A = t'_A$.

The preceding definition enables us to say that an event at A is simultaneous with an event at B if $t_A = t_B$.

The problem of extending time throughout space does not seem to have been realized by the classical physicists. In many instances they undoubtedly transported their clocks with the tacit assumption that the rate remained the same; in other cases they used the method of signaling and assumed that the speed of the signal was infinite. Thus when Galileo measured the time in which a ball rolls down an inclined plane he had only one clock, his water clock. He started the clock when he released the ball and stopped it when he observed that the ball had arrived at the foot of the plane. For the small space in which the experiment took place he assumed that the time was determined by his single clock—or, if one wishes to be more rigorous, one may say that the time at the top of the plane and the time at the bottom of the plane were obtained from the time at the clock by light signals whose speed was so great that it was assumed to be infinite. It seems clear that the classical physicists did not have a standardized procedure for the extension of time throughout space; they thought in terms of a metaphysical concept of absolute time, and transported clocks and used signals in an uncritical fashion.

The problem of time was made central in physics by Einstein. His method of extending time throughout a space resulted in the conception of time as relative to a frame of reference.

CHAPTER IV

KINEMATICS

1. KINEMATICS OF A POINT

In the preceding chapters we have considered the concepts of geometrical properties, especially the concepts of geometrical quantities such as the measure of distance, angle, etc., and further the concepts of temporal properties and of temporal quantities such as the measure of time interval. In the general sense these are all physical quantities: they are measure numbers which are assigned to measurable characters. The quantities are referred to as: the distance, the angle, the x coordinate, the time. In the present chapter we shall take up the concepts of quantities which serve to describe **motion.** The subject matter of this chapter is kinematics, which may be characterized as the description of motion. Inasmuch as motion is change of position in space with time, the quantities of kinematics are to be defined in terms of geometrical and temporal quantities.

The fundamental condition for the description of motion is the choice of a **frame of reference.** The representation of the motion of any body depends upon the frame of reference employed, as can be seen from a simple example. Suppose that a train is moving along a straight track, and a man in the train drops a stone out of the window. What is the path of the stone? The man in the train would use the car as a frame of reference, and relative to this frame the stone moves in a straight line. An observer on the ground would choose the surface of the earth as a frame of reference, and relative to it the stone moves in a parabolic path. Relative to a frame attached to the sun the path would be still different. Thus the description of motion presupposes the selection of a frame

of reference. The most frequently employed frame consists of three mutually perpendicular planes like the three sides of a box, or sometimes it is convenient to use curved surfaces, and in the general theory of relativity flexible surfaces are permissible. Frames of reference are physical coordinate axes. The frame must be thought of as constituted so that it does not interfere with the motion of bodies. A system of reference is a frame provided with measuring rods and clocks; we may also think of a system of reference as a lattice system of coordinate lines to the points of which clocks are attached.

In order to introduce the concepts of kinematics I shall define them for the simplest possible case, the motion of a point in a straight line. We begin with the concept of the position of a point on the line; the relative position is determined by assigning the measure of distance, the distance from a fixed point, the origin. In this case the frame of reference reduces to a line. I shall denote the distance from the origin by s. The result of a motion of the point is a change in the distance from the origin: let us represent the distance passed over by Δs, and the time of the motion by Δt. Δs and Δt are increments of s and t, respectively. Average speed is now defined as

$$\bar{v} = \frac{\Delta s}{\Delta t}$$

The question now arises, what is the significance of this definition? From the traditional point of view the symbol of division represents a rate. Average speed is average rate of change of distance with respect to time. Average speed is a magnitude which is a correlation of the magnitudes distance and interval of time. According to the point of view adopted in this book, the preceding equation defines the measure of average speed as the measure of distance divided by the measure of time. We express this by the statement that the average speed is equal to the distance divided by the time.

Thus far we have considered only characters of the motion as a whole. In order to describe the instantaneous characters of a motion we define the concept of instantaneous speed at time t. Let Δt be a time interval extending from time t, and

Δs be the distance passed over during the time Δt. The average speed is $\frac{\Delta s}{\Delta t}$. Now let Δt approach zero as a limit; then Δs also approaches zero as a limit. $\frac{\Delta s}{\Delta t}$ approaches a finite limiting value. The instantaneous speed at t is defined as the limit of $\frac{\Delta s}{\Delta t}$ when Δt approaches zero as a limit.

Here again we are confronted by the difference between the traditional and contemporary interpretation of equations. There is a difficulty in the conception of the limiting process. Whether we define instantaneous speed as the limit of a relation between two magnitudes, or as the limit of a ratio between two measures, we are confronted by the difficulty that we can not actually arrive at the limit for a concrete motion. There is a finite limit to the qualities which we apprehend and the numbers which we assign. This suggests the distinction between physical kinematics, ideal kinematics, and pure kinematics. Physical kinematics is a description of the motions of physical points, etc., in terms of physical measures made with material measuring rods and clocks. Like physical geometry physical kinematics is inexact. Ideal kinematics is an exact description of the motion of ideal points in terms of ideal measures. The problem of the reality of these ideal entities renders this sort of kinematics unsatisfactory. In pure kinematics we deal with variable numbers which may be interpreted as the physical measures of geometrical and temporal quantities. Thus from the pure point of view we have given a variable number s as a function of another variable t:

$$s = f(t).$$

We assume that the function is continuous and has a derivative

$$\frac{ds}{dt} = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t}.$$

The differential calculus is the study of derivatives independently of whether they can be employed to represent characters of motion.

Classical physical kinematics is now characterized by the following physical assumption. If s is the measured coordinate of a physical point at the measured time t, an instantaneous speed at t is defined by $\frac{\Delta s}{\Delta t}$ where the interval is made as small as is technically possible. A functional relation $s = f(t)$ is approximately satisfied by the measured values of s and t, and the number $\frac{ds}{dt}$ is the more closely approximated to by the instantaneous speed the smaller Δt is taken.

Thus I concede the difficulty in the conception of a physical derivative, but employ a mathematical representation of physical objects and processes which admits of exact deductions.

In the remainder of the book we shall use the symbolism of analysis.

Finally I define the acceleration of a point as the rate of change of speed,

$$a = \frac{dv}{dt} = \frac{d^2s}{dt^2}.$$

I now repeat the preceding discussion for the general curvilinear motion of a point. The position of a point may be determined with respect to a Cartesian frame of reference by the three Cartesian coordinates of the point (x, y, z). Consider a line drawn from the origin of coordinates to the point; this directed line, then, determines the position of the point. The components of this line, x, y, z, are the components of a vector, hence the position of a point is determined by the vector (x, y, z) which we symbolize by $\mathbf{r}$. We can now describe the motion of the point in terms of the changes of the position vector. This is the vectorial representation of kinematics.

The displacement of a point is its change of position; it is a directed line running from the initial point to the final point. I here use "displacement" for a particular process. In the traditional physics a displacement would have been called a vector. However, I am using the term vector for a set of three physical quantities. The displacement in the above sense may be represented by three numbers, which are the respective changes in the components of the position

vector. Thus $\Delta\mathbf{r}$ is the physical quantity displacement in the sense of three physical quantities. The physical quantity displacement is to be distinguished from the distance measured along the path of the motion.

Suppose that the displacement is $\Delta\mathbf{r}$ in time Δt. Then we can define the average velocity by

$$\bar{\mathbf{v}} = \frac{\Delta\mathbf{r}}{\Delta t}.$$

The instantaneous velocity is defined by

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \lim_{\Delta t \to 0} \frac{\Delta\mathbf{r}}{\Delta t}.$$

The acceleration is defined by

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d^2\mathbf{r}}{dt^2}.$$

A vector equation $\mathbf{v} = \dfrac{d\mathbf{r}}{dt}$ stands for three scalar equations

$$v_x = \frac{dx}{dt}$$

$$v_y = \frac{dy}{dt}$$

$$v_z = \frac{dz}{dt}.$$

The application of these concepts in physical kinematics is based upon physical assumptions similar to the one previously formulated for motion in a straight line.

I remark that we frequently write $\dot{x}$ for $\dfrac{dx}{dt}$ and $\ddot{x}$ for $\dfrac{d^2x}{dt^2}$.

2. APPLICATIONS OF KINEMATICS

Thus far I have merely stated definitions; we must now examine some processes to which the concepts are applicable. As a first example of the applicability of the concepts of kinematics for the description of motion we may consider motion in a straight line with constant acceleration, such a motion as is exemplified approximately by a freely falling body and by a ball rolling down an inclined plane. The latter motion was the subject of Galileo's classical experiments. The motion is described by the differential equation:

$$\frac{d^2s}{dt^2} = a,$$

that is, the absolute value of the acceleration is constant. Now acceleration is an instantaneous property of the motion. We do not measure the acceleration directly, but determine for a whole motion the character which flows from the instantaneous property. Thus we assume that the differential equation holds throughout the motion. By integration we obtain

$$s = s_0 + v_0 t + \tfrac{1}{2}at^2.$$

Assuming that the body starts at rest from the origin at time $t = 0$,

$$s = \tfrac{1}{2}at^2.$$

Thus if the instantaneous property expressed by the differential equation holds throughout the motion, the distance passed over varies as the square of the time.

The preceding example illustrates how an instantaneous property of a motion is described by a differential equation. The hypothesis is made that the differential equation holds throughout an interval of time, and a solution expressing s as a function of t is obtained by integration. The quantities s and t can be measured directly, hence the hypothesis can be tested by testing deductions from it.

An important kind of motion is simple harmonic motion, that is, the vibration of a point along a straight line such that the acceleration is directed towards the center of the path and

is proportional to the distance from the center. The differential equation is

$$\frac{d^2x}{dt^2} = -kx.$$

An important kinematical problem is the description of the motion of the heavenly bodies. The concepts for the motion of a point are sufficient for a description to the first approximation. It is clear that the description will depend upon the frame of reference employed. In the Ptolemaic astronomy the origin of the frame of reference was the earth, whereas in the theory of Copernicus the origin was the sun. Upon the basis of the Copernican frame of reference Kepler discovered three laws which determine the motions of the planets in elliptical orbits around the sun. It is of interest to note that, like the solution of the differential equation for falling bodies, these laws are expressed in terms of quantities which characterize the motion as a whole. Newton subsequently discovered the differential equation which describes the instantaneous character of the motions of the heavenly bodies.

From the standpoint of pure theory the Ptolemaic and Copernican frames of reference are equally permissible. The Copernican frame, however, rendered possible a simpler kinematics and furnished the basis for a simple celestial mechanics.

3. TRANSFORMATION OF SYSTEM OF REFERENCE

The description of motion depends upon the system of reference. Now if a motion is referred to a given system we may wish to find the description relative to some other system. I shall therefore discuss briefly the transformation of systems of reference in the classical kinematics; by a system of reference I mean a frame of reference provided with measuring rods and clocks.

Given a system of reference (K), in which one can determine the Cartesian coordinates x, y, z of a point and the time t of an event. Let K' be a second system in which one can determine the Cartesian coordinates x', y', z', and the time t'. With respect to K the coordinates of the origin of K' are x_0, y_0, z_0.

I shall further assume that the x', y', z' axes are parallel to the x, y, z axes, respectively.

Assume first that K' is at rest with respect to K. Then the transformation equations are

$$x' = x - x_0$$

$$y' = y - y_0$$

$$z' = z - z_0$$

$$t' = t.$$

Thus if a motion is described in terms of the variables x, y, z, t we can find its description in terms of the variables x', y', z', t'.

Let us now suppose that K' moves with uniform rectilinear motion in the positive direction of the x axis. For simplification I assume that the directions of the x and x' axes coincide. Then

$$x_0 = ut$$

$$y_0 = 0$$

$$z_0 = 0.$$

It is assumed that at $t = 0$ the origins of K and K' coincide. In classical kinematics it was assumed that

$$x' = x - ut$$

$$y' = y$$

$$z' = z$$

$$t' = t.$$

These equations define a Galilean transformation.

The classical equations are based upon physical assumptions of which physicists did not become aware until Einstein. Before stating these assumptions it must be postulated that the time of an event relative to K is t as determined by a clock at rest in K, and the time of the same event relative to K' is t' as determined by a clock at rest in K'. Furthermore, the coordinates x, y, z are determined by measuring rods at rest in K, and x', y', z' are determined by measuring rods at rest in K'. The clocks of both systems are similar in the sense that

if they are all at rest in a given system they run synchronously. Similarly the rods are all equal in length in a given system. Now, the Galilean transformation was based upon two assumptions, of which the first was that $t' = t$; the second, that the length of a rigid segment is the same relative to K and K'. Thus suppose that there is at rest in K' a rigid segment, of which one point coincides with the origin of K' and the other point has coordinates x, y, z at time t relative to K. The length of the segment relative to K is $x - ut$. The length relative to K' is x'. In classical kinematics it was assumed as self evident that $x' = x - ut$. The criticism in terms of the theory of relativity, however, has shown that the classical conclusion is not true. The classical result undoubtedly was obtained by an uncritical use of the geometrical imagination, which may be traced to the belief that the propositions of metrical geometry express relations between properties which are independent of the observer instead of functional relations between measures assigned by an observer. However, the Galilean transformation is correct to the first approximation.

The Galilean transformation is fundamental in classical dynamics. A set of these transformations constitutes a group. The time t of an event and the length l of a rigid rod are invariant under the Galilean group.

One may interpret the transformation equations as defining time and space to be common to the two systems. It is an empirical fact that the measures made with rods and clocks satisfy the relations to the first approximation.

4. WAVE MOTION

Thus far I have defined the fundamental concepts for the description of the motion of a point. In order to give a complete account of the subject it would be necessary to define additional concepts for the kinematics of rigid bodies and of continuous media such as fluids and elastic bodies, but it is beyond the scope of this book to expound these further details. However, I shall explain an important concept for the kinematics of continuous media, the concept of wave motion.

Let φ be a function of x, y, z and t. I suppose that φ satisfies the equation

$$\frac{\partial^2\varphi}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2}.$$

A particular solution of this equation is

$$\varphi = r \sin 2\pi\left(\frac{t}{T} - \frac{x}{\lambda}\right).$$

An interpretation of φ is that it is the y component of the displacement of a point having the coordinate x at the time t in an elastic medium. At a given point, say $x = 0$

$$\varphi = r \sin \frac{2\pi}{T} t.$$

This equation states that the y component of the displacement is a periodic function of the time. That is, at a given point on the x axis a point in the medium executes a simple harmonic vibration with period T. At a given time, say $t = 0$,

$$\varphi = -r \sin \frac{2\pi}{\lambda} x.$$

This equation states the φ is a periodic function of the coordinate x with period λ. This means that if we take an instantaneous view of the process we see that values of φ are represented by a sine curve. The distance between two crests is λ. The original particular solution of the differential equation describes the propagation of a plane sine wave in the positive direction of the x axis with velocity c.

If the medium starts vibrating at a point the motion is communicated to neighboring parts of the medium, which in turn transmit the process. The propagation takes time so that a certain phase of displacement at a given point is attained by a distant point at some later time.

5. CONCLUSION

One may say that the equations of pure analytical kinematics constitute an implicit definition of the formal properties of space, time, and motion. In order to endow the equations with physical significance one must point to a standard rod and a standard clock which are employed in assigning numbers to spatial and temporal properties.

CHAPTER V

DYNAMICS

1. INTRODUCTION

WE have seen how one can describe motion in terms of the concepts of space and time. Now the sort of motion that occurs depends on the characters of the moving body and of external bodies. Dynamics is the study of the conditions of motion. In order to describe the conditions of motion it is necessary to add to the concepts of the geometrical and temporal quantities employed in kinematics two new fundamental concepts: those of force and mass. The physical quantity **mass** is assumed to represent some intrinsic character of a body. **Force** is often defined as the cause of changes of motion, and dynamics is said to be the study of the action of force upon material bodies. If kinematics is characterized as the description of motion, dynamics may be characterized as the description of the conditions of motion. Thus one may say that dynamics offers an explanation of motion, meaning by explanation, however, the determination of empirical conditions.

The first step in the construction of dynamics is the selection of a **frame of reference**. In classical dynamics one chooses as a first approximation a rigid frame attached to the earth. One may then define concepts and discover laws which very accurately determine the motions of bodies with respect to the earth. However, there are motions with respect to the earth which do not conform to the laws of dynamics, such as the rotation of the plane of vibration of a pendulum, the deflection of the trade winds to the right in the northern hemisphere, the difference in the weight of a body at the poles and the equator other than that due to the difference in distance from the center of the earth, etc. One may account for these facts

either by changing the laws or by changing the frame of reference. In classical dynamics the laws were retained and a more satisfactory frame of reference was chosen. As a second approximation one chooses a frame with origin at the center of mass of the solar system and having axes with directions fixed with respect to the fixed stars. The phenomena on the earth which do not conform to the laws of dynamics are then attributed to the rotation of the earth with respect to the proper frame of reference for classical dynamics. Newton held that the proper frame of reference was absolute space. However, when one applies dynamics to actual motions one always chooses a frame consisting of observable bodies.

Sometimes it is convenient to use the earth, or some other body rotating with respect to a permissible frame, as a frame of reference even for precise description. In that case one modifies the laws of dynamics by adding " fictitious " forces.

The field of dynamics is usually divided into two parts, statics and kinetics: the study of the conditions of equilibrium, and the study of the conditions of changes of motion, respectively. Correspondingly two methods may be employed in constructing the foundations of dynamics: one may first build up statics, thereby choosing the concept of **force** as fundamental, and then go on to kinetics; or one may start with kinetics, in which the concept of **mass** is fundamental, define force, and then obtain statics as a limiting case. In order to illustrate the construction of physical concepts I shall start with statics, and by the method of successive definition finally arrive at a kinetical formulation, my method of exposition corresponding approximately to the historical order of development.

2. THE STATICAL CONCEPT OF FORCE

Statics may be defined as the study of bodies in equilibrium. When the forces acting upon a body balance each other, the body is in a state of equilibrium. The concept of **force**, therefore, is fundamental in statics. The first stage of the exposition will achieve a statical definition of force. The concept of rigid body as defined by the postulates of Euclidean geometry, and the designation of a standard of length, form the foundation of the discussion.

An exposition of the physical concept of force presupposes an understanding of the primitive concept of force. Now the primitive concept arises from experiences of our own exertions; hence the concept of force is anthropomorphic in origin. Observing that when we exert a force upon a body its state of rest or motion tends to be changed, we conceive of force as the cause of changes of motion. From the direction in which bodies move when we exert forces upon them one recognizes the directional aspect of force. One also has experience of forces which annul each other, for example, if I press upon a body with both hands the body may remain at rest because two oppositely directed forces balance each other. The primitive concept of force is of something which produces changes of motion and which may be balanced by an equal and opposite force. The propositions that equal and opposite forces maintain a body in equilibrium and that an unbalanced force produces motion constitute an implicit definition of the primitive concept of force. I shall choose the **principle of equilibrium** as the basis for the building up of a physical concept of force in statics.

Suppose that I lift a body and hold it at rest; then I know from experience that I am exerting a force upon the body. Since the body is at rest the principle of equilibrium must be satisfied. Apparently, however, the only force acting upon the body is the force which I exert with my hand, hence it seems as if the principle of equilibrium were violated. However, I retain the principle of equilibrium and extend the concept of force. I conceive of weight as a force which acts downwards upon a body and is balanced by the force which I exert. Thus I have the concepts of muscular force and of weight. I have qualitative estimates of muscular force, and weight may be estimated by the muscular force required to balance it. One is thus able to estimate the equality of different bodies with respect to weight by lifting them.

One may now perform experiments with a lever. Suppose I have given a lever in the form of a straight rigid rod supported at the middle by a fulcrum. If at each end of the lever I attach bodies whose weights are the same, as estimated by my muscular sensations, I then discover that the lever remains

horizontal, that is, it is in rotational equilibrium. I further discover that if two bodies maintain the lever in equilibrium at one time, they will continue to do so at all other times and places, and also that the balancing of a lever by two bodies is independent of the total length of the lever, the material, the color, etc., the only requirement being that the two bodies be at the same distance from the point of support. Now the balancing of a lever can be determined accurately by the method of coincidences, hence an exact relation between two bodies is definable in terms of the lever. The relation is symmetrical and transitive. If A balances B, then B balances A. If A balances B, and B balances C, then A balances C. We therefore define a physical concept of equality in weight by the principle of the lever with equal lever arms. The principle of abstraction permits us to assert that two bodies which maintain an equal-arm lever in equilibrium are equal with respect to weight. Thus equality with respect to weight is similarity which is definable in terms of the similarity of coincident points.

Thus the special principle of the lever defines the equality of bodies in weight. It is a fact of experience that bodies determined to be equal in weight by the principle of the lever are approximately equal when estimated by muscular experiences.

Inasmuch as the definition of equality in weight by the principle of the lever recognizes **weight** as the typical force, we proceed to build up a scale for measuring weights by use of the equal arm lever. Let us suppose that we have given a set of equal weights, that is, a set of bodies equal in weight. A weight which, attached to the end of one arm, balances two equal weights attached to the end of the other arm will be defined to have twice the weight of either of the other two weights. Similarly we can construct weights three, four, etc., times as great as the individual given weights. Thus if we select a given weight as a standard to which we assign the number 1, we can construct weights to which we assign the numbers 2, 3, 4, etc. The method of assigning the numbers defines the weight of a body to be an extensive quantity. We may define the unit of weight as the weight of a standard body at an assigned point on the surface of the earth.

Assuming the definition of weight by the principle of the lever, the ratio of two weights is independent of time and place. It is now natural to postulate that the weight of a single body is unchanged upon displacement. This assumption was made in the early history of the subject, and even now it is abstractly possible. We shall not make the assumption because the laws of dynamics would then depend upon the gravitational field; instead we preserve the generality of dynamical laws by the assumption that weight depends upon position.

One may now extend the concept of force by introducing a spring balance. I observe that when a weight is replaced by a spring balance, for a certain extension of the spring the lever remains in equilibrium. I therefore extend the concept of force and conceive of the spring balance as exerting a force which balances the weight at the end of the other arm. With the aid of my set of weights I discover that the force exerted by a spring is proportional to the extension; this is Hooke's law. I now postulate the law of force of the spring to be independent of position and this gives me a means of extending a given measure of force throughout space.

With the aid of the concepts which have been explained we are able to express laws. An important law is the parallelogram law for forces. If two forces F_1 and F_2 act on a point, they can be replaced by a single force, the resultant of the two forces. Experiment shows that if F_1 and F_2 are represented by the adjacent sides of a parallelogram such that the lengths of the sides are proportional to the absolute values of the forces, respectively, and the angle between the sides is equal to the angle between the forces, then the resultant of F_1 and F_2 is represented by the included diagonal of the parallelogram. In the traditional terminology, force is a vector quantity. In accordance with the symbolic point of view of this book one would say that the concept of force is the concept of a method of assigning a set of three numbers to certain physical conditions. The three numbers are the components of a vector.

We are able to formulate general conditions of equilibrium for bodies. The condition of equilibrium of a point subject to several forces is that the resultant of the forces be zero. In

view of our definitions this principle can be verified experimentally.

We have used the principle of the lever for equal arms as a definition of force; it is now possible to verify experimentally the principle of the lever for unequal arms. The general principle of the lever states that if F_1 has lever arm l_1, and F_2 has lever arm l_2, then in equilibrium

$$F_1 l_1 = F_2 l_2.$$

One may formulate a general principle of moments for rigid bodies; I shall state the principle for forces in a plane. The moment of a force about a point is the value of the force multiplied by the perpendicular lever-arm, that is, the perpendicular distance of the line of action of the force from the point. Moments which tend to produce a counter-clockwise rotation are positive; those which tend to produce a clockwise rotation are negative. The principle of moments states that the condition of rotational equilibrium is that the sum of the moments about any point be zero.

In general, there are two conditions of equilibrium for a rigid body: the condition that there be **translational** equilibrium is that the resultant of the forces be zero; the condition that there be **rotational** equilibrium is that the sum of the moments about any axis be zero.

The conditions of equilibrium for special cases are all contained in a very general principle, the principle of virtual displacements. In order to state this principle we must first define the **work** of a force. Suppose that a point moves a distance s in a straight line under the action of a constant force. If the direction of the force makes an angle θ with the direction of motion, then the work done by the force is defined by

$$W = Fs \cos \theta.$$

If the motion is along a curve and the force is variable the element of work done in an elementary displacement is

$$\Delta W = F \Delta s \cos \theta.$$

Using the vectorial representation the work done is the scalar product of the force and the elementary displacement:

$$\Delta W = \mathbf{F} \cdot \Delta \mathbf{r}.$$

We may consider that a finite displacement is made up of a large number of successive steps, $\Delta \mathbf{r}$. We then form the sum of the products:

$$\Sigma F \cdot \Delta \mathbf{r}.$$

As the subdivisions $\Delta \mathbf{r}$ approach zero as a limit the sum approaches a limit. We now make the physical assumption that we can express $\mathbf{F}$ as a continuous function of $\mathbf{r}$, and that the smaller the physical intervals $\Delta \mathbf{r}$, the more accurately is the sum obtained from measurements given by

$$\int \mathbf{F} \cdot d\mathbf{r}$$

Thus one expresses the definition of work by

$$\int \mathbf{F} \cdot d\mathbf{r}$$

In Cartesian coordinates the work is defined by

$$\int (X dx + Y dy + Z dz)$$

where X, Y, Z are the components of $\mathbf{F}$. It is to be noted that we use the concepts of pure analysis to express a physical definition.

In order to introduce the principle of virtual displacements I shall consider a simple case, the equilibrium of a small body resting on a frictionless inclined plane of angle of elevation θ. In order to keep the body in equilibrium we apply a force P parallel to the plane. Then we say that the body is acted upon by the given forces, P and Q, where Q is the weight. If we investigate the equilibrium of this body by the previously stated condition for a particle, we assume that a third force acts on the body, the reaction of the plane. Thus the body is in equilibrium under the action of P, Q and the normal reaction. By application of the principles we find

$$P = Q \sin \theta.$$

This result is verifiable by experiment.

Let us now suppose that the body is given an elementary virtual displacement, that is, one which is consistent with the constraints. In the present case a virtual displacement is a motion of the body up or down the plane. Now a calculation, based upon the previous experimental result and a definition, shows that in the present example the work done by the given forces is zero. This result is an instance of the principle of virtual displacements. Let δs be the value of the displacement up the plane. Then the work done by P is

$$P\delta s.$$

The work done by Q is

$$-Q\delta s \sin\theta.$$

The total work done by the given forces is

$$P\delta s - Q\delta s \sin\theta.$$

Since $P = Q \sin\theta$, the work done by the given forces is equal to zero. It is important to note that the work done by the normal reaction is zero, because the direction is at right angles to the direction of motion.

The principle of virtual displacements, which has been demonstrated for the given simple case, may now be expressed in the following general form. The necessary and sufficient condition for the equilibrium of a system which is subject to internal connections and constraints and to the actions of given forces $\mathbf{F}_1$, $\mathbf{F}_2$, etc., is that, in a virtual displacement whose value is a small quantity of the first order, the total virtual work done by the given forces is equal to zero to the first order of small quantities. A virtual displacement is one which is consistent with the connections and constraints at a given time. Analytically

$$\Sigma\mathbf{F}\cdot\delta\mathbf{r} = 0$$

or

$$\Sigma(X\delta x + Y\delta y + Z\delta z) = 0.$$

From this principle one can deduce all the conditions of equilibrium for special cases, such as the equilibrium of points, rigid bodies, etc. Accordingly we shall accept it as the fundamental principle of statics.

Relative to our definition of force the principle of virtual displacements is a well-founded generalization from experience. We may now transform the concept of force. We may transform the principle of virtual displacements from a generalization, relative to certain definitions of force, into a definition of force, i.e., force is now to be defined by the principle of virtual displacements. This is a physical definition since we have a physical concept of rigid body. I remark that the definition is of the concept of given force; we can include forces due to connections and constraints by removing the condition that the displacement be compatible with the connections and constraints.

Our new concept of force is applicable to experience, since weights, the tension in spring balances, etc., are instances of the physical quantity force. We may express our result this way. It is possible to assign numbers F_1, F_2, etc., such that the principle of virtual displacements is satisfied.

The principle of virtual displacements gives a definition of force which is not bound to weights, spring balances, etc. We thus achieve complete emancipation from the original anthropomorphic definition of force. Muscular force is an approximate example of our present definition.

We may summarize statics as follows.

(1) *The postulates of Euclidean geometry define the concept of rigid body.*

(2) *A designated standard of length gives physical significance to the concept of rigid body.*

(3) *On the basis of the physical concept of rigid body the principle of virtual displacements defines the concept of force.*

(4) *A definite number is assigned to the standard of force, for example, a weight in a given position. It is an experimental fact that weights, the tension of spring balances, and muscular forces satisfy the definition of force approximately.*

3. THE KINETICAL CONCEPT OF FORCE

In the preceding section we have defined force in terms of its character as exhibited in equilibrium. We recall, however, that the primitive concept of force implies the power to pro-

duce motion. We must therefore complete our discussion by considering the **kinetical aspects of force.**

In this exposition we adopt the statical definition of force and then assume that when a force is not balanced it retains its character as a force. Thus I attach a spring balance to a body and measure the tension in the spring by the weight which it balances. Next I attach the spring balance to a body resting on a frictionless plane and produce an acceleration, and assume that although the spring balance is now in motion the same extension indicates the occurrence of the same force as in the statical case. This process involves an extension of the physical concept of force which is expressed by the propositions set forth in the following discussion.

Let us suppose that an unbalanced force acts on a body. As long as the force acts the body is accelerated, that is, the acceleration is a function of the force.

$$\frac{d\mathbf{v}}{dt} = f(\mathbf{F}). \tag{1}$$

This proposition expresses Newton's First Law of Motion, which states that every body continues in its state of rest or of uniform motion in a straight line unless compelled by impressed force to change that state. The significance of equation 1 is that momentum is a function of velocity. We shall say that when a body is accelerated it acquires a momentum which is a function of the velocity. The vector momentum is therefore expressed by

$$\mathbf{I} = f(\mathbf{v}).$$

The concept of momentum is now further defined by

$$\mathbf{I} - \mathbf{I}_0 = \int_{t_0}^{t} \mathbf{F}dt \tag{2}$$

where $\mathbf{I}_0$ is the initial momentum at the time t_0 and $\mathbf{I}$ is the momentum at the time t.

$\int \mathbf{F}dt$ is called the impulse of the force. Thus change of momentum is defined to be equal to the impulse of the force.

If the force is constant:

$$\int_{t_0}^{t} \mathbf{F}dt = \mathbf{F}(t - t_0).$$

If the body starts from rest, $\mathbf{I}_0 = 0$, hence

$$\mathbf{I} = \int \mathbf{F}dt.$$

Since the measures of $\mathbf{F}$ and t have been assumed, $\mathbf{I}$ can be calculated by the preceding equations. Thus the concept of force is of something which generates momentum. If we assume that the process is continuous

$$\frac{d\mathbf{I}}{dt} = \mathbf{F}. \tag{2'}$$

Equation 2 expresses Newton's Second Law of Motion.

We need further to consider the forces between interacting bodies, for example, the forces exerted upon each other by two colliding billiard balls. Suppose that the hooks of two spring balances are attached together and the springs are extended. The one balance then indicates the force exerted upon the other, and the other balance indicates the force exerted upon the first one. Experiments based upon the statical definition of force show that the two forces are equal and opposite. We assume that this property of force holds when the bodies are accelerated. Hence

$$\mathbf{F}_1 = -\mathbf{F}_2. \tag{3}$$

This is Newton's Third Law of Motion.

From this law we can deduce for two interacting bodies

$$\mathbf{I}^1 - \mathbf{I}^1_0 = -(\mathbf{I}^2 - \mathbf{I}^2_0)$$

$$\mathbf{I}^1 + \mathbf{I}^2 = \mathbf{I}^1_0 + \mathbf{I}^2_0. \tag{4}$$

that is, the total momentum after the interaction is equal to the momentum before the interaction. This is the principle of the conservation of momentum, which is thus explained as a deduction from an experimental law and a definition.

The preceding discussion defines the concept of **momentum.** The equations also extend the concept of force; force, which

was initially defined statically, is exhibited as generating momentum. We now transform our concept of force; force and momentum are implicitly defined by the fundamental equations of dynamics (1 to 3). When $\mathbf{F} = 0$ the momentum and hence the velocity is constant. Rest is a special case of constant velocity. Now the essential feature of equilibrium being that the velocity of a body is constant, the statical concept of force is a limiting case of the complete concept as defined by the fundamental equations.

It is not precise to say that we may begin with a statical definition of force, for force and momentum are defined completely by the three fundamental equations. However, in order to make measurements we must start somewhere. The spirit of the preceding discussion is that we obtain a measure of force in the limiting case of statics; then we calculate momentum. However, we may make momentum the first measured quantity and then force is calculated as the rate of change of momentum.

In order to use this new method I begin with the statement that a moving body has momentum; the momentum of the body is a vectorial function of the velocity.

$$\mathbf{I} = f(\mathbf{v}).$$

Let us now suppose that two bodies (1) and (2) move towards each other in the same straight line with constant velocities such that upon collision they interlock and both come to rest. By definition the momentum of the two bodies is equal in value and opposite in direction, so that the total momentum of the two bodies is zero. The equality of bodies with respect to momentum is a symmetrical and transitive relation. We can readily define equality in momentum in the same direction. Let two bodies having the same momentum in the same direction collide with another body moving with constant velocity in the opposite direction so that upon collision all the bodies are brought to rest. By definition, the value of the momentum of the latter body is twice that of either of the two bodies having the same momentum in the same direction. In this way one can determine when a body has a momentum whose value is two, three, etc., times as great as another. We

thus have a definition of momentum in terms of the principle of the conservation of momentum; a more detailed discussion would explain how one takes account of the vectorial character of momentum. We may state our first equation

$$\mathbf{I}^1 + \mathbf{I}^2 = \text{constant.} \tag{1}$$

It is possible to assign vectors $\mathbf{I}^1, \mathbf{I}^2$, such that the principle of conservation of momentum is satisfied.

The second proposition states that force is a function of acceleration

$$\mathbf{F} = f\left(\frac{d\mathbf{v}}{dt}\right). \tag{2}$$

This expresses Newton's First Law. We now express the character of force more definitely by the following equation

$$\int_{t_0}^{t} \mathbf{F}dt = \mathbf{I} - \mathbf{I}_0. \tag{3}$$

Assuming continuity, we get

$$\mathbf{F} = \frac{d\mathbf{I}}{dt}. \tag{3'}$$

The principle of conservation of momentum for two bodies is

$$\mathbf{I}^1 + \mathbf{I}^2 = \text{constant.}$$

Assuming continuity, we deduce

$$\frac{d\mathbf{I}^1}{dt} + \frac{d\mathbf{I}^2}{dt} = 0 \tag{4}$$

$$\frac{d\mathbf{I}^1}{dt} = -\frac{d\mathbf{I}^2}{dt}$$

$$\mathbf{F}_1 = -\mathbf{F}_2.$$

This is the third law of motion. The present development explains the third law as a deduction from an experiment and a definition. The equations 1 to 3 constitute an implicit definition of force and momentum. We do not avoid the concept of force in starting with momentum, for the experiments by which we define bodies with equal momentum require the interaction of the two bodies, and this is merely another name for force.

Thus a proper description of the experiments employed to define momentum employs the concept of force.

It is frequently stated that

$$\mathbf{F} = \frac{d\mathbf{I}}{dt}$$

is the kinetical definition of force. In my opinion it is only part of it; one must also have $\mathbf{F}_1 = -\mathbf{F}_2$. The statement that force is defined by the Newton's Second Law alone is equivalent to the statement that the commutative law alone is adequate to define addition.

The fundamental equations define force and momentum implicitly. We may use either a statical unit of force or a kinetical unit of force.

As stated thus far our equations apply only to free bodies; we must also consider systems which are subject to internal connections and constraints. For example, if two particles are connected by a rigid rod and also are constrained to move in a circular groove, the problem is to set up a dynamical principle for the motion of such a system when certain forces are given as impressed forces. In order to solve the problem we think of the connections and constraints as giving rise to forces of constraint, or reactions, $\mathbf{N}$. Assuming that the system is constituted of n material particles, each of which is acted upon by given forces and forces of constraint, then for each particle we can write the equation of motion

$$\mathbf{F} + \mathbf{N} = \frac{d\mathbf{I}}{dt}$$

$$\mathbf{F} - \frac{d\mathbf{I}}{dt} = -\mathbf{N}.$$

However, we do not know the forces of constraint. Now, we saw that statics could be based upon the principle of virtual displacements, which states that, if the system is in equilibrium, the virtual work of the given forces in an elementary virtual displacement is zero, the principle depending upon the assumption that the work done by the forces of constraint is zero. We now assume that even when the system is not in equilibrium, the work done by the forces of constraint is zero in an

elementary virtual displacement carried out at a certain time. Then

$$\Sigma\left(\mathbf{F} - \frac{d\mathbf{I}}{dt}\right) \cdot \delta\mathbf{r} = -\Sigma\mathbf{N} \cdot \delta\mathbf{r} = 0$$

where the summation sign indicates the addition of the elements of work for all the particles. In Cartesian coordinates the equation is

$$\sum\left\{\left(X - \frac{dI_x}{dt}\right)\delta x + \left(Y - \frac{dI_y}{dt}\right)\delta y + \left(Z - \frac{dI_z}{dt}\right)\delta z\right\} = 0.$$

This equation reduces to the principle of virtual displacements when the momentum of each particle is constant. Under certain conditions the equation yields the principle of conservation of momentum.

4. CLASSICAL DYNAMICS

Experiment reveals that the absolute value of the momentum of a body is approximately proportional to the absolute value of the velocity. Thus

$$\frac{|\mathbf{I}|}{|\mathbf{v}|} = m.$$

This factor of proportionality is called the mass of the body. We then have

$$\mathbf{I} = m\mathbf{v}.$$

Hence

$$\mathbf{F} = \frac{d\mathbf{I}}{dt} = m\frac{d\mathbf{v}}{dt} = m\frac{d^2\mathbf{r}}{dt^2}.$$

In Cartesian coordinates the equations are

$$X = m\frac{d^2x}{dt^2}$$

$$Y = m\frac{d^2y}{dt^2}$$

$$Z = m\frac{d^2z}{dt^2}.$$

These are called the equations of motion for a particle.

Now experiment shows that if a body is allowed to fall freely the acceleration is constant. The force acting, $F = mg$, is the weight of the body. Furthermore, g is the same constant for all bodies. Thus we arrive at the result that the ratio of the weights of two bodies is proportional to the ratio of the masses, which is the basis of the measurement of mass by means of the beam balance. We postulate that the mass of the standard is constant and independent of position. In physics the standard of mass is the kilogram. This is by definition the mass of a particular material body which is kept in the vaults of the International Bureau of Weights and Measures. A gram is 1/1000 of a kilogram. The kinetical unit of force, called the **dyne**, is defined as the force which gives a mass of one gram an acceleration of one centimeter per second per second.

The mass of a body is the fundamental invariant of classical dynamics; in traditional physics mass is the measure of a property, **inertia**. The classical physicist thought that he had a clear conception of inertia as an attribute of material substance. However, the status of inertia as a property is very dubious; it seems evident that we are not acquainted with inertia in the sense in which we are acquainted with redness, extension, etc. It is true that we appear to have direct acquaintance with inertia when we touch a body and start it in motion. However, my interpretation of this experience is that it consists of sensations of effort. The property of inertia is ascribed to a body in order to represent the results of experiments on the body. We are acquainted with the phenomena in the experiment, and know the measure which we assign to the body, but we do not appear to be acquainted with the property ascribed to the body. The late introduction of the concept of inertia by Newton further indicates that inertia is not a character with which we are directly acquainted.

5. THE CONCEPT OF ENERGY

Suppose that a force **F** acts on a particle through a displacement **dr**. The work done by the force is equal to

$$\mathbf{F}\cdot d\mathbf{r} = m\frac{d\mathbf{v}}{dt}\cdot d\mathbf{r} = d(\tfrac{1}{2}mv^2)$$

$$\int \mathbf{F}\cdot d\mathbf{r} = \tfrac{1}{2}mv^2 - \tfrac{1}{2}mv_0^2.$$

$\frac{1}{2}mv^2$ is called the **kinetic energy** in classical dynamics. The gain in kinetic energy is defined as equal to the work done on the body. Conversely, if a body is brought to rest by the resistance of an obstacle, the work done against the obstacle is equal to the loss in kinetic energy, in this case, the initial kinetic energy. The kinetic energy of a body is thus equal to the work which the body can do in virtue of its motion.

Suppose that there is a function of position $V(x,y,z)$ so that

$$X = -\frac{\partial V}{\partial x}$$

$$Y = -\frac{\partial V}{\partial y}$$

$$Z = -\frac{\partial V}{\partial z}.$$

The force is then called a conservative force, for we have

$$\int \mathbf{F}\cdot d\mathbf{r} = -\int dV = \tfrac{1}{2}mv^2 - \tfrac{1}{2}mv_0^2.$$

Then

$$V_0 - V = T - T_0$$

where $T = \frac{1}{2}mv^2$.

Thus we get

$$T + V = T_0 + V_0.$$

In this case the quantity $T + V$ is a constant of the motion. Although the kinetic energy T changes, the change in V is such the sum of the two remains constant. In order to have

a theorem of the conservation of energy we call V the **potential energy.** Hence when a body moves in a conservative field of force the total energy remains constant. The earth's gravitational field, an electrostatic field, etc., are examples of a conservative field of force; the tension in a spring is a conservative force.

Suppose, for example, that a body is projected upwards from the surface of the earth with a given initial kinetic energy. As it rises it loses kinetic energy but gains an equal amount of potential energy. Neglecting friction, the total energy remains constant.

6. APPLICATIONS OF CLASSICAL DYNAMICS

The equations of classical dynamics constitute an implicit definition of force and mass, so that dynamics furnishes us with a general schema. In order to apply dynamics to physical phenomena, one must know F as a function of other physical quantities. Now it is possible to determine the motion of interesting systems by comparatively simple assumptions about F, such as:

(1) $F = -kx$, for a particle attached to the end of a spring.
(2) $F = G \cdot m'm/r^2$ for a particle of mass m in a gravitational field due to a particle of mass m' at a distance r from the given particle.

An equation such as $-kx = ma$ is the law of motion for the given system.

In the study of elastic bodies it is possible to express stress as a function of strain, and by substituting this function in the differential equations of the theory of elasticity one obtains equations which describe the propagation of elastic waves. The concepts of dynamics serve to define elastic properties of bodies. The empirically discovered constant ratios of stress to strain are called **moduli of elasticity.**

7. THE PRINCIPLE OF RELATIVITY

In the preceding account of dynamics I have assumed that a proper system of reference has been employed, for example, a system with origin at the center of mass of the solar system

and oriented with respect to the fixed stars. In such a system, which may be called fixed, the laws of dynamics have a simple form; moreover, they have the same form in a system which is moving with uniform rectilinear velocity u relative to the fixed system. That is, the form of the equations is invariant under a transformation from the fixed system to a system moving with uniform rectilinear motion. If x, y, z, t are the coordinates and the time in the fixed system, and x', y', z', t' are the coordinates and the time in the moving system, then the transformation equations of the Galilean group are, as previously set forth,

$$x' = x - ut$$

$$y' = y$$

$$z' = z$$

$$t' = t.$$

Now

$$\ddot{x}' = \ddot{x}, \quad X' = X.$$

$X = m\ddot{x}$ is transformed into $X' = m\ddot{x}'$. (The symbol $\ddot{x}$ is written for $\frac{d^2x}{dt^2}$.)

Thus the laws of classical dynamics are invariant under the Galilean group; this is the principle of relativity of uniform motion in classical dynamics, and it is based upon the following assumptions:

(1) *The distance between two points on a rigid body is the same in the fixed system and in the moving system. Thus the several systems have the same space in common.*
(2) *The time of an event is the same in the two systems. Thus the several systems have the same time system in common.*
(3) *The mass of a body is the same in all frames.*

The experimental import of the principle of relativity in classical dynamics is that it is impossible to detect, by a dynamical experiment, a uniform rectilinear motion of a system of reference.

Thus classical dynamics presupposes the existence of a set of privileged systems of reference, moving relatively to each

other, in which the laws of dynamics have a simple form. In accelerated systems, a rotating system, for example, there are characters of motion which do not conform to classical dynamics. Newton assumed that such phenomena revealed acceleration with respect to absolute space. Sometimes it is convenient to use an accelerated system of reference, in which case one changes the laws of dynamics by the addition of fictitious forces, for instance, centrifugal forces.

8. THE FUNDAMENTAL PRINCIPLE OF DYNAMICS

We have seen that the fundamental principle of dynamics can be written

$$\sum\left\{\left(X - \frac{dI_x}{dt}\right)\delta x + \left(Y - \frac{dI_y}{dt}\right)\delta y + \left(Z - \frac{dI_z}{dt}\right)\delta z\right\} = 0.$$

In classical dynamics, $I_x = m\dot{x}$, etc. Hence in Cartesian coordinates the fundamental principle of classical dynamics is

$$\Sigma\{(X - m\ddot{x})\,\delta x + (Y - m\ddot{y})\,\delta y + (Z - m\ddot{z})\,\delta z\} = 0.$$

This is Lagrange's form of d'Alembert's principle.

An important step in advance was made by Lagrange, who transformed the fundamental principle from Cartesian coordinates into generalized coordinates. Thus far we have used $3n$ Cartesian coordinates to specify the configuration of a system of n particles. Now on account of the connections and constraints these Cartesian coordinates are not in general independent. It is then possible, however, to specify the configuration of the system by a set of f independent quantities of which the $3n$ Cartesian coordinates are functions. Thus

$$x_1 = f_1(q_1 \ldots q_f, t), \text{ etc.}$$

The independent quantities $q_1 \ldots q_f$ are called generalized coordinates, and for so-called holonomic systems, to which we shall restrict ourselves, the number of generalized coordinates is equal to the number of degrees of freedom of the system. The employment of generalized coordinates introduces a great simplification. For example, the configuration of a rigid body, which consists of countless particles, can be specified by six coordinates: three to determine the position of a

point, two to determine the direction of an axis through the point, and one to specify an azimuth with respect to rotation about the axis. Lagrange succeeded in transforming d'Alembert's principle from Cartesian coordinates to generalized coordinates. Lagrange's equations in generalized coordinates are

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_k}\right) - \frac{\partial T}{\partial q_k} = Q_k \quad k = 1 \ldots f.$$

T is the kinetic energy.

Q_k is the generalized component of force.

In dynamics one is most interested in conservative systems for which the given forces are derivable from a potential V,

$$Q_k = -\frac{\partial V}{\partial q_k}.$$

It is then possible to define a function

$$L = T - V$$

such that Lagrange's equations become

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k} - \frac{\partial L}{\partial q_k} = 0.$$

A new form of the equations for a conservative system was found by Hamilton. Introducing

$$p_k = \frac{\partial L}{\partial \dot{q}_k}$$

$$H = \Sigma p_k \dot{q}_k - L$$

we get

$$\dot{q}_k = \frac{\partial H}{\partial p_k}$$

$$\dot{p}_k = -\frac{\partial H}{\partial q_k}.$$

where H is a function of q_k, p_k and t. These are Hamilton's canonical equations.

For a conservative system we have the interesting result that if a single function is known for the system the motion can be determined. We may choose as the single function the

Lagrangian function $L(q_k, \dot{q}_k, t)$, or the Hamiltonian function $H(q_k, p_k, t)$.

Another form of the fundamental dynamical principle can be stated in terms of the calculus of variations. We may represent the momentary configuration of a conservative system by a point in an f dimensional space of which the coordinates are $q \ldots q_f$. A motion of the system is represented by a curve traced in this space. Consider the integral

$$\int_{t_1}^{t_2} L dt$$

for the actual motion of the system. We may consider this integral evaluated along the path described by the point during an actual motion of the system. Suppose that we compute the integral for an infinitely near motion which has the same initial and final configuration and the same initial and final time. Then the form of L is such that to the first order of small quantities the integral is the same for the varied path as for the actual path. This is Hamilton's principle of stationary action; it is written

$$\delta \int L dt = 0.$$

The principle may also be expressed by the proposition that for the actual motion the integral

$$\int L dt$$

is an extremum, that is, a maximum or a minimum.

The most elegant form of the fundamental dynamical principle is the statement: It is possible to find a function $L(q_k, \dot{q}_k, t)$ such that

$$\delta \int L dt = 0.$$

9. REVERSIBILITY

The principles of dynamics are especially designed to describe the motion of conservative systems, in which the given forces are functions of position and are derivable from a potential energy which is a function of position. The differential

equations of the system are ordinary linear differential equations of the second order with the time as the independent variable, and their solution describes a reversible motion, that is, if the time is changed in sign the solution still satisfies the differential equation. For example, for small vibrations the differential equation for the motion of a simple pendulum is

$$\ddot{\theta} + \frac{g}{l}\theta = 0.$$

The solution is

$$\theta = A \sin\left(\sqrt{\frac{g}{l}}t - \delta\right).$$

Another solution is obtained if one replaces t by $-t$. Thus there is no distinction between the direct motion of the pendulum and the reverse motion, and therefore for a conservative system there is no distinction between the forward and backward direction of time.

10. CONCLUSIONS

The equations of dynamics are functional relations between the physical quantities, the time, the distance, the mass, the force. We may view the analytical equations as constituting an implicit definition of the formal properties of space, time, mass and force. The definition becomes physical if we point out a rigid rod, a clock, and a material body, which respectively embody units of length, time and mass. A common space and time for a set of priviliged systems is defined by the Galilean transformation. It is an experimental fact that the concepts thus defined are applicable, that is, it is possible to assign t, x, y, z, m, F, such that the equations of dynamics are satisfied.

REFERENCES

APPELL, P. Traité de Mécanique Rationelle. Paris, 1919.
HAAS, A. E. Die Grundgleichungen der Mechanik. Leipzig, 1914.
WHITTAKER, E. T. Analytical Dynamics. Cambridge, 1916.

CHAPTER VI

ELECTRODYNAMICS

1. ELECTROSTATICS

We shall build the theory of electricity and magnetism upon the definition of force given in dynamics.

As early as the time of Thales it was known that if a rod of amber is rubbed it will attract light bodies, such as small bits of paper. Bodies consisting of other substances also possess this property; if one rubs the hard rubber barrel of a fountain pen on a piece of cloth it will attract light bodies. Bodies which exhibit these properties are said to be electrified, or to have the property of electrification. Thus we ascribe **electrification** to a body when it is the condition of an acceleration, in some other body, which is not directly describable in terms of gravitation, mechanical pressure, etc. We say that an electrified body is the origin of a field of force which is ascribed to the vicinity of the body in virtue of the accelerations experienced by bodies in the field. Speaking uncritically, we may say that the property of electrification is defined in terms of the effects of electric action. Electrification is not a property which is immediately apprehended like color, sound, hotness, etc.; the state of electrification is attributed to bodies in order to describe observable interactions with other bodies.

Simple experiments reveal that electrified bodies exert forces on each other which are either forces of attraction or repulsion. From experience we obtain laws of the attraction and repulsion of electrified bodies, and then transform these laws into postulates which implicitly define the concepts of similarly electrified and dissimilarly electrified bodies. By definition two bodies are similarly electrified if they repel each other; dissimilarly electrified, if they attract each other. It is obvious that these concepts are applicable to experience: an electrified glass rod is said to be **positively electrified; a body**

which is attracted by an electrified glass rod is said to be **negatively** electrified.

Some bodies have the capacity to conduct an electrical condition from one point to another, whereas other bodies do not have this property. Thus one may distinguish between **conductors** and **non-conductors.**

The early investigators attributed the electrified state of a body to the presence of a substance, called electricity, which, in view of the fact of conduction, was thought of as a fluid. There were two theories of the nature of electrical fluid: according to the two-fluid theory there were two kinds of fluid corresponding to the two kinds of electrification; Franklin proposed the one-fluid theory, ascribing positive electrification to an excess of fluid and negative electrification to a deficiency of fluid. We need, however, only the concept of a state of electrification which is ascribed to bodies on the basis of observable phenomena.

We must now introduce concepts of physical quantities to be assigned to electrified bodies. The initial definitions presuppose that we have an estimate of the constancy of a state of electrification, which is based upon an estimate of constancy of effects.

Let us suppose that we have given an electrified body at a point in air (strictly speaking in free space). We assume that the body has constant electrical properties. Let a small electrified body (1) be placed at the point Q; then there is a force of attraction or repulsion upon (1), $\mathbf{F}_1$. If we replace (1) by another similarly electrified body (2), the force upon it is $\mathbf{F}_2$. $\mathbf{F}_1$ and $\mathbf{F}_2$ are in the same line and the ratio of their values may be written F_1/F_2. Suppose that the bodies (1) and (2) are such that the forces are equal; then if we vary the position of Q the forces upon the bodies vary, but are always equal. Thus the two bodies are similar with respect to the action of force; the relation between them is symmetrical and transitive. We express the facts by the statement that the bodies are equal with respect to electrification. Since the force is a physical quantity we assign to the two bodies the same number, the electric charge. Then the ratio F/F_s, where F_s is the force on a unit charge, will serve to define the charge of any body in terms

of the unit. We may define the unit charge as the charge which repels an equal like charge with a force of one dyne at a distance of one centimeter in free space. In this way we define the physical quantity, electric charge. The electric charge may be positive or negative. In traditional physics electric charge is the measure of quantity of electricity; in our terminology quantity of electricity is electric charge which is a number assigned on the basis of measurements of force. We postulate the unit of charge to be invariable upon displacement.

Suppose that at a given point in the field of a charged body one brings up charges e_1 and e_2 in succession. The ratio of the forces on these charges is F_1/F_2. The forces vary with the position of the test bodies, but the ratio remains the same. We may describe the facts by analyzing the force into two factors, $\mathbf{E}$ and e, so that

$$\mathbf{F}_1 = \mathbf{E}e_1$$
$$\mathbf{F}_2 = \mathbf{E}e_2.$$

$\mathbf{E}$ depends upon the charge of the body which excites the field and also upon the position of the test body. The electric field $\mathbf{E}$ is defined as the force on a unit charge.

The electric flux through a surface over which the field is uniform is the product of the normal component of the field and the area of the surface. If the field is variable the flux is $\int \mathbf{E} \cdot d\mathbf{S}$. The flux through a surface may be represented by lines through the surface.

We introduce a new physical quantity, φ, called the **potential.** The potential at a point is the work which must be done against the forces of the field in order to bring a unit positive charge from infinity to the point. The electric field is the negative maximum rate of change of the potential function with respect to change of position. Thus if, φ is a function of x, y, z,

$$E_x = -\frac{\partial \varphi}{\partial x}$$
$$E_y = -\frac{\partial \varphi}{\partial y}$$
$$E_z = -\frac{\partial \varphi}{\partial z}.$$

One can experimentally verify Coulomb's law of force between point charges:

$$F = \frac{e_1 e_2}{r^2}.$$

For a point charge the value of the field at a distance r is $E = e/r^2$, and the potential is $\varphi = e/r$.

From Coulomb's law of force one can deduce Gauss' theorem, which states that the electric flux through a closed surface is equal to $4\pi\Sigma e$, where Σe is the sum of the charges within the surface,

$$\int \mathbf{E} \cdot d\mathbf{S} = 4\pi\Sigma e.$$

If there is a continuous distribution of charge we use the charge density ρ, which is the electric charge per unit volume. Then Gauss' theorem becomes

$$\int \mathbf{E} \cdot d\mathbf{S} = 4\pi \int \rho d\tau.$$

Gauss' theorem is an integral law.

If we assume that it holds as we allow a volume element enclosing a point, x, y, z to approach zero as a limit, we obtain

$$\text{div } \mathbf{E} = 4\pi\rho$$

that is, the electric flux which diverges from an element is $4\pi\rho$ per unit volume. Thus the space rate of change of the field is functionally related to the charge density at the point.

In Cartesian coordinates:

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 4\pi\rho.$$

In empty space

$$\text{div } \mathbf{E} = 0.$$

From the inductive point of view we begin with Coulomb's law, an integral law. It states a functional relation between finite forces, charges, and distances. On the assumption that properties throughout a region hold as we proceed to the limit, we deduce a differential equation which expresses a functional relation at a point. Physically we are not able to obtain exact differential equations. However, we express the differential

equations of physics in the language of pure analysis for which the concept of limit is precisely defined, and we employ the operations of pure analysis. Since the definiteness of analysis outruns that of physical quantities we are satisfied with a sufficiently approximate solution.

Again, we may carry a charged body around a closed path in an electrostatic field and discover that the total work done is zero.

$$\oint \mathbf{E}e \cdot d\mathbf{r} = 0.$$

By a transformation and the assumption of homogeneity as we proceed to the limit we obtain

$$\text{curl } \mathbf{E} = 0.$$

The fundamental equations of electrostatics are

$$\text{div } \mathbf{E} = 4\pi\rho$$
$$\text{curl } \mathbf{E} = 0.$$

These differential equations may be made the basis of a deductive theory from which one can deduce the observable properties of the electrostatic field. Now a theory based upon differential equations partakes largely of the nature of hypothesis, but it gives a more detailed analysis of physical states and processes. On the assumption that conditions at a point hold throughout a finite region one can deduce directly verifiable laws, for example, Coulomb's law.

Thus far we have considered electric charges and electric fields in free space. Let us now suppose that the medium is a **dielectric**, that is, an insulator. Coulomb's law of force between two point charges becomes

$$F = \frac{e_1 e_2}{\epsilon r^2}$$

where ϵ depends upon the medium. Gauss' theorem no longer holds for the electric field in a dielectric. Hence we introduce a new vector $\mathbf{D}$ such that

$$\int \mathbf{D} \cdot d\mathbf{S} = 4\pi\Sigma e = 4\pi \int \rho d\tau.$$

$\mathbf{D}$ is called the electric displacement.

We deduce

$$\text{div } \mathbf{D} = 4\pi\rho.$$

It is an experimental result that when the dielectric is isotropic

$$\mathbf{D} = \epsilon\mathbf{E}.$$

ϵ is the dielectric constant of the medium.

2. PERMANENT MAGNETISM

It was known to the ancients that certain minerals have the power to attract iron; bodies having this property are called lodestones. If a piece of steel is brought in contact with a lodestone the steel acquires the property of attracting iron. Bodies which exhibit this power of attracting iron are called permanent magnets. A **magnet** is thus characterized by the fact that it is the condition of an acceleration, in a piece of iron, which is not directly describable in terms of gravitation, mechanical pressure, etc. We may say that a magnet is the origin of a field of force which is ascribed to the vicinity of the body in virtue of the acceleration experienced by a piece of iron within the field. The essential property of a magnet may be called a state of magnetization. Speaking uncritically, we may say that the property of magnetization is defined in terms of the effects of magnetic action. Magnetization is not a property which is immediately apprehended, like color, sound, hotness, etc. The state of being magnetized is attributed to bodies in order to describe specific observable interactions with iron.

If a piece of iron is placed successively near different parts of the magnet one finds that the parts exert different attractive forces, i.e., the property of magnetization is possessed most strongly by certain regions of the magnet. These seats of magnetic action are called the **poles** of the magnet; a magnet always has at least two poles. Out of a long slender rod of steel one can construct a magnet with the special property that it has two poles which are located approximately at the two ends. In the present discussion of magnets we shall assume that they are of this type.

One may next study the mutual action of magnetic poles. Assuming as a first approximation that the poles may be

viewed as isolated, we find that two poles exert forces on each other which are either forces of attraction or repulsion. Laws which are inductions from experience are then transformed into postulates which constitute implicit definitions of like and unlike poles. By definition, two poles are similar with respect to magnetization if they repel each other and dissimilar if they attract each other. If a magnet is pivoted so that it can swing in a horizontal plane it takes a position in which the axis is in a north and south line. The pole in the direction of the north pole of the earth is called the **north** or **positive** pole of the magnet; the other pole is called the **south** or **negative** pole.

We must now introduce concepts of physical quantities. I first define pole strength, or quantity of magnetism, on the assumption that the magnets are in free space. By a procedure which is exactly the same as that employed in electrostatics, one defines the physical quantities pole strength m, magnetic field **H**, magnetic flux, magnetic potential, and verifies Coulomb's law of force. If m is the pole strength, then Coulomb's law is

$$F = \frac{m_1 m_2}{r^2}.$$

3. THE ELECTRIC CURRENT

It has already been pointed out that some material substances, notably metals, are able to conduct an electrical condition from one point to another. This process is called the flow of an electric **current** through the conductor. It is obvious that the conventional terminology is a relic of the theory that electric action is caused by a fluid, called electricity. We must now define a physical quantity which will serve to describe an electric current.

Suppose that we have given a condenser, the plates of which carry electric charges $+ e$ and $- e$, respectively. If we connect the two plates with a wire, the charges on the plates decrease to zero; we then say that a current has flowed through the wire. I define the strength of the current by

$$J = \frac{de}{dt}.$$

Suppose that the current which discharges the condenser is allowed to flow through an electrolytic cell which consists of a solution of silver nitrate and two silver plates. Then we find that silver is removed from the plate at which the current enters and is deposited on the plate at which the current leaves the cell, the mass of the silver deposited being proportional to e. If we perform this experiment for different values of e, we find that the proportionality of mass to charge always holds. The deposition of silver by a current thus offers a simple and accurate method of measuring the electric charge which flows through the cell. If equal masses are deposited in equal times the current is stationary. Suppose that a circuit is made by joining two wires of different materials. If the two junctures are kept at different constant temperatures a stationary current flows in the circuit.

We now have the physical quantities which are needed to set up Ohm's law. Suppose that a stationary current flows through a homogeneous wire from a point at potential φ_1 to a point at potential φ_2. Let J be the strength of the current. Then it is an experimental result that the difference of potential is proportional to the current:

$$\varphi_1 - \varphi_2 = RJ.$$

The factor of proportionality is constant for a given length of wire, and is defined to be the measure of the property of electrical resistance which is ascribed to the wire. Thus the physical quantity electrical **resistance** is defined by the above equation. In this way Ohm's law is explained as an experimental result which is the basis of the definition of a new physical quantity.

The foregoing statement of Ohm's law expresses it as an integral law. One can derive a differential form of the law on the assumption that it continues to hold as the wire becomes vanishingly small in length and area. For an isotropic conductor

$$\mathbf{i} = \sigma \mathbf{E},$$

where σ is the conductivity and $\mathbf{i}$ is the current density, the absolute value of which is the current per unit area of cross-section.

In an isolated, perfectly homogeneous, closed circuit no current flows. However, if there is a battery in the circuit, or junctures of different metals maintained at different temperatures, a current flows. We say that there is an impressed electromotive force in the circuit which is defined by

$$E^e = JR.$$

We thereby extend Ohm's law to the complete circuit and use it to define electromotive force.

4. THE FIRST EQUATION OF THE ELECTROMAGNETIC FIELD

We are now prepared to describe the relation between electricity and magnetism. In 1820 Oersted discovered that if a wire carrying a current is placed above a compass needle and parallel to it, the needle is deflected. This experiment indicates that an electric current excites a magnetic field. A pictorial description is that a magnetic field encircles a conductor carrying a current. Biot and Savart discovered that if a constant current flows through a straight wire the magnetic field is determined by a law which can be expressed by

$$H = \frac{2J}{cr},$$

where c is the velocity of light and r is the distance from the wire. This is an integral law between quantities which have been defined independently. The integral law correlating the magnetic field and the current can be expressed in the form

$$\oint \mathbf{H} \cdot d\mathbf{r} = \frac{4\pi J}{c}$$

for a closed path encircling the conductor. The integral is the work done by the field on a unit positive pole which is carried around the conductor in a closed path. If the path does not encircle the conductor the work done is zero.

We can now derive a differential equation on the assumption that the law continues to hold as the circuit becomes vanishingly small. By a transformation one obtains

$$\text{curl } \mathbf{H} = \frac{4\pi \mathbf{i}}{c}.$$

This is an hypothesis which is justified by the verification of deductions from it based upon the assumption of homogeneity.

We initially defined **H** in terms of permanent magnets. We now transform our concept. **H** is to be defined as a vector such that

$$\text{curl } \mathbf{H} = \frac{4\pi \mathbf{i}}{c}.$$

Thus an hypothesis based upon experiment is transformed into a definition. The foregoing equation is called the **first fundamental equation of electromagnetism.** However, the equation applies only to a stationary current in a closed conducting circuit. Such a current is characterized by the equation

$$\text{div } \mathbf{i} = 0.$$

This is the equation of continuity, which states that the rate at which charges are entering an element of volume is equal to the rate at which they are leaving it. In the traditional terminology one would say that div $\mathbf{i} = 0$ expresses the conservation of electricity.

The preceding discussion has been restricted to stationary currents in a closed conductor. Another type of current is that which flows during the discharge of a condenser. For this current the equation of continuity does not hold. For example, electric charge is flowing from the positively charged plate and nothing is flowing in to take its place. Before Maxwell a fundamental distinction was drawn between closed and open currents. Maxwell extended the concept of current by introducing the **displacement** current. We have previously defined the electric displacement by

$$\text{div } \mathbf{D} = 4\pi\rho.$$

The density of the displacement current is defined by

$$\frac{1}{4\pi}\frac{\partial \mathbf{D}}{\partial t}.$$

Now, if the charged plates of a condenser are connected by a wire, a conduction current flows in the wire from the positively charged plate to the negatively charged plate. The current is

$$J = \frac{de}{dt}$$

where e is the changing charge on the plates. Now, when the charge on the plate of a condenser is changing, the displacement also is changing. Hence during the flow of the current the electric displacement within the medium between the plates is changing. The variable displacement is correlated with the variable density of charge on the plates. Maxwell's hypothesis was that the conduction current in the conductor is continued by the displacement current in the dielectric between the plates. The displacement current is a current in the sense that it excites a magnetic field. If a medium is partly a conductor and partly an insulator, the density of the total current is

$$\mathbf{k} = \mathbf{i} + \frac{1}{4\pi}\frac{\partial \mathbf{D}}{\partial t}.$$

It is true for the total current that div $\mathbf{k} = 0$. It is as if there were an indestructible fluid whose rate of flow through a surface per unit area is measured by the total current. The equation of continuity expresses the fact that the charge which enters an element of volume replaces the total electric displacement outwards through the boundary.

The first fundamental equation of the electromagnetic field now becomes

$$\text{curl } \mathbf{H} = \frac{4\pi\mathbf{k}}{c} = \frac{4\pi\mathbf{i}}{c} + \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t}$$

for bodies at rest. For Maxwell this equation was an hypothesis which was later verified by the discovery of electromagnetic waves.

5. THE SECOND EQUATION OF THE ELECTROMAGNETIC FIELD

Thus far we have seen that an electric current gives rise to a magnetic field; we have now to describe the phenomenon of electromagnetic **induction** whereby a changing magnetic field induces a current. It was Faraday who first discovered that if a conducting circuit is in a changing magnetic field a current is induced in the conductor. The variable magnetic field may be excited by a variable current in a neighboring

conductor, by relative motion between the conductor and a magnet, etc. A good experimental arrangement is one in which a secondary coil in which the current is induced is wound around a primary coil which is in an electric circuit which may be made or broken. In order to investigate the induced current, one places in the secondary circuit a ballistic galvanometer, whose throw depends upon the total electric charge which flows in the circuit. We have

$$e = \int J dt.$$

If a bar of soft iron is inserted in the primary coil the effect is greatly increased.

We define a new quantity $\mathbf{B}$, the magnetic induction, in terms of quantities which measure observable effects. Assuming that initially there was no magnetic field we set

$$-\int \mathbf{B} \cdot d\mathbf{S} = cRe$$

where e is the total electric charge that flows through the circuit and R is the resistance of the circuit. The quantity on the left-hand side is the sum of the products of the normal components of $\mathbf{B}$ and the elements of area of a surface bounded by the circuit, whereas the right-hand side of the equation is independent of the surface. Hence the total induction is the same through all surfaces with the same boundary. We may express this by

$$\operatorname{div} \mathbf{B} = 0.$$

Thus $\mathbf{B}$ is defined by these two equations. Experiment shows that except for ferromagnetic substances

$$\mathbf{B} = \mu \mathbf{H},$$

where μ is a constant which is called the permeability.

The equation by which we introduced magnetic induction is an integral law. From it one can obtain a differential equation. We put

$$\Phi = \int \mathbf{B} \cdot d\mathbf{S}$$

Then $$-\frac{1}{c}\frac{d\Phi}{dt} = R\frac{de}{dt} = RJ.$$

$-\frac{1}{c}\frac{d\Phi}{dt}$ is called the induced electromotive force. Using Ohm's law in the differential form, and the formula for the resistance in terms of the conductivity, one obtains

$$-\frac{1}{c}\frac{d\Phi}{dt} = \oint \mathbf{E}\cdot d\mathbf{r}.$$

For a conductor at rest one obtains

$$-\frac{1}{c}\int \frac{\partial \mathbf{B}}{\partial t}\cdot d\mathbf{S} = \oint \mathbf{E}\cdot d\mathbf{r}.$$

This integral law is initially based on experiments with linear conductors. It is verified for any circuit within a bodily conductor. Upon the assumption of homogeneity one obtains the differential equation.

$$-\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = \text{curl } \mathbf{E}.$$

This is the second fundamental equation of electromagnetism. A pictorial representation is that a variable magnetic induction is encircled by an electric field. Maxwell assumed that this equation holds not only in conductors, but also in insulators. His hypothesis was verified by the discovery of electromagnetic waves.

6. CONCLUSIONS

The fundamental equations of the electromagnetic field are

$$\text{I} \qquad \text{curl } \mathbf{H} = \frac{4\pi \mathbf{i}}{c} + \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t}$$

$$\text{II} \qquad -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = \text{curl } \mathbf{E}.$$

These equations constitute a general schema.

Now it is an experimental fact that in an isotropic body with dielectric constant ϵ and conductivity σ, for the electrostatic field

$$\mathbf{D} = \epsilon\mathbf{E},$$

and for a stationary electric current Ohm's law holds,

$$\mathbf{i} = \sigma\mathbf{E}.$$

Furthermore,

$$\mathbf{B} = \mu\mathbf{H},$$

except for ferromagnetic bodies.

Maxwell's theory of electricity and magnetism assumes that these relations, which are verified for statical and stationary fields, hold also for rapidly alternating fields. Introducing these experimental assumptions into the equations, one obtains

$$\text{I} \qquad \frac{\epsilon}{c}\frac{\partial \mathbf{E}}{\partial t} = \text{curl } \mathbf{H} - \frac{4\pi\sigma}{c}\mathbf{E}$$

$$\text{II} \qquad -\frac{\mu}{c}\frac{\partial \mathbf{H}}{\partial t} = \text{curl } \mathbf{E}.$$

If the momentary distribution of the magnetic and electric field is given, the two differential equations determine the rates of change of the fields throughout space. One can therefore calculate the sequence of states which originates in a given initial state.

I remark that arguments may be adduced for the use of the vector **B**, instead of **H**, in the equations.

REFERENCES

ABRAHAM, M., and FÖPPL, A. Theorie der Elektrizität I. Leipzig and Berlin, 1918. The preceding chapter is based primarily upon this work.

HERTZ, P. Magnetostatik. Handbuch der Physik, edited by H. Geiger and K. Scheel, Vol. 14. Berlin, 1927.

MAGIE, W. F. Principles of Physics. New York, 1911.

MASON, M., and WEAVER, W. The Electromagnetic Field. Chicago, 1929.

CHAPTER VII

THERMODYNAMICS

1. TEMPERATURE

In the chapter on dynamics we introduced the physical quantity mass, which serves to represent many things in the world given in experience. Things which have mass are called material bodies. In the chapter on electrodynamics we saw that material bodies are also described as electrified, magnetized, conducting, etc. In thermodynamics we introduce another character of material bodies—that which we express when we call them **hot** or **cold.** Bodies are thus characterized in terms of hotness and coldness. Frequently occurring phenomena are connected with variations in hotness and coldness, for example, expansion, change of state, transfer of heat, performance of work, etc. We must fashion concepts of physical quantities in order to describe the phenomena correlated with the thermal properties of bodies.

I first consider the concept of **temperature** and shall begin with a qualitative analysis. The original qualitative basis of this concept is the sensation of hotness or coldness: we touch a body and declare it to be hot, warm, cold, etc. I first explain the concept of **equality** of two bodies with respect to hotness. At first glance it might appear that the concept of equality with respect to intensity of heat could be based upon the experience that two bodies appear equally hot to touch. This basis is unsatisfactory because the sensation depends upon the kind of material, hence the criterion would lead to contradictions. Our concept of temperature will be defined in such a way that, of two bodies at the same temperature, one may appear colder to the touch than the other, as for example, steel and wood.

We shall base the definition of equality with respect to hotness upon the empirical fact that it is possible to bring two

bodies into contact and experience no change in hotness in either body. The relation between the bodies is symmetrical, if body A is as hot as B, then B is as hot as A. The relation is also transitive, for, if A is as hot as B, and B is as hot as C, then A is as hot as C. In virtue of this symmetrical and transitive relation between the two bodies we ascribe to them a common character, the same temperature. Thus the concept of temperature is defined on the basis of generalization from experiences of equally hot bodies. By definition, two bodies have the same temperature if, when they are brought into contact, no change in the hotness of either body is experienced.

One body is hotter than another, if on contact the first cools and the second becomes warmer. This process will continue until both bodies are at the same temperature. We can arrange bodies according to a scale of hotness: very hot, hot, warm, lukewarm, cool, cold, very cold. Temperature is a non-additive property: if one joins two bodies equal in temperature, one obtains a body of the same temperature.

The discussion thus far has yielded a concept of temperature, but one which is not satisfactory because of the impossibility of making discriminating and reproducible estimates of temperature by our temperature senses. A physical concept of temperature is based upon the empirical fact that as a body changes in hotness certain correlated measurable properties change; for example, the volume of a body generally increases upon heating. Hence one can define **changes** of temperature in terms of the changes of some more accurately measurable property. An instrument which embodies a definition of temperature in terms of some measurable property is called a **thermometer**. The first thermometer was Galileo's air thermometer, which consisted of an inverted bulb with a stem whose open end was in water. The air in the bulb was at less than atmospheric pressure, so that the water rose in the stem above the level in the outer vessel. Changes in hotness of the air in the bulb were indicated by a change in level of the water in the stem. Such a thermometer defines the physical quantity temperature in terms of changes in volume of air.

There are various possible definitions of temperature in terms of different thermometers. In order to construct a

thermometer one must choose: (*a*) a substance, (*b*) a property which varies with hotness and coldness, (*c*) the zero of a scale, (*d*) the standard unit of a scale. The international standard thermometer is the constant volume hydrogen thermometer. Temperature is defined in terms of the pressure of hydrogen gas at constant volume. Defining the temperature of a mixture of water and ice under a pressure of one atmosphere as 0° Centigrade, and the temperature of steam over boiling water under a pressure of one atmosphere as 100° Centigrade, fixes the zero and the standard unit of the scale. One calibrates the thermometer by noting the pressure of the gas for a given volume at these two fixed points. In determining the fixed points of the thermometer one must wait until temperature equilibrium is established between the gas and the medium in terms of which the fixed points are defined. The meaning of this procedure is based upon our initial definition of temperature and the empirical law that if a hot body and a cold body are placed in contact they gradually come to the same temperature. The definition of the physical quantity temperature is thus based upon a law the meaning of which depends upon a qualitative definition of temperature. Furthermore, the definition of a fixed point depends upon the fact that the pressure of the gas is always the same at the ice-point and steam-point. The definition of the ice-point as 0° C and the steam-point as 100° C is possible in virtue of the fact that ice has a definite melting-point and water a definite boiling-point. Temperature as a physical quantity may now be defined by the equation

$$t = \frac{P_t - P_0}{\frac{P_{100} - P_0}{100}}$$

where P_0 is the pressure of the hydrogen gas at 0° C,
P_{100} is the pressure at 100° C,
P_t is the pressure at some temperature t which is to be determined.

In order to measure temperature we must establish temperature equilibrium between the hydrogen gas and the body whose temperature is to be determined. From the initial dis-

cussion of temperature we know that we must wait until equality in temperature is established. The pressure of the hydrogen gas is then measured. From the pressure coefficient of hydrogen one can determine the temperature at which the pressure of hydrogen gas would be zero, if the properties of the gas did not depend upon the conditions. This temperature is approximately $-273°$ C and is called the **absolute zero** of temperature. If a temperature is assigned relative to the absolute zero it is symbolized by T.

Thermometers based upon other properties and other substances are also used, but they must be calibrated in terms of the standard thermometer. One can discover laws correlating the temperature and other quantities, for example, the law of radiation according to which the rate at which energy is radiated from a black body depends upon the fourth power of the absolute temperature. Very high temperatures are defined by the extrapolation of this law beyond the limits within which it has been verified. We shall later employ the concept of temperature to set forth the second law of thermodynamics; thereupon the second law will be transformed into a definition of temperature.

2. QUANTITY OF HEAT

The second fundamental concept in thermodynamics is that of quantity of heat. In the first part of the discussion of this concept I shall assume that substances which are heated or cooled remain in a given state, that is, remain in the solid, liquid, or vapor state. Let us now suppose that we have two bodies of different materials and at different temperatures, for example, a bar of iron and a vessel of water. If we bring the bodies into contact, for example, plunge the iron into the vessel, the hotter body is cooled and the colder body is heated until the two bodies are at the same final temperature. Let Δt_1 be the decrease in temperature of the hotter body, and Δt_2 the increase in temperature of the colder body. If the masses of the bodies are m_1 and m_2, respectively, then we find by experiment that

$$\frac{m_1 \Delta t_1}{m_2 \Delta t_2} = k.$$

Assuming heat insulation, which will be explained later, this constant is independent of time, place, and the masses of the materials entering into the thermal interaction. It does depend upon the temperature, but as a first approximation we may neglect this fact. We may write the equation

$$\frac{m_1\Delta t_1}{m_2\Delta t_2} = \frac{s_2}{s_1}.$$

A similar experiment with bodies of substances (1) and (3) yields

$$\frac{m_1\Delta t_1}{m_3\Delta t_3} = \frac{s_3}{s_1}.$$

We may arbitrarily choose a value for s_1, and thus our experiments yield a set of constants s_1, s_2, s_3, . . . which characterize the substances (1), (2), (3), etc. These constants are defined to be the specific heats of the different substances. One arbitrarily posits the specific heat of water as 1. For the best possible precision one must specify the temperature range for which the specific heat of water is 1.

Returning now to the equation

$$\frac{m_1\Delta t_1}{m_2\Delta t_2} = \frac{s_2}{s_1}$$

we obtain

$$m_1s_1\Delta t_1 = m_2s_2\Delta t_2.$$

Thus, whenever two bodies at different temperatures are brought into contact, the final common temperature is such that the decrease in $ms\Delta t$ for one is equal to the increase in $ms\Delta t$ for the other. It is therefore useful to introduce the concept of **quantity** of heat which is defined by

$$q = ms\Delta t.$$

Thermal interaction between two bodies which results in a common final temperature may be described by the proposition that the heat lost by one body is equal to the heat gained by the other body. In general, if a body of mass m and specific heat s experiences a change of temperature Δt, the heat gained or lost is given by $ms\Delta t$.

We thus define quantity of heat so that in experiments on bodies in thermal contact the heat lost by the one body is equal to the heat gained by the other body.

The concepts of specific heat and quantity of heat were fashioned during the eighteenth century, when the theory was held that heat is an imponderable, indestructible substance, called caloric, the physical quantity $ms\Delta t$ being a measure of the amount of caloric which entered or left a body. The *a priori* concept of heat substance undoubtedly guided the selection of $ms\Delta t$ as a measure of the quantity of this substance. However, in the preceding account we discovered $ms\Delta t$ to be a quantity which satisfied the equation

$$\Sigma ms\Delta t = 0$$

in a thermal exchange between two bodies. Although it is not necessary to interpret this empirical result as grounded in the indestructibility of a substance, yet our terminology reveals the influence of the concepts of the caloric theory, and so we continue to speak of the heat gained or lost by a body as if there were some substance which entered or left the body. The empirical fact, however, is that it is possible to assign numbers s_1, s_2, etc., such that $\Sigma ms\Delta t = 0$ under certain specified conditions. We shall see later that in contemporary theory $ms\Delta t$ is interpreted as change in energy.

Our definition of quantity of heat was a first approximation based upon crude experiments. We may use this concept to explain lack of agreement in repeated trials as a consequence of loss of heat by radiation, conduction, etc. We recognize then that if two bodies interact thermally we must insulate the system carefully or else correct our calculations for losses. These precautions render it possible to define more precisely the conditions of the experiments on the basis of which the fundamental quantities are defined for the description of thermal phenomena.

In the discussion of exchanges in heat between two bodies we have thus far presupposed that each body is unchanged in state by the process. Let us now perform the experiment of placing 1 gram of ice at 0° C in 1 gram of water at 80° C. The result will be 2 grams of water at 0° C. In accordance with

our definition of quantity of heat the 1 gram of water has lost heat in being cooled from 80° to 0° C. We should expect that the 1 gram of ice would gain an equal amount of heat. However, this case is not included in our definition, because the temperature of the ice remains unchanged. Thus our principle that heat lost is equal to heat gained appears to be violated. The problem is solved by generalizing the concept of quantity of heat so that changes of heat are not restricted to changes in temperature. Accordingly, we interpret the process of fusion as one accompanied by an absorption of heat: the heat required to change unit mass of a solid at the melting-point into a liquid at the same temperature is called the **heat of fusion.** In our experiment, in which 1 gram of ice at 0° C and 1 gram of water at 80° C resulted in 2 grams of water at 0° C, the heat lost by the water in cooling from 80° C to 0° C is equal to the heat absorbed by the ice in melting without change of temperature. Conversely, upon solidification heat is given out. Thus the principle, heat lost is equal to heat gained, is a defining character of the physical quantity heat. One can readily see that the concept of heat of fusion would be required by the demand to interpret thermal phenomena in terms of an *a priori* concept of indestructible substance. In the light of the mechanical theory of heat, heat of fusion measures the change in potential energy per unit mass of a substance when it changes from the solid to the liquid state or *vice versa.*

Again, in changing from the liquid state to the vapor state, heat is absorbed without change of temperature. The concept of **heat of vaporization** is then introduced on the basis of the defining principle that heat lost is equal to heat gained.

An alternative method of defining heats of fusion and vaporization is the following. Suppose that a given mass of water is in contact with a flame. The temperature of the water will gradually rise and we describe the process as one in which the flame supplies heat to the water. We now replace the water by a mass of ice and there occurs melting of the ice without change of temperature. We assume that the flame now supplies heat to the ice and that the absorption of heat manifests itself in a change of state. With this method of defining

heat of fusion we can verify experimentally the principle that heat lost is equal to heat gained by mixing water and ice and measuring the masses and temperatures.

The unit of quantity of heat in physics is the heat required to raise 1 gram of water 1 degree Centigrade. In precise work one must specify the degree.

3. THE MECHANICAL THEORY OF HEAT

In describing thermal interaction between bodies which do not change in state we found that a physical quantity defined by $ms\Delta t$ was of general application. This physical quantity was called quantity of heat and has the property that the heat lost by one body is equal to the heat gained by the other body. In order to describe thermal phenomena involving change of state, we extend the concept of heat to include quantities such as heat of fusion which is applicable in the description of change of state.

The caloric theory of heat was based upon the assumption that the ground of these quantitative relations was a weightless, indestructible substance. The permanent character of heat substance was expressed by the principle that the heat lost by one body is equal to the heat gained by the other. The caloric theory was rendered unplausible by a series of experiments which have been interpreted in terms of a mechanical theory of heat.

Some pioneer experiments for the mechanical theory of heat were performed by Count Rumford in 1798. While engaged in the boring of cannon he noticed that there was a continued and apparently unlimited production of heat. This result was contrary to the principle of the conservation of heat as expressed by the concept of an indestructible substance. Rumford expressed the inadequacy of the caloric theory in the following words: " It is hardly necessary to say that anything which any insulated body, or system of bodies, can continue to furnish without limitation cannot possibly be a material substance; and it appears to me to be extremely difficult, if not quite impossible, to form any distinct idea of anything capable of being excited and communicated in the

manner the heat was excited and communicated in these experiments, except it be motion."

About the same time Sir Humphry Davy rubbed two blocks of ice together and thereby melted the ice. This experiment and other experiments in which heat was produced by friction were interpreted in terms of the mechanical theory of heat.

However, the mechanical theory of heat was not generally accepted until the middle of the nineteenth century. The decision in favor of the mechanical theory was largely determined by Joule's measurement of the mechanical equivalent of heat. He discovered experimentally that a given amount of mechanical work always produces the same amount of heat. A typical experiment was one in which water was heated by stirring it up by paddle wheels. Joule measured the work done and the heat produced, and from a set of experiments obtained a fairly consistent set of values for the mechanical equivalent of heat, that is, the mechanical energy required to produce a unit of heat. It is essential to distinguish the experimental result that heat and mechanical work are proportional from its interpretation in terms of the mechanical theory. Einstein has stated the principle that when two physical quantities are always proportional they are essentially identical. Thus the theory that heat consists of mechanical energy affords a simple interpretation of the experimental discovery of the equivalence between work and heat. Then the thermal quantities such as $ms\Delta t$, heat of fusion, etc., measure, not substance, but **mechanical energy**.

Now the heating of a body does not appear to be accompanied by an increase in motion. Hence the mechanical theory of heat requires the assumption that a visible and tangible body is constituted of smaller particles whose energy constitutes heat. Thus the mechanical theory introduces atomistic, microscopic concepts into physics.

We have now been introduced to the notion of the transformation of mechanical work into heat. Conversely, heat may be transformed into mechanical work by a heat engine. Furthermore, the application of dynamical concepts in electrodynamics is the basis of the concepts of electrical and magnetic energy. Electrical and magnetic energy can be trans-

formed into heat, and *vice versa*. The general theory of transformations in which the thermal properties of bodies are considered constitutes thermodynamics.

4. THERMODYNAMICS

Thermodynamics is the study of transformations of systems in which thermal properties participate. A system consists of material bodies and other entities which may be the object of physical investigation. Thus electromagnetic radiation within a space unoccupied by matter may constitute a system. The state of the system is defined by the values of certain physical quantities. In thermodynamics the physical quantities employed are restricted to macroscopic quantities, which are directly measurable or definable in terms of directly measurable quantities. For example, in the thermodynamic study of a gas the quantities which determine the state of the gas are the pressure, volume, and temperature. The temperature can be directly measured with a thermometer, the pressure can be measured by a manometer, and the volume is readily calculable from direct measurements. A relation between the variable physical quantities which characterize a system is called an equation of state. Thus for a gas the equation of state is approximately $PV = RT$. A system is said to be in equilibrium if the variables which define the state of the system are constant in value. A change in the state of a system is called a transformation. The transformation takes place by means of a process during which there is a change in the values of the variables which define the state. We distinguish between reversible and irreversible processes. A reversible process is one in which the system is always infinitely near to equilibrium conditions. If a certain mass of gas at a given pressure and temperature is enclosed in a cylinder by a piston, subject to an external pressure equal to that of the gas, the system consisting of the gas and piston is in equilibrium. In order to expand the gas reversibly one must keep this external pressure on the piston infinitesimally less than the equilibrium pressure. The gas will then expand infinitely slowly and the state of the system will always be infinitely near to equilibrium. The process is reversible in the sense that an infinitesimal

change in pressure will cause the process to go in the opposite direction through states infinitely near to equilibrium. Thus in a reversible process the direct and reverse processes pass through the same series of states of the system. In this respect reversible processes in thermodynamics are similar to reversible processes in dynamics. There is the difference, however, that reversible processes in thermodynamics do not involve the time, or in other words involve the time to the zero order, whereas the reversible processes in dynamics are governed by differential equations of the second order with respect to the time. An irreversible process is one in which the state is always at a finite distance from equilibrium. Thus, in the illustration of the gas enclosed by a piston, if the external pressure were kept equal to one-half of the equilibrium pressure, the expansion would proceed at a finite rate and the system would pass through states which differ by a finite amount from equilibrium conditions. A finite change in conditions would be necessary to change the direction of the process. The direct and reverse processes would not pass through the same series of states.

In thermodynamics use is made of special processes. An isothermal process is one carried out at constant temperature. An adiabatic process is one in which no heat is added to, or subtracted from, the system. A cycle is a process in which the final state is the same as the initial state. An important cycle is the Carnot cycle, which consists of the following series of processes: reversible isothermal expansion with absorption of heat and the performance of external work, reversible adiabatic expansion with performance of external work and drop in temperature, reversible isothermal compression in which work is done on the system and heat is rejected by the system, reversible adiabatic compression in which work is done on the system and the temperature is raised. The Carnot cycle is the basis of an ideal heat engine and is useful in deducing consequences from the principles.

Classical thermodynamics is based upon two principles, the First and Second Laws of thermodynamics. In recent years a third law has been set forth, but I shall confine my attention to the first and second laws.

5. THE FIRST LAW OF THERMODYNAMICS

I begin the exposition of the First Law with a statement of presuppositions. We have a definition of the physical quantity heat. In the standard case the heat received or given out is $ms\Delta t$. Then we have a definition of energy or work; if a force acts through a distance work is done by the force. I shall employ the term mechanical work for the work done by gravitational forces and elastic forces such as occur in the contact of two bodies.

The experimental basis of the First Law is the proportionality between heat on the one hand, and mechanical work or other forms of work or energy, on the other. For example, if mechanical energy is used up in doing work against friction, a definite amount of heat is produced. Accordingly, Q, quantity of heat, can be expressed in energy units, and in the following I shall assume that this has been done.

The functional relation between heat and work suggests the mechanical theory of heat, according to which Q is a measure of molecular mechanical energy. This in turn offers a basis for a general principle of the conservation of energy. Energy which is lost is transformed into molecular energy which is manifested as heat. The principle of the conservation of energy gave rise to the First Law of thermodynamics. Indeed, the First Law is often said to express the conservation of energy. However, in the present discussion I shall formulate the First Law independently of the mechanical theory of heat. In the following formulation of the First Law I use only the **proportionality of heat and work.**

I begin with the experimental fact that if energy in any form is lost, a corresponding amount of heat is gained. Thus in Joule's experiment a body loses mechanical potential energy and a mass of water gains heat. If a current flows in a wire electrical energy is lost and there is an evolution of heat. The experimental fact is the functional relation between heat and energy.

Suppose, to consider a concrete case, we have a gas enclosed in a cylinder by a piston. The gas is in a given state (1) which is determined by the pressure, volume and temperature. If

we add heat Q to the gas and do work W upon it the gas is changed to state (2). Now it is an experimental result that

$$Q + W$$

is a function of the initial and final states alone. Since Q and W have been defined we are able to verify this result. A general verification consists in the fact that if $Q + W$ were not independent of the path, it would be possible to construct a perpetual motion nachine. The past failures to construct such a machine are evidence of the truth of our proposition. Thus the experimental result is that there is a function of the state, E, such that

$$E_2 - E_1 = Q + W.$$

E is called the internal energy of the system. Thus the increase in the internal energy is equal to the dynamical equivalent of the heat added plus the work done on the system. In differential form our equation becomes

$$dE = q + w,$$

where q and w are infinitesimal. Our experimental result is that $q + w$ is a perfect differential.

We may express the First Law by the statement that $q + w$ is a perfect differential. An equivalent statement is that there is a function of the state, the internal energy, such that the increase in internal energy is equal to the dynamical equivalent of the heat added plus the work done on the system. Conversely, the decrease in internal energy is equal to the dynamical equivalent of the heat given out plus the work done by the system. The total internal energy, which is defined except for an arbitrary constant, is the sum of the dynamical equivalent of the heating effect plus the total external work which a system can do.

If in the equation $dE = q + w,$

we interpret q as molecular energy, then we arrive at a general principle of conservation of energy. In anticipation of this result the function of the state was called internal energy. Then the external environment loses or gains energy as the system gains or loses energy.

It is desirable to compare the general principle of conservation of energy with the principle of energy in dynamics. Suppose that we have given a mechanical system of particles subject to the action of forces which are functions of the positions of the particles. We can then derive an equation, called the equation of energy, which states that the kinetic energy plus the potential energy is a constant. Now in real systems there is a gradual decrease in the mechanical energy of the system on account of frictional forces. For example, a vibrating pendulum will gradually be brought to rest by the resistance of the air. Thus a general principle of the conservation of energy seemed impossible in physics. However, the mechanical theory of heat rendered possible the formulation of a general principle of energy. Experiment shows that there is a function of the state E such that

$$dE = q + w.$$

If we interpret q as molecular energy, E becomes the internal energy of a system and we obtain the principle that in a transformation the total change in energy is zero.

$$dE - (q + w) = 0.$$

The interpretation of the First Law as the principle of the conservation of energy involves an extension of the concept of energy. To macroscopic energy we add molecular energy.

Thus the First Law has been explained as the definition of internal energy upon the basis of experimental results. We may now transform the law. Suppose that in a problem we know the change in internal energy and W. Then the First Law can be interpreted as a definition of heat. It is from this point of view that heat is, to borrow a phrase from Professor Bridgman, a waste-basket concept.

6. THE SECOND LAW OF THERMODYNAMICS

The First Law expresses a functional relation between heat and energy, but it gives no clue to the direction of isolated processes. If in a machine mechanical energy is transformed into heat by friction, the First Law states that the

decrease in mechanical energy is proportional to the increase in heat, but the First Law does not indicate whether the process goes in the direction of transforming mechanical energy into heat or *vice versa*. The **direction** in which processes go when left to themselves is prescribed by the Second Law.

Suppose that a hot body and cold body are joined by an iron rod. Then we know from experience that heat will be conducted through the rod from the hot to the cold body until a state of equilibrium is reached in which both bodies are at the same temperature. The process is irreversible; at the beginning the conditions vary by a finite amount from equilibrium and the process goes on at a finite rate; as equilibrium is approached the process goes more and more slowly. Now an essential characteristic of irreversible processes is that the system can not be returned to the initial state without some change in the surroundings. It is possible to transfer heat from one body to another so that their temperatures are the same as at the beginning of the process of conduction which has just been described, but this reestablishment of the initial state of the two bodies requires the employment of an engine the operation of which involves a change in the surroundings. The Second Law of thermodynamics can be expressed by the statement that natural processes are irreversible.

Specific formulations of the principle were first given by Clausius (1850) and Kelvin (1851). Clausius stated: It is impossible for a self-acting machine to convey heat from one body to another at a higher temperature. Kelvin's statement was: It is impossible by means of any continuous inanimate agency to derive mechanical work from any portion of matter by cooling it below the lowest temperatures of its surroundings. Thus one can not propel a ship across the ocean by transforming the heat of the water into mechanical work.

From either of the preceding statements of the Second Law one can deduce a theorem which was first stated by Carnot. One must first remark that heat is transformed into work by means of an engine, which may be assumed to be a Carnot engine, which takes in heat Q_1 at a given temperature T_1 and gives out heat Q_2 at a lower temperature T_2. By the First Law the heat transformed into work is $Q_1 - Q_2$. The work

done divided by the heat taken in is called the efficiency of the engine,

$$e = \frac{Q_1 - Q_2}{Q_1}.$$

Carnot's theorem states that all reversible engines working between the same temperatures have the same efficiency, and no irreversible engine working between the same temperatures can have a greater efficiency. Thus the efficiency of a reversible engine is independent of the working substance and the pressures; it is a function of the temperatures alone. That is,

$$\frac{Q_1 - Q_2}{Q_1}$$

depends only on the temperatures T_1 and T_2.

Q_2/Q_1 therefore depends only on the temperatures. Kelvin then created the thermodynamic scale of temperature by the definition,

$$\frac{Q_2}{Q_1} = \frac{\theta_2}{\theta_1},$$

where Q_1 is the heat taken in by a Carnot engine at temperature θ_1, and Q_2 is the heat given out at temperature θ_2. We have therefore transformed the Second Law of thermodynamics for reversible processes into a definition of temperature. This definition of temperature is not dependent upon the properties of a special substance. Experiment shows that the temperature T as determined by the constant-volume hydrogen thermometer agrees very closely with the thermodynamic definition.

For the Carnot cycle

$$\frac{Q_2}{Q_1} = \frac{\theta_2}{\theta_1}$$

by definition.

Then

$$\frac{Q_2}{\theta_2} = \frac{Q_1}{\theta_1}.$$

Let us now introduce the physical quantity $\frac{q}{\theta}$, where q is an infinitesimal quantity of heat which is reversibly added to the

system at temperature θ. Let us further consider the summation of the values of this quantity in a Carnot cycle. The result is

$$\frac{Q_1}{\theta_1} - \frac{Q_2}{\theta_2},$$

which is zero, from the definition of temperature. This result is an instance of the fact that q/θ is a perfect differential. Thus there is a function of the state S such that

$$dS = \frac{q}{\theta}.$$

S is called the **entropy** of the system. In the Carnot cycle it follows that during the absorption of heat Q_1 at temperature θ_1 there is an increase in entropy Q_1/θ_1; during the adiabatic expansion there is no change in entropy because no heat is added or subtracted; during the rejection of heat Q_2 at the lower temperature θ_2 the loss in entropy is Q_2/θ_2; and during the final adiabatic compression there is no change in entropy. The total change in entropy of the working substance during the cycle is zero. Similarly, the source of heat at the higher temperature θ_1 loses entropy Q_1/θ_1 and the source of heat at the lower temperature θ_2 gains an equal amount Q_2/θ_2. In general a reversible process is characterized by the fact that the entropy of the whole system involved does not change. An irreversible process in an isolated system is characterized by an increase in entropy.

We may then express the Second Law by the statement that natural processes in an isolated system result in an increase in entropy. The state of equilibrium is that in which the entropy is a maximum.

The Second Law expresses the tendency of natural processes to equalize temperatures throughout the universe. Now in order to transform heat into mechanical or other kinds of energy it is necessary to have an engine working between a higher and a lower temperature. Thus the equalization of temperature renders heat unavailable for the performance of work. Hence the Second Law has been expressed as the principle of degradation of energy, or better as the degradation of systems.

The Second Law renders it possible to discriminate between the forward and backward direction of time. The increase in entropy which results from irreversible processes is correlated with the forward direction of time.

REFERENCES

LEWIS and RANDALL. Thermodynamics. New York, 1923.

MENDENHALL, C. E. Heat, *in* A Textbook of Physics, edited by A. W. Duff. Philadelphia, 1921.

PLANCK, M. Thermodynamik. Berlin und Leipzig, 1921.

PART II

UNITARY SYSTEMS OF PHYSICAL THEORY

CHAPTER VIII

THE MECHANICAL CONCEPTION OF NATURE

1. INTRODUCTION

THE mechanical conception of nature was an attempt at a unitary representation of the physical order.

A physical theory is the representation in concepts of selected characters and processes of the physical order. In dynamics we have a set of physical quantities which serve for the general description of the motion of bodies and the conditions of changes of motion. In electrodynamics we represent electrical and magnetic properties of bodies and correlated phenomena. In thermodynamics we have a very general description of the physical order in which thermal properties and phenomena are the principal objects of interest. In setting forth these systems of physical theory I have sought to retain the phenomenological point of view as much as possible. Now one seeks a unitary representation of the physical order; one endeavors to interpret the concepts of one theory in terms of the concepts of a fundamental theory. From the inductive point of view this is the explanation of a concept, for example, of heat; from the deductive point of view interpretation is definition.

The first, most sustained, and most detailed attempt at a unitary system of physics was the mechanical conception of nature. I distinguish between the mechanical conception of nature and materialism. Materialism is the extension of mechanical concepts to all characters of reality.

According to the mechanical conception of nature the fundamental physical phenomenon is that of a body in motion. The primary physical character of a body is represented by its mass. The acceleration of bodies is attributed to the action

of force. The concept of force expresses the fact that acceleration is dependent upon conditions outside the body. Now in dynamics the concept of force is general, so that instances of force are weight, the stress in an elastic body, electric attraction, magnetic attraction, etc. The mechanical conception of nature is characterized by the choice of certain kinds of force as fundamental in terms of which other forces are to be explained. Thus the mechanical theory is based upon the kinematical concepts, the concept of mass and the concept of mechanical force. Physical phenomena are described or explained in terms of bodies having mass and acted upon by mechanical forces.

What we understand by the mechanical conception of nature depends upon our concept of mechanical force. In the historical development of the theory two types of force have been especially considered; first, action at a distance, which is exemplified by the Newtonian theory of gravitation; second, action by contact between material bodies, which is exemplified in the macroscopic theory of elasticity. Speaking generally, one may say that the typical mechanical forces are those operative in machines like the pulley, the wheel and axle, etc.

Let us now review **physical phenomena** from the standpoint of the mechanical conception of nature. It is fitting to begin with those phenomena whose adequate description by a physical theory inspired the mechanical theory.

2. GRAVITATION

The early history of dynamics was molded to a considerable extent by the program of finding a theory of the motions of the heavenly bodies. The system of celestial mechanics was constructed by Newton upon the laws of motion and the law of gravitation. Newton's law of gravitation is that if m and m' are the masses of two bodies and r is the distance between them, then the force upon either mass is given by

$$F = Gm'm/r^2.$$

The motion of m is determined by

$$Gm'm/r^2 = ma,$$

In general, the acceleration of a given body is a function of the masses of surrounding bodies and their distances from the body. The gravitational field surrounding the earth may be taken as constant throughout a small region; hence a freely falling body has a constant acceleration which is the same for all materials.

However, the preceding description of the motion of a freely falling body is only a first approximation; one observes that a feather and a piece of lead fall with different accelerations. In order to describe the different behavior one introduces the concept of resistance of the air. The theory of gravitation thus provides a schema for the description of motion under gravity, and deviations from the schema are attributed to disturbing influences.

In classical celestical mechanics the force of gravitation is conceived of as an action at a distance. However, there is evidence that Newton did not consider this the final account of the phenomenon. In a letter to Bentley, Newton says:

"That gravity should be innate, inherent and essential to matter, so that one body may act upon another at a distance through a vacuum, without the mediation of anything else . . . is to me so great an absurdity that I believe no man, who has in philosophical matters a competent faculty of thinking, can ever fall into."

A theory by LeSage explains the force of gravitation in terms of the impacts of corpuscles; this is an example of action by contact.

The success of celestical mechanics inspired the attempt to interpret all forces as actions at a distance. The history of the theory of electricity and magnetism from the middle of the eighteenth to the middle of the nineteenth centuries was influenced by the attempt to find laws of force similar to the inverse square law in the theory of gravitation. Thus the quantitative theory of electricity and magnetism was initiated by Coulomb's law for the attraction between electric charges or magnetic poles, Laplace's law for the dependence of the element of magnetic field upon the current in an element of the electric circuit, etc. In the early history of the theory of elasticity it was attempted to explain stresses in terms of forces

at a distance between the molecules constituting an elastic body. Thus gravitational action at a distance gave rise to the consideration of central forces. In a very general sense a mechanical conception of nature can be based upon a theory of central forces.

Such a conception of nature has been based upon the notion of central forces which are functions of position. However, the development of physical theory has been guided by the attempt to reduce action at a distance to action by contact. Thus the mechanical theory is best exemplified by the classical theory of elasticity.

3. ELASTICITY

Solid bodies are characterized by the fact that if a body is deformed within certain limits by the application of some force, upon the removal of the force the body resumes its original size and shape; this property of solids is called elasticity. The elastic properties of bodies can be described with the aid of the concept of force. As an example, consider a wire one end of which is attached to a fixed support and the other end of which is supporting a load; under the action of the load the wire is stretched. If L is the original length of the wire and ΔL is the increment in length, the measure of stretch is defined by $\Delta L/L$. In addition to the geometrical concept of stretch, or extension, we need a dynamical concept. Let us imagine a dividing plane at right angles to the length of the wire. The part of the wire below the plane is a body in equilibrium. Now the load exerts a downward force upon it in virtue of the weight of the load; therefore, in accordance with the principle of equilibrium we must conceive of a force exerted by the upper part of the wire upon the lower part across the dividing plane between them. This force per unit area is called the stress in the wire. Stress may be characterized as action through contact in contrast with action at a distance.

In the simple case of the wire the application of the principle of equilibrium leads to the result that the stress is the load per unit area. Now it was discovered by Hooke that the stretch of a wire varies directly with the load, within the elastic limit. Hence stress and stretch are proportional. Further-

more, it is an experimental fact that this ratio of stress to stretch is a constant for a given material. On the basis of these experiments we ascribe to a given material a physical quantity M which is defined by

$$M = \frac{\text{the stress}}{\text{the stretch}}.$$

M is called Young's modulus of elasticity.

In the theory of elasticity one defines different kinds of stress and strain. In general, bodies undergo strains upon the application of force. The strains are accompanied by stresses which act across surfaces separating contiguous parts of a body; the stresses are linear functions of the strains. The concepts of these physical quantities enable one to describe the equilibrium and motion of elastic bodies.

Thus in the theory of elasticity we have the concept of an action through contact; this is the typical mechanical force. There have been attempts at a more detailed mechanical theory of these contact stresses in elastic bodies. In the present chapter I shall assume these stresses as fundamental.

Stresses are also assumed in hydrodynamics, which is the study of the motion of fluids.

One may say that the most typical mechanical force is an action by contact. In the mechanical conception of nature the action of stress is fundamental and not to be further explained. The statement that stress is action by contact is expressed analytically by the proposition that the components of stress are functions of the coordinates and the time which satisfy certain partial differential equations.

The driving force of the mechanical conception of nature has been the attempt to employ the concepts of motion, mass and stress, which have been found applicable in the description of elastic and fluid bodies, for the explanation of all physical phenomena.

4. SOUND

The phenomenon of sound is readily explained on the mechanical theory.

I assume that we have a concept of sound which is abstracted from sensations of sound. Investigation of our surroundings

when we hear a sound reveals that the sensation of sound is correlated with the vibrations of material bodies; in particular, a simple harmonic vibration of a sounding body is the source of a sound called a pure tone. We observe that the vibration precedes the sensation of sound. We conceive of the vibration as the cause of a disturbance which travels with a finite velocity; the sensation of sound is correlated with the arrival of the disturbance at the position of the observer. The propagation of the disturbance requires the existence of some material medium, which is usually air. When a sound is emitted by a sounding body the vibrations of that body set up longitudinal vibrations in the medium which are transmitted to the immediate neighborhood and thus a wave motion is set up in the medium. The theory of the propagation of sound is the theory of those waves in a continuous medium which depend upon volume elasticity. The velocity of a wave is a function of the density, the bulk modulus and the ratio of the specific heats.

The theory of sound is a macroscopic theory and holds even if one derives the mechanical properties of matter from a microscopic electrodynamics.

5. LIGHT

In the section on geometrical optics I have already given a partial description of the phenomenon of light. The concept of the propagation of light is there represented as an abstraction from observations of the visual appearances of certain bodies. To the first approximation light travels in straight lines and is further characterized by the laws of reflection and refraction. Our concept of the propagation of light is made definite by the measurement of its velocity.

Further experiments indicate that light has a periodic character; if one illuminates two parallel slits and allows the light from the slits to fall upon a screen one observes a set of parallel light and dark bands. This experiment, which was performed by Thomas Young, exhibits the interference of light and indicates that the light from the two slits has a periodic character such that for certain distances from the slits

the two beams of light reinforce each other and for other distances they counteract each other.

Another important characteristic of light is capacity for polarization. If a beam of light is allowed to pass through a crystal of calcite, the beam is broken up into two rays which have different properties as tested by allowing them to pass through another crystal of calcite. The partial beams of light are said to be polarized.

Thus far our description of light has been in terms of concepts which are seen to arise by abstraction from experience. There has been a minimum of hypotheses.

Let us now turn to the interpretation of optical phenomena in terms of the mechanical conception of nature. It is clear that the phenomena do not involve the motion of observable material bodies. Hence the application of the mechanical conception of nature requires an hypothesis that visible qualities are correlated with not directly observable mechanical processes.

There have been two principal mechanical theories of light, the corpuscular theory and the theory of ether waves. According to the corpuscular theory, which was held by Newton, a visible body emits small bodies, called corpuscles, which have mass and are perfectly elastic. In accordance with the principle of inertia the corpuscles travel in straight lines unless they meet an obstacle; thus the corpuscular theory affords a simple explanation of the rectilinear propagation of light. The phenomena of interference and polarization are difficult to explain on the corpuscular theory.

The wave theory of light is that light consists of vibrations in a medium called the ether; the state of vibration is transmitted through the ether by wave motion. In the seventeenth century Huyghens expounded a theory according to which the vibrations were longitudinal, but the authority of Newton brought about the acceptance of the corpuscular theory. At the beginning of the nineteenth century, however, the experiments of Young and Fresnel revived the wave theory. Foucault's measurement of the velocity of light in a liquid was a crucial experiment which decided in favor of the wave theory.

In the wave theory of the nineteenth century the vibrations constituting light were conceived to be transverse vibra-

tions in an ether with the properties of an elastic solid. The wave theory furnished an adequate basis for the periodic properties of light as revealed by Young's experiment. In that experiment light waves are propagated from two parallel slits; at a point on the screen whose distance from one slit differs from the distance from the other slit by an even number of half wave lengths the two waves meet in phase and there is reinforcement and hence light. At points on the screen whose distances from the two slits differ by an odd number of half wave lengths, the two waves meet in opposite phase and there is interference and darkness.

The concept of transverse vibrations furnishes an adequate basis for the phenomena of polarization.

The intensity of illumination is correlated with the square of the amplitude of vibration, and the color is correlated with the wave length.

We have no theory as to how a mechanical vibration gives rise to a specific seen quality.

After the development of Maxwell's theory of electromagnetic radiation the mechanical theory of light was abandoned.

6. THE KINETIC THEORY OF MATTER

The most important thermal phenomena have been described in the chapter on thermodynamics. The processes whereby mechanical and other forms of energy are transformed into heat and heat into energy are governed by the laws of thermodynamics. On the assumption of the mechanical theory of heat the first law becomes the principle of conservation of energy. However, the principle of energy does not presuppose a specific theory of the form of energy which we call heat.

The thermodynamic quantities are macroscopic, that is, they are directly measurable or calculable in terms of directly measurable quantities. Now a mechanical theory of heat demands the employment of atomistic concepts, for the addition of heat to a body does not increase its macroscopic energy. Hence the detailed development of the mechanical theory of heat led to a kinetic molecular theory of matter; on the basis of such a microscopic theory the laws of thermodynamics are

derivable from the laws governing the motion of the molecular processes.

The hypothesis that observable bodies are composed of many small particles originated with the Greek atomists, of whom Democritus constructed the most significant theory. In modern times the theory of atomism was revived by Gassendi and developed by Daniel Bernoulli. But the rapid quantitative development of the atomic hypothesis occurred in the nineteenth century. In order to provide a quantitative basis for chemistry, Dalton and others introduced the concepts of atoms and molecules which are composed of atoms. The discovery of the mechanical equivalent of heat, which seems to demand the interpretation that heat is a manifestation of molecular motion, furnished the basis for the rapid development of the kinetic molecular theory of matter which took place after the middle of the century.

The kinetic molecular theory is a microscopic theory. Now a microscopic theory is characterized by the assumption of the existence of microscopic bodies, such as molecules and atoms, which are not observable in the usual sense. In a macroscopic theory one studies relations between observable quantities: thus in thermodynamics the quantities which occur are volume, pressure, temperature, quantity of heat, etc., all of which are directly measurable or are defined in terms of directly measurable quantities. According to the kinetic molecular theory of matter, the observable quantities of the macroscopic theory of heat depend upon the average values of many molecular quantities. For example, the pressure of a gas, which is measurable with a manometer, is the time average of the instantaneous force exerted by all the molecules striking an enclosing surface of unit area. Thus in order to build a microscopic theory it is necessary to assume the laws governing the microscopic processes. The program of microscopic theory is the interpretation of the macroscopic quantities in terms of microscopic quantities and the deduction of macroscopic laws from the laws governing the elementary processes. The evidence for a microscopic theory is thus indirect.

7. THE KINETIC THEORY OF GASES

The kinetic molecular theory of matter has been most successful in the theory of gases. In the elementary theory a homogeneous gas is supposed to be constituted of a very large number of exactly similar, smooth, elastic spheres, between which the forces are negligible except during collisions. It is assumed that the laws of dynamics, which have been derived from observable bodies, determine the motion of the unobservable molecules. During impact the momentum and energy of the system of colliding particles are conserved. From the microscopic point of view, the motion of the system is represented by the solutions of the equations of motion. The solutions contain arbitrary constants which are determined by the initial conditions, that is, the positions and velocities of all the molecules at any given time. Now it is evident that an actual microscopic description of the motion is impossible on account of the mathematical complexity of the problem and because the crudeness of our senses does not permit the determination of the positions and velocities of the individual molecules. Thus, although the motion of the molecules may be causally determined, we are not able to obtain knowledge of particular functional relations which express causal connection. Recourse is therefore had to **statistical methods**.

The methods of the kinetic theory of gases may be illustrated by an elementary derivation of the expression for the pressure of a gas. Suppose that we have a cubic centimeter of gas consisting of n similar molecules each of mass m. I assume that three of the faces of the cube are in three mutually perpendicular planes which constitute a Cartesian frame of reference. Let q be the velocity of a molecule with components u, v, w in the x, y, z directions, respectively. Then

$$q^2 = u^2 + v^2 + w^2.$$

I now introduce the fundamental statistical assumption. On account of the random motion of the very large number of molecules it seems reasonable to assume that the average values of u^2, v^2 and w^2 are equal. Hence we may put

$$\bar{u}^2 = \bar{v}^2 = \bar{w}^2 = (\tfrac{1}{3})\bar{q}^2.$$

Let us now calculate the average force which a molecule exerts when it impinges on a face perpendicular to the x axis. Only the x component of the velocity enters into this calculation, and hence we may assume that the molecule is traveling in the x direction with velocity u. Its momentum is mu. On striking the wall the molecule is reflected with a velocity $-u$ in the opposite direction, and hence a momentum $-mu$. The change in momentum is

$$-mu - mu = -2mu.$$

The molecule then travels to the opposite face, is reflected, and again impinges on the wall in a time $2/u$ after the first impact; hence the number of impacts per second is $u/2$. Consequently the average change of momentum per second for the molecule is

$$-2mu(u/2) = -mu^2.$$

But by definition the average rate of change of momentum is the average force acting on a body; therefore the average force acting on a molecule is $-mu^2$. By the law that action and reaction are equal and opposite, the average force exerted by the molecule against the wall is

$$mu^2.$$

There are n molecules in the volume and hence the sum of the average forces exerted by the molecules is

$$\Sigma mu^2,$$

where Σ indicates the sum of these quantities for all molecules. The result is

$$\Sigma mu^2 = mn\bar{u}^2 = (\tfrac{1}{3})mn\bar{q}^2.$$

One may interchange the operations of taking an average throughout time and finding a sum. Hence the time-average of the total molecular force exerted on a face of one square centimeter area is

$$(\tfrac{1}{3})mn\bar{q}^2.$$

This is the macroscopic pressure exerted by the gas and hence

$$P = (\tfrac{1}{3})mn\bar{q}^2.$$

Now the number of molecules in one cubic centimeter multiplied by the mass of a molecule is equal to the density

$$\rho = mn.$$

Therefore

$$P = (\tfrac{1}{3})\rho\bar{q}^2.$$

Thus in the kinetic theory of gases we explain the pressure of a gas as the time-average of the instantaneous force exerted by all the molecules striking a boundary surface of unit area. Consider further a cylinder of uniform cross-section closed at both ends and containing two masses of gas separated by a closely fitting, weightless, movable piston. Then from the macroscopic point of view one asserts that the piston remains at rest if the pressures of the two masses of gas are the same. From the standpoint of kinetic theory this means that the time-average of the total molecular force on the piston must be the same on both sides. The instantaneous values of the total molecular force will vary, but this will produce only unobservable fluctuations in the position of the piston. In order that there be observable motion the time-average of molecular force on one side must be greater than that on the other side.

Returning to our expression for the pressure let us multiply both sides of the equation by the volume. Then

$$PV = (\tfrac{1}{3})Vnm\bar{q}^2.$$

Now $\frac{1}{2}m\bar{q}^2$ is the average kinetic energy per molecule, and $(\frac{1}{3})$ $(\frac{1}{2}m\bar{q}^2)$ is the average kinetic energy per degree of freedom. Now another assumption of the kinetic theory is that the macroscopic temperature of a gas is correlated with the average kinetic energy per degree of freedom. Thus we may put

$$\tfrac{1}{3}(\tfrac{1}{2}m\bar{q}^2) = \tfrac{1}{2}kT.$$

Then

$$PV = VnkT.$$

Let V be the volume of one gram molecule of gas. Then $Vn = N$, the number of molecules in one gram molecule. Hence

$$PV = NkT$$

$$PV = RT.$$

This is the general gas law, which expresses the relation between the macroscopic volume, pressure and temperature of a gas. R is the universal gas constant. Thus from the assumptions that a gas consists of molecules which are subject to the laws of classical dynamics, and that pressure is due to molecular impacts and temperature is correlated with the average kinetic energy per degree of freedom, we have deduced a macroscopic law.

In the preceding discussion only the average square of the velocity of the molecules was employed. In other problems we need to take account of the difference in velocities. Owing to collisions, however, the molecules will have different velocities. We can not determine the velocity of an individual molecule, but we can determine the statistical distribution of velocities. That is, one can specify that a certain fraction of the molecules will have velocities which fall within certain limits. Thus one could calculate the fraction of the molecules whose velocities are between 100 and 101 meters per second. In the following I shall give a rough sketch of the methods by which the statistical distribution of molecular velocities is obtained, for example, by J. H. Jeans.

The gas is supposed to be constituted of a very large number of exactly similar, smooth, elastic spheres, between which the forces are negligible except during collisions. It is assumed that the laws of dynamics, which have been derived from observable bodies, determine the motion of the unobservable molecules. During impact the momentum and energy of the system of colliding molecules are conserved. Let us suppose that the components of velocity of a molecule are u, v, w. The law of distribution of velocities can be described with the aid of a velocity space. Suppose a set of rectangular coordinate axes to be chosen and let the coordinates of a point represent the components of velocity of a molecule. Then the different points of the space represent different velocities of a molecule. One selects an element of volume between u and $u + du$, v and $v + dv$, w and $w + dw$, that is, a small cube $dudvdw$. The velocities of the molecules may be described statistically by stating the number of points in the given volume element. Thus $Nf(u, v, w)\, dudvdw$, where N is the number of molecules

and $f(u, v, w)$ is the law of distribution of velocities, gives the number of molecules whose velocities have components between the assigned limits.

An assumption which is fundamental in the whole investigation is that of molecular chaos: the molecules having velocity components lying within any small specified range are, at any moment, distributed at random throughout the gas.

Consider a certain class of molecules (class A) which are represented by points that fall in a definite element of volume in the velocity space. The first step consists in finding an expression for the change, caused by collisions, in the number of molecules belonging to this class. It is possible to set up an expression for the number of collisions of a certain type in a small interval of time per unit volume, such as collisions in which one molecule (of class A) has velocity components within given intervals, while the other molecule (of class B) has velocity components in different intervals, and the line joining the centers of the two colliding molecules at the moment of impact falls in an elementary cone whose axis has an assigned orientation. The center of a molecule of class B which collides with a given molecule of class A in a small element of time must be situated in a small cylinder, the volume of which can be expressed in terms of the given data. On multiplying by the number of molecules of class B per unit volume one obtains the number of collisions for a given molecule of class A within the given time interval. On multiplying by the number of molecules of the class A per unit volume one obtains the number of collisions of the assigned type through which molecules of class A leave that class in the given time interval per unit volume.

One next considers a type of collision after which one of the molecules is of class A, the second is of class B, and the direction of the line of centers satisfies the same condition as for a collision of the previous type. One can then express the number of collisions per unit volume in a given time through which molecules enter the class A. In setting up this expression one makes use of the relations between the velocities before and after impact given by the conservation laws. By integrating the two expressions over all possible classes of collisions one

obtains, first, the total number of molecules of class A which enter into collision in the given time; second, the total number of molecules which enter class A during the same interval. The net gain to class A in the time interval is the difference of these integrals. Again, the gain to class A can be expressed in terms of the rate of change of the distribution function. When the two expressions for the net gain to class A are equated there results an equation for the rate of change of the distribution function with respect to the time. The problem is to find the distribution function which does not change with the time. After Boltzmann this is attacked by considering a function H which depends upon the law of distribution of velocities. Investigation shows that dH/dt is either negative or zero. The necessary and sufficient condition for constancy of the distribution function is that

$$\frac{dH}{dt} = 0.$$

From this equation there follows Maxwell's law for the distribution of velocities

$$f = Ae^{-hm(u^2 + v^2 + w^2)}.$$

If the initial distribution does not satisfy this law the collisions between the molecules change the statistical distribution until the steady state is realized.

Thus, by applying the laws of dynamics to the motion of a system of unobservable particles, it has been possible to obtain statistical knowledge of the distribution of molecular velocities. From the distribution law it is possible to calculate average values of molecular quantities upon which observable properties depend. Furthermore, a function H has been discovered which is either constant or decreases. If the initial distribution is not Maxwellian, H decreases until the latter distribution is realized. In this property the function H corresponds to the thermodynamic function entropy which increases in consequence of irreversible processes. The theorem concerning the decrease of H suggests the general principle which demands the reduction of irreversible macroscopic phenomena to microscopic reversible mechanical processes. The

second law of thermodynamics expressed in the form that natural processes are irreversible is thereby explained mechanically.

The theorem expressing the decrease of H during the transformation of any arbitrary distribution of velocities into the Maxwellian distribution in an irreversible manner was announced by Boltzmann in 1872. Boltzmann's result was soon subjected to criticism, for it was felt to be paradoxical that a reversible gas model should be able to explain irreversible processes. Loschmidt drew the consequences from the reversibility of the motion of the gas. Since the gas model is a conservative mechanical system, if it is capable of a given motion it is also capable of the reverse motion. From the definition of H it follows that, if in the one motion the values of H decrease from H to H_n, in the reverse motion the values of H increase from H_n to H. The possibility of a motion in which H increases is inconsistent with the fact of irreversibility, and further indicates a self-contradiction in dynamics, because the decrease of H follows from the dynamics of the collisions. A second criticism was made by Zermelo by appliton of a dynamical theorem proved by Poincaré. If the gas model passes through a series of phases during which H decreases, after a finite time there will occur in the course of the undisturbed motion a series of phases which are as near as one pleases to the first series. Consequently the values of H in the second series will be nearly equal to the values for corresponding phases in the first series. Thus in the course of the motion from the end of the first series to the beginning of the second series there is an increase in the function H. In response to these criticisms Boltzmann developed the statistical conception of the H theorem. This was the origin of the contemporary view that macroscopic laws express the probable sequence of phenomena. The statistical distributions of atomic processes are probable results which are subject to fluctuations. Thus macroscopic phenomena, which depend on the average values of molecular quantities, may exhibit irregularities. The further analysis of statistical method may be carried out by an exposition of the method of statistical mechanics.

8. THE METHOD OF STATISTICAL MECHANICS

In the classical dynamics one studies the motion of a system of a few degrees of freedom, such as a particle or rigid body, by solving the equations of motion and determining the arbitrary constants from the initial conditions. This procedure is not applicable to a system consisting of many parts. In statistical mechanics one has a method for the derivation of the macroscopic properties of a very general mechanical system.

In the statistical mechanical study of a complex system the state of the system is represented with the aid of n-dimensional space. Consider a system of n degrees of freedom. The momentary phase of the system is the configuration and state of motion which are defined by $2n$ data, n independent generalized coordinates and n generalized momenta. The phase is represented by a point in a $2n$-dimensional space; the Cartesian coordinates of the point are the constitutents of the phase. As the system moves, the phase point describes a path in the representative space.

In accordance with the statistical method we shall describe the momentary phase by defining the element of volume in which the phase point falls, for example, the element bounded by the hyperplanes q_1 and $q_1 + dq_1$, etc.

The fundamental problem of statistical mechanics is: How is the observable state correlated with the variable microscopic states? We have previously seen that macroscopic phenomena are average resultants of atomic processes; for example, the pressure of a gas is the time-average of the total force exerted by the molecules in collision with a unit surface of the containing vessel. It is evident that those microscopic states which occur with the greatest frequency have the most significance for the observable states. In order to express the frequency of occurrence of a given microscopic state consider the motion of the system during a long interval of time T. Suppose that during the time τ the system is in the state P. Then the probability that at any moment chosen at random the system is in P may be expressed as τ/T. The succession of phases P_1, P_2 . . . and the respective probabilities that the

system is in those phases constitute the time-ensemble of the individual system.

The method of statistical mechanics is based upon the fact that the time-ensemble of a single system can be represented by a fictitious ensemble of systems (virtual ensemble) all of which are like the given real system in structure but differ from it in phase. Thus all possible phases of the given real system are represented by an ensemble of systems. The possibility of representing the time-ensemble of a single system by an ensemble may be exemplified by the following example which is due to P. Hertz. Consider a roulette game in which there are 37 equal sectors through which a ball passes with constant speed; during $\frac{1}{37}$ of the time the ball will be in sector 0. Thus the probability of finding the ball in sector 0 at any random time is $\frac{1}{37}$. If there were an ensemble of 37,000 independent roulette games in progress at a given time, one would then expect that in $\frac{1}{37}$ of them, that is 1000, the ball would be in the sector 0 at any time. If one selects a game from the system at random the probability that the ball is in sector 0 is $\frac{1}{37}$. The distribution in phase of the ensemble of many systems represents the distribution in phase of the time-ensemble of the single system. By the method of statistical mechanics the probable phases and behavior of a single real system are inferred from the statistical properties of an ensemble of systems. It should be added, however, that there is lacking a completely satisfactory logical justification for the use of virtual ensembles in the investigation of the properties of a single system.

A type of ensemble which has been widely used is constructed as follows. Given a real system of constant energy E_0. In the course of its motion the phase point of the system will describe a path on the energy surface $E = E_0$. Phase points are distributed with uniform density in the region of the phase space between the surfaces $E = E_0$ and $E = E_0 + dE_0$. One investigates the properties of a single system by studying the statistical properties of the ensemble. From the equations of motion it follows that in the course of the motion the density of the phase points remains uniform. It is therefore possible to express the probability, of finding in an assigned state a

system chosen at random, as proportional to the volume element in which the points represent the given state. By a geometrical analysis of the phase space occupied by the ensemble one discovers that in all but a relatively small portion of the space the points represent systems having certain properties. In selecting a point at random it is highly probable that it will represent a system which possesses this property. Since the statistical behavior of the ensemble represents the probable behavior of a single system it is highly probable that the real system will possess the given property at a random time.

The method may be illustrated by an investigation of the density of a gas. We shall employ an ensemble of systems in which the points are uniformly distributed throughout the phase space. In this problem we shall ignore differences of velocities of molecules and assume that the phase of the system is defined by the distribution of the molecules throughout the container. In order to define the density of the gas one supposes the container to be subdivided into equal elements of volume which may be labeled 1, 2, 3, . . . n. It is assumed that each volume-element contains a large number of molecules. The state of the gas may now be considered from two points of view. The macroscopic state of the gas is defined by the assignment of the number of molecules in the different elements of volume, for example, a macroscopic state is defined by the statement that there are N_1 molecules in cell 1, N_2 molecules in cell 2, etc. The microscopic state of the gas is defined by the assignment of particular molecules to the different volume-elements. Thus if the molecules are given labels a, b, c, d, etc., a microscopic state is defined by the statement that a and b are in cell 1, c and d in cell 2, etc. It is evident that to a given macroscopic state there are many microscopic states, for if two molecules in two different cells exchange places there is a change in the microscopic but not in the macroscopic state.

The relation between macroscopic and microscopic states may be explained with the aid of a simple example. Consider a gas of four molecules, labeled a, b, c, d. Suppose that the container is subdivided into two equal parts. The different possible macroscopic states are:

(1) *Four molecules in cell* 1; *zero molecules in cell* 2.
(2) *Three molecules in cell* 1; *one molecule in cell* 2.
(3) *Two molecules in cell* 1; *two molecules in cell* 2.
(4) *One molecule in cell* 1; *three molecules in cell* 2.
(5) *Zero molecules in cell* 1; *four molecules in cell* 2.

There is only one microscopic state corresponding to macroscopic state (1). There are four microscopic states corresponding to (2), that is,

a, *b*, *c* in cell 1; *d* in cell 2.
b, *c*, *d* in cell 1; *a* in cell 2.
c, *d*, *a* in cell 1; *b* in cell 2.
d, *a*, *b* in cell 1; *c* in cell 2.

There are six microscopic states corresponding to macroscopic state (3):

a, *b* in cell 1; *c*, *d* in cell 2.
a, *c* in cell 1; *b*, *d* in cell 2.
a, *d* in cell 1; *b*, *c* in cell 2.
b, *c* in cell 1; *a*, *d* in cell 2.
b, *d* in cell 1; *a*, *c* in cell 2.
c, *d* in cell 1; *a*, *b* in cell 2.

Similarly there are four microscopic states corresponding to macroscopic state (3), and one corresponding to macroscopic state (4). From this example it is clear that the macroscopic state which is realized by the greatest number of microscopic states is that of uniform density.

Now from thermodynamics we know that a non-equilibrium condition tends to change into a state of equilibrium. Observation on gases in equilibrium reveals that the density is uniform throughout the gas. Thus the example of the distribution of molecules suggests a relation between the equilibrium state and the macroscopic state which is realized by the greatest number of microscopic states. The relation is based on the concept of the probability of a macroscopic state. The equilibrium state is that of maximum probability, and the probability of a macroscopic state is proportional to the number of microscopic states which realizes it.

Continuing the exposition in terms of the density of gas,

I shall now express the probability of a given macroscopic distribution of molecules. We have seen that the probability of a given state is proportional to the volume of the phase space which represents the given state. Consider first a given microscopic state which is defined by the assignment of particular molecules to the various volume-elements into which the container is subdivided. Since in a given microscopic state the molecules may occupy any position in their respective cells without changing the state, there is a corresponding element of phase volume for freedom of motion of the phase point. Hence the phase point representing a given microscopic state occupies an element of volume, say V. Keeping the macroscopic state fixed, new microscopic states may be generated by exchanging molecules between cells. Corresponding to each microscopic state there is a particular element of volume in the phase space. The total volume of the phase space corresponding to a given macroscopic state is equal to the volume corresponding to a microscopic state multiplied by the number of microscopic states. The complete expression is

$$\frac{N!}{N_1!N_2!N_3!\ldots}V.$$

The probability of a given macroscopic state is proportional to the foregoing expression. The most probable state is the one which is represented by the greatest volume of the phase space. Analysis reveals that the overwhelmingly major portion of the phase space represents a state of uniform density, that is, a macroscopic state in which $N_1 = N_2 = N_3 \ldots$ etc.

We may now consider the predictions of the behavior of the system from the properties of the phase space. A uniform distribution of phase points exists permanently throughout the phase space. If a phase point is selected at random it is highly probable that it represents a system in which the macroscopic density is uniform. From this one infers that if one observes the single real system at any moment of time chosen at random it is highly probable that the density is uniform. If a phase point which has been selected at random should happen to be in a region of the space for which the density is non-uniform, in the course of its motion it will enter the region of

uniform density. Hence if a real system is observed in a non-uniform state it is probable that it will pass directly into a uniform state. Further, since phase points pass from the region of uniform density into that of non-uniform density, one may expect in the real system departures from the state of uniform density.

From the conclusions for the foregoing special case we may now pass to a general statement. It has been explained that the impetus to the development of statistical mechanics was given by the need of providing a self-consistent mechanical basis for the macroscopic laws of thermal phenomena. From the standpoint of the kinetic-molecular theory of matter, a gas is a reversible mechanical system. But the Second Law of thermodynamics, which is derived from macroscopic observation, asserts the occurrence of irreversible processes. The reconciliation of the two points of view is achieved through the application of the concept of probability. According to the Second Law, if a system is left to itself non-equilibrium states are succeeded by a condition of equilibrium in which the entropy is a maximum. According to statistical mechanics it is probable that a system will pass from less probable states to one of maximum probability. The thermodynamic quantity entropy is proportional to the logarithm of the statistical mechanical quantity probability. On this basis one achieves a mechanical derivation of a statistical form of the Second Law. It is probable that natural processes will go in the direction of maximum probability, that is, maximum entropy. There is also possible a passage from the more probable to the less probable states, that is, an increase in departure from equilibrium. If a system is initially in a state of equilibrium, fluctuations from the equilibrium state will occur. For the most part these fluctuations will not be observable, but if the system be observed for a long enough time it is certain that observable departures from equilibrium will occur. In general, macroscopic laws are the manifestation of statistical regularities among microscopic processes.

The statistical interpretation of macroscopic laws was developed in order to satisfy the theoretical demand for a self-consistent mechanical basis for irreversible processes, but it

was later verified by experiments rendered possible by more refined methods of observation. In the Brownian movement, the irregular motion of small particles suspended in a liquid, one is able to observe the fluctuations predicted by molecular theory.

We may summarize the preceding analysis with a comparison between causal and statistical laws. A causal law is exemplified by the differential equations of dynamics. The causal law is independent of the initial conditions. Now we have seen that macroscopic phenomena are dependent upon microscopic processes. In the classical statistical mechanics, the elementary processes are assumed to be governed by causal laws, from which the statistical laws of macroscopic phenomena are then derived. Since statistical regularities are subject to fluctuations, sequences in observable phenomena theoretically are relative to the conditions of observation. The observable macroscopic process is dependent upon the initial conditions; if the period of observation were long enough, the exceptional states predicted by the theory of probability would occur. However, on account of the enormous number of elementary processes which constitute a macroscopic phenomenon, very refined methods of observation are necessary to observe irregularities in the sequence of observable events. For all ordinary purposes the most probable sequence of events may be assumed to be invariable. Theoretically, however, there exists a distinction between the strict determinism of causal laws and the fluctuating determinism of statistical laws.

In the classical statistical theory, which has been expounded in this chapter, it is assumed that the laws of microscopic processes are causal laws. But in contemporary quantum theory the view has been adopted that the elementary quantum processes are subject to statistical laws.

9. THE ELECTROMAGNETIC FIELD

In the chapter on electrodynamics we learned that there is a force of attraction or repulsion between electrified bodies and between magnetized bodies. In the early history of the theory of electricity and magnetism electric and magnetic

actions were interpreted as actions at a distance. Faraday, however, conceived of the medium as the seat of electric and magnetic forces. Maxwell then found the equations which represent electromagnetic phenomena as continuous processes in a medium. The mechanical conception of nature requires that one interpret electric and magnetic action in terms of mechanical stresses, strains and motions of the medium.

Electric and magnetic fields may be represented by lines of force which begin and end on charged or magnetized bodies respectively. Faraday pictured these lines as real ropes of molecules which tend to shorten and to expand laterally; thus electric and magnetic action was viewed as the manifestation of a tension along the lines and a pressure at right angles to them. Maxwell also was guided in his electromagnetic theory by mechanical models. He imagined the magnetic field as the seat of rotations of molecules about the magnetic lines of force. In this model the strength of the magnetic field is determined by the angular velocity of the molecules, electromagnetic induction is the effect of the force called into play when the angular velocity of the rotating elements changes, electromotive force arises from the stress on the connecting mechanism, electric displacement arises from the elastic yielding of the connecting mechanism. With the aid of this model Maxwell discovered the laws of the electromagnetic field. During the latter half of the nineteenth century a great effort was made to construct a satisfactory mechanical model of the electromagnetic field out of gyroscopes, pulsating spheres, etc. All these attempts have now been practically abandoned. However, traces of the mechanical point of view are to be found in British books on electricity and magnetism. Thus R. A. Houstoun says of the vector **D** that it is supposed to measure a state of strain.

One valuable result remains of the attempt to construct a mechanical theory of the electromagnetic field. Maxwell succeeded in expressing in the Lagrangian form the law of induction and the law of force on a circuit due to other circuits. In the chapter on dynamics the term coordinate applies to geometrical coordinates. Maxwell extended the concept to include electric charge; the corresponding generalized velocities are

electric currents. He then found a function of the currents and the configuration of the system which plays the rôle of kinetic energy in Lagrange's equations. Thus the laws of electrodynamics, in the restricted sense of the mutual action between currents, may be expressed in the Lagrangian form. This is said to indicate the possibility of mechanical explanation. However, as Abraham remarks, it proves equally the possibility of reducing mechanics to electrodynamics.

REFERENCES

EHRENFEST, P. and T. Begriffliche Grundlagen der statistischen Auffassung in der Mechanik. Leipzig und Berlin, 1912.

GLAZEBROOK, R. T. James Clerk Maxwell. London, 1896.

HERTZ, P. Statistische Mechanik. Ergebnisse der exakten Naturwissenschaften, Vol. 1, 1922.

—— Statistische Mechanik. In Weber and Gans, Repertorium der Physik, I, 2. Leipzig und Berlin, 1916.

JEANS, J. H. The Dynamical Theory of Gases. Cambridge, 1916.

LEWIS, E. P. Light, *in* A Textbook of Physics, edited by A. W. Duff. Philadelphia, 1921.

MAGIE, W. F. Principles of Physics. New York, 1911.

PLANCK, M. Physikalische Rundblicke. Leipzig, 1922.

TOLMAN, R. C. Statistical Mechanics. New York, 1927.

CHAPTER IX

THE ELECTRODYNAMIC CONCEPTION OF NATURE

A. MICROSCOPIC ELECTRODYNAMICS

1. INTRODUCTION

In the mechanical conception of nature the fundamental physical phenomenon is that of a material body in motion. The fundamental concepts are of space, time, mass and mechanical force. Mechanical force has been conceived of as an action at a distance between bodies, or as an action by contact of material bodies. The mechanical theory was the canonical system of the nineteenth century; contemporary physics, however, is largely guided by an electrodynamic conception of nature.

Electrodynamics is the study of **electric charges** and of the **electromagnetic field** which is excited by the charges. Like mechanical phenomena, electrodynamic processes occur in space and time. An electrically charged body has position in space and is surrounded by an electric field. Electric charges in motion constitute an electric current which excites a magnetic field. An electromagnetic field is a physical condition which is propagated throughout space with a finite velocity in accordance with laws expressed by Maxwell's equations. Thus the fundamental concepts of electrodynamics are of space, time, electric charge, electric field, etc. In place of mechanical force we have the concept of a force which arises from the action of the electromagnetic field upon the electric charges of bodies. The program of an electrodynamic conception of nature is the interpretation of all physical properties and phenomena in terms of the concepts of electrodynamics.

A first step in the development of a unitary electromagnetic theory of physics was the recognition of the fundamental reality of the electromagnetic field. We have seen that an

attempt was made to find a mechanical basis for this field. During the second half of the nineteenth century, the British physicists in particular sought to construct mechanical models for the interpretation of electromagnetic phenomena. However, all these mechanical models proved to be unsatisfactory. Gradually, there occurred a change in the point of view of physicists; the concept of the electromagnetic field began to be recognized in its own right, to be considered as fundamental as the concepts of mass and mechanical force. The recognition of the fundamental character of electromagnetic theory was expressed by H. Hertz in the statement that Maxwell's theory of electricity and magnetism consists of Maxwell's equations. However, as we shall see, a relic of the mechanical point of view remained in the concept of a substantial ether as the seat of the electromagnetic field.

2. THE ATOMISTIC THEORY OF ELECTRICITY

The systematic extension of electrodynamics to the entire physical order was rendered possible by the discovery of the atomic constitution of electric charges. We have previously seen that a material body may be conceived of as built up of atoms, so that the mass of the body is the sum of the masses of the atoms. Similarly, the macroscopic electric charge of a body is the sum of elementary charges.

The first clue to the atomic nature of electric charge was given by the laws of **electrolysis** which were discovered by Faraday. Suppose, for example, that two platinum plates are partially immersed in a solution of hydrochloric acid (electrolyte) and the plates (electrodes) are inserted in a circuit containing a source of current. A current will flow in the circuit and decompose the hydrochloric acid; chlorine will be released at the plate (anode) at which the current enters the solution, hydrogen will be released at the plate (cathode) at which the current leaves the solution. This process is called electrolysis and is characterized by two laws which were discovered by Faraday. The first law is that

The mass of the substance liberated at an electrode is proportional to the quantity of electricity which passes through the electrolyte.

The second law is that

The mass of any substance liberated for a given quantity of electricity is proportional to the chemical equivalent of the substance.

Helmholtz in 1881 drew the conclusion that the facts of electrolysis implied an atomic constitution of electricity. Thus he said, " If we accept the hypothesis that the elementary substances are composed of atoms, we cannot avoid concluding that electricity is also divided into elementary portions which behave like atoms of electricity."

Our conception of the process of solution is that molecules are dissociated into electrically charged constituents, called ions. During electrolysis the positively charged ions move toward the negative electrode (cathode), and the negatively charged ions move toward the positive electrode (anode).

Further evidence in favor of the atomic conception of electricity was obtained from the study, during the second half of the nineteenth century, of the **discharge of electricity through gases.** In these experiments one employs a vacuum tube—a glass tube containing gas which can be removed by means of a pump, and having sealed into it electrodes between which there is a difference of potential arising from some electrical machine. When the pressure of the gas within the tube is sufficiently low, cathode rays are emitted from the cathode. The behavior of the rays in magnetic and electric fields indicated that the rays consist of particles charged with negative electricity. It was possible to measure the ratio of the charge to the mass, e/m, and the value was found to be about 1800 times of the value for the hydrogen ion in electrolysis. It was assumed that the charge is the same as on the hydrogen ion in electrolysis and hence the mass of the negatively charged particle was then inferred to be 1/1800 of the mass of the hydrogen atom. The value of e/m was found to be the same, regardless of the kind of gas in the vacuum tube. This result suggests the hypothesis that the negatively charged particles of the cathode rays are universal constituents of atoms.

In 1895 Roentgen discovered that, when cathode rays strike a solid, there is an emission of radiation which is able to pene-

trate an opaque screen and affect a photographic plate. The investigation of these Roentgen, or X-rays, led to the discovery of **radio-activity.** The salts of certain elements have the property of spontaneously emitting radiations. A part of this radiation consists of β rays, which consist of negatively charged particles like the cathode rays. This was additional evidence in favor of the view that negatively charged particles are the building stones of matter.

During this period Zeeman discovered that, if a sodium flame is placed in a strong magnetic field, a certain yellow spectral line is split into a number of components. Lorentz was able to derive this effect from the assumption that the emission of light is due to the motion of negatively charged particles, like the cathode rays.

Thus the phenomena of electrolysis, electric discharge in vacuum tubes, radio-activity, and the action of a magnetic field upon a gas emitting light demand the hypothesis that matter is constituted of negatively charged particles. Positively charged particles were also discovered within vacuum tubes and in radio-active processes. Hence the hypothesis arose that matter consists of positively and negatively charged particles. As we shall see later, the final outcome of this development was that there are two kinds of electrified particle, **electrons** and **protons**; the electron is the constituent of cathode and beta rays and its charge is the ultimate unit of negative electricity; the proton is the nucleus of the hydrogen atom and is the atom of positive electricity. The nuclei of other atoms are built up of protons and electrons. A neutral atom, that is, one without an electric charge, consists of a positively charged nucleus and electrons, so that the total charge is zero. A molecule consists of a union of atoms. The kinetic molecular theory of matter thus becomes a first approximation to a microscopic theory. The electrical theory of matter gives a still more detailed analysis.

The atomistic theory of electric charge is verified relatively directly by Millikan's method of determining the unit of charge. In Millikan's experiment an oil drop is electrified and its motion is observed under the action of gravity and of an electric field. The measurements enable one to calculate the charge on the

oil drop on the assumption that the electric force on the drop is equal to the product of the electric field and the charge. Measurements of the electric charge on many oil drops show that the charge is always an integral multiple of a unit charge, which is therefore the atomic unit of electric charge. In Millikan's method one uses macroscopic measurements and macroscopic laws in order to determine a microscopic quantity. Microscopic electrodynamics thus has a more direct experimental basis than classical kinetic theory.

We have now seen that electricity is atomistic in structure; there are elementary negatively charged particles, electrons, of mass 1/1800 that of the hydrogen atom, and positively charged particles of at least atomic mass. Thus one has the material for a microscopic theory of electromagnetic phenomena within material media. The electromagnetic theory of Maxwell gave a representation of the macroscopic properties of electric charges, currents, electromagnetic fields, etc. One is impelled to develop a microscopic electrodynamics for the purpose of obtaining a more detailed representation of electromagnetic processes. The program of constructing a microscopic theory of electrodynamics was undertaken by Lorentz.

3. THE ELECTRON THEORY

In the theory of electrons it is assumed that ponderable bodies contain small positively and negatively charged particles, the term electron usually being applied to both types of charged particle. A force is exerted upon electrons by an electromagnetic field. A conductor is characterized by the presence in it of free electrons which move under the action of electric force and thereby constitute a macroscopic electric current. In a non-conducting material the electrons are bound to positions of equilibrium; upon the application of a field these electrons are displaced until the electric force is balanced by an elastic restoring force—a process which is called polarization of the dielectric. Again, electrons can vibrate about their positions of equilibrium and thereby become the centers of electromagnetic waves. Conversely, an incident electromagnetic wave sets electrons into vibration and this process is responsible for the dispersion and absorption of radiation.

This rough qualitative sketch shows that the microscopic theory renders possible a detailed analysis of macroscopic phenomena. The quantitative microscopic theory is based upon a set of equations due to Lorentz. Like mechanical kinetic theory, microscopic electrodynamics is based upon the assumption that the equations which hold for macroscopic processes can be extended to microscopic processes.

A fundamental assumption of Lorentz's theory is the existence of an ether as the medium of electromagnetic actions; electric and magnetic fields are states of the ether. The ether is assumed to be at rest and is the proper frame of reference for the description of electromagnetic phenomena. Unlike the classical dynamics, in which there is a set of privileged frames of reference, the electron theory restricts the permissible frames of reference to those at rest with respect to the fixed ether. The ether pervades electrons, which are assumed to be finite in extension.

The microscopic equations are as follows:

$$\operatorname{div} \mathbf{e} = 4\pi\rho. \tag{1}$$

This equation is obtained from the equation

$$\operatorname{div} \mathbf{D} = 4\pi\rho.$$

The small letter **e** indicates that it symbolizes the field of an electron, and therefore is a microscopic quantity.

$$\operatorname{div} \mathbf{h} = 0. \tag{2}$$

This equation is obtained from

$$\operatorname{div} \mathbf{B} = 0.$$

Again the small letter **h** symbolizes a microscopic quantity. I remark that Mason and Weaver make the valuable suggestion that the symbol **b** would be more appropriate.

$$\operatorname{curl} \mathbf{h} = \frac{1}{c}\frac{\partial \mathbf{e}}{\partial t} + \frac{4\pi}{c}\rho\mathbf{v}. \tag{3}$$

This equation is derived from

$$\operatorname{curl} \mathbf{H} = \frac{4\pi\mathbf{k}}{c},$$

where **k** is the total current.

We obtain the equation of the electron theory by the assumption that there are two kinds of current. There is the displacement current in the ether

$$\frac{1}{4\pi}\frac{\partial \mathbf{e}}{\partial t}.$$

The motion of an electron constitutes a convection current

$$\rho \mathbf{v}$$

where ρ is the density of charge and $\mathbf{v}$ is the velocity of the electron.

$$\text{curl } \mathbf{e} = -\frac{1}{c}\frac{\partial \mathbf{h}}{\partial t}. \tag{4}$$

This equation is obtained from

$$\text{curl } \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}.$$

Lastly we have an equation which expresses the force acting on an electron,

$$\mathbf{f} = \mathbf{e} + \frac{1}{c}[\mathbf{v} \times \mathbf{h}]. \tag{5}$$

$\mathbf{f}$ is the electromagnetic force per unit charge. The velocity of light and the velocity of electrons are to be measured with respect to the ether.

I shall now sketch the method whereby the principles of macroscopic electrodynamics for bodies at rest can be derived from the microscopic equations of the electron theory. The deduction is based upon the assumption that the macroscopic quantities are average values of microscopic quantities.

From the microscopic equations we obtain by averaging

$$\text{div } \bar{\mathbf{e}} = 4\pi\bar{\rho}. \tag{1a}$$

$$\text{div } \bar{\mathbf{h}} = 0. \tag{2a}$$

$$\text{curl } \bar{\mathbf{h}} = \frac{1}{c}\frac{\partial \bar{\mathbf{e}}}{\partial t} + \frac{4\pi}{c}\overline{\rho \mathbf{v}}. \tag{3a}$$

$$\text{curl } \bar{\mathbf{e}} = -\frac{1}{c}\frac{\partial \bar{\mathbf{h}}}{\partial t}. \tag{4a}$$

In order to show that the foregoing equations may be interpreted as the macroscopic equations it is necessary to indicate how the electric and magnetic properties of bodies depend upon the existence of electrons. The conduction of electricity is based upon conduction electrons, which in metals are free electrons and in electrolytes may be attached to neutral atoms or molecules. In a dielectric, negative electrons are bound to positive electrons; upon the application of an electric field the negative electrons are displaced with respect to the positive and together with them form dipoles. This is called polarization of the dielectric, and these electrons may be called polarization electrons. In a magnetizable body, the negative electrons revolve in orbits about the positive electrons, and may be called magnetization electrons.

Suppose in the first place that the body contains only conduction electrons. Then the observed density of electricity ρ is the average value of the density of conduction electrons throughout a volume which is physically infinitesimal.

$$(\overline{\rho})_c = \rho.$$

Similarly, the observed conduction current is the average value of the convection current

$$(\overline{\rho \mathbf{v}})_c = \mathbf{i}.$$

Suppose next that the body contains dipoles which consist of equal positive and negative charges separated by the distance l. The electric moment of the dipole is a vector whose absolute value is equal to the product of the absolute value of the charges and the distance between them ($p = el$), and whose direction is that of a line running from the positive charge to the negative charge. If we add up the electric moments of all the dipoles in a unit volume we obtain the polarization $\mathbf{P}$. During the process of polarization, electrons are being displaced and hence there is an electric current. The current density due to the polarization electrons is

$$(\overline{\rho \mathbf{v}})_p = \frac{\partial \mathbf{P}}{\partial t}.$$

The electric density arising from the polarization electrons is

$$(\bar{\rho})_p = -\operatorname{div} \mathbf{P}.$$

The process of magnetization is dependent upon electrons which move in orbits within the molecules. Suppose, then, that the body contains molecules of which the total charge and the electric moment are zero, but for which

$$\mathbf{m} = \frac{1}{2c}\int d\tau \rho[\mathbf{r} \times \mathbf{v}]$$

is not equal to zero. The molecule has the properties of a magnetic dipole and the time-average of $\mathbf{m}$ is the moment of the equivalent magnetic dipole. If we add up the magnetic moments of all the molecules in a unit volume we obtain the magnetization $\mathbf{M}$. According to the electron theory the average value of the electric current which arises from such molecules is

$$(\overline{\rho\mathbf{v}})_m = c \operatorname{curl} \mathbf{M}.$$

Combining all the preceding results we find that the average electric density is given by

$$\bar{\rho} = (\bar{\rho})_e + (\bar{\rho})_p = \rho - \operatorname{div} \mathbf{P}.$$

The average value of the electric current is given by

$$\overline{\rho\mathbf{v}} = (\overline{\rho\mathbf{v}})_e + (\overline{\rho\mathbf{v}})_p + (\overline{\rho\mathbf{v}})_m.$$

We assume

$$\mathbf{B} = \bar{\mathbf{h}}$$
$$\mathbf{E} = \bar{\mathbf{e}}.$$

Thus the vectors $\mathbf{B}$ and $\mathbf{E}$ are the average values of the microscopic field strengths $\mathbf{h}$ and $\mathbf{e}$.

On substituting in 3*a* we obtain

$$\operatorname{curl} \mathbf{B} = \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi}{c}\frac{\partial \mathbf{P}}{\partial t} + \frac{4\pi \mathbf{i}}{c} + 4\pi \operatorname{curl} \mathbf{M}$$

$$\operatorname{curl} (\mathbf{B} - 4\pi\mathbf{M}) = \frac{1}{c}\frac{\partial}{\partial t}(\mathbf{E} + 4\pi\mathbf{P}) + \frac{4\pi \mathbf{i}}{c}.$$

If we determine

$$\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}$$
$$\mathbf{H} = \mathbf{B} - 4\pi\mathbf{M},$$

we obtain from equations $1a$–$4a$,

$$\operatorname{div} \mathbf{E} = 4\pi\bar{\rho} = 4\pi\rho - 4\pi \operatorname{div} \mathbf{P}$$

$$\operatorname{div}(\mathbf{E} + 4\pi\mathbf{P}) = 4\pi\rho$$

$$\operatorname{div} \mathbf{D} = 4\pi\rho \tag{1b}$$

$$\operatorname{div} \mathbf{B} = 0 \tag{2b}$$

$$\operatorname{curl} \mathbf{H} = \frac{4\pi \mathbf{i}}{c} + \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t} \tag{3b}$$

$$\operatorname{curl} \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}. \tag{4b}$$

These are the macroscopic equations.

One may raise the problem of the significance of the microscopic quantities, such as the vectors **e** and **h**. The microscopic field is assumed to have a value at points within an electron, and it is obvious that this is not verifiable by direct measurement. In my opinion the point of view of traditional physics was that of ideal theoretical physics. That is, the microscopic quantities were thought of as ideally measurable in terms of ideal measuring rods and clocks. We have previously considered the objections to this interpretation. It appears to me that a satisfactory explanation of the significance of the microscopic equations has been given by Professor Swann.

There are experiments in which the observable results depend upon the electromagnetic force on the electron. Thus the connection of the microscopic equations with experiment may be through equation 5 of the microscopic theory. The force is a function of the electric and magnetic fields. Now the first four microscopic equations can be solved for **e** and **h** as functions of ρ and $\rho\mathbf{v}$. Thus one may view the equations as definitions of **e** and **h**, in terms of ρ and $\rho\mathbf{v}$. In a theoretical problem, in which one uses the vectors **e** and **h**, the starting point is the assumption of the charges and their motions. The applicability of the equations depends on this: It is possible to assume ρ and $\rho\mathbf{v}$ such that the substitution of the calculated values of **e** and **h** in the force equation yields the experimental result. Thus we interpret the equations in terms of the hypothetical assignment of numbers to microscopic bodies.

4. THE ELECTROMAGNETIC THEORY OF MASS

Thus far it has been assumed that an electron is characterized by the physical quantities mass and electric charge. The concept of mass is the basic concept of a mechanical theory, and it would be an important step toward an electrodynamical conception of nature if the concept of mass could be defined in terms of the electromagnetic concepts.

In 1881 J. J. Thomson calculated that it would require more work to give a charged body a given kinetic energy than an uncharged body. The motion of the body sets up a field which opposes the motion. The effect of the motion can be described by attributing an electromagnetic mass to the body in addition to its material mass.

In the theory of electrons each part of the electron experiences a force due to the field of the electron. If the electron moves with constant velocity, the resultant force due to its field is zero; if the electron is accelerated, its field exerts an opposing force. We may resolve the acceleration $\mathbf{a}$ into two components, the one $\mathbf{a}'$ in the direction of motion and the other $\mathbf{a}''$ at right angles to it. Then the force exerted upon the electron by its own electromagnetic field may be expressed by

$$-m'\mathbf{a}' - m''\mathbf{a}''.$$

Suppose that the electron is acted upon by some force $\mathbf{F}$ which is not due to its field. Then the resultant force is

$$\mathbf{F} - m'\mathbf{a}' - m''\mathbf{a}'',$$

and the equation of motion is

$$\mathbf{F} - m'\mathbf{a}' - m''\mathbf{a}'' = m_0(\mathbf{a}' + \mathbf{a}'').$$

This can be written

$$\mathbf{F} = (m_0 + m')\mathbf{a}' + (m_0 + m'')\mathbf{a}''.$$

Thus the electron moves as if it had two different masses, $m_0 + m'$ and $m_0 + m''$. m' is called the longitudinal electromagnetic mass since it comes into play when the acceleration is along the path of the motion, and m'' is called the transverse electromagnetic mass since it comes into play when the acceleration is normal to the path.

Experiments on beta rays, which are rapidly moving electrons, indicate that e/m decreases, and hence the mass increases, with the velocity. The experimental results indicate that one may interpret all the mass of the electron as electromagnetic. However, the experimental result is inconclusive in view of the fact that the special theory of relativity yields the same formula for the dependence of mass upon velocity as does the electron theory.

Taking the point of view of microscopic electrodynamics, however, one may say that the mechanical mass of the electron is a derivative quantity in electrodynamics. This result was generalized in the idea of an electrical theory of matter, the motives of which are summarized by Lorentz. He says, " In a more general sense, I for one should be quite willing to adopt an electromagnetic theory of matter and of the forces between material particles. As regards matter, many arguments point to the conclusion that its ultimate particles always carry electric charges and that these are not merely accessory but very essential. We should introduce what seems to me an unnecessary dualism, if we considered these charges and what else there may be in the particles as wholly distinct from each other."

A serious shortcoming in the electromagnetic theory of matter has been the difficulty in explaining the existence of the electron. In order that an electron be held together, there must be cohesive forces to counterbalance the self-repulsive tendency of its charge. These cohesive forces are not deducible from the Maxwell-Lorentz theory. Accordingly, Poincaré introduced a cohesive pressure of unknown character; Mie generalized the classical theory so that within the elementary electrified particles the repulsive forces would be counterbalanced by electrical forces which become negligible outside the charge; Einstein assumed that elementary particles are held together by gravitational forces. All these theories proved to be unsatisfactory. We shall not be interested in further details, for the classical microscopic electrodynamics is merely the first approximation to a quantum electrodynamics which is in process of creation.

5. THE ELECTRODYNAMIC THEORY OF MECHANICAL FORCE

Having discussed the electromagnetic theory of mass, I turn to the electromagnetic theory of mechanical force. The typical mechanical force is that exerted across a surface between two contiguous parts of an elastic body. In the mechanical conception of nature electrical forces between charged bodies have been interpreted as stresses in an elastic ether. In the electrodynamic theory, however, the electromagnetic field is assumed to be a fundamental reality; hence the mechanical stresses in elastic solids are to be derived from electromagnetic forces.

The possibility of an electromagnetic theory of mechanical forces is demonstrated by the electrical theory of crystals of M. Born. A crystal consists of ions arranged in the form of a lattice; for example, rock salt is built up of positive sodium ions and negative chlorine ions alternately situated at the corners of a cubical lattice system. As a first approximation one assumes that the ions are at rest under the action of electrostatic forces. Now the electrical forces between two ions must be partly attractive and partly repulsive: a force of attraction between the ions is necessary to prevent disintegration; a force of repulsion is required to allow compression. It is assumed that the force of attraction is determined by Coulomb's law; the law of repulsion is assumed to vary inversely as some power of the distance. The resultant force between the ions is described in terms of the potential energy of the two ions

$$\varphi = \pm \frac{e^2}{r} + \frac{\beta}{r^n},$$

where r is the distance apart and β is a constant. From the potential energy of two ions one is able to calculate the potential energy of the entire lattice. One obtains

$$\Phi = -\frac{a}{\delta} + \frac{b}{\delta^n},$$

where δ is the distance between an ion and the next similar ion on the edge of the cube. The condition of equilibrium is that the potential energy be a minimum. One can then express

the equilibrium potential energy in terms of the equilibrium value of δ.

The compressibility $1/k = VdP/dV.$

The pressure $$P = -\frac{d\Phi}{dV},$$

$$dP/dV = -\frac{d^2\Phi}{dV^2}.$$

It follows that

$$k = \frac{9\delta_0^4}{a(n-1)}.$$

We thus see that macroscopic compressibility is expressed in terms of microscopic constants. Assuming the truth of the theory one substitutes the macroscopic value of the compressibility and calculates n. For the alkali halides the value of n is approximately 9.

The electrical theory of crystals has been employed to explain the various properties of crystals. Although further development of the theory awaits the solution of the problem of the structure of the atom, nevertheless important progress has been made in the reduction of macroscopic mechanical forces to electromagnetic forces acting on electrified particles.

From this point of view a mechanical explanation of the electromagnetic field in terms of an ether with elastic properties would involve a logical circle, for the elastic, mechanical properties are to be derived from the field. Indeed, the electrodynamic conception of nature requires the elimination of the ether. This task was accomplished by the special theory of relativity, to which we must now turn.

REFERENCES

ABRAHAM, M. Theorie der Elektrizität II. Leipzig und Berlin, 1920.
BORN, M. Der Aufbau der Materie. Berlin, 1920.
JOFFÉ, A. The Physics of Crystals. New York, 1928.
LORENTZ, H. A. The Theory of Electrons. Leipzig, 1916.
SWANN, W. F. G. Relativity and Electrodynamics. *Reviews of Modern Physics*, Vol. 2, p. 243, 1930.

CHAPTER X

THE ELECTRODYNAMIC CONCEPTION OF NATURE

B. THE SPECIAL THEORY OF RELATIVITY

1. THE CONTRADICTION IN CLASSICAL ELECTRODYNAMICS

THE electrodynamic conception of nature expresses the endeavor to interpret all physical phenomena in terms of electrodynamics. In the preceding chapter we have seen that a basis for such a program is furnished by microscopic electrodynamics, which is based upon the concepts of space, time, electric charge, electromagnetic field, etc. Now the electron theory of Lorentz presupposed the same kinematics as did classical dynamics—indeed, in all the preceding discussion we have presupposed the classical theories of space, time and motion. In the present chapter I shall explain how the requirements of electrodynamics brought about a reconstruction of the concepts of space, time, and kinematics.

Space is a form of relational structure and is bound to a frame of reference. In the classical kinematics different frames have the same space; length is invariant under a transformation of frames; the geometrical properties of figures are independent of the frame of reference—all of which is expressed by the statement that, in the classical theory, space is objective. Again, the time system of events is the same for all frames of reference; the time interval between events is an invariant under a transformation of frames. Thus classical kinematics presupposes a single space and a single time. The classical theory of space and time is summed up in the transformation from one system of reference to a system moving in the x direction with respect to the first. We have

$$x' = x - vt$$
$$y' = y$$
$$z' = z$$
$$t' = t.$$

These equations define a common space and time for systems in motion with respect to each other; if v is constant, the transformation is called Galilean. The classical transformation equations were accepted in the electron theory.

The electron theory of Lorentz further contained a relic of the mechanical conception of nature, the concept of a substantial ether. The equations of electrodynamics were expressed relative to a frame of reference which was at rest with respect to the ether. However, classical kinematics and the hypothesis of an ether proved to be inadequate for a satisfactory expression of the results of observation. In the special theory of relativity we achieve the elimination of the concept of the ether and a reconstruction of the theory of space and time. Definitions which are more explicit than those in the classical theory render time and space relative to a frame of reference. Thus classical kinematics becomes a first approximation to a new kinematics which has been molded by electrodynamics.

We have seen that in order to describe motion it is necessary to choose a system of reference, that is, a frame of reference together with rods and clocks. In classical dynamics one employs a set of privileged systems, of which an example is the system of fixed stars. From classical kinematics it follows that the equations of dynamics also hold with respect to systems having a uniform rectilinear motion with respect to the fixed stars. The proposition that a set of privileged systems in uniform rectilinear motion with respect to each other are equivalent is called the principle of relativity of uniform motion; the principle of relativity expresses the fact that a uniform motion of a system of reference can not be detected by mechanical experiments in the system.

The principle of relativity was abandoned in Lorentz's electron theory; in this theory the substantial ether is a uniquely privileged frame of reference. The kinematical quantities which enter into the equations of electrodynamics are supposed to be measured with respect to the ether. Thus the principle that the velocity of light is a constant independent of the motion of the source presupposes that the velocity is measured with respect to the ether. Now it was assumed that classical kinematics was applicable to the electron theory.

If, from the equations of electron theory expressed with respect to the ether, one deduces, by means of classical kinematics, the equations relative to a moving frame, one finds that the electromagnetic equations are changed in form. The equations of Lorentz are not invariant under the Galilean transformation; thus in Lorentz's electron theory the principle of relativity of uniform motion does not hold.

As an example of the change in electromagnetic phenomena to be expected with a change in the system of reference we may consider the principle that the velocity of light with respect to the ether is a constant regardless of the velocity of the source. By classical kinematics we can calculate the velocity of light in a system moving with velocity v relative to the ether. The result is that the velocity of a ray moving in the same direction as the moving system is $c - v$ in this system; for a ray traveling in the opposite direction the velocity is $c + v$. Hence in a system moving with respect to the ether the velocity of light depends upon the direction. This consequence of electrodynamics offers the possibility of determining by optical means the velocity of a system moving with respect to the ether. As early as 1881 the Michelson-Morley experiment had been performed to detect the velocity of the earth with respect to the ether; but the outcome of this experiment, and of other experiments designed for a similar purpose, was negative. As in dynamics, it has not been possible to devise an experiment which would enable one to detect a uniform motion of a system of reference. Thus negative experimental evidence offered a basis for extending the principle of relativity of uniform motion to electrodynamics as well as to dynamics.

Physical theory was thereby confronted by a contradiction. On the one hand, the electron theory of Lorentz had furnished a satisfactory basis for the explanation of electromagnetic and optical phenomena. Now electrodynamics may be represented by the principle of the constancy of the velocity of light—a principle which presupposed a fixed ether and thus contradicted the principle of relativity. On the other hand, the principle of relatively had been confirmed by negative experimental results.

Fitzgerald and Lorentz sought to explain the failure to obtain evidence of the motion of the earth relative to the ether by the hypothesis that bodies contract in the direction of motion; Lorentz deduced this assumption from the electrical theory of matter. Ritz and others entertained the hypothesis that the principle of the constancy of the velocity of light does not hold.

Einstein then proposed the radical solution of reconstructing the fundamental concepts of space, time and kinematics. In the special theory of relativity the concepts of space and time are so refashioned that electrodynamics conforms to the principle of relativity of uniform motion. In classical physics concepts were constructed with a view to the mechanical conception of nature; in the special theory of relativity electrodynamics becomes the legislative system of physical theory.

The problem of the special theory of relativity is the reconciliation of

(1) *The principles of kinematics*;
(2) *The principle of relativity of uniform motion*;
(3) *The principle of the constancy of the velocity of light with respect to at least one system.*

If one assumes classical kinematics, principles (2) and (3) are incompatible. Einstein assumed principles (2) and (3) and modified kinematics.

2. THE THEORY OF TIME

The basis of the kinematics of the theory of relativity is the application of electromagnetic processes in the definition of physical time.

We have previously seen that time is an order of events. Events may be simultaneous or related by succession: simultaneous events occur at the same time; events which are successive are separated by an interval of time. In order to measure an interval of time, or better, a duration, we select a standard interval as a unit and count the number of units which elapse during the given duration. This standard unit is ordinarily realized in the periodic motion of some mechanism, called a clock, such as the balance wheel of a watch or the earth

rotating on its axis. One may measure the time of an event by the observation that the event is simultaneous with the tick of a clock. However, the usual method of measuring time involves the observation of the space-time coincidences of points. Ordinarily the clock is provided with a pointer which moves over a scale division during each unit of time. One may then determine the duration of a given process by subtracting the reading of the pointer simultaneous with the beginning of the process, from the reading simultaneous with the end of the process. Thus the estimate of simultaneity is fundamental in the measurement of time.

We must distinguish between local time and extended time. The time system at a given place is based upon the simultaneity and succession of aspects given in immediate experience. The judgment that two events at a given place are simultaneous is a judgment of a specific kind of similarity of events; simultaneity at a place is one of the ultimate qualitative notions of physics. The physical time of a local event is the number of the scale division with which the pointer coincides simultaneously with the event. In the theory of relativity the theory of local time is the same as in classical kinematics.

The special contribution of the theory of relativity concerns the **extension of a time system throughout space.** It is evident that this is necessary for physics, for in order to measure the velocity of a body we measure the time at which it leaves a certain point and the time at which it arrives at some other point. We assume that the two places have the same time system and therefore measure the duration of the motion by subtracting the initial time from the final time. We thus meet the problem, how do we define a time system for the whole of space?

Before Einstein no one considered the problem critically. It was assumed that there was a single time system for the physical order. The concept of physical time, however, has no meaning unless it indicates how the times of events at different places are to be compared. In particular, it is necessary to have a definition of the simultaneity of events at different points. The classical kinematics did not have explicit defini-

tions for extended time. The practice of the classical physicist, however, is exemplified by the behavior of a student who is performing an elementary experiment in mechanics. He may measure time with a watch which he carries in his pocket; as he walks around his laboratory table his behavior indicates that he unreflectively assumes that the watch always gives the same physical time. Or he may look at a clock on the wall; the time of an event on the table is the time indicated by the hand of the clock. Thus the student tacitly assumes that the displacement of a watch is a method of extending time throughout a space, or he assumes that the time required for light to reach him from the clock on the wall is to be neglected. For a precise concept of time it is necessary to lay down an explicit definition.

In the electron theory the principle of the constancy of the velocity of light was an experimental law. In the theory of relativity the principle constitutes a definition of a time system.

3. THE POSTULATES OF THE THEORY OF RELATIVITY

In order to explain the theory of relativity we select a system of reference K in which the classical dynamics holds to the first approximation. The system consists of a frame of reference and a set of unit measuring rods and similarly constructed clocks. Let K' be any other system moving with uniform rectilinear motion relatively to K in the positive direction of the x axis. K' is provided with rods and clocks which agree with those of K when all are at rest in the same system. One now assumes two postulates: the first is the principle of relativity which states that the laws of physics are the same with respect to K and K'; the second is the principle of the **constancy of the velocity of light**, which states that the velocity of light is constant independently of the motion of the source in at least one system of reference. From the principle of relativity it follows that the principle is to hold with respect to all systems K and K'. Thus, regardless of the motion of the source, or uniform motion of the system of reference, the measured velocity of light is a constant. The velocity of light is the fundamental invariant of the special theory of relativity.

One may build up the kinematics of relativity in the following manner. Let us start with the system of reference K, in which there is a set of unit measuring rods. We assume on the basis of experience that the positional relations of the rods satisfy the postulates of Euclidean geometry. Thus we are able to specify the position of a point in terms of Cartesian coordinates x, y, z, which are the distances measured by a unit rod from the three mutually perpendicular planes of the frame of reference. We further assume that the system is provided with a set of periodic mechanisms which may be used as clocks.

I assume that the principle of the velocity of light is a definition of time. Let Δs be the distance passed over by a ray of light as measured by a unit rod, c the velocity of light, and Δt the time. Then

$$\Delta s = c\Delta t$$

defines Δt. We first define local time. Suppose that at a given point A a source of light emits a ray which travels a distance r, and is then reflected back to A. If t_A is the time of emission and t'_A is the time of return, then

$$t'_A - t_A = 2r/c.$$

We now choose r and c so that $t'_A - t_A = 1$; this defines the unit of time. Upon the arrival of the signal at t'_A another signal is emitted and returns at t''_A. Then $t''_A - t'_A = 1$. Thus a complete journey of the light from A to the distant point and back defines the unit of time. It is an experimentally verified hypothesis that the clocks of classical physics indicate the time at a point as thus defined. From the present point of view an experimental determination of the velocity of light is a proof of the applicability of our definition.

Let us now extend the time of A throughout space. At the time t_A we send a light ray to any point B. Then by definition

$$t_B = t_A + r/c.$$

If the light ray is reflected and returns to A at t'_A, then by definition

$$t'_A = t_B + r/c.$$

It follows that

$$t'_A - t_B = t_B - t_A.$$

It is an experimental fact that similar clocks at A and B can be regulated so that the indicated times are related by the preceding equation; such clocks are said to run synchronously. Two events at distant places are simultaneous if they occur at the same time as indicated by synchronous clocks at the respective places. Thus we have defined a time system for the system K. The position and time of an event relative to K are denoted by x, y, z, t.

Let K' be a system which is moving along the positive direction of the x axis with constant velocity v. The system contains a set of rods which agree with those in K when the rods of both systems are at rest in one system. We postulate that the rods in the moving system have unit length in that system and that the positional relations of the rods in K' satisfy the laws of Euclidean geometry. We are therefore able to specify the position of a point in terms of Cartesian coordinates x', y', z'.

I now define a time system in the moving system by the equation

$$\Delta s' = c\Delta t'.$$

It is an experimental hypothesis that this time can be indicated by synchronized clocks in K', which clocks can be synchronized with those of K when all are at rest in a single system. The position and time of an event relative to K' are denoted by x', y', z', t'.

4. THE LORENTZ TRANSFORMATION

We have next to answer this question: How are the space and time of configurations and processes relative to one system related to the space and time of the same configurations and processes relative to a system which is moving uniformly with respect to the first? The problem is to determine the equations which define the transformation from x, y, z, t to x', y', z', t'. If an event is described in terms of x, y, z, t, what is the description in terms of x', y', z', t'? The transformation equations can be obtained from the principle of the velocity of light which holds in both systems.

Suppose that at time $t = t' = 0$ the origins of the two systems coincide. At this moment a light wave is emitted from the common origin. Relative to K the points x, y, z in a wave front at the time t satisfy the equation

$$x^2 + y^2 + z^2 - c^2t^2 = 0.$$

This expresses the principle of the velocity of light. Relative to K' the points on the wave front satisfy the equation

$$x'^2 + y'^2 + z'^2 - c^2t'^2 = 0,$$

where x', y', z' are the coordinates of a point in a wave front at the time t'. The equations connecting x, y, z, t and x', y', z', t' can be obtained from the condition that if

$$x^2 + y^2 + z^2 - c^2t^2 = 0,$$

$$x'^2 + y'^2 + z'^2 - c^2t'^2 = 0.$$

Thus

$$x^2 + y^2 + z^2 - c^2t^2 = \lambda(x'^2 + y'^2 + z'^2 - c^2t'^2).$$

On the assumption of the homogeneity of space and time λ is a constant, and on the assumption of the symmetry of space with respect to direction $\lambda = 1$. The condition yields the equations:

$$\begin{aligned} x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \\ t' &= \gamma(t - vx/c^2), \end{aligned}$$

where

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

These equations are called the Lorentz transformation.

The transformation equations clearly reveal that time is relative to a system of reference. The equations are expressed so that upon coincidence of the origins the clock at the origin of K and the clock at the origin of K' read 0. For all other points and times, however, the reading of a clock in K differs

from that of the coincident clock in K'. Thus if the clock at $x = a$ in K reads 0, the coincident clock in K' reads

$$t' = \gamma(-va/c^2).$$

If the clock at $x = a$ in K reads t_1, the coincident clock in K' reads

$$t'_1 = \gamma(t_1 - va/c^2).$$

The concept of simultaneity is relative to the system of reference. Thus if relative to K an event at x_1 is simultaneous with an event at x_2, then the difference in time of the two events relative to K' is

$$\gamma\left(t - \frac{vx_1}{c^2}\right) - \gamma\left(t - \frac{vx_2}{c^2}\right) = \frac{\gamma v}{c^2}(x_2 - x_1).$$

Thus according to the special theory of relativity there is no single universal time system in which events have temporal position.

The transformation equations further reveal that space is relative to a system of reference. In order to measure the length of a rod at rest relative to the system of reference, one counts the number of times a unit rod can be laid off on the rod. The rod is at rest and hence the measuring rod can be directly applied to it. If the rod is in motion with respect to the system of reference we have two modes of determining the length. We may measure the length relative to a system in which the rod is at rest; this may be called the rest length of the rod. Or we may measure the length relative to the system with respect to which the rod is moving; this result may be called the length of the moving rod in the system at rest. In order to determine the latter length we mark on the frame of reference the points which coincide with the endpoints of the moving rod at the same time. Then the distance between these points on the frame can be measured directly and is the length of the moving rod in K. The transformation equations of the theory of relativity reveal that the length of a moving rod in the system at rest is less than the rest length in the moving system. Suppose, for example, that the endpoints of a rod in the x axis have x' coordinates, x'_1 and x'_2, respectively,

in the moving system K'. The length of the rod relative to K' is $x'_2 - x'_1$. Then the length of the rod relative to K is found from

$$x'_2 - x'_1 = \gamma (x_2 - vt) - \gamma (x_1 - vt),$$

$$x_2 - x_1 = \frac{x'_2 - x'_1}{\gamma} = \sqrt{1 - \frac{v^2}{c^2}} (x'_2 - x'_1),$$

$$l = \sqrt{1 - \frac{v^2}{c^2}} l'.$$

The length of the moving rod in the system at rest is less than the length of the rod in the moving system. Thus the relativity of time entrains the relativity of space to the system of reference.

The relativity of space and time is a consequence of the union of the finite velocity of light and the principle of relativity. If we allow c to become infinite, the Lorentz transformation reduces to the Galilean transformation.

From an abstract point of view it is possible to define a single time system by means of imaginary signals with an infinite velocity. However, assuming that the theory of relativity has been verified, there are no known natural mechanisms which would serve as clocks to measure absolute time. The theory of relativity is more than a theory based upon the limitations of our methods of measurement. According to the theory clocks in motion have characters which seem paradoxical to our inherited modes of thinking.

The principle of relativity implies the abandonment of the concept of a substantial ether in electrodynamics, for motion relative to an ether can not be measured and therefore has no physical significance. The relativity of time to a system of reference implies that one can not conceive of time as a single flowing stream. Similarly one can not conceive of space as a single substance.

In the special theory of relativity the transformation of the coordinates and the time on changing from one allowable system of reference to another is given by the Lorentz transformation. The Lorentz transformation defines a group under which the fundamental equations of electrodynamics are

invariant. Thus the outcome of the theory of relativity is that electrodynamics satisfies the principle of relativity. If a uniform velocity is impressed upon an allowable system of reference there is no change in electromagnetic phenomena within the system. However, since the equations of classical dynamics are not invariant under the Lorentz group, it was necessary to reconstruct dynamics in order to render it compatible with the principle of relativity. It is possible to build a relativistic dynamics upon the assumption that the principles of the conservation of momentum and energy are invariant with respect to the Lorentz transformation. From the conservation of momentum one obtains a definition of mass as a quantity which is a function of the absolute value of the velocity:

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}},$$

where m_0 is the mass relative to the system in which the body is at rest. In classical dynamics the mass of a body was assumed to be invariant; in the theory of relativity the physical quantity mass is relative. On the other hand, electric charge is invariant under the Lorentz transformation.

All classical physics is a first approximation to relativistic physics. Thus the reconstruction of physical theory, on the basis of a kinematics which was designed to render possible a satisfactory theory of electrodynamics, is a further indication of the regulative character of the electrodynamic conception of nature.

5. SPACE-TIME

The mathematical representation of the theory of relativity is greatly simplified by the employment of a four-dimensional space-time. Suppose that in a permissible system of reference the position of an event is determined by the Cartesian coordinates x, y, z and the time by t. The quantity

$$x^2 + y^2 + z^2 - c^2t^2$$

is an invariant with respect to the Lorentz transformation. That is, upon transformation to another permissible system

the foregoing function of the coordinates and the time remains unchanged. Let us now substitute for the time the imaginary quantity $u = ict$. Then

$$x^2 + y^2 + z^2 - c^2t^2$$

becomes $x^2 + y^2 + z^2 + u^2$. This suggests that x, y, z, u may be interpreted as the coordinates of a point in a four-dimensional manifold. Then $x^2 + y^2 + z^2 + u^2$ will be the four-dimensional distance of the point from the origin of coordinates. The set of numbers x, y, z, u will be said to determine a world point and the continuum of such sets will constitute the four-dimensional space-time world. The expression

$$(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 + (u_2 - u_1)^2$$

will be called the space-time separation of the points

$$(x_1, y_1, z_1, u_1) \quad \text{and} \quad (x_2, y_2, z_2, u_2).$$

The Lorentz transformation may be interpreted as an imaginary rotation of the space-time axes.

An event may be viewed as a point in a space-time manifold. The motion of a particle consists in a series of space-time events which form a space-time line, the world line of the particle. A particle moving with uniform velocity describes a straight world line; if the motion is accelerated the world line is curved.

In another mode of representation a world point is determined by x, y, z and ct. For purposes of illustration let us suppose that there is only one space coordinate x. Then we may construct a graphical representation of the space-time manifold: draw two lines at right angles to each other, and let points on the horizontal line represent values of x and points on the vertical line represent values of ct. A world point is determined by the coordinates x and ct. $s^2 = x^2 - c^2t^2$ is the square of the separation between the world point (x, ct) and the world point represented by the origin. The equation

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

is the equation of a hyperbola. Thus in accordance with our definition all events which are represented by points on the

hyperbola have the separation 1 from the event represented by the origin. The equation

$$x^2 - c^2t^2 = 0$$

is the equation of the asymptotes. Since the propagation of light in the x-direction satisfies the equation $x^2 - c^2t^2 = 0$ it follows that the asymptotes represent the world lines of a ray of light. The Lorentz transformation leaves $s^2 = x^2 - c^2t^2$ invariant, and is represented by a change to oblique axes.

Lines through the origin which represent the motion of light constitute a light cone of which the axis is the vertical line. For points within the cone the value of x is less than ct; hence lines within the cone which pass through the origin represent the motions of particles that are traveling with less than the velocity of light. The points within the cone have the property that by a proper choice of axes the separation between the event and the event at the origin can be made to consist of an interval of time only; by no choice of axes can the events be made simultaneous. Thus the points in the lower part of the cone represent events which are past with respect to the event represented by the origin. The points in the upper part of the cone represent events that are in the future with respect to the event represented by the origin. Thus we see that some events are definitely related as earlier and later.

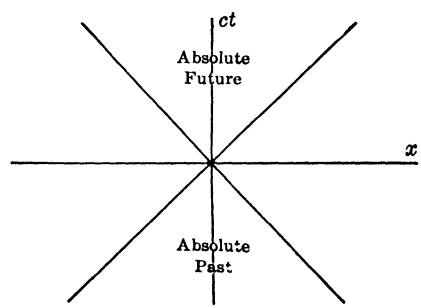


FIG. 3.

Again, the points in the intermediate region outside the cone represent events the separation of which from the event at the origin can be made spatial by a proper choice of axes. Events in the intermediate region may be either past or future with respect to the event represented by the origin.

The distinction between events in the absolute past and absolute future indicates that the distinction between time and space is not abolished in the theory of relativity. The

difference between time and space is symbolized by the occurrence of a minus sign in the expression for the separation of two events. The statement that the world is four-dimensional means that the values of four variables are required to determine an event. One may characterize the kinematics of the special theory of relativity by the statement that kinematics is reduced to geometry. This means that the relations between the space-time variables are the same in form as relations between geometrical variables. Thus the functional relations of the kinematics of the theory of relativity may be represented by points in a four-dimensional space-time diagram.

REFERENCES

EINSTEIN, A. Zur Elektrodynamik bewegter Körper. *Annalen der Physik*, 17 (1905).

—— Vier Vorlesungen über Relativitätstheorie. Braunschweig, 1923.

PAULI, W. JUN. Relativitätstheorie. Leipzig und Berlin, 1921.

THIRRING, H. Die Relativitätstheorie. *Ergebnisse der exakten Naturwissenschaften*, Vol. 1, 1922.

CHAPTER XI

THE GEOMETRICAL CONCEPTION OF NATURE

A. RIEMANNIAN GEOMETRY

1. INTRODUCTION

In the entire preceding discussion of physical theory we have presupposed that physical space is Euclidean. This means that the spatial properties of configurations of rigid bodies, which are measured by a rigid rod, are described by Euclidean geometry. We have now to consider the hypothesis that Euclidean geometry is only a first approximation.

Our initial physical problem was the description of the positional relations of rigid bodies independently of the dynamical, electrical and other specific physical properties of bodies. As a second approximation lengths must be corrected for changes in temperature, external forces, etc. The changes in geometrical properties with the temperature, external forces, etc., depend upon the material, and therefore one can correct for such changes. Hence the statement that space is Euclidean means that configurations of rigid bodies at a constant temperature, subject to no external forces, etc., are described by Euclidean geometry.

Thus we first explained the concepts of space and time and then introduced concepts which are especially characteristic of classical physics. Bodies in space and time were described in terms of mass, electric charge, electric field, etc. The definition of the specific dynamical and electrodynamical concepts presupposed the concepts of distance, interval of time, etc.

The special theory of relativity gave rise to the concept of a four-dimensional space-time in which physical events have spatial and temporal position. The functional relations

in kinematics were thereby given a form similar to that of the relations of geometry. If one employs the term geometry in the sense of analysis or abstract geometry one may say that in the special theory of relativity kinematics was reduced to geometry.

Like classical physics, the theory of relativity was based upon the assumption that one could ascribe a definite structure to space-time which is independent of specific physical characters and processes. One introduced the characteristic physical quantities by way of addition. In the general theory of relativity, however, Einstein has shown that it is desirable to adopt a variable space-time. The general metrical relations of rigid bodies and clocks are correlated with the gravitational field; hence the gravitational field can be expressed in terms of a variable structure of space-time. Now the structure of the field, and therefore that of space-time, depends upon the distribution of stress, momentum and energy of matter. This has suggested the program of interpreting all physical concepts in terms of geometrical concepts, so that we have the possibility of a geometrical conception of nature. The basis of this type of physical theory is a more general geometry than that of Euclid. I shall begin the discussion with an exposition of Riemannian geometry.

2. EUCLIDEAN SPACE

We shall initially presuppose a Euclidean space of three dimensions, by which we mean that there is a class of solid bodies which have certain positional properties. On one of these bodies we make two marks and define the segment between the two marks to be the standard which embodies the unit of distance. It is an aspect of the definition that the distance between the two marks remains the same upon displacement of the solid. Out of solid bodies one may construct rigid lines which extend between two points; a straight line is defined as the line along which the distance between any two of its points is a minimum. The distance, more precisely the measure of distance, between the endpoints of a straight line is called the length of the line. One can readily extend the

notion of length to a curved line. We also introduce the concept of angle between two lines. We construct rigid surfaces and regular solids. In terms of the synthetic method Euclidean space is then characterized by the following properties. In the first place, Euclidean space has a metrical structure: the metrical character of space depends upon the definition that the length of a line does not change upon displacement. The proposition that Euclidean space is metrical means that two lines are comparable in length. This property of lines is also assumed in Riemannian geometry; it does not hold in topological geometry.

In the second place, Euclidean space is isotropic and homogeneous. Isotropy means that a figure can be rotated about a fixed point and brought into coincidence with a congruent figure. Homogeneity means that a figure can be translated and brought into coincidence with a congruent figure. That is, the properties of Euclidean figures are invariant under rotation and translation.

In the third place, it is possible to construct similar figures in Euclidean space. This property is connected with the proposition that the sum of the angles of a triangle is equal to two right angles.

The properties of figures constructed out of rigid lines, surfaces and solids are described to the first approximation by Euclidean geometry.

In terms of the analytic method Euclidean space is characterized by the fact that one can construct in it a Cartesian coordinate system based upon rigid rods. The Cartesian system may be thought of as consisting of three sets of mutually perpendicular planes which are spaced with the aid of a rigid measuring rod. However, the fundamental place of the measuring rod is brought out more clearly if we think of the Cartesian system as a three-dimensional lattice built up of cubical frames constructed out of unit rods. We may select one point of the lattice as an origin; the coordinate axes consist of three mutually perpendicular lines extending from the origin; the axes determine three mutually perpendicular planes. Then every point of the lattice may be characterized by three numbers, the Cartesian coordinates of the point, which are the per-

pendicular distances of the point from the three coordinate planes. The Cartesian coordinates of any point in space may be determined from the coordinates of the nearest lattice point. Now the properties of the Cartesian lattice may be expressed analytically by means of the Pythagorean theorem. Consider two lattice points with coordinates (x_1, y_1, z_1) and (x_2, y_2, z_2); then the distance between these points is expressed by

$$s^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2.$$

It is an experimental fact that one can build a cubical lattice out of unit rods such that the Pythagorean theorem holds approximately. Analytic Euclidean geometry can be based upon the expression for the square of the distance.

3. GAUSS' THEORY OF SURFACES

On the assumption that three-dimensional space is Euclidean we now proceed to the theory of the space of surfaces.

Our first example of a surface is a plane. We postulate that lines in the plane are comparable with respect to length and hence the plane is in a metrical two-dimensional space, plane space. Plane space being homogeneous and isotropic, the relation of congruence holds between plane figures. Also it is possible to construct similar figures in the plane. Plane figures constructed out of rigid lines are described by Euclidean geometry. From the analytic point of view the plane is characterized by the fact that one can construct in it a two-dimensional rectangular lattice system. If (x_1, y_1) and (x_2, y_2) are the coordinates of two points in the plane, the distance between the points is given by

$$s^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2.$$

The analytic geometry of the plane may be based upon this expression for the square of the distance between two points.

We next consider a curved surface, like the surface of a sphere or an ellipsoid. Our fundamental presupposition is that lines in the surface are comparable with respect to length, and hence the space of the surface is a two-dimensional metrical space. We shall think of the lines in the surface as consisting of inextensible flexible cords, for it is an essential part of the

definition of length that a line can be bent without changing its length. Or we may think of a cord as constructed out of a number of small rigid rods. In any event, the standard of length in the three-dimensional Euclidean space can be employed to measure the length of a line in the surface.

It is necessary to give a precise definition of the curvature at a point. Consider on the surface any curve which passes through the given point. Let the tangent to the curve and the normal to the surface at the point determine a plane which intersects the surface in a curve called the normal section. The radius of this section is called the radius of normal curvature for the given direction. Now at every ordinary point of a surface there is a direction for which the radius of normal curvature is a maximum and a direction for which it is a minimum, and they are at right angles to each other. The maximum and minimum values of the radius of curvature are called the principal radii of curvature at the point. The reciprocal of the radius of curvature is called the curvature. We now define the Gaussian curvature of the surface at a point as the product of the principal curvatures:

$$K = \frac{1}{\rho_1 \rho_2}.$$

In a surface of variable curvature there is no relation of congruence between figures. A figure at one point of the surface can not be displaced in the surface and brought into coincidence with a figure at another point without distortion.

In a surface of constant curvature there is a relation of congruence between figures at different points. A plane is characterized by zero curvature; the surface of a sphere has a positive curvature; a saddle surface has a negative curvature.

On a curved surface there are no Euclidean straight lines. One defines the geodesic line between two points as the line whose length is an extremum, usually a minimum. One can then construct figures out of geodesic lines on curved surfaces. The curvature of a surface may then be characterized in the following way: in a region of zero curvature the sum of the angles of a triangle is two right angles, in a region of positive curvature the sum is greater than two right angles, and in a

region of negative curvature the sum is less than two right angles.

We shall now review the properties of curved surfaces in terms of the analytic method. In general one cannot construct a Cartesian coordinate system on a surface; in a relatively small area one can build a Cartesian lattice out of small rods, but upon extending the system one finds that the structure deviates more and more from the Cartesian form. Hence in order to describe the properties of a surface analytically we employ Gaussian, that is, curvilinear coordinates. We shall assume that we are studying a finite, but limited, region of a surface without properties that would cause difficulties for a formal theory.

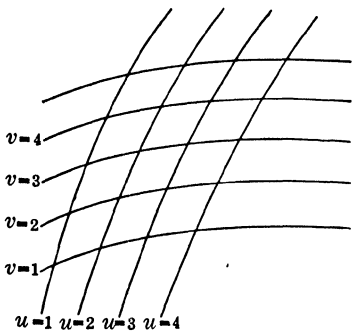


FIG. 4.

A Gaussian system of coordinates consists of two families of intersecting lines which cover the region. In order to identify these lines we label one family $u = 1$, $u = 2$, $u = 3$, etc., and the intersecting family $v = 1$, $v = 2$, $v = 3$, etc. Lines of the same family do not intersect, and no two lines of different families intersect more than once. The lines constitute a lattice the points of which are specified by coordinates u and v. The position of any point in the surface is given by the coordinates of the corner of the lattice element in which the point falls. For example, a point may have coordinates $u = 2$, $v = 3$. Now it is essential to emphasize that Gaussian coordinates are not distances which have been determined by a standard rod. Gaussian coordinates are identification numbers; if we are given only the differences between the Gaussian coordinates we can not calculate the distance between the points. In order to express the metrical properties of a surface we need additional concepts.

One method of expressing the metrical properties of a surface is based upon the fact that it is embedded in a three-dimensional Euclidean space. The position of a point on the surface can be specified by its Gaussian coordinates u, v. The

position can also be represented by Cartesian coordinates x, y, z in the Euclidean space. We have

$$x = f_1(u, v)$$
$$y = f_2(u, v)$$
$$z = f_3(u, v).$$

Now to the first approximation a small region of a curved surface has the properties of a plane; a short line element is to the first approximation a straight line. We can therefore express the length of such a line element in terms of the differences of the Cartesian coordinates of the endpoints in the three-dimensional Euclidean space. Let ds be the length of a small line element, and let dx, dy, dz be the differences between the coordinates of the endpoints. Then by the Pythagorean theorem

$$ds^2 = dx^2 + dy^2 + dz^2.$$

Now, making use of the fact that x, y and z are functions of the Gaussian coordinates u and v, we obtain an expression of the form,

$$ds^2 = g_{11}du^2 + g_{12}dudv + g_{21}dvdu + g_{22}dv^2.$$

This equation expresses the length of a line element relative to a unit rod, which is the standard of length on the surface as well as in the three-dimensional Euclidean space, in terms of the differences in the Gaussian coordinates of its endpoints. The formula for the length of the line element expresses the metrical geometry of the surface. The coefficients g_{11}, $g_{12} = g_{21}$, g_{22} are called the components of the fundamental metrical tensor. In general the g's are functions of the Gaussian coordinates and vary over the surface. If the g's are known, the metrical geometry of the surface is determined; one can then calculate the lengths of lines, the area of figures, etc.

In a plane we can construct a Cartesian coordinate system so that

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

We may also use curvilinear coordinates (u, v) in the plane and thus express ds^2 in terms of du and dv. However, we can transform the formula for the line element in curvilinear coor-

dinates into the Cartesian form. On a curved surface the formula for the line element is

$$ds^2 = g_{11}du^2 + 2g_{12}dudv + g_{22}dv^2.$$

In general we can not transform this into the Cartesian form. We can, however, find coordinates x, y, z which are functions of u and v, such that

$$ds^2 = dx^2 + dy^2 + dz^2.$$

This is the analytical expression of the fact that we have presupposed that the surface is embedded in a Euclidean space of three dimensions. This remark is important for the later discussion of curved spaces of three dimensions.

In the preceding discussion we have supposed that the curved surface is embedded in a three-dimensional Euclidean space. I shall now define the g's by an intrinsic method, that is, without referring to the third dimension.

We assume that the length of a line element is given by

$$ds^2 = g_{11}du^2 + 2g_{12}dudv + g_{22}dv^2.$$

Consider a point on the surface having coordinates u, v. To the first approximation a small region in the neighborhood of the point is plane. Hence we may consider the lattice element bounded by lines u and $u + du$, v and $v + dv$, as a plane parallelogram. Now one side of the lattice element extends from the point u, v to the point $u + du$, v. The difference in the u coordinates of the endpoints is du. We assume that this coordinate difference is directly proportional to the length of the line. Thus the length of one side of the lattice element is

$$k_1du.$$

One can measure the length of the side and thus calculate k_1. Similarly the length of the side of the lattice element extending from u, v to u, $v + dv$ is

$$k_2dv.$$

The angle between the coordinate lines is θ. Then by the Pythagorean theorem for oblique triangles,

$$ds^2 = k_1^2du^2 + 2k_1k_2 \cos\theta\, dudv + k_2^2dv^2.$$

Thus we see that $g_{11} = k_1^2$, or $\sqrt{g_{11}}$ is the ratio of distance, in the direction of a v coordinate line, to the increase in the u coordinate. Similarly $\sqrt{g_{22}}$ is the ratio of distance, in the direction of a u coordinate line, to the increase in the v coordinate. g_{12} involves the angle between the coordinate lines. Thus the g's of a given coordinate system can be determined by measurements made in the surface. Given a coordinate system of which the lines are numbered; at a given point one can measure the sides of a lattice with a unit rod and the angle between the lines and then calculate the values of the g's at the given point. For our purposes we need consider only the geometrical properties of a surface which are determinable from measurements in a surface.

We have thus seen that the metrical geometry on a surface is determined by the formula for the square of the length of a line element. The g's are called the components of the fundamental metrical tensor. The g's depend on the coordinate system and also on the character of the surface. A given surface allows only certain coordinate systems. One thus seeks a relation between the g's which holds for all systems on a given surface. Such a relation would constitute an invariant description of the surface, that is, it would be unchanged in form by a transformation of coordinates.

Now it is possible to construct a tensor R_{ijkl} out of the g's and their derivatives. From this one derives R_{ij}, and lastly one obtains an invariant R. Now $R = -2K$. One thus obtains the important result that the Gaussian curvature, which we defined with the aid of three-dimensional Euclidean space, can be expressed in terms of the g's. The Gaussian curvature is thus calculable from measurements made within the surface.

In order to obtain an invariant description of a surface we need merely recall that the curvature is describable in terms of K and hence in terms of the invariant R. $R = 0$ expresses a relation between the g's for coordinate systems in a plane. The plane is thus described by $R = 0$. Similarly we may describe surfaces of constant curvature for which R is different from zero.

A spherical surface is characterized by the invariant propo-

sition that the Gaussian curvature is constant and positive. The formula for the length of a line element on a spherical surface can be put in the form[1]

$$ds^2 = \frac{d\theta^2 + \sin^2\theta d\varphi^2}{k}.$$

The geometry on the sphere is a non-Euclidean geometry. The geodesic lines are great circles; there are no parallel lines. However, since the curvature is constant there are relations of congruence between spherical figures.

In what follows I shall initially assume that surfaces are embedded in a three-dimensional Euclidean space. The unit rod in Euclidean space is the standard of length on the surface. Relative to our three-dimensional Euclidean space there is a definite distinction between the surface of a sphere and a plane.

In describing a sphere it is convenient to think of it as having two poles and an equator. Let us now suppose that we have given a sphere of radius $\frac{1}{\sqrt{k}}$ and a plane tangent to it at one of its poles. Then, by projection from the center of the sphere, we can map half of the spherical surface upon the tangent plane. That is, from the center of the sphere one draws a straight line which intersects the surface of the sphere and the plane. The point in the plane then represents the corresponding point on the sphere. Thus lines and figures in the plane represent lines and figures on the sphere, for example, great circles on the sphere are represented by straight lines in the plane.

Let x, y be the Cartesian coordinates of a point in the plane. Then we may express the coordinates of a point on the sphere in terms of the coordinates of its representative point in the plane, and the length of a line element on the sphere is given by

$$ds^2 = \frac{dx^2 + dy^2 + k(xdy - ydx)^2}{[1 + k(x^2 + y^2)]^2}.$$

Now if the distance between two points in the plane is determined by $dx^2 + dy^2$, the distance between the corresponding

[1] v. LAUE, M. Die Relativitätstheorie, II, chap. IV.

points on the sphere is given by the preceding expression in dx and dy. Equal distances in the plane do not correspond to equal distances on the sphere.

We now make the geometry in the plane identical with the geometry on the sphere by the postulate that the length of a line element in the plane is also to be given by the preceding formula. That is, by definition, equal line elements on the sphere are mapped into equal line elements on the plane. We thereby obtain the elliptic geometry of the plane, which is thus seen to be identical with spherical geometry. The elliptic plane is characterized by a positive constant value of the Gaussian curvature.

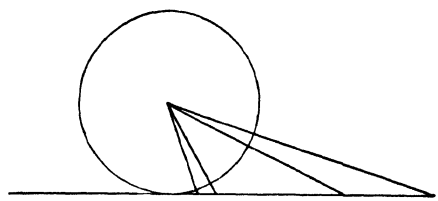


FIG. 5.

The preceding result may be restated with the aid of a distinction between Euclidean and non-Euclidean rods. A rod which can be employed to construct a Cartesian lattice system in three-dimensional space is a Euclidean rod; in the study of spherical geometry we employ a Euclidean rod. On account of the curvature of the spherical surface in the third dimension we obtain a spherical geometry, the propositions of which differ from those of plane Euclidean geometry.

In the elliptic geometry of the plane we employ a non-Euclidean rod. The divergence of elliptic geometry from plane Euclidean geometry arises from the non-Euclidean behavior of the unit rod: rods in the plane which are congruent from the non-Euclidean point of view are not congruent from the Euclidean point of view.

Suppose that an observer is confined to a surface of two dimensions and discovers that geometrical figures do not satisfy Euclidean geometry. On the assumption of a three-dimensional Euclidean space there are two explanations for the non-Euclidean geometry.

(1) *The observer employs a Euclidean rod, but the surface is curved in a three-dimensional Euclidean space.*

(2) *The two-dimensional surface is a plane, but the observer employs non-Euclidean rods.*

To us the distinction between these two cases is significant because we have an intuition of the distinction between two and three dimensions. However, an observer confined to a surface of two dimensions would know nothing about three-dimensional space. He would be able to deduce an invariant, which would enable him to describe the structure of the surface as defined by the relations between equal measuring rods. He could distinguish between surfaces of zero, positive and negative values of the invariant. Our intuitive picture of the distinction between plane and curved would be unknown to him. If from our three-dimensional point of view we saw that the surface is plane, we could call it curved because the geometry is the same as on a curved surface.

4. RIEMANNIAN GEOMETRY

Let us now consider the curvature of a space of three dimensions. We start out with a set of congruent unit rods. In order to employ the analytic method we construct a coordinate system x_1, x_2, x_3, just as on a surface we constructed Gaussian coordinates u and v. If it is possible to construct a Cartesian lattice out of unit rods the space is Euclidean. Then

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2.$$

Suppose we have to introduce coordinates such that

$$ds^2 = g_{11}dx_1^2 + g_{22}dx_2^2 + g_{33}dx_3^2 + 2g_{12}dx_1dx_2 + 2g_{13}dx_1dx_3 + 2g_{23}dx_2dx_3.$$

In expressing the element of length in this form one presupposes that a small region is Euclidean to the first approximation. ds is an invariant which is embodied in a standard measuring rod.

Hypothetically we may interpret the deviation from Euclidean geometry in two ways:

(1) There is a Euclidean space of six dimensions in which the unit rods can be combined to form a Cartesian lattice. The unit rods in the three-dimensional space are Euclidean.

A Cartesian lattice can not be constructed in three dimensions because the three-dimensional space is curved in the higher dimensions.

(2) There are only three dimensions. The geometry is non-Euclidean because the measuring rods do not fit into a Cartesian lattice. Thus the unit rods are non-Euclidean from the standpoint of rods in a six-dimensional Euclidean space. The term curvature can be used only in a figurative sense. However, if

$$ds^2 = \Sigma g_{ik} dx_i dx_k,$$

where $i, k = 1, 2, 3$, it is possible to find six variables $y_1 \ldots y_6$ which are functions of x_1, x_2, x_3, such that

$$ds^2 = dy_1^2 + dy_2^2 + \ldots dy_6^2.$$

Thus although only three dimensions exist one can employ the formal theory for a Euclidean space of six dimensions. One can express the length of a line element as if the three-dimensional space were embedded in a six-dimensional space. We do not know if six dimensions exist, and even if we did we could not tell from measurements within the three-dimensional space whether it is really curved in the higher dimensions. It seems clear that for empirical reasons one should accept the second interpretation.

The differential geometry of curved surfaces was initiated by Gauss in 1829. He defined the Gaussian measure of curvature; he introduced curvilinear coordinates in the surface and expressed the curvature in terms of coefficients determinable from measurements in the surface. However, Gauss' investigations were restricted to two-dimensional space.

Riemann extended the method to n dimensions. Given a finite, but limited, region of a space of n dimensions. We construct an n-dimensional lattice which defines Gaussian coordinates $x_1 \ldots x_n$ in the region. The fundamental assumption is that the space is metrical, that is, the lengths of lines are comparable. As a standard of length we have a rigid rod, or better, an inextensible flexible cord. Another fundamental assumption is that a small region of the space is Euclidean to the first approximation. Thus in an infinitesimal region the

Pythagorean theorem holds. and the metric is defined by the formula

$$ds^2 = \Sigma g_{ik} dx_i dx_k,$$

where i and k each take on the values 1 to n, and then the products are summed. ds is an invariant and $ds = 1$ is physically embodied in the standard of length. The g's are the components of the fundamental metrical tensor. One can derive a tensor whose components are symbolized by R_{ijkl}, and which is a function of the g's and their derivatives; this tensor is called the Riemann-Christoffel tensor. One can further derive a tensor R_{ik}.

Euclidean space is determined by the condition that

$$R_{ijkl} = 0.$$

This is the condition which must be satisfied by the g's of all coordinate systems which can be introduced into Euclidean space. If $R_{ijkl} = 0$, then the formula for ds can be transformed into $ds^2 = dx_1^2 + dx_2^2 + \ldots dx_n^2$. In a Euclidean space one can construct a Cartesian coordinate system.

If $R_{ijkl} \neq 0$ the space may be said to be curved in at least a figurative sense. It is then possible to find $\frac{n(n+1)}{2}$ variables y_1, y_2, etc., etc., which are functions of the $x_1 \ldots x_n$ so that

$$ds^2 = dy_1^2 + dy_2^2 + \ldots \text{ to } \frac{n(n+1)}{2} \text{ terms.}$$

The analysis is as if the n-dimensional non-Euclidean space were curved in a Euclidean space of $\frac{n(n+1)}{2}$ dimensions.

5. THE PARALLEL DISPLACEMENT OF A VECTOR

The character of Riemannian space can be rendered more intuitive with the aid of the concept of parallel displacement of a vector. I here use the term vector in the sense of a directed segment of a straight line, and I assume that we understand what it means to displace a vector parallel to itself in Euclidean space. The direction of a line at a point is determined by a vector which is tangent to the line at that point. Now a Euclid-

ean straight line has the property that the direction vector at any point on the line can be obtained from that at a neighboring point by parallel displacement; thus the straight line has the same direction at all points. If a vector is displaced parallel to itself around a closed path in Euclidean space it returns to itself. If a vector is transferred from one point to another by parallel displacement the result is independent of the path.

Let us now characterize a curved surface, for example, the surface of a sphere. An infinitesimal region of the surface is plane. We can then describe the curvature of the surface in terms of the parallel displacement of an infinitesimal vector by infinitesimal steps at a time.

A geodesic line is defined by the proposition that the direction vector at a point is related to the one at an infinitely near point by parallel displacement. Suppose that a vector is displaced parallel to itself by infinitesimal steps around a closed circuit. Then on a curved surface the final position of the vector will not coincide with the initial position. For example, suppose that a vector is given at the north pole of the earth. It is displaced parallel to itself along a meridian down to the equator, it is then displaced parallel to itself along the equator for 90 degrees, and is then displaced parallel to itself along a meridian back to the north pole. The final direction of the vector will differ from the initial direction by 90 degrees. Again, if a vector is carried from one point to another by parallel displacement, the result will depend upon the path. Thus the curvature of a surface is indicated by the change in an infinitesimal vector upon parallel displacement by infinitesimal steps. The length of the vector is an invariant, but the direction is changed. In general, a Riemannian space is characterized by the change in an infinitesimal vector upon parallel displacement. Given a coordinate system $x_1 \ldots x_n$. If A^h represents the components of a vector, the change in the vector upon parallel displacement around a circuit is represented by

$$\Delta A^h = \tfrac{1}{2} R^h_{ijk} A^i dS^{jk}.$$

In Euclidean space

$$R^h_{ijk} = 0.$$

Hence $\Delta A^h = 0$, that is, there is no change in the vector.

REFERENCES

BORN, M. Die Relativitätstheorie Einsteins. Berlin, 1922.
EDDINGTON, A. S. The Mathematical Theory of Relativity. Cambridge, 1923.
EINSTEIN, A. Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, Vol. 49 (1916).
—— Vier Vorlesungen über Relativitätstheorie. Braunschweig, 1922.
EISENHART, L. Differential Geometry. Boston, 1909.
v. LAUE, M. Die Relativitätstheorie, II. Braunschweig, 1923.
PAULI, W., Jr. Relativitätstheorie. Leipzig und Berlin, 1921.
WEYL, H. Raum-Zeit-Materie. Berlin, 1923.

Note: The author has also explained the concept of curvature in a paper, "World Geometry," in *The Monist*, October, 1931.

CHAPTER XII

THE GEOMETRICAL CONCEPTION OF NATURE

B. THE THEORY OF GRAVITATION

1. ACCELERATED FRAMES OF REFERENCE

THE geometrical conception of nature is the outcome of the endeavor to reconstruct physical theory so that it will satisfy a general principle of relativity.

Classical dynamics presupposes a set of priviliged systems of reference. A privileged frame which is employed in practice is one whose origin is at the center of mass of the solar system and whose axes are oriented with respect to the fixed stars. The equations of dynamics also hold with respect to frames in uniform rectilinear motion with respect to the foregoing privileged frame. The special theory of relativity presupposes the same privileged frames as does the classical dynamics. The proposition that the laws of physics have the same form with respect to a set of privileged frames is called the special principle of relativity.

Classical dynamics and the special theory of relativity did not admit the use of accelerated frames of reference. This may be illustrated by the following example from dynamics. Suppose that a car is moving with constant velocity relative to the surface of the earth; then mechanical phenomena within the car are described with respect to the car approximately by classical dynamics. For example, suppose that a body of negligible size is attached to the lower end of a spring balance which is attached to the roof of the car. The body will be in equilibrium in a position directly below the point of support of the spring. Now the condition for the equilibrium of a particle is that the resultant force be zero. Neglecting the resistance of the air, the only forces acting on the body are its weight

which acts downwards, and the tension in the spring which acts upwards. The spring balance will register a force equal to the weight of the body. Hence the resultant force on the body is zero, and thus the principle of equilibrium of classical dynamics is satisfied.

Let us now accelerate the car with respect to the earth. The body will swing backwards and the position of equilibrium will be such that the axis of the spring is inclined at a definite angle to the vertical. The force indicated by the balance will be greater than before and is inclined at an angle with the vertical. The resultant of the force exerted by the balance and the weight is a force whose direction is in the direction of the acceleration of the car. Now with respect to the car the acceleration of the body is zero; hence relative to the car the action of a force does not produce an acceleration. But this experimental result is contrary to classical dynamics. The explanation is that a frame which is accelerated with respect to a privileged frame is not allowable. However, sometimes it is convenient to employ frames that are admittedly accelerated. One then preserves the laws of motion by adding fictitious forces, called inertial forces. Thus in the preceding illustration we may think of the resultant force as balanced by an inertial force which acts in the direction opposite to that of the acceleration of the car. The addition of the inertial force rescues the principle that the condition of equilibrium is that the resultant force be zero. However, the inertial force does not satisfy the princple that action and reaction are equal and opposite. Hence accelerated frames can not be employed systematically in classical dynamics. The principle of relativity holds only for uniform rectilinear motion.

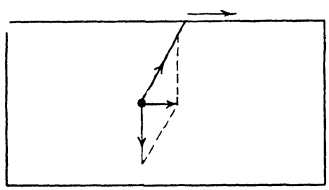


FIG. 6.

2. THE PRINCIPLE OF EQUIVALENCE

Einstein found a basis for a general theory of relativity in the law that all bodies have the same acceleration in a gravitational field. This may be interpreted in the proposition that

inertial mass and gravitational mass are proportional, and therefore essentially the same. Let m_i represent the inertial mass, m_g the gravitational mass, and G the absolute value of a gravitational field. The equation of motion is

$$F = m_i \cdot a.$$

The concept of gravitational field is defined so that the force due to the field is

$$F = m_g G.$$

Hence

$$m_g G = m_i a$$

$$a = (m_g/m_i)G.$$

It is clear that if a is the same for all bodies in a constant field, m_g/m_i is constant. Ordinarily in classical dynamics one symbolizes both quantities by the single letter m, which we have employed for inertial mass.

In a constant electric field the acceleration of a body depends upon its mass and electric charge, which are not proportional. Hence the acceleration of all bodies is not the same in an electric field. The gravitational field thus has a unique character which is not shared by other fields of force which are real from the point of view of classical dynamics.

The nature of gravitation is significant for a general theory of relativity because gravitational fields are similar to the fictitious fields relative to accelerated frames of reference. Let us suppose that in a region free from gravitation there is a laboratory at rest with respect to an inertial, that is, privileged, frame K. A body acted upon by no forces will remain at rest or in uniform motion with respect to K. Suppose that the laboratory is accelerated upwards with an acceleration of 980 centimeters per second per second; then with respect to the laboratory all bodies will experience the same downward acceleration of 980 cm./sec.^{2} Relative to the laboratory all bodies are in an inertial field of force which arises from the acceleration of the frame of reference and acts like a gravitational field.

The foregoing illustration is an instance of a general principle, the principle of **equivalence**, which states that the effect upon phenomena of an acceleration of the frame of reference

is indistinguishable from that of a gravitational field. The principle of equivalence thus denies the possibility of detecting an acceleration of a frame of reference and hence the inertial frame is deprived of its privileged position. Any frame may be chosen as at rest, but relative to some frames there will be gravitational fields.

We have seen that a gravitational field can be induced by the acceleration of a frame of reference. Conversely, a homogeneous gravitational field can be transformed away. Natural gravitational fields, however, have the property that the field can not be transformed away throughout a finite region. It is, therefore, necessary to express the principle of equivalence in the following way: In an infinitesimal space-time region, in which the gravitational field may be viewed as homogeneous and constant, it is possible to introduce a local system of reference relative to which there is no gravitational field. Thus in a region near the surface of the earth the frame of such a local system would be constituted by a freely falling box. That the principle of equivalence holds for mechanical phenomena is verified by the proportionality between mass and weight. The extension of the principle to all physical phenomena, such as electromagnetic phenomena, is an hypothesis to be confirmed by experiment.

The principle of equivalence expresses a relation between gravitation and geometry.

3. VARIABLE SPACE-TIME

The exposition thus far has presupposed classical kinematics. Classical kinematics presupposes that relative to a given system geometrical configurations have the same properties whether at rest or in motion; the geometrical properties of figures are invariant under transformations in process, such as translation and rotation. The fundamental assumption is that length is an invariant. One may express the invariance of geometrical properties by the concept of a single space, which in classical theory was endowed with Euclidean properties. Now the mind has a natural tendency to hypostatize invariants; thus Newton conceived of this single Euclidean space as an

absolute reality which serves as a frame of reference for motion. The acceleration of a body is the same relative to all inertial frames; hence acceleration reveals motion relative to absolute space.

The concept of a single space was abandoned in the special theory of relativity. The geometrical properties of configurations of rigid bodies are relative to a frame of reference. We have described space as the totality of possibilities of relative position of things and phenomena. Hence, since the positional relations of bodies are relative we must conclude that space is attached to a frame of reference.

The special theory of relativity introduced the concept of a single four-dimensional space-time. The separation between two events is an invariant with respect to a Lorentz transformation. The structure of space-time is defined in terms of the relations between rigid rods and by the principle of the constancy of the velocity of light. The structure is described by the formula for the separation between two events,

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2dt^2.$$

This expresses the fact that space as determined by rigid bodies at rest relative to a frame is Euclidean, and that time relative to a frame is defined by the principle of the velocity of light. The special theory of relativity retained a relic of the absolute theories of space and time in the assumption that space-time has a single type of structure. In connection with this absolute structure there is a set of privileged frames for the description of physical phenomena. As we shall see, the unique structure of space-time is abandoned in the reconstruction of physical theory with a view to a general theory of relativity.

We now return to the discussion of the **gravitational field.** By means of arguments based upon the special theory of relativity Einstein concludes that solid bodies and clocks are influenced by a gravitational field. Hence we can not construct a system of reference consisting of a Cartesian lattice and normal clocks in a finite region containing a gravitational field. Therefore we introduce Gaussian coordinates x_1, x_2, x_3, x_4; the system of reference may be realized by a three-dimensional lattice of flexible cords at the points of which are located

mechanisms with recurrent states. It is assumed that analytically space-time has the characteristics of a Riemannian continuum. Then the metric is determined by

$$ds^2 = \Sigma g_{ik} dx_i dx_k,$$

where ds is the invariant separation between two events.

The physical significance of ds is defined with the aid of the principle of equivalence. It is assumed that in an infinitesimal space-time region a local system of reference can be constructed relative to which the special theory of relativity holds. Then

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2dt^2.$$

where dx, dy, dz are the differences in the Cartesian coordinates of two events and dt is the difference in time as determined by clocks synchronized by the velocity of light. Thus ds can be determined by measurements made with rigid rods and normal clocks in a local system. If we transform to Gaussian coordinates the formula for ds^2 becomes

$$ds^2 = \Sigma g_{ik} dx_i dx_k.$$

ds is invariant; it is unchanged by the transformation. The unit of separation ds between events corresponds to the unit rod or measuring cord in Riemannian geometry. The invariant separation, however, is composed of space and time and therefore can not be visualized as easily as can the invariant length.

If a region is referred to a system of reference relative to which the special theory of relativity holds, then,

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2dt^2.$$

There is no gravitational field relative to the given system of reference. If one introduces Gaussian coordinates x_1, x_2, x_3, x_4, the formula for the invariant becomes

$$ds^2 = \Sigma g_{ik} dx_i dx_k.$$

The general transformation includes the change to systems of reference which are accelerated with respect to the original system. By the principle of equivalence the transformation to Gaussian coordinates in general induces a gravitational field. The g_{ik}'s determine the geometry of the region and likewise the gravitational field. Conversely, suppose we have

given a region in which a Gaussian system of coordinates has been constructed so that

$$ds^2 = \Sigma g_{ik} dx_i dx_k.$$

The g's describe the geometry and also the gravitational field. In the case of natural gravitational fields like that of the earth, it is not possible to transform to a system of coordinates relative to which

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2 dt^2$$

holds at all points. However, one can make the transformation for a given point. This means that in an infinitesimal space-time region one can construct a system of reference relative to which there is no gravitational field.

4. THE LAW OF GRAVITATION

We have seen that gravitational fields can be induced by transformations of coordinates. Now in physics one is interested in the natural gravitational fields which are excited by matter. In the geometrical theory of gravitation the field is determined by the metrical structure of the region. The system of coordinates in a space-time region is specified by the formula for the invariant ds in terms of the coordinates. The g's are components of the fundamental metrical tensor and likewise potentials of the gravitational field. The metrical structure of space-time is defined by the sets of g's which characterize the various systems of reference which can be introduced into the region. The law of gravitation expresses the relations which must be satisfied by the g's. In empty space the law is

$$R_{ik} = 0.$$

The general law, which holds for points within matter, is

$$R_{ik} + \lambda g_{ik} = -\kappa(T_{ik} - \tfrac{1}{2} g_{ik} T).$$

T_{ik} is the stress-momentum-energy tensor of matter. The equations of the gravitational field are tensor equations; they are covariant with respect to all transformations of coordinates which leave $ds^2 = \Sigma g_{ik} dx_i dx_k$ invariant.

The fundamental law of motion is the principle that the world line of a particle or ray of light is a geodesic line; this is

a generalization of Newton's principle of inertia. As a first approximation Einstein's law of motion yields the classical equation of motion for a particle in a gravitational field.

5. GEOMETRICAL PHYSICS

The geometrical theory of gravitation characterizes gravitational force in terms of a curvature of space-time. We mean by the curvature of space that unit measuring rods will not serve to build a Cartesian lattice system. The curvature of space-time means that the separation of two world points can not be expressed as

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2dt^2,$$

where dx, dy, dz, dt are determined by unit rods, and clocks which are regulated by the principle of the constancy of the velocity of light. In curved space-time one introduces a Gaussian coordinate system which is characterized by the components of the fundamental metrical tensor, g_{11}, g_{12}, etc.

In the Newtonian theory the motion of a body is determined by the principle of inertia and a force of gravitation; the force explains the departure of the motion from that in a straight line with constant velocity. In the geometrical theory the motion is such that the world line is a geodesic in space-time. The geodesic is determined by the g's, and thus the g's play the rôle of potentials of a force. The geometry of a space-time region is determinable from space-time coincidences of all sorts of bodies. Thus the geometrical theory of gravitation defines it in terms which are independent of the special properties of bodies.

Now, assuming that we have a concept of matter, the geometry of a region is found to be determined by the distribution of matter. The geometry is determinable from the observation of coincidences; hence the law of gravitation may be transformed into a definition of matter. It is then an experimental fact that what we have previously called matter satisfies this definition. The transformation of the experimentally verified law of gravitation into a definition has been made by Eddington. Einstein has expressed this result in a popular fashion by the statement that space is eating up matter.

The g's which define the geometry of a space-time region are the potentials of the gravitational field. The thought occurs that the coefficients of a still more general geometry might serve to characterize the electromagnetic field.

The first proposal in this direction was made by Weyl. In Riemannian geometry the length of a line is invariant; Weyl invented a geometry in which length depends upon position. He then correlated the potentials of the electromagnetic field with the coefficients which determine change in length. This was the beginning of the program for a unitary geometrical theory of physics.

Inasmuch as Weyl's theory leads to consequences which do not agree with experience, a distinction has therefore been drawn between natural geometry and world geometry. Natural geometry is descriptive of the behavior of measuring rods and clocks; world geometry is a graphical representation of physical theory. Eddington has also developed a system of world geometry. However, since world geometry is merely of mathematical interest I shall not expound it in further detail.

Einstein has continued the search for a unitary theory which would enable one to express electromagnetic phenomena in terms of the behavior of real rods and clocks, but the development has not yet been brought to a conclusion.

The conception of a geometrical system of physics was envisaged by Descartes. In his analysis of the concept of matter he concluded that the essence of matter is extension, for only of extension did he have a clear and distinct idea. Descartes thus discovered the epistemological basis for a geometrical theory of physics. But the accomplishment of this goal required the construction of a space-time geometry and the invention of several branches of mathematics, and so physics was developed in terms of the physical quantities mass and force by Newton and his followers. In recent years, however, the mathematical technique has been developed which offers hope to the prospect of constructing a geometrical system.

The geometrical conception of nature is based upon the notion of space-time coincidence—a concept which is subject to criticism in quantum mechanics.

PART III

THE QUANTUM THEORY

CHAPTER XIII

THE CLASSICAL QUANTUM THEORY

1. INTRODUCTION

PHYSICAL theory originates in the quantitative representation of the macroscopic physical order. We have seen, however, that in order to obtain a unitary theory recourse is had to microscopic concepts. We have a first approximation to a microscopic theory of matter in the kinetic theory of matter; an even more detailed microscopic theory which is based on electrodynamics is the electron theory of Lorentz; the culmination of a microscopic theory of matter is promised by a theory of the structure of the atom. In this chapter I resume the exposition of microscopic electrodynamics with an introduction to the quantum theory of atomic structure and radiation.

By way of introduction to the radical concepts of the quantum theory I shall first review the general character of physical theory as thus far set forth. The entire preceding discussion has presupposed the concept of a physical **order** existing in space and time. Space and time have been represented as continuous manifolds. Bodies in space and time were characterized in terms of the concepts of particle and of continuous medium. Physical processes have been described in terms of the motion of particles and the propagation of a state of vibration by wave motion. The fundamental principles have been stated as differential equations, which expresses the assumption of the continuity of physical processes.

In the mechanical conception of nature the fundamental physical property of particles and media was represented by the concept of mass; mechanical force was conceived of as action at a distance or action by contact. In the electrodynamic conception of nature the fundamental concepts were those of electric charge and electromagnetic field. In the geometrical

conception of nature gravitation was interpreted in terms of metrical concepts, and the program calls for a similar interpretation of electromagnetism. However, the mechanical conception of nature no longer guides the program of detailed physical theory, and the geometrical theory presents an abstract ideal which attracts only a few. The active frontier of contemporary physics is the electrodynamic conception of nature. The search for a dynamics and electrodynamics of microscopic electrified particles has led to the quantum theory.

The quantum theory has introduced radically new concepts into physical theory. In opposition to the previously discussed theories, the quantum theory assumes **discontinuity** in elementary processes. This has led to the conception of the inadequacy of our classical concepts of space, time and causality for the description of microscopic processes. The concepts of particle and wave motion are limited in their application to elements of the physical order.

The starting point of the electrodynamic conception of nature was the formulation of the laws of the electromagnetic field. From the equations of the field Maxwell deduced an equation which represents the propagation of an electromagnetic vibration and formed the basis of the hypothesis that light consists of electromagnetic vibrations. Upon the failure to find a mechanical explanation of the electromagnetic field it acquired the status of a fundamental reality. The concept of a substantial ether as the bearer of the field was eliminated by the special theory of relativity.

The development of electromagnetic theory was paralleled by the experimental investigations which led to an atomistic theory of electricity. The laws of electrolysis, the discharge of electricity through gases, the phenomena of radio-activity, were interpreted in terms of particles carrying atomic electric charges. Thus the hypothesis was at hand that material bodies contain electrified particles, and that the motions of the particles are the source of electromagnetic waves, for example, the waves which constitute light. In the electron theory the concept of electrified particle served as a basis for a theory of electromagnetic phenomena in material media. There arose further a detailed theory of the atom as a system of positively

and negatively charged particles. Although in the development of the theory of the structure of the atom the chemical properties as exhibited in the periodic table have furnished important clues, the principal method of attack has been the spectroscopic study of the light emitted by the atom. Thus from the program of an electrodynamic conception of nature arises the problem of the interaction between **matter** and **radiation.** Matter consists of molecules and atoms which are built out of positively and negatively charged particles; radiation consists in the propagation of a state of vibration of an electromagnetic field. Thus we have electric charges and electromagnetic fields. The problem of the interaction between matter and radiation thus involves the constitution of the atom, the nature of the electromagnetic field, and the mode of connection between them.

The first theories of the structure of the atom and of radiation were based upon the classical concepts of space, time, motion, electric charge, electromagnetic field. As we have seen, considerable success was attained in the electron theory; but this classical theory proved to be inadequate. The adequate representation of the nature of atoms and of their interaction with radiation has led to the development of a quantum mechanics for the motion of electrified particles and a quantum electrodynamics. However, the final quantum theory was preceded by a preliminary theory, the classical quantum theory, which is the subject matter of the present chapter.

2. THE QUANTUM HYPOTHESIS

The quantum theory was first proposed by Planck in 1900 in a theoretical discussion of the distribution of energy in the spectrum of black body radiation.

In the theory of light it is assumed that a luminous body is the origin of a periodic disturbance which is propagated through space to other bodies. Similarly a hot body emits temperature radiation which is detected by its heating effect. Temperature radiation can be reflected, refracted, and polarized like visible light; we assume that temperature radiation consists of electromagnetic waves of wave length longer than that

of visible light. A space enclosed by solid walls is called a hohlraum; the material of the walls emits radiation into the hohlraum and absorbs incident radiation. When the radiation within the hohlraum is in equilibrium with the material it is called black body radiation. Now if a beam of black body radiation is allowed to pass through a prism it will be spread out into a continuous spectrum. By means of sensitive instruments for detecting heat one can determine the density of radiant energy whose frequency falls within assignable limits. Theoretical investigations of Kirchhoff revealed that the density of energy is a function of the temperature and frequency. During the last decade of the nineteenth century theoretical investigations were carried on for the purpose of finding the form of the function, but laws deduced from classical dynamics and electrodynamics were in conflict with the experimental results. Then Planck found it possible to deduce the experimental law by assuming the existence of discrete quanta of energy and discontinuity in microscopic processes.

It was an hypothesis of electrodynamics that electromagnetic radiation is emitted and absorbed by electrified particles which vibrate with the same frequencies as the radiation. Planck assumed that the source of temperature radiation consists of a set of harmonic oscillators, each of which consists of an electron which vibrates with simple harmonic motion of a charactertistic frequency relative to an equal positive charge at the position of equilibrium. According to classical dynamics an oscillator could vibrate with a continuously varying energy. Planck's quantum hypothesis was that the energy of an oscillator is always an integral multiple of a discrete quantum which is proportional to the frequency. That is, the oscillator may vibrate with values of the energy, 0, ϵ, 2ϵ, 3ϵ, etc., where $\epsilon = h\nu$. Thus the oscillator may exist in one of a discrete series of energy states determined by

$$E = nh\nu.$$

The emission or absorption of energy by an oscillator takes place discontinuously, by discrete steps. Thus h was introduced hypothetically as a microscopic constant, and its value was first calculated from the fact that the form of the experi-

mental curve was expressed by the radiation law which Planck deduced with the aid of the hypothesis of quanta.

The Planck oscillator was the first example of a microscopic system subject to a quantum dynamics. In 1907 Einstein applied the quantum theory to the vibrations which constitute the heat content of a solid. Whereas, according to the classical kinetic theory, the specific heat should be independent of the temperature, experiment shows that the specific heat decreases with the temperature. But the assumption that the vibrations of a solid are subject to quantum dynamics yielded an approximate agreement with experiment.

The molecules of a diatomic gas have rotational energy which contributes to the specific heat. The extension of the quantum hypothesis to rotation requires a discrete series of rotational energy states and a consequent decrease in the specific heat with the temperature. This has been verified for hydrogen.

In its original form the hypothesis of quantum dynamics assumed the existence of quanta of energy. Planck then gave a more general formulation in terms of atomicity of action. Action is defined by

$$\int 2T dt,$$

that is, action is the integral of twice the kinetic energy with respect to the time.

In order to state the more general form of the quantum hypothesis we use the variables q and p which occur in the Hamiltonian equations of motion. For a system of one degree of freedom q is the generalized coordinate which specifies the position of the system; for example, q may be the linear coordinate of an electron vibrating along a straight line, or q may be the angle which a line in a rotating molecule makes with a fixed line. p is the generalized momentum which is correlated with q; thus in the case of the linear oscillator p is linear momentum $m\dot{q}$; in the case of a rotating molecule p is angular momentum. The values of the variables q and p determine the momentary phase of the system, that is, the instantaneous configuration and state of motion.

In order to study the motion of a system it is convenient

to represent it graphically by the motion of a point in a phase space. We assume that the Cartesian coordinates of a point in a plane are equal to a pair of values of q and p, respectively; the momentary phase of a system is thus represented by a phase point or representative point. During the motion of the system the phase point describes a curve, the phase curve of the system—for example, the phase curve of a Planck oscillator is an ellipse. The area under a phase curve is given by $\int pdq$ which is the action during the motion. The fundamental principle of classical quantum dynamics for the periodic motion of systems of one degree of freedom is

$$\oint pdq = nh.$$

The action taken over a period is an integral multiple of Planck's quantum of action. We may express this by the statement that the area between consecutive phase curves is h. The classical dynamics had admitted the possibility of a continuous series of phase curves; the quantum principle restricts the possible phase curves to a discrete series which is defined by an integral condition.

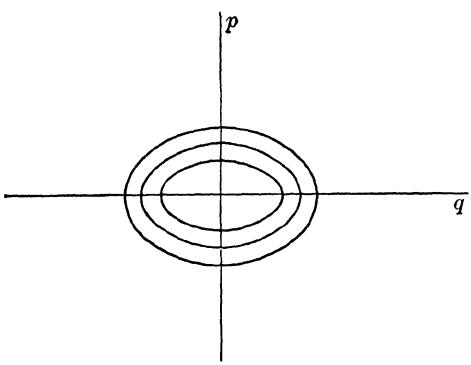


FIG. 7.

Thus the initial form of quantum dynamics did not imply the abandonment of classical dynamics—indeed it was assumed that the motion of a system in an allowed state is determined by the equations of classical dynamics—but the quantum dynamics adds the restriction that, of all solutions of the classical equations, only those which conform to a certain integral principle are physically realizable. Hence the quantum condition is supplementary to classical dynamics.

The principle that the action taken over a period is an integral multiple of the quantum of action was the basis of

later developments. From this principle one can deduce Planck's original assumption of atomicity of energy for the Planck oscillator.

The application of classical dynamics to microscopic systems subject to the quantum condition presupposes that the system is in one of its allowed states of motion. The transition from one state to another is accompanied by the emission or absorption of energy. The detailed description of this process was lacking in the theory.

3. THE HYPOTHESIS OF LIGHT QUANTA

We have seen thus far that the problem of the distribution of energy in the spectrum of temperature radiation in equilibrium with matter led to the hypothesis that the motion of microscopic dynamical systems is subject to a quantum condition. Planck had assumed, however, that electromagnetic radiation in free space is governed by the differential equations of classical electrodynamics. A quantum theory of radiation was initiated by Einstein in 1905 with the hypothesis of light quanta.

The foundation of the quantum theory of radiation was the interpretation of experiments on the interaction between matter and radiation of high frequency. Suppose that light of high frequency is incident upon the surface of an alkali metal, such as potassium. When the frequency of the radiation attains a certain critical value, electrons are emitted from the surface. Einstein proposed the hypothesis that the process is determined by the equation

$$\tfrac{1}{2}mv^2 = h\nu - P_0,$$

where $\frac{1}{2}mv^2$ is the kinetic energy of the electron, ν is the frequency of the radiation, and P_0 is the work done in separating the electron from the surface. This equation has been verified experimentally, for example, by Millikan.

The interpretation of Einstein's equation appears to require the view that electromagnetic energy exists in the form of corpuscles, called light quanta, each possessing energy $h\nu$. If electromagnetic energy is concentrated within a relatively small region it is plausible that the total energy of a light quan-

tum can be transferred to an electron in a single process. In order to account for the phenomenon of light pressure one must then ascribe momentum to the light quantum

$$\frac{h\nu}{c} = \frac{h}{\lambda}.$$

The theory of light quanta marks a partial return to the corpuscular theory of the mechanical conception of nature. The hypothesis presents itself that the interaction between a free electron and a light quantum is like a collision between material particles, and is subject to the principles of conservation of momentum and energy. The hypothesis has been confirmed by experiments on the Compton effect, which is the change in frequency of X-rays upon scattering by free electrons. In particular, the experiments of Bothe and Geiger (1924-25), and Compton and Simon (1925), in which the behavior of an individual electron was observed, indicate that the principles of conservation hold for the individual interaction.

4. THE BOHR THEORY OF THE ATOM

A great impetus was given to the quantum theory upon its application to the theory of the atom by Bohr. We have previously seen that the study of electric discharges through gases and radio-activity gave rise to the hypothesis that the atom is complex and that the electron is the universal building-stone of matter. Positively charged particles of at least atomic mass were also discovered, and thus the material was at hand for a theory of the structure of the atom. The problem was to represent the atom as a system of particles carrying positive and negative charges, and to correlate the observable properties of matter with the structural characteristics of atoms. The ultimate goal is to express material constants in terms of atomic constants.

An early theory of the atom was that of J. J. Thomson (1904), according to which the atom consists of electrons which are symmetrically arranged throughout a sphere containing a uniform distribution of positive electricity. In the electrically neutral atom the positive charge is equal to the negative charge.

The theory which furnished the basis for the application of the quantum theory was proposed by Rutherford in 1911. On the basis of experiments on the scattering of alpha particles when shot through thin sheets of metal Rutherford conceived of the atom as consisting of a positive nucleus, which is relatively small in size, but furnishes practically all the mass of the atom, and electrons revolving in orbits around the nucleus. The law of force is the inverse square law for electric charges. Thus the atom was conceived of as similar in structure to the solar system.

Inasmuch as an atom consisting of electrified particles in motion subject to classical dynamics and electrodynamics would be unstable, for the electrons would be accelerated and the system would lose its energy in a continuous manner by radiation, the theory did not provide a basis for the observed ordinary stability of matter and the occurrence of spectral lines. In 1913 Bohr introduced stability into the nuclear atom by means of the quantum theory.

The Bohr theory of the atom was initially formulated for hydrogen. The hydrogen atom is assumed to consist of an electron revolving in a circular orbit around a positive nucleus, and the stability of the atom is secured by Planck's quantum condition. Bohr's first postulate was that the atom can exist in one of a discrete series of stationary states, in which, contrary to classical electrodynamics, it does not radiate energy. The stationary state is characterized by the fact that the electron moves in accordance with classical dynamics in a circular orbit selected by the quantum condition that the moment of momentum of the electron is an integral multiple of $h/2\pi$. This quantum condition is a special case of

$$\oint p dq = nh,$$

for p is moment of momentum which is constant, and q is the angle which increases by 2π in one period.

$$\oint p dq = p \oint dq = p2\pi = nh.$$

Therefore

$$p = nh/2\pi.$$

Bohr's second postulate was an adaptation of Einstein's equation for the interaction between an electron and radiation. Bohr's postulate, called the frequency condition, states that upon transition from one stationary state to another there is emission or absorption of radiation of frequency determined by

$$h\nu = E_1 - E_2.$$

This assumption is contrary to the classical assumption that the frequency of radiation is the same as the frequency of the vibrating electrified particle. It may be noted that the frequency condition gives no detailed information regarding the nature of emission or absorption and does not presuppose the hypothesis of light quanta.

By design the Bohr theory was successful in explaining the origin of spectral lines emitted by incandescent hydrogen. A spectral line is due to the radiation which is emitted when an atom jumps from a given state to one of lower energy, thus the lines constituting the Balmer series of hydrogen arise when an electron jumps to the second orbit from the third, fourth, etc. If the electron is in the nth orbit the energy of the system is

$$E_n = -\frac{2\pi^2\mu e^4}{h^2n^2}.$$

The energy emitted by a jump from the nth orbit to the mth orbit is

$$h\nu = \frac{2\pi^2\mu e^4}{ch^2}\left(\frac{1}{m^2} - \frac{1}{n^2}\right),$$

$$\nu = R\left(\frac{1}{m^2} - \frac{1}{n^2}\right)$$

where $R = \dfrac{2\pi^2\mu e^4}{ch^3}$

μ = the mass of the electron,
e = the charge on the electron,
h = the quantum of action,
c = the velocity of light.
ν is expressed in waves per centimeter = $1/\lambda$.

For $m = 2$ the formula becomes

$$\nu = R\left(\frac{1}{2^2} - \frac{1}{n^2}\right)$$

where $n = 3, 4, 5 \ldots$. The calculated value of R is 109,000. The derived formula for the frequency agrees within the limits of experimental error with that for the Balmer series of hydrogen in which the measured value of R is 109,678.

This expression of a measurable constant in terms of atomic constants gave an enormous impetus to the quantum theory.

A fundamental principle in spectroscopy is the Ritz combination principle, according to which the frequency of any spectral line is the difference of those of two other spectral lines for a given substance. This is readily explained in terms of the Bohr theory. Let

$$\nu(nk) = \frac{E_n - E_k}{h}$$

and

$$\nu(mk) = \frac{E_m - E_k}{h}.$$

Then another possible frequency is

$$\nu(nk) - \nu(mk) = \frac{E_n - E_k}{h} - \frac{E_m - E_k}{h} = \frac{E_n - E_m}{h} = \nu(nm).$$

That is, if frequency $\nu(nk)$ is emitted when the electron jumps from the nth to the kth orbit, and frequency $\nu(mk)$ is emitted when the electron jumps from the mth to the kth orbit, then frequency $\nu(nm) = \nu(nk) - \nu(mk)$ is emitted when the electron jumps from the nth to the mth orbit. I have explained the Ritz combination principle in some detail because it has played an important part in the development of the new quantum mechanics.

The normal state of the hydrogen atom is that in which the electron is in the innermost orbit. The ionization potential depends upon the work required to remove the electron from the atom against the electrostatic force due to the nucleus. The energy necessary to ionize an atom may be supplied by an electron which is shot into the gas and gives up its energy

to an atom upon collision. The occurrence of ionization is indicated by the occurrence of the spectral line which is produced when the electron is again captured by the nucleus. Franck and Hertz shot electrons into a gas and measured the ionization potential. This type of experiment is said to give direct evidence of the existence of the stationary states. Another experiment which reveals the existence of stationary states is the Stern-Gerlach experiment. A stream of silver atoms, passing through a non-homogeneous magnetic field, splits into two parts, one part consisting of atoms whose magnetic moment is parallel to the field, and the other of atoms whose magnetic moment is anti-parallel.

To summarize, the occurrence of spectral lines, the existence of discontinuous ionization potentials, the discrete separation of a beam of atoms in a non-homogeneous magnetic field, require the hypothesis that an atom possesses **discrete stationary states.** In hydrogen there is one electron revolving around a positive nucleus. In helium there are two electrons, in lithium three and in uranium ninety-two. Corresponding to each electron in a complex atom there is a discrete series of stationary states which may be determined by spectroscopic methods or electron bombardment. Molecules possess stationary states for the motion of the electrons, vibrations of atoms, and rotations of the molecule as a whole. Thus the fundamental assumption of the quantum theory of the atom is the existence of discrete stationary states, or energy levels, in atoms and molecules. Radiation is emitted or absorbed in a transition between stationary states in accordance with Bohr's frequency condition

$$h\nu = E_1 - E_2.$$

5. THE QUANTUM CONDITIONS FOR MULTIPLY PERIODIC SYSTEMS

The quantum condition which was formulated by Planck, and applied by Bohr to the hydrogen atom, was restricted to the periodic motion of systems of one degree of freedom. In order to apply the quantum theory to a hydrogen atom in an electric or magnetic field and to complex atoms and molecules it was necessary to have quantum conditions for a system of

several degrees of freedom. This extension of the quantum theory was first achieved by Planck, but Wilson and Sommerfeld devised another method which was employed in atomic theory.

In the following statement of the quantum conditions for a system of f degrees of freedom we assume that the system is conservative. The configuration of the system is specified by f generalized coordinates $q_1 \ldots q_f$. The system is characterized by a Hamiltonian function $H(q, p)$ in which the time does not occur explicitly. Thus $H = E$, where E is a constant. The systems which we shall consider are called multiply periodic, which characteristic Born defines by the following conditions.

We assume that one can introduce a new set of variables w_k, J_k such that the following conditions are satisfied.

(A) The configuration of the system depends periodically upon the w_k with the period 1. This means that if a given w_k increases by 1, the q_k return to their original values, or in the case of an angle increase by 2π. Thus the systems to which the quantum conditions are applied have periodic characters.

(B) The Hamiltonian function goes over into a function E which depends only on the J_k. From this it follows that the w_k are linear functions of the time and the J_k are constant.

There is another condition which I shall omit.

The w_k are of the form $\nu_k t + \delta_k$, where the ν_k are the fundamental frequencies. If there are relations of the form

$$\tau_1 \nu_1 + \ldots \tau_f \nu_f = 0,$$

where the τ's are integers, the system is called degenerate, if not, it is called non-degenerate.

If the system is degenerate one can transform to new variables w_k and J_k such that

$$\nu_\alpha \text{ are incommensurable, } \alpha = 1 \ldots s,$$
$$\nu_\rho = 0, \qquad \rho = s + 1, \ldots f,$$

and the Hamiltonian function has the form $E(J_\alpha)$. The number s of independent frequencies ν_α is called the degree of periodicity of the system. The J_α have the property that they are uniquely determined, except for a trivial transformation.

This property is essential in order that they may be employed in the quantum conditions.

Now one knows from experience that slowly acting influences do not produce quantum jumps. This furnishes a basis for the adiabatic hypothesis of Ehrenfest; this principle, also called the principle of mechanical transformability, asserts that in the limiting case of infinitely slow changes of a system the laws of classical dynamics remain valid. Quantities which remain unchanged during such a process are called adiabatic invariants. Now the J_α are adiabatic invariants and are therefore to be chosen as the quantities to be quantized.

Thus in view of the uniqueness and adiabatic invariance of the J_α it is reasonable to assume that the stationary states of a system of several degrees of freedom are determined by the conditions

$$J_\alpha = n_\alpha h \qquad (\alpha = 1 \ldots s).$$

For special systems, which are called separable,

$$J_\alpha = \oint p_\alpha dq_\alpha.$$

Thus the quantum conditions become

$$\oint p_1 dq_1 = n_1 h$$

$$\oint p_2 dq_2 = n_2 h$$

etc.

It was in this form that the quantum conditions for a system of several degrees of freedom were stated by Sommerfeld in 1915.

6. THE CORRESPONDENCE PRINCIPLE

The quantum conditions determine the stationary states of the system and thus enable one to express the discrete values of the energy in terms of the quantum numbers. The frequency of radiation emitted or absorbed is determined by the frequency condition. The quantum theory of the atom as thus far set forth enables one to calculate the frequency of light emitted or absorbed by an atom. The principles give no information, however, about the intensity, phase, and state of polarization of the light. These details are determined in the

classical theory of radiation. However, Bohr discovered that one could supplement the quantum conditions and the frequency principle with classical methods by means of the Correspondence Principle. I shall explain this principle for a system of one degree of freedom.

Suppose we have given a periodic system of one degree of freedom. An example is a particle which executes the same vibration along a straight line in equal intervals of time. Let q be the coordinate which determines the position of the particle. We can then find an angle variable $w = \nu t + \delta$ such that q is a periodic function of w with the period 1. Corresponding to w there is an action variable J which remains constant during the motion. Since q is a periodic function of w it is possible to express q as a Fourier series,

$$q = a_0 + a_1 \cos 2\pi w + a_2 \cos 2 \cdot 2\pi w + \ldots b_1 \sin 2\pi w + \ldots$$

$$q = C_0 + C_1 \cos (2\pi w + \theta_1) + C_2 \cos (2 \cdot 2\pi w + \theta_2) + \ldots$$

$$q = C_0 + C_1 \cos (2\pi \nu t + \epsilon_1) + C_2 \cos (2 \cdot 2\pi \nu t + \epsilon_2) + \ldots$$

The meaning of this representation can be more easily understood if one uses a graphical interpretation. If a point has Cartesian coordinates (w, q), then $q = f(w)$ represents a curve whose form is repeated at regular intervals. Now it is possible to approximate to this curve by superposing a set of component sine and cosine curves. A special case is

$$q = \sin 2\pi w + \tfrac{1}{2} \sin 2 \cdot 2\pi w + \tfrac{1}{4} \sin 3 \cdot 2\pi w \cdot + \ldots$$

The fundamental sine curve repeats itself as w increases by 1, the others, called harmonics, repeat themselves as w increases by $\frac{1}{2}$, $\frac{1}{3}$, etc.

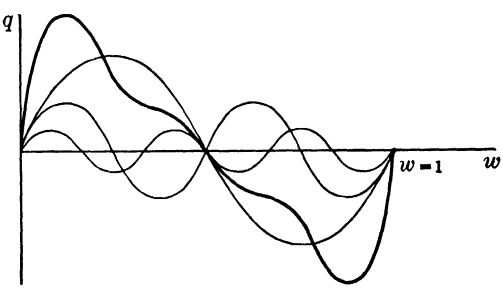


FIG. 8.

The expansion of a periodic function of w in a series of sines and cosines is the analytical correlate of the geometrical fact that a periodic curve can be built up by superposition of sine and

cosine curves. Of course the function must have certain properties which have not been mentioned.

Thus the position of the vibrating particle can be represented by

$$q = \Sigma C_\tau \cos(\tau 2\pi\nu t + \epsilon_\tau).$$

Another form is

$$q = \sum_{\tau=-\infty}^{\tau=+\infty} q_\tau e^{2\pi i \tau \nu t}$$

where q_τ and $q_{-\tau}$ are conjugate complex quantities.

Now according to classical electrodynamics the vibrator will emit light of frequencies ν, 2ν, 3ν, etc. Each harmonic in the Fourier expansion is the origin of radiation of the same frequency. The intensity of radiation is proportional to the square of the coefficient C, the phase is determined by the factor ϵ, and the state of polarization is determined by the direction of the line along which the vibration occurs.

The Bohr quantum theory of the atom enables one to calculate only the frequency. However, the correspondence principle justifies the employment of the classical theory in order to obtain approximate results. For quantum numbers n which are large with respect to 1, the frequency calculated by the quantum theory is approximately equal to the value calculated by the classical theory. The Correspondence Principle is the assumption that the intensity and state of polarization would be given by the classical theory. The principle further involves the extrapolation that even in the realm of small quantum numbers the classical theory may be employed to obtain approximate information about the intensity and state of polarization of the radiation. One correlates the radiation emitted upon a given change Δn in the quantum number n with the term in the Fourier expansion whose τ is equal to Δn. In a system of several degrees of freedom and several degrees of periodicity, each coordinate may be expanded as a multiple Fourier series. The term in the expansion characterized by a set of numbers $\tau_1 \ldots \tau_s$ is the basis for inferring the properties of radiation emitted by a quantum jump in which the quantum number n_1 changes by τ_1, n_2 by τ_2, etc. The corre-

spondence principle served as an explanation of the occurrence or non-occurrence of spectral lines corresponding to quantum jumps that appeared to be abstractly possible.

REFERENCES

BIRTWISTLE, G. The Quantum Theory of the Atom. Cambridge, 1926.

BORN, M. Vorlesungen über Atommechanik I. Berlin, 1925.

JORDAN, P. Die Erfahrungsgrundlagen der Quantentheorie. *Die Naturwissenschaften*, Vol. 17, p. 498, 1929.

CHAPTER XIV

SYMBOLIC QUANTUM MECHANICS

1. INTRODUCTION

THE classical quantum theory was a microscopic theory based upon the concepts of space, time and causality. Thus in the Bohr theory of the atom electrons are assumed to revolve in orbits around a positive nucleus in accordance with determinate laws. The classical quantum theory was further based upon an unsystematic union of classical and quantum principles. The state of an atomic system was described in terms of classical quantities such as the coordinates and momenta of an electron at a given time. The quantum theory introduced the notion of stationary states of motion; motion in the stationary states is determined by classical dynamics subject to the restrictions imposed by the quantum conditions. While the atom is in a stationary state there is, contrary to classical electrodynamics, no emission or absorption of radiation; radiation is emitted or absorbed during a transition between stationary states, and the frequency is determined by the frequency condition. The classical quantum theory of radiation did not supply a basis for the intensity, phase, and state of polarization of the radiation, but in this respect the theory was supplemented by classical methods in accordance with the correspondence principle. Thus the classical quantum theory was a system based upon inconsistent principles and points of view. Moreover, apart from the unsatisfactory nature of the foundations, the classical theory was open to the criticism that it had not succeeded in giving a satisfactory theory beyond the hydrogen atom; thus there was need of a quantum theory based upon a new point of view. The new quantum theory was initiated by Heisenberg in 1925.

2. MATRIX MECHANICS

The classical quantum theory had been a microscopic theory; Heisenberg undertook to substitute for it a macroscopic theory. He adopted the principle that only observable quantities should be employed in physical theory. Now the classical theory employed the concepts of the position, velocity, and period of an electron—quantities which are of questionable observability. The observable characters of the atom are the energies of the stationary states, and the frequency, intensity and state of polarization of the emitted radiation. In effect, the atom is known through the waves which it emits, and so Heisenberg proposed to construct a quantum mechanics in terms of the properties of the radiation emitted by the atom.

It should be noted that, strictly speaking, the characters of the waves are interpretations of observable phenomena in terms of a wave theory of light. What is observed is a spectral line (better, the aspects which are correlated in the concept of spectral line), whose position is measured and frequency calculated. However, we may admit that, compared with the characters of an electron, the properties of waves are directly observable. Thus Heisenberg adopted the phenomenological point of view which was proposed by Mach; Sommerfeld has remarked that the creation of the new quantum mechanics is the exceptional case in which Mach's point of view has proved to be fruitful.

The data of the new quantum mechanics consisted of the characters of the waves emitted by the atom; the construction of the fundamental quantities and principles was guided by the correspondence principle; and the methodological principle was that the quantum theory was to be fashioned so as to constitute a rational generalization of classical physics.

The employment of the correspondence principle is based upon the representation of a classical periodic motion by a Fourier series. Supposing that a particle has a periodic motion of one degree of freedom, the position of the particle as a function of the time is represented by

$$q = \Sigma q_\tau e^{2\pi i \tau \nu t}.$$

Thus the classical motion is represented as the superposition of a series of harmonic motions of frequencies ν, 2ν, 3ν, etc., and amplitudes q_1, q_2, etc. According to classical electrodynamics the frequencies emitted are ν, 2ν, 3ν, etc. Although this is contradicted by the quantum frequency principle, nevertheless the correspondence principle justified the employment of the terms of the expansion for the purpose of obtaining information about the character of the radiation.

Now the atom is to be represented by quantities which characterize the waves emitted. In accordance with the correspondence principle the quantities belonging to the waves are to correspond to the terms in the Fourier expansion. The classical coordinate q is the sum of a series of terms. The quantum coordinate q is chosen as the set of corresponding quantities belonging to the waves. In order to accomplish this the classical term

$$q_\tau e^{2\pi i\tau\nu t}$$

is changed into

$$q(nm)e^{2\pi i\nu(nm)t}.$$

The occurrence of terms $e^{2\pi i\tau\nu t}$ in the classical theory means that the frequencies of the emitted light are integral multiples of a fundamental frequency. The frequencies observed do not obey this rule and are of the form

$$\nu(nm) = \frac{1}{h}(E_n - E_m).$$

Thus the quantum theoretical quantities are constructed to accord with the frequency principle. $q(nm)$ denotes the complex amplitude. Thus the quantum coordinate q is the set of quantities

$$\mathbf{q} = \begin{vmatrix} q(11)e^{2\pi i\nu(11)t} & q(12)e^{2\pi i\nu(12)t} & \cdot & \cdot \\ q(21)e^{2\pi i\nu(21)t} & \cdot & \cdot & \cdot \\ \cdot & & & \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \end{vmatrix}$$

The coordinate $\mathbf{q}$ is a two-dimensional array which represents the two-dimensional set of waves emitted by the atom; hence in accordance with our point of view $\mathbf{q}$ represents the atom.

Now in classical physics it is frequently necessary to multiply two quantities or to find the square of a quantity. For example, we form qp or q^2 in accordance with the laws of ordinary algebra. The quantum mechanical problem thus presents itself: What is the operation on quantum quantities corresponding to multiplication of classical quantities? Again the correspondence principle furnished the clue. Suppose that q and p are each expressed as a Fourier series. Then the product qp can be expressed as a Fourier series in which the terms are obtained from the terms of q and p by multiplying each term of one series by every term of the other. No new frequencies appear. In accordance with the correspondence principle the product of the quantum quantities $\mathbf{q}$ and $\mathbf{p}$ is to consist of an array, the members of which are obtained from the members of $\mathbf{q}$ and $\mathbf{p}$ in a prescribed way. It is assumed that the frequencies in $\mathbf{q}$ and $\mathbf{p}$ are the same. The method must be such that the product has the same frequencies as the factors.

Now the Ritz combination principle states that a frequency emitted by an atom can be expressed as the difference of two terms. The explanation of this was one of the triumphs of the original Bohr theory. The term is the energy of a state divided by h, E/h. Upon transition from state n to state m the frequency emitted is determined by

$$\nu(nm) = \frac{E_n}{h} - \frac{E_m}{h}.$$

Now the mathematician has created an algebra for arrays which are called matrices. In the algebra of matrices various operations such as addition and multiplication are defined by certain laws. It now turns out that if the quantum quantities $\mathbf{q}$ and $\mathbf{p}$, which are arrays, are multiplied in accordance with the laws of matrices, the product contains the same frequencies as the factors, for in order to form the product one has to multiply terms of the form

$$e^{2\pi i\nu(nk)t} \text{ and } e^{2\pi i\nu(km)t}.$$

This is $e^{2\pi i[\nu(nk) + \nu(km)]t}$.

But $$\nu(nk) + \nu(km) = \frac{1}{h}(E_n - E_k) + \frac{1}{h}(E_k - E_m)$$

$$= \frac{1}{h}(E_n - E_m) = \nu(nm).$$

Hence we get terms of the form

$$e^{2\pi i\nu(nm)t},$$

in which the same frequencies occur as in the factors. Thus the correspondence principle in conjunction with the Ritz combination principle suggests that corresponding to algebraic multiplication we should multiply the quantum quantities **q** and **p** in accordance with the law of multiplication for matrices. In other words, the arrays **q** and **p** which represent the atom are matrices. The law of multiplication of **q** and **p** is

$$\mathbf{qp} = (\Sigma_k q(nk)p(km)e^{2\pi i\nu(nm)t}),$$

where the parenthesis contains the symbol of one member of the two-dimensional matrix, which is the product of the matrices **q** and **p**. It is to be noted that the multiplication of matrices is not commutative, that is, **pq** is not in general equal to **qp**.

Now in classical dynamics we describe the instantaneous configuration and state of motion of a system by the coordinates q and momenta p. We form the Hamiltonian function H as a function of q and p; then we determine the motion by the canonical equations,

$$\dot{q} = \frac{\partial H}{\partial p}$$

$$\dot{p} = -\frac{\partial H}{\partial q}.$$

In accordance with the correspondence principle one builds the quantum mechanics upon similar laws for matrices.

The matrix **q** is called the coordinate, and correlated with it is a momentum matrix **p**. One then forms the Hamiltonian function **H** by the addition and multiplication of the matrices **q** and **p**. One introduces the quantum condition by the commutative law

$$\mathbf{pq} - \mathbf{qp} = \frac{h}{2\pi i}\mathbf{1}.$$

The basis for the law was the fact that for large quantum numbers it becomes identical with the quantum condition for periodic systems

$$\oint p dq = nh.$$

One defines the differentiation of matrices. Then corresponding to the canonical equations of motion one assumes the canonical equations for matrices,

$$\dot{\mathbf{q}} = \frac{\partial \mathbf{H}}{\partial \mathbf{p}}$$

$$\dot{\mathbf{p}} = -\frac{\partial \mathbf{H}}{\partial \mathbf{q}}.$$

The definition of the equality of matrices is that each element of one is equal to the corresponding element of the other. Hence the canonical equations give rise to an infinite number of algebraic equations in an infinite number of variables.

A diagonal matrix is one in which only the diagonal terms are different from zero. From the canonical equations one can deduce that $\dot{\mathbf{H}} = 0$, which is the principle of the conservation of energy. $\mathbf{H}$ is thus a diagonal matrix

$$\mathbf{H} = \begin{vmatrix} H_1 & 0 & 0 & \cdots \\ 0 & H_2 & 0 & \cdots \\ 0 & 0 & H_3 & \cdots \\ \vdots & \vdots & \vdots & \cdots \end{vmatrix}$$

Bohr's frequency condition can be deduced from the principles.

A typical system to which the theory is applied is the harmonic oscillator. In classical dynamics

$$H = \frac{p^2}{2\mu} + \frac{kq^2}{2}.$$

In quantum mechanics one assumes that $\mathbf{H}$ is the same function of the matrices $\mathbf{q}$ and $\mathbf{p}$. One applies the canonical equations and the commutative law and obtains

$$H_n = h\nu(n + \tfrac{1}{2}).$$

In this way one finds the stationary states of the harmonic oscillator. As we have previously seen, Planck first introduced the quantum theory for the harmonic oscillator.

The quantum mechanics of matrices was invented by Heisenberg, and the development of the theory was shared by Born, Jordan and Heisenberg. The development of matrix mechanics was guided by the phenomenological point of view. However, as we shall see presently, matrix mechanics has been reinterpreted in the light of the completed system of quantum mechanics.

I have sketched the development of matrix mechanics in some detail because it exhibits the creation of a system of mathematics for the representation of physical phenomena.

3. WAVE MECHANICS

Shortly after the invention of matrix quantum mechanics Schrödinger discovered a new method in quantum theory. Schrödinger's wave mechanics was suggested by de Broglie's wave theory of matter, concerning which it is desirable to say a few words.

In our discussion of radiation we have seen that both the concepts of wave and of particle are applicable to the phenomenon of radiation: in describing interference and diffraction we employ the concept of wave; in describing the interaction between radiation and matter we use the concept of particle. According to the hypothesis of light quanta, radiation exists in the form of corpuscles of energy $E = h\nu$ and momentum $p = \frac{h\nu}{c}$.

Thus far in our theory of matter we have exclusively employed the concept of particle, regarding the atom as a system of positively and negatively charged particles. Now de Broglie conceived the idea of applying the concept of wave to particles, such as electrons, atoms, etc. The dualistic use of the concepts of particle and wave in the theory of radiation is thereby extended to the theory of matter.

de Broglie associated a wave with a material particle by means of the following hypothesis. If a material particle has energy E relative to a given system of reference, then there exists in this system a periodic process of frequency ν, where $E = h\nu$. If the particle is at rest in the system of reference

the frequency will be that of a periodic process extending throughout the space surrounding the particle. This may be expressed by the statement that the particle is at the singular point of a vibration which at a given time is in the same phase throughout space. Thus the periodic process can be represented by

$$\psi = A \sin 2\pi\nu t.$$

By the Lorentz transformation one can obtain the representation of the process in a system of reference K' with respect to which the particle is moving with velocity v. One obtains

$$\psi = A' \sin 2\pi\nu'\left(t' - \frac{x'}{u}\right)$$

where

$$u = c^2/v.$$

Thus with respect to K' the periodic process is a wave which travels in the positive direction of the x' axis with speed u. This wave is called the phase wave of the particle, and its speed is greater than that of light. In the new system the energy of the particle is related to the frequency of the phase wave by $E' = h\nu'$.

In the theory of wave motion one considers the motion of a set of waves which differ slightly in frequency. If one takes an instantaneous view of such motion one observes that the waves alternately reinforce and annul each other. The maxima and minima of the disturbance travel with a constant velocity given by

$$\frac{1}{v} = \frac{d}{d\nu}\left(\frac{\nu}{u}\right),$$

where u is the velocity of the waves. The waves are said to form groups or packets, and the group velocity v is given by the above equation. Now in de Broglie's theory it turns out that the classical velocity of the particle is equal to the group velocity of a set of associated phase waves. Thus one may interpret the particle as a group of waves.

If material particles have a wave aspect one would expect particles to exhibit diffraction phenomena. The existence of this phenomenon has been revealed by experiments of Davisson

and Germer, G. P. Thomson, etc., so that the applicability to matter of the concept of wave is confirmed by experiment.

de Broglie was the first to propose a wave theory of matter; Schrödinger found the differential equation of the wave function. Although he discovered the equation by another method, his principal justification for proposing it as an hypothesis was furnished by optics.

We recall that geometrical optics is the study of the propagation of light on the assumption that light travels in rays. Geometrical optics may be based upon Fermat's principle of least time, which states that the path of a ray of light between two points is such that the time required is a minimum. One may also describe the propagation of light in terms of the motion of surfaces whose normals are the geometrical rays, and then the propagation of light is determined by a certain differential equation of the first order and second degree. Now geometrical optics is adequate only as a first approximation for the description of optical phenomena in which distances are involved which are large in comparison with the wave length of light. It does not afford an explanation of phenomena such as diffraction. For a detailed representation one must have recourse to physical optics in which the propagation of light is determined by the wave equation, a partial differential equation of the second order and first degree,

$$\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}+\frac{\partial^2\varphi}{\partial z^2}=\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2}.$$

Now the classical dynamics of a particle is analogous to geometrical optics; corresponding to the principle of least time in optics there is Maupertuis' principle of least action in dynamics. The motion of a particle between two given points is that for which the integral

$$\int 2T dt$$

is a minimum or maximum, assuming that the energy remains constant. The motion of the system is also described by the same partial differential equation of the first order and second degree as in geometrical optics. This is the Hamilton-Jacobi

equation, and in classical dynamics one can determine the motion of those systems for which the equation can be solved.

We have seen, however, that classical dynamics is inadequate for the description of atomic processes. The breakdown of classical dynamics for fine-scale processes is analogous to the failure of geometrical optics for diffraction phenomena. Hence the hypothesis presents itself that, just as it was necessary to pass from geometrical optics to physical optics, so it is necessary to make the transition from the classical dynamics of particles to a wave mechanics in order to obtain a representation of atomic phenomena. Schrödinger therefore replaced the Hamilton-Jacobi equation of classical dynamics by a wave equation for microscopic processes.

The Hamilton-Jacobi equation for conservative systems is

$$\frac{\partial W}{\partial t} + H\left(q, \frac{\partial W}{\partial q}\right) = 0.$$

The complete solution is of the form

$$W = -Et + S(q).$$

W is the integral of the Lagrangian function with respect to the time; the symbol q represents the set of generalized coordinates

$$q_1, q_2, \ldots .$$

Let us suppose that the instantaneous configuration of a system is represented by a point in a configuration space, that is, the coordinates of position of the system are equal to the Cartesian coordinates of the point. Then the above solution represents the propagation in configuration space of surfaces over which the phase W is constant. The velocity of a W surface is determined to be

$$u = \frac{E}{\sqrt{2\mu(E - V)}}.$$

Now Schrödinger assumes that W determines the phase of a wave function which satisfies the wave equation. He assumes that the wave function which characterizes a material particle is of the form

$$\psi = A \sin (kW + \text{const.}).$$

Since W is action and the phase of a sine must be dimensionless, k must be a reciprocal action. He assumes that

$$k = \frac{2\pi}{h}.$$

Then

$$\psi = A \sin\left(-2\pi \frac{E}{h} t + \frac{2\pi}{h} S(q) + \text{const.}\right).$$

Then the frequency of the waves is

$$\nu = \frac{E}{h};$$

the velocity of the waves is

$$u = \frac{h\nu}{\sqrt{2\mu(h\nu - V)}}.$$

The wave function ψ is determined by the equation for wave motion,

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} - \frac{1}{u^2}\frac{\partial^2\psi}{\partial t^2} = 0.$$

The solution is of the form

$$\psi = ae^{2\pi i\nu t}$$

Then

$$\frac{\partial^2\psi}{\partial t^2} = -4\pi^2\nu^2\psi,$$

so that the equation becomes

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} + \frac{8\pi^2\mu}{h^2}(E - V)\psi = 0.$$

This is Schrödinger's equation in Cartesian coordinates for the motion of a single particle. E is the total energy and V is the potential energy of the particle. E is called the parameter of the differential equation.

The next step consists in finding solutions ψ of the equation. The existence of solutions depends upon the parameter E. If the solutions are to be finite, single-valued, and continuous throughout space, then the parameter E must assume any

positive value or a discrete series of negative values. It turns out that the discrete series of negative values of E agree with the energies of the stationary states as calculated by the classical quantum theory. The values of the parameter E are called the characteristic values of the differential equation and the solutions ψ corresponding to the different characteristic values are called characteristic functions. The quantum condition $\oint pdq = nh$ of the classical quantum theory is then replaced by the following: The stationary states of a system consisting of a particle moving in a field of force are specified by the discrete negative values of the parameter E such that the corresponding solutions ψ of Schrödinger's equation are finite, single-valued and continuous throughout all space. The problem of quantization is thus reduced to a problem in the classical theory of differential equations.

The classical dynamics of continuous media can be expressed in terms of a similar mathematical theory. Indeed, for one degree of freedom the wave equation reduces to the equation of a vibrating string. The solution of the problem of the vibrating string may be described by the statement that the string can vibrate as a whole, in two equal segments, in three equal segments, etc. Thus we see that the classical theory of elasticity gives rise to integral numbers in the representation of motion. The reduction of the problem of quantization to a boundary value problem similar to those in the classical theory of continuous media thus held out the prospect that the quantum theory could be encompassed by the classical physics of the continuum.

As an illustration of the method of wave mechanics we may apply it to the determination of the stationary states of a system consisting of a particle moving in a Coulomb field of force —this is the problem of the hydrogen atom. We take from the classical theory the expression for the potential energy

$$V = -e^2/r.$$

The wave equation becomes

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} + \frac{8\pi^2\mu}{h^2}\left(E + \frac{e^2}{r}\right)\psi = 0.$$

The negative characteristic values are

$$E_n = -\frac{2\pi^2\mu e^4}{h^2}\frac{1}{n^2}\cdots$$

These are the energies of an electron moving in an elliptical orbit about the nucleus of the hydrogen atom.

The general method of obtaining Schrödinger's equation in Cartesian coordinates is as follows. One sets up the Hamiltonian function as for a classical problem; then in place of p_k one writes

$$\frac{h}{2\pi i}\frac{\partial}{\partial q_k}, \text{ etc.}$$

The Hamiltonian function is thereby transformed into an operator. The symbolic product of this operator and the wave function ψ is equal to the product $E\psi$.

$$\left\{H\left(\frac{h}{2\pi i}\frac{\partial}{\partial q}, q\right) - E\right\}\psi = 0.$$

This is the general Schrödinger equation. The energies of the stationary states of the system are the characteristic values of the differential equation.

Wave mechanics and matrix mechanics have been proved to be equivalent. The characteristic values E_n of Shrödinger's equation agree with the diagonal terms H_n of the matrix which represents the energy. From the characteristic functions one calculates the matrices which represent the coordinates and the momenta. A matrix element $q(nm)$ is given by

$$q(nm) = \int q\psi_n\bar{\psi}_m d\tau$$

where $d\tau$ is the product of the differentials of the Cartesian coordinates and the integral is extended over the whole configuration space. $\bar{\psi}$ is the conjugate of ψ.

4. THE INTERPRETATION OF QUANTUM MECHANICS

We have seen that the stationary states of an atomic system are determined by the discrete negative values of a parameter in a partial differential equation for a wave function ψ.

Corresponding to each characteristic value of the parameter there is a characteristic function ψ. The symbol ψ represents the state of the atomic system. A solution of the equation may have the form

$$\psi = ae^{\frac{2\pi t}{h}L:t}$$

We may need to build up a solution of the form

$$\psi = c_1\psi_1 + c_2\psi_2 + \ldots$$

The problem now arises: What is the physical significance of ψ?

Schrödinger adopted the hypothesis that $\psi\bar{\psi}$ is the expression for the density of electricity. This implies that one may think of the electron as spread throughout the whole of space. Schrödinger further sought to work out a wave theory of particles on the hypothesis that a particle is represented by the superposition of a set of characteristic solutions,

$$\psi = \Sigma c_k\psi_k.$$

For the representation of a harmonic oscillator he was able to construct a wave packet which is concentrated in a relatively small region and whose center of mass vibrates with simple harmonic motion. However, Heisenberg showed that in general such a packet of waves would disperse in time. Schrödinger thus sought to explain quantum phenomena in terms of a physics of continuity. His theory of the significance of the wave function has not been accepted by physicists; in its stead there has been developed a statistical interpretation of quantum mechanics.

The statistical interpretation is as follows. In

$$\psi = c_1\psi_1 + c_2\psi_2 + \ldots$$

$|c_1|^2$ is the probability that the system is in state ψ_1, $|c_2|^2$ that it is in state ψ_2, etc.

For a single particle, such as an electron or light quantum, $\psi\bar{\psi}dxdydz$ is the probability that the particle is in the element of volume $dxdydz$.

The statistical interpretation of quantum mechanics implies an **indeterminism** in atomic processes. The values of the physical quantities which characterize a system in a given state can not all be determined simultaneously with perfect accuracy.

Consider, for example, a free particle with constant energy for which the amplitute of the solution of Schrödinger's equation is given by e^{ikx}, so that ψ represents a plane wave. In this case the momentum of the system is sharply determined. The probability that the particle falls in an element of space dx is $\psi\bar{\psi}dx$. This is equal to 1 independently of x. That is, the position of the particle with this characteristic function is fully undetermined. The probability of finding it in an element dx is the same for all points. Conversely, we can build a solution ψ by superposition of an infinite series of waves e^{ikx} with variable k. We can determine the coefficients so that, at $t = 0$, ψ is appreciably different from zero at only one point in space. In this way the position of the particle is accurately determined as at this point. But it now follows that the momentum depends on x and may assume all possible values between $-\infty$ and $+\infty$ with different probability. A single term e^{ikx} corresponds to a determinate momentum. This single term renders $\psi\bar{\psi}$ independent of x and produces indeterminateness in the coordinate. Conversely, the accurate determination of the coordinate requires infinitely many terms e^{ikx}, to each of which corresponds a different momentum. In general, if one of two canonically conjugated quantities is determinate the other is completely indeterminate.

In the preceding discussion I have assumed that one of them was completely determinate; the more general case is that in which both are partly determinate. We are then able to deduce the following result. If Δq is the uncertainty in q, and Δp is the uncertainty in p, then

$$\Delta p \Delta q \geqq h.$$

The classical assumption was that all quantities could be measured independently and as accurately as one pleases. In quantum mechanics some of the quantities—those which are commutable—can be measured simultaneously with perfect accuracy. However, it is not possible to measure all physical quantities with perfect accuracy independently of each other. This result is called the principle of indeterminacy; it implies that the observation of an atomic system disturbs it. A physical basis for the principle, and therefore for the statistical

interpretation of quantum mechanics, is provided by examples of ideal experiments which have been discussed by Heisenberg and Bohr.

The statistical interpretation is based upon the fact that the observation of an atomic system involves an interaction with the agency of observation. The consequence is that the concepts of space, time and causality of classical and relativistic physics are shown to be inadequate for the description of atomic phenomena. We can obtain an exact space-time description of an atomic process, but in so doing we are unable to determine values of the momenta and energy, and are thus unable to satisfy the principle of causality which implies the conservation of momentum and energy. Conversely, we are able to determine the momentum and energy precisely, but then the position and time become indeterminate. Consequently the concepts of particle and wave are reciprocally limited in their application to atomic systems.

The quantum mechanics suggests a modification in the definition of physical quantity.

We denote the state of an atomic system by the symbol ψ.

In classical physics we characterize a system in terms of physical quantities t, q, p, E, etc. These are numbers which are assigned to the system on the basis of certain operations and calculations. The physical quantities represent characters of the system.

In quantum mechanics we employ the symbols of the classical quantities for the quantum quantities; the quantum quantities t, q, p, E, are symbols which denote the system. The symbols express the possibility of assigning numbers to a system in a certain state. In classical physics the physical quantities were definitely determined by the system; in quantum mechanics the physical quantities are symbols which may be represented by numbers with a certain probability. The physical quantity is represented by a matrix. Now in an appropriate coordinate system the matrix is diagonal; then the elements of the matrix are the possible values of the physical quantity. They are the numbers which may be assigned to the system on the basis of certain probabilities.

Quantum mechanics may be summarized as follows. It

is possible to denote the characters of an atomic system in a given state by symbols, such that the symbols may be represented by matrices which exhibit the possible results of measurement. The elements of the matrix are the numbers which express the results of a measurement with a calculable probability. We may say that quantum mechanics initiates a symbolic conception of nature.

REFERENCE

BIGGS, H. F. Wave Mechanics. Oxford, 1927.

BOHR, N. The Quantum Postulate and Atomic Theory. *Nature*, Vol. 121, 1928.

BORN, M. Problems of Atomic Dynamics. Cambridge, Mass., 1926.

BORN, M., and JORDAN, P. Elementare Quantenmechanik. Berlin, 1930.

CONDON, E., and MORSE, P. Quantum Mechanics. New York, 1929.

DE BROGLIE, L. Wave Mechanics, translated by H. T. Flint. London, 1930.

DIRAC, P. A. M. Principles of Quantum Mechanics. Oxford, 1930.

HALPERN, O., and THIRRING, H. Die Grundgedanken der neueren Quantentheorie. *Ergebnisse der Naturwissenschaften*, Vol. 8, Berlin, 1929.

HEISENBERG, W. The Physical Principles of the Quantum Theory. Chicago, 1930.

SCHRÖDINGER, E. Abhandlungen zur Wellenmechanik. Leipzig, 1927.

SOMMERFELD, A. Einige grundsätzliche Bemerkungen zur Wellenmechanik. *Physikalische Zeitschrift*, Vol. 30, 1924.

THIRRING, H. Die Grundgedanken der neureren Quantentheorie. *Ergebnisse der Naturwissenschaften*, Vol. 7, Berlin, 1928.

WEYL, H. Gruppentheorie und Quantenmechanik. Leipzig, 1928.

PART IV

METHODOLOGICAL PRINCIPLES

CHAPTER XV

THE CONCEPT OF SUBSTANCE

1. THE CRITERION OF TRUTH

THE problem of physical theory is the systematic conceptual representation of the physical order. The first stage in the development of physical theory consists in the immediate experience of given reality, and the formation of the concepts of qualities and relations. Fundamental concepts are formed by abstraction from experience, and are applicable to experience in virtue of their origin. Concepts may also be invented and then their applicability must be tested by experience. The development of quantitative physical theory begins when it is found possible to represent bodies by physical quantities which are the results of measurement. The essential feature of measurement is the observation of space-time coincidences. The laws of physics are then expressed by functional relations between variable physical quantities. In the most recent developments of physical theory physical laws are expressed by propositions about abstract symbols.

In the preceding chapters we have reviewed the principal concepts, principles and systems of physical theory. I propose now to present some general conclusions on methodology.

The construction of a physical theory requires the acceptance of certain data. The initial data are particular aspects, given in **experience**, which one represents by the **concepts** of the qualities of which the aspects are instances. With the aid of concepts one can express **propositions** about the characters of the given. Inasmuch as the basis of a physical theory consists of true propositions about the nature of given aspects, a fundamental methodological problem is the criterion of the truth of propositions which express the characters of immedi-

ately given aspects. The truth of these propositions is tested in a simultaneous experience of the concepts and the given. The process of verification consists in the observation of instances of the characters expressed in a proposition. One may say that such propositions correspond to, or agree with, experience. This relation of a true proposition to experience is fundamental and can be explained only in terms of synonyms; the fundamental fact is that we can determine the correspondence to experience of propositions about experience. The assumption of the applicability of a concept to experience implies the truth of certain propositions about the given.

Propositions about characters of given aspects provide a basis for the concepts of bodies which are said to be given in experience. The fundamental concept of physical theory is that of body, and of the space-time coincidence of special bodies called points. The concept of body expresses a correlation between aspects which we can represent by concepts. Thus the recognition of the applicability of the concept of body to experience presupposes the experience of the correspondence of propositions about aspects to experience.

The special work of the physicist takes for granted the applicability to experience of the concepts of body and of space-time coincidence. In a restricted sense, physical science begins with the construction of concepts of macroscopic physical quantities. Thus a fundamental physical quantity is length. Now the concept of length is that of the method of assigning a number which expresses the ratio of the extension of a body to the extension of some standard body. Ultimately the operation of assignment depends upon the observation of space-time coincidences. Hence the determination of directly measurable quantities depends upon the test of truth by correspondence. It is to be noted that measurement is approximate; however, the reproducibility of coincidences is so much greater than other similarities that a science based upon measurements is relatively exact.

The experimental data of physics are expressed by propositions which are subject to the criterion of correspondence. However, the construction of a systematic physical theory requires propositions which can not be directly tested for cor-

respondence with experience. Thus the fundamental principles may be expressed in the form of differential equations, or they may be assumptions about microscopic reality. The test of such fundamental assumptions must be indirect; the assumptions must be the premises for deductions of propositions which correspond to experience. Therefore the construction of a system requires the employment of such criteria as coherence, self-consistency, logical stability.

2. METHODOLOGICAL PRINCIPLES

Now several sets of assumptions may yield the same experimental results. Thus the problem arises as to the principle of selection of the assumptions. The criterion of coherence is negative; the necessity for a positive criterion can be seen from an examination of the nature of measurement. Consider for example the concept of length. One has to postulate at the outset that the number which constitutes the length of a body is unchanged by a displacement. Now it may be desirable to change the number in view of certain phenomena; for example, if the standard measuring rod becomes hot we assign a different number to it. Classical physics, including the theory of relativity, presupposes that the concept of constancy of conditions has a meaning. As we succeed in specifying the conditions more and more precisely we are able to define physical quantities to a higher order of approximation; this process is well illustrated by the physical quantity length. Again we may change the very definition of a concept; a principle which is based upon a certain concept may be transformed into a definition of the concept. For example, the second law of thermodynamics initially presupposes the concept of temperature, and then becomes the definition of temperature. Furthermore, definitions may be changed so that a fundamental concept is transformed into a derived concept. Thus the concept of mass is fundamental in dynamics; it is definable in terms of the electromagnetic field in electrodynamics. In general, the process of development is as follows: There is an abstraction of concepts from experience, the discovery of laws expressed in terms of the concepts, the definition of the original

concepts to a higher order of approximation in view of the greater accuracy in the definition of conditions, the redefinition of concepts in terms of laws, the reinterpretation of original concepts in terms of the new. Thus we come upon the problem of the criteria for the choice of fundamental concepts and systems.

I remark at the outset that every concept may be the basis of a methodological principle. However, we are concerned with the concepts which express the most general methodological principles.

The fundamental concept is that of **unity**. The discovery of similarities between things and phenomena suggests the general program of interpreting physical phenomena in terms of a few fundamental concepts and principles. This has been illustrated by the several unitary systems of physical theory. The tendency has been to select one system of physical theory as canonical; thus we have had the mechanical conception of nature, the electrodynamic conception of nature, the program of a geometrical conception of nature; and lastly the achievement of a symbolic quantum mechanics offers the prospect of explaining all phenomena as manifestations of statistical elementary processes.

The concept of unity has been expressed primarily in the concepts of substance and causality. In the present chapter I examine the function of a concept of substance in physical theory.

3. SUBSTANCE AS SUBSTRATUM

In the exposition of the concepts which serve to describe the physical order I began with particular aspects which may be given in experience. I then showed how knowledge about aspects is expressed in the concepts of characters and relations. The concept of thing is built up out of the concepts of characters and relations of aspects. Thus we think of a thing as a certain class of aspects. Now things have been called substances. Is there any need to specify things more precisely than we have done by calling them substances?

The concept of substance has been employed in various senses. One meaning of substance is that of substratum. The

thing as substance is the self-identical unity of the aspects. Substance is the ineffable substratum of the aspects of a thing which inhere in it. The concept of substance thus supplies an element of permanence in the manifold of changing aspects.

I shall not employ the concept of substance in the sense of substratum. Our concept of thing is adequate to express the unity of aspects. We combine aspects in one thing on the basis of characteristics of the aspects and not in virtue of a relation of inherence. A thing is constituted by observable correlations between aspects. Thus a succession of similar correlations of similar aspects is the basis of the concept of thing. Some of the aspects may change, but we experience many fairly constant correlations which are instances of the concept of thing—and even though aspects change they may still be viewed as belonging to the same thing if the changes are regular.

Things have measurable properties which are the object of physical investigation. A thing which is characterized in terms of measurable properties is called a body. The changes of the measurable properties of a given body are determined by the laws of physics. In a completed system of physics we could define a body as a class of aspects whose properties change in accordance with the laws of physics, as was pointed out by Bertrand Russell.

4. SUBSTANCE AS THE PERMANENT

Substance has also been defined as the permanent in phenomena in the sense of the embodiment of invariable attributes. On the basis of observation of bodies we describe them as possessing certain attributes. One or more attributes may be invariable; if we view a body as something having an invariable attribute the body is substance.

In physics an invariable attribute of a body is represented by a constant measure. In classical physics a constant measure expresses the property (magnitude) of a quantity, such as the quantity of matter, which may also be called quantity of material substance and was called the mass of the body. In the mechanical conception of nature the essential property of a body is inertia which is the invariable property of matter.

Material substance was the hypostatization of the invariable property of matter. Matter was also viewed as objective and independent.

Substance in the foregoing version of the concept is the permanent, the objective, and the self-determining or independent. In seventeenth century physics the concept was interpreted in terms of a dualistic metaphysics. Galileo, Descartes, Locke, etc., made a distinction between primary qualities, which are objective, and secondary qualities which are subjective, from the physical point of view the distinction being essentially that between measurable and non-measurable properties. However, seventeenth century physicists and philosophers viewed the primary qualities as objective and the essence of a permanent and independent substance. As Professor Lamprecht has remarked, "Descartes and Locke turned this methodological dualism of scientific thought into a metaphysical dogma." Thus the concept of substance in classical physics combined the Scholastic concept of substance with the distinction between primary and secondary qualities.

Material substance was considered to be an independent reality which produces the aspects directly experienced by the mind. The concept of material substance thus furnishes a realistic interpretation of the invariable mechanical properties of bodies. The preceding method of interpretation was then extended to the theory of electricity and the theory of heat. Electricity, or electrical substance, is an indestructible fluid which is the cause of observed electrical actions; it is the hypostatization of constancy of electric charge. Similarly heat, or caloric, is an indestructible fluid whose presence in a body is the cause of the sensation of hotness; it is the hypostatization of the constancy of the measure of heat in thermal interactions. However, in the mechanical theory of heat the concept of heat substance was abandoned. Nevertheless, the concepts of material substance and electrical substance were retained in classical physics. During the nineteenth century the mechanical conception of nature was the principle of unity in physics; it was assumed that material substance dwelt in a real absolute space and moved in the course of a real absolute time. The mechanical theory was superseded by the electro-

dynamic conception of nature according to which electrical substance is the fundamental reality.

In contemporary physics there is a movement to abandon the classical concept of substance, which has expressed itself partly in the attempt to construct a pure field theory, in which the properties of electricity and therefore of matter are definable in terms of the electromagnetic field. The self-determination of substance therefore vanishes. Physical reality is not a set of independent elements, but a variable field. However, the attempts by Mie, Weyl, etc., to construct a satisfactory theory of matter on the basis of classical and relativistic ideas have not succeeded, and the ultimate solution must await the completion of quantum electrodynamics.

Even though one gave up the self-determination of substance one could still retain the concept of substance as the expression of an objective reality which is independent of mind. One may say, for example, that there is a physical quantity which has a measurable property which is represented by a measure. Constant quantities are substances.

In this book I have adopted an empirical point of view. A body is a class of aspects. The essential things are bodies and the assignment of numbers by the operation of measurement. The physical quantity is a number which is assigned by measurement or is assumed. The laws of physics are expressed by relations between variable physical quantities. Bodies and processes are represented by numbers, or by symbols which may be represented by matrices. Hence the search for substance becomes the search for constants and invariants. There are functional relations between numbers. Thus the permanence, objectivity, and self-determination of substance are replaced by the constancy, invariance and functional relationship of numerical measures. In the present chapter I consider some instances of constants and invariants.

5. CONSTANTS

I shall first consider some of the constants of classical dynamics. The length (number) of a standard rigid rod is constant under the same conditions; similarly the time of a vibration of a standard clock is constant; also the mass of a

standard body is constant. In classical dynamics the foregoing constants are assumed to be independent of time and place provided that the conditions of temperature, gravitational force, etc., are the same. I recall that the specification of the conditions proceeds by successive approximation.

What are the constants of a dynamical system consisting of particles? In the first place, the total mass of the system is constant. For certain systems, principles of the conservation of energy, momentum and moment of momentum, respectively, hold. Consider, for example, two billiard balls which are moving towards each other along a line joining their centers. The principle of the conservation of momentum is that if no forces act, except those called into play by the impact, then the momentum after the impact is equal to the momentum before impact. This principle may be used as a definition of momentum or it may be derived from the laws of motion if they are set up independently of the principle. In general, if no external forces act upon a system the total momentum remains constant.

The principle of conservation of energy in dynamics may be obtained by considerations illustrated by the following. Suppose that a particle is falling freely in the earth's gravitational field. The equation of motion is

$$m\frac{d^2s}{dt^2} = -mg.$$

This may be written

$$mv\frac{dv}{ds} = -mg.$$

Upon integration one obtains

$$\frac{mv^2}{2} = -mgs + \text{cons.}$$

Suppose that when $s = s_0$, $v = v_0$. Then the constant is

$$\frac{mv_0^2}{2} + mgs_0.$$

Hence we obtain

$$\frac{mv^2}{2} + mgs = \frac{mv_0^2}{2} + mgs_0.$$

Now $mv^2/2$ is called kinetic energy and mgs is called potential energy. We have the result that kinetic energy plus potential energy is a constant. $mv^2/2 + mgs$ is called the integral of energy. We see by this example that functions of the coordinates and the velocities which are constant during the motion arise in the integration of the equations of motion, and we fix attention upon these constants by giving them special names.

Dynamical systems such as we have been considering consist of a fixed number of particles each of constant mass. In hydrodynamics we describe the motion of fluids and deal with elements of changing mass. It is useful to employ a quantity called the divergence of a vector, div **A**, where **A** represents a vector. Suppose that **v** is the velocity of an element of fluid. Consider a point enclosed by an element of volume; then the divergence of the velocity at the point is the time rate at which the volume of fluid is flowing out through the surface per unit volume. If ρ is the density, the rate at which the mass of fluid passes through an element of surface is $\rho\mathbf{v}$ per unit area. Then div $\rho\mathbf{v}$ is the rate of flow of mass outwards through the boundary surface per unit volume. The rate at which the mass per unit volume is decreasing at the point is $-\partial\rho/\partial t$. The conservation of matter states that the rate at which mass is leaving the volume element per unit volume is equal to the rate at which the density is decreasing,

$$\text{div } \rho\mathbf{v} = -\partial\rho/\partial t$$

$$\text{div } \rho\mathbf{v} + \partial\rho/\partial t = 0.$$

This equation, which states that the rate of change of the density of fluid at a point plus the rate of flow outwards per unit volume is zero, is called the equation of continuity. We may thus define material substance as something which is represented by a ρ and a **v** such that the equation of continuity is satisfied. In steady motion $\partial\rho/\partial t = 0$, and hence div $\rho\mathbf{v} = 0$, which means that the rate at which fluid is leaving a volume element is equal to the rate at which it is entering the volume element.

The equation of continuity is also applicable to electricity. An electric current is thought of as a motion of electrified par-

ticles so that the current density is the electric charge which passes through unit area in unit time. In the electron theory $\mathbf{i} = \rho\mathbf{v}$. Now if a steady current flows in a closed circuit

$$\operatorname{div} \mathbf{i} = 0,$$

that is, the rate at which electric charge is leaving the volume element is equal to the rate which electric charge is entering it. But if a current flows in an open circuit, such as the current which momentarily flows in a wire connecting the two plates of a condenser, div $\mathbf{i}$ is not equal to zero. div $\mathbf{i}$ is the rate at which electric charge is leaving a volume element per unit volume. It is

$$\operatorname{div} \mathbf{i} = -\frac{\partial \rho}{\partial t},$$

then

$$\operatorname{div} \mathbf{i} + \frac{\partial \rho}{\partial t} = 0.$$

Now in the chapter on electrodynamics we introduced the electric displacement $\mathbf{D}$, for which

$$\operatorname{div} \mathbf{D} = 4\pi\rho.$$

Then

$$\frac{\partial \rho}{\partial t} = \frac{1}{4\pi} \operatorname{div} \frac{\partial \mathbf{D}}{\partial t}.$$

Hence

$$\operatorname{div} \left(\mathbf{i} + \frac{1}{4\pi} \frac{\partial \mathbf{D}}{\partial t} \right) = 0$$

Maxwell introduced the concept of displacement current $\frac{1}{4\pi} \frac{\partial \mathbf{D}}{\partial t}$. The total current is the conduction current plus the displacement current, $\mathbf{i} + \frac{1}{4\pi} \frac{\partial \mathbf{D}}{\partial t}$. Thus Maxwell redefined the concept of electric current so that the current density satisfied the equation of continuity. This offers the possibility of conceiving of electricity as an indestructible substance. In general the concept of divergence is fundamental for the expression of laws of conservation in a theory of continuous phenomena.

The mind has a tendency to hypostatize constants. In the mechanical conception of nature the total mass of the universe is constant; the discovery of the mechanical equivalent of heat furnished the basis for a general principle of the conservation of energy. The conservation of matter and energy, respectively, have been interpreted in terms of an uncreatable, indestructible substance.

6. INVARIANTS

We have seen that the character of permanence of substance is expressed by the concept of constant quantity. In building up a physical theory one takes especial interest in constant functions and constructs definitions so that the quantities satisfy laws of conservation. The classical concept of substance also expressed objectivity. From our point of view the objectivity of substance is expressed by the invariance of physical quantities.

The representation of physical properties and phenomena requires a system of reference, that is, a frame of reference together with measuring rods and clocks. In general, the values of physical quantities will depend upon the system of reference—thus a body which is at rest in one system of reference will have a velocity with respect to another system which is moving with respect to the first system. A quantity which assumes the same value in both systems is called an invariant.

In Euclidean geometry the distance between two points is invariant with respect to a translation or rotation of axes. Similarly the angle between two lines is an invariant.

In classical dynamics one employs inertial systems of reference. With respect to transformations from one inertial system to another the fundamental invariants are: the distance between two points, the interval of time determined by a standard clock, the mass of a body. Two observers in uniform motion with respect to each other would employ the same numbers for lengths, times, and masses. Thus these quantities represented objective characters of the physical order. Invariant distance was the basis of the single order of space, and invariant interval of time was the basis of the single order of time. As we have already seen, constant and invariant mass was interpreted as

material substance in space and time. The energy of a conservative system was constant but not invariant.

The special theory of relativity removed the basis for objective space, objective time and material substance. In the theory of relativity, length, interval of time, and mass are not invariant. The fundamental invariant is the space-time separation between two point-events. Electric charge is invariant and the charge of an electron is constant and invariant.

In the general theory of relativity the fundamental invariant is the space-time separation of two world-points,

$$ds^2 = \Sigma g_{ik} dx_i dx_k.$$

The constant and invariant value at a space-time point is determined as

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2 dt^2$$

by means of a standard rod and the principle of the constancy of the velocity of light in a system of reference relative to which there is no gravitational field.

7. THE CHARACTER OF PHYSICAL REALITY

The concept of substance is an expression of the rational demand for unity in the representation of the physical order. In the development of physics the attempt has been made to derive all properties of bodies from the properties of a single substance. From our point of view this means that certain physical quantities are assumed to be fundamental for the representation of the physical order: in the mechanical conception of nature the physical quantity mass was assumed to be fundamental; in the electrodynamic theory electric charge and electromagnetic field are fundamental. The physical quantities, however, are representations of physical reality, and so, in order to grasp the nature of physical reality which is represented by physical concepts, we must try to experience what bodies having mass or electric charge really are, and what an electromagnetic field is.

With regard to the question as to how the property represented by mass differs from the property represented by charge, I remark that this appears to me to be an unanswerable question. I assign a mass or charge to a body on the basis of cer-

tain operations, observations of coincidences, assumptions, etc., but I have no immediate experience of the essence of matter or electricity. The difference between matter and electricity is that the number called the mass of a body is assigned by methods which differ from those by which the number called the charge is assigned, and the mass and electric charge enter into different functional relations with other quantities.

However, I believe that we have some immediate knowledge of space and time. Thus it would appear that the geometrical conception of nature should be the canonical theory, since the fundamental metrical quantities are directly correlated with experienced properties.

The geometrical point of view must first be reconciled with quantum mechanics.

In the light of contemporary theory it appears that electrodynamic concepts will be fundamental in the unitary microscopic theory, for a relatively final theory appears to be promised by quantum mechanics and electrodynamics. Now electrodynamics employs the concepts of particle and wave. Although in classical physics particles and waves permitted a perfectly definite description in terms of space-time quantities, the quantum mechanics has revealed a limitation in these concepts when applied to atomic phenomena. Electrons and electromagnetic radiation exhibit in a limited sense the characters of both particles and waves. The classical physical quantities can not be assigned simultaneously with ideal definiteness to an atomic element. This result is symbolized by the reciprocal employment of the concepts of particle and wave.

We thus arrive at a symbolic representation of physical reality. Physical quantities are symbols which may be represented by matrices. It is possible to represent a symbol by a diagonal matrix, the values of which are the possible results of a measurement. The assignment of a classical physical quantity, a number, to an atomic body or process is subject to an inaccuracy on account of the disturbance of the system by our operations. The quantum mechanics thus limits the concept of space-time coincidence which is the foundation of the geometrical conception of nature.

The quantum mechanics is a symbolic representation of the physical order. It thus carries us still further from the classical concept of substance.

REFERENCES

LAMPRECHT, S. P. Sense Qualities and Material Things. *Philosophical Review*, Vol. 38, 1929.

SCHLICK, M. Naturphilosophie, *in* Die Philosophie in ihren Einzelgebieten, edited by M. Dessoir. Berlin, 1925.

WEYL, H. Was ist Materie? Berlin, 1924.

CHAPTER XVI

THE CONCEPT OF CAUSALITY

1. THE PRIMITIVE CONCEPT OF CAUSALITY

AN essential feature of the classical concept of substance was the assumption of the self-determination and independence of substance. It has been pointed out, however, that physical characters are interdependent. In consequence, the concept of causality becomes the primary expression of unity in physical theory.

The concept of causality represents a relation between successive phenomena, such that the antecedent phenomenon is the cause of the subsequent phenomenon which is the effect. More precisely, the cause is some state or process which is the condition of some succeeding state or process. For example, force is the cause of acceleration. The state or process which is called the cause presupposes some body or system which is called the agent, and likewise the effect presupposes some body or system which is called the patient. The concept of causality thus implies a relation between a state of the agent and a state of the patient. Sometimes the term cause is applied to the agent itself—thus a body is the cause of the motion of another body. I shall first consider the primitive concept of causality and then the functional concept.

The most striking characteristic of the primitive concept is the implication of **production**. The cause is thought of as something which produces the effect. The origin of this concept undoubtedly is the consciousness in the individual of his own power to produce effects, for we have the experience of exerting a force and thereby producing an effect. In the primitive concept the state of activity of the agent is the cause which produces an effect in the patient. However, the term cause is

sometimes applied to the agent, for example, the Creator has been spoken of as the First Cause. Similarly the term effect may be applied to some body which is created by the causal agent. In this discussion of the primitive concept of causality I shall assume that the concept applies to a relation between states or processes of bodies. The primitive concept implies efficiency, production, creation.

Hume mentions **contiguity** as a character of causal connection. It seems clear that in the first instance cause is thought of as contiguous to effect, the origin of this notion being doubtless the observation that a part of our body is in contact with some body upon which we are exerting a force. A typical example of physical causality occurs in the collision of two bodies. A moving billiard ball A collides with a ball B at rest and sets it in motion. The cause is the force exerted by A and the effect is a state of motion of B. The cause produces the effect when the bodies are in contact; hence we may say that the cause as state is contiguous to the effect as state.

There are, however, apparent exceptions to the contiguity of cause and effect. Thus the luminosity of the sun is the cause of the illumination of the moon, and, again, we say that a body falls to the earth because the earth exerts a gravitational force upon it and thereby produces an acceleration—in neither of which examples do cause and effect seem to be contiguous.

Three methods have been employed to interpret causality between distant bodies:

(1) The causal process has been interpreted as pure action at a distance. During the eighteenth and early nineteenth centuries the success of the Newtonian theory of gravitation was the basis of an attempt to characterize gravitational, electrical and magnetic actions as actions at a distance. However, this point of view is completely neglected at the present time.

(2) Distant action has been interpreted as contiguous action after transport. Thus in a corpuscular theory of light a luminous body sends out corpuscles which travel through empty space and then act upon some body by contiguous action.

(3) The favorite interpretation of action at a distance has been to explain it as action through a continuous medium. The cause produces a process in a contiguous medium which

transmits the process from point to point until it affects a distant body. Thus action through a medium is contiguous causal action.

Hume also mentions **succession** as a characteristic of causality; thus the cause precedes the effect. This is evident where the causal process is transmitted through a medium with a finite velocity. It may be contended that when an action is transmitted through a medium the term causality should be applied only to the relation between the contiguous processes. I shall show, however, that even where the causal process is thought of as contiguous there is a basis for thinking of cause as preceding effect.

Consider the physical process in which a force acts upon a body. We describe this by stating that the force causes an acceleration. Now inasmuch as the force and the acceleration are simultaneous, it appears that this example shows that the cause does not precede the effect. And yet the experience of force as a cause seems to be the original basis for the concept of causality. I have the following explanation of how one could arrive at the view that the force, the cause, precedes its effect, the acceleration.

I shall assume that the body is initially at rest and that a constant force acts upon it. Then from the force equation

$$F = ma$$

we find that the acceleration a is constant. The distance passed over in time t is

$$s = \tfrac{1}{2}at^2.$$

Now if a force acts upon a body and produces motion we recognize that motion has occurred from the distance passed over. It would seem to be the primitive idea that if no distance is observed to have been passed over then there was no effect. Suppose that the force is allowed to act for 1/100 second. Then the distance passed over is

$$s = \tfrac{1}{2}a(1/100)^2 = \tfrac{1}{2}a(1/10{,}000).$$

If the time is a small quantity of the first order then the distance passed over is a small quantity of the second order. In other words, a measurable time may elapse in which no observ-

able distance is passed over; the effect lags behind the force. Hence the primitive mind would gain the impression that force is exerted before the effect, that the force has to act for some time before the effect occurs. Thus we can see why one would be led to the notion of cause as preceding the effect. In all causal processes there is an inertial factor on account of which the building up of the effect lags behind the building up of the cause. As we shall see, the causal process is precisely described by certain differential equations in which the time is at least one independent variable.

The question occurs whether the temporal element in causality would apply if we think of force as an action at a distance, for action at a distance is thought of as the instantaneous propagation of force from one body to another. However, the displacement would still lag behind the force in the way previously described. Thus even in the case of action at a distance there is significance in the statement that cause precedes the effect.

The concept of causality contains the notion that cause **equals** effect. This has been expressed by the proposition that nothing can come out of nothing; in particular M. Meyerson has stressed this aspect of the causal relation. In order to give an interpretation of the equality of cause and effect let us consider the collision of two billiard balls. At impact, A exerts a force upon B and accelerates it, and likewise B exerts an equal and opposite force upon A and accelerates it. During this interaction the total momentum of the two balls is conserved, and assuming perfectly elastic balls the total kinetic energy is also conserved. These laws of conservation express the principle that the cause equals the effect.

The regularity of physical processes is expressed by the **principle of causality** which states that the same cause produces the same effect. If a particular phenomenon produces a particular effect then, if a phenomenon with the same physical characters as the cause occurs at some other time or place, a phenomenon will occur with the same physical characters as the effect. Thus a force of one dyne will give a mass of one gram an acceleration of one centimeter per second per second independently of the position and time of the process. The

principle of causality, thus implying that causality is a relation between recurrent characters, asserts the existence of **causal laws.**

Now the physical order is very complex; for every phenomenon there is a plurality of causes. It appears that every particular phenomenon arises from a unique set of conditions. Since it seems clear that a particular set of conditions can not recur, it appears to be impossible to set up a principle of causality. One could say that the complete state of the universe at one time is the cause of the state at a later time—the causal relation would be particular since each state of the universe is unique.

The answer to the foregoing objection is that the total physical process can be analyzed into approximately independent processes that recur. Consider for example the motion of a falling body. At time t_1 it is released from a certain height and falls to the ground in a describable manner. At some later time t_2 the experiment can be repeated. The conditions are not exactly the same, but we can ignore the rest of the universe and its changes and consider the motion of the body as a process that is repeated. By a method of successive approximation we can fix the conditions of an experiment so that the course of the process is the same with respect to the relevant characters. Thus a process can be considered in isolation. There are causal strands in the physical order and therefore recurrent processes which satisfy the principle of causality.

2. THE FUNCTIONAL CONCEPT OF CAUSALITY

In the preceding analysis of causality we have seen that the primitive concept contains the anthropomorphic element of production, efficiency, creation. In the functional concept one expresses causality by means of differential equations.

The functional concept presupposes the possibility of isolation of physical processes. The notion of production is eliminated; in accordance with Hume's criticism one does not think of the cause as something which produces the effect. The functional concept expresses a correlation of characters which are represented by numerical measures. Thus causal

connection is expressed by a functional relation between measures. However, it should be noted that one still employs the language of efficient causality. In this analysis of the functional concept I shall consider the characteristics of the functional relations which express causality in the dynamics of particles, in field physics, and in quantum physics.

3. CLASSICAL DYNAMICS

A typical example of a causal process in classical dynamics is the motion of a particle projected vertically upwards in the earth's gravitational field. Let

m = mass of the particle;
s = distance measured from the surface of the earth;
t = time measured from a fixed origin.

The equation of motion is

$$m \frac{d^2s}{dt^2} = - mg,$$

where g is a constant.

Thus the functional relation which describes the causal process is a differential equation of the second order in the time as independent variable. It is to be noted that m, g, s, t are numbers, and thus we have entirely eliminated the concept of efficiency.

The applicability of a differential equation in which time is the independent variable presupposes that the motion is continuous in time. From the position and state of motion at a given instant one can calculate the position and state of motion at an infinitely near earlier or later time. The preceding statement is not satisfactory from the standpoint of modern logic, but it is the accepted mode of expression in mathematical physics.

On integrating the equation we obtain

$$s = s_0 + v_0 t - \tfrac{1}{2} g t^2.$$

Assuming $v_0 = 0,$

then $s_0 - s = \tfrac{1}{2} g t^2.$

If the time is a small quantity of the first order the distance gone from rest is a small quantity of the second order. The inertial mass induces a time lag which offers a basis for the view that cause precedes the effect. Thus the occurrence of time as the independent variable in the differential equation expresses the succession of cause and effect.

The differential equation gives a detailed analysis of the process. The equation expresses an instantaneous property of the motion. We are not able, however, to verify the differential equation directly. We therefore assume that the equation holds for a certain finite period of time and integrate the equation. Now we can measure t and s and thus experimentally verify the integral law.

In the primitive concept of causality the cause was thought of as that which produces the effect, so that the discovery of the causes of a phenomenon constituted an explanation. The question arises whether the restriction of causality to a functional relation renders the concept of explanation impossible.

The distinction between explanation and description is possible on the functional view and depends upon the definition of the system under consideration. Consider, for example, a freely falling body in the earth's gravitational field. To the first approximation the equation of motion is

$$\frac{d^2s}{dt^2} = -g.$$

Suppose that our problem is the explanation of the acceleration of a given body. Our system is then defined to be this body which has position s at time t. The variable quantities which represent the body are functionally related to g, which depends upon the mass of the earth and the distance of the given body from the center of the earth. That is, the physical quantities which describe the body depend upon a body outside the system; hence the differential equation may be said to offer an explanation of the acceleration of a body in terms of the position and mass of some other body.

But now consider that the earth and the falling body are included in a single system; then all the variables belong to one system. The differential equation may then be said to

describe the process of interaction between the parts of the system. Thus the complete analysis of the causal process shows that it is a mutual action. This was first expressed by Newton's third law of motion: action and reaction are equal and opposite.

In classical dynamics the typical system is one in which the forces acting on the particles of the system are derivable from a potential energy which is a function of the positions of the particles. Under these conditions there is an integral of energy and the system is called conservative. If the system is isolated there is also conservation of momentum and of moment of momentum. These principles of conservation express the content of the principle that cause equals effect.

The motion of conservative systems is determined by linear differential equations of the second order in which the time is the independent variable. Thus

$$\frac{d^2s}{dt^2} = -g.$$

For a particle vibrating with simple harmonic motion

$$\frac{d^2x}{dt^2} + \omega^2 x = 0.$$

The solution of the second equation is

$$x = A \cos(\omega t - \delta).$$

Now this solution has the property that if one replaces t by $-t$ the equation of motion is still satisfied—the solution which is a function of $-t$ describes the reverse motion. We have then illustrated the fact that a reverse motion is determined by the same equation as the direct motion, which is expressed by the statement that the processes of classical dynamics are reversible. If at a given moment one should reverse all the velocities of the particles of the system, the motion would proceed in the reverse direction according to the same equations. This result may also be expressed by the proposition that in the conservative systems of classical dynamics there is no distinction between the forward and backward direction of time.

In classical dynamics one also considers systems in which

there are dissipative forces which depend upon the velocities. One then obtains differential equations in which occur derivatives of the time to the first order. In consequence the motion is not reversible. In the chapter on thermodynamics it was pointed out that all macroscopic physical processes are irreversible. However, in the mechanical conception of nature irreversible macroscopic processes were analyzed into totalities of reversible microscopic process. Thus reversibility is an essential characteristic of classical dynamical causality.

Professor G. N. Lewis has proposed the general postulate of the symmetry of time for all physical processes.

We have now to formulate the functional form of the principle of causality which in qualitative terms states that the same cause produces the same effect.

In order to describe the motion of a dynamical system it is necessary to have a frame of reference with respect to which one determines position in space, and an origin of time with respect to which one determines position in time. The equations of motion are differential equations in the coordinates with respect to the time. Assuming f degrees of freedom there will be f differential equations of the second order. The solutions of these equations will express the coordinates as functions of $2f$ arbitrary constants and the time; the arbitrary constants are determined by the initial conditions, that is, the values of the coordinates and the corresponding velocities (or momenta) at a given time t_0. The initial conditions are referred to the chosen frame of reference and the origin of time.

The principle of causality now states that if the frame of reference is transported in space and the origin of time is transported in time, then if the same initial conditions are realized relative to the new frame of reference and origin of time, the same motion occurs relative to the new frame and origin of reference. That is, space and time are homogeneous with respect to the laws of classical dynamics. In other words, given the same relative initial conditions the same motion will occur at some other place and time.

The principle of causality is expressed by the invariance of the representation of physical laws with respect to a transformation in space and time. In accordance with the principle

of relativity the representation of laws is invariant with respect to the Galilean transformation from an inertial system to one moving with constant velocity with respect to the first. One may formulate the principle of relativity so that it includes what has been called the principle of causality.

3. THE PHYSICS OF FIELDS

I turn now to the concept of causality in the physics of fields. A field may be defined as a region of space such that at every point there are conditions which are described by physical quantities which are functions of the position and the time, examples being the gravitational field, the electromagnetic field, etc. The quantities which determine the field are scalars, vectors and tensors. The rate of change in space and time of the quantities which describe the field is determined by certain partial differential equations, such as are exemplified by the differential equations of the theory of elasticity, of hydrodynamics, and of the electromagnetic field. These equations describe a process which is continuous in space and time. From the state at a given point at a given instant one can determine the state at an infinitely near point after an infinitesimal lapse of time; thus the equations express the contiguity of the causal process in space and time. From the equations of the theory of elasticity, or of the electromagnetic field, one can deduce a wave equation which describes the propagation of a disturbance with a finite velocity. This expresses the primitive idea that cause precedes the effect. The solution of the partial differential equation is subject to specified initial conditions and boundary conditions.

The principle of causality can now be expressed as follows. If the frame of reference is transported in space and the origin of time is transported in time, then if the same initial conditions and boundary conditions are realized relative to the new frame and origin, the same processes occur. Suppose, for example, that at a certain place in space there is a box which is equipped with a physical laboratory. Suppose that at one o'clock the physical conditions within the box are completely known, and suppose that the physical conditions on the sur-

face are controlled from one to two o'clock. Then the physical processes within the box are uniquely determined from one to two o'clock. If the same conditions are reproduced at any time and at any place the same series of events will occur.

From the mathematical point of view the principle of causality expresses the fact that the initial conditions and boundary conditions determine a unique solution of the differential equation.

4. QUANTUM PHYSICS

Lastly I consider the concept of causality in quantum physics.

As we have seen, quantum physics deals with microscopic, that is, elementary processes. The classical kinetic theory assumed determinism for elementary processes. The invariable causal laws of macroscopic dynamics were extended to the microscopic realm. In quantum physics, however, a statistical causality is assumed to hold for elementary processes. The initial conditions of an atomic system are not precisely determinable. The elementary quantum processes are symbolized by a probability function. However, the probability function is determined by a differential equation of the type of the classical physics of fields. Hence if the initial values of the probability function are known throughout a region and if the values over the boundary are controlled during a certain interval of time, then the probability funtion is determined throughout the region during the given time interval.

The same sequence of values of the probability function is determined if the same relative initial conditions and values over the boundary are given for some other region at some other time. Thus the mode of determination of the probability function expresses a principle of causality.

REFERENCES

JORDAN, P. Kausalität und Statistik in der modernen Physik. *Die Naturwissenschaften*, Vol. 15, p. 105, 1927.

LEWIS, G. N. The Symmetry of Time in Physics. *Science*, Vol. 71, 1930.

PAINLEVÉ, P. Les axiomes de la mécanique. Paris, 1922.

WEYL, H. Philosophie der Mathematik und Naturwissenschaft. Munich und Berlin, 1927.

INDEX OF PROPER NAMES

(The numbers refer to the pages)

ABRAHAM, M., 132, 177, 191
ABRAHAM and FÖPPL, 132
APPELL, P., 118
ARCHIMEDES, 47

BALMER, J., 244
BERKELEY, 5, 73
BERNOULLI, DANIEL, 161
BIGGS, H. F., 268
BIOT and SAVART, 127
BIRTWISTLE, G., 251
BLISS, G. A., 49
BOHR, N., 242, 267, 268
BOLTZMANN, L., 167
BORN, M., 190, 191, 247, 251, 258, 268
BOTHE and GEIGER, 242
BRIDGMAN, P. W., 30, 49, 146
BROAD, C. D., 65, 74

CARNAP, R., 4, 12, 26, 49
CARNOT, S., 147
CLAUSIUS, R., 147
COMPTON, A. H., 242
COMPTON and SIMON, 242
CONDON, E., and MORSE, P., 268
COPERNICUS, 91
COULOMB, C., 155
COUTURAT, L., 49

D'ALEMBERT, J., 115
DALTON, J., 161
DAVISSON and GERMER, 259
DAVY, SIR HUMPHRY, 141
DE BROGLIE, L., 258, 268
DEMOCRITUS, 51, 161
DESCARTES, 54, 231, 276
DIRAC, P. A. M., 10, 12, 268
DUHEM, P., 49

EDDINGTON, A. S., 29, 49, 230
EHRENFEST, P., 177, 248
EHRENFEST, P. and T., 177
EINSTEIN, A., 84, 92, 141, 189, 195, 206, 208, 224, 230, 239, 241
EUCLID, 58, 64

FARADAY, M., 129, 176, 179
FERMAT, P., 73
FITZGERALD, G., 195
FOUCAULT, J., 159
FOURIER, J., 249
FRANCK and HERTZ, 246
FRANKLIN, B., 120
FRESNEL, A., 159

GALILEO, 5, 78, 84, 90, 134, 276
GASSENDI, P., 161
GAUSS, K., 210, 219
GLAZEBROOK, R. T., 177

HAAS, A. E., 118
HAMILTON, W., 116
HEISENBERG, W., 10, 12, 252, 267, 268
HELMHOLTZ, H., 5, 180
HERTZ, H., 179
HERTZ, P., 132, 170, 177
HILBERT, D., 66
HOOKE, R., 100
HOUSTOUN, R. A., 176
HUME, D., 5, 286
HUNTINGTON, E. V., 24, 74
HUYGHENS, C., 159

JACOBI, C., 261
JEANS, J. H., 165, 177
JOFFÉ, A., 191
JORDAN, P., 251, 258, 268, 295
JOULE, J., 141

KANT, 11, 44, 51, 65, 67, 75
KELVIN, 147, 148

KEPLER, 91
KIRCHHOFF, G., 238
KLEIN, F., 74

LAGRANGE, 115
LAMPRECHT, S. P., 276, 284
LAPLACE, 155
LAUE, M. VON, 216, 222
LEIBNIZ, 51, 54, 75
LESAGE, 155
LEWIS, C. I., 11, 12
LEWIS, E. P., 177
LEWIS, G. N., 150, 293, 295
LEWIS and RANDALL, 150
LOCKE, 276
LORENTZ, H. A., 181, 182, 189, 191
LOSCHMIDT, J., 168
LOVEJOY, A. O., 9, 12

MACH, E., 4, 5, 253
MAGIE, W. F., 132, 177
MASON and WEAVER, 132, 183
MAUPERTUIS, 260
MAXWELL, J. C., 128, 167, 176, 182, 280
MENDENHALL, C. E., 150
MEYERSON, E., 288
MICHELSON and MORLEY, 194
MIE, G., 189, 277
MILLIKAN, R. A., 181, 241

NEWTON, 5, 47, 91, 97, 111, 154, 159, 226

OERSTED, H., 127
OSGOOD, W. F., 49
OSTWALD, W., 6

PAINLEVÉ, P., 295
PAULI, W. JUN., 206
PLANCK, M., 5, 19, 150, 177, 237
POINCARÉ, H., 168, 189

REICHENBACH, H., 56, 74
RIEMANN, B., 219
RITZ, W., 195
ROENTGEN, W., 180
RUMFORD, COUNT, 140
RUSSELL, B., 21, 48, 49, 65, 274, 275
RUTHERFORD, E., 243

SCHLICK, M., 284
SCHRÖDINGER, E., 258, 268
SOMMERFELD, A., 247, 253, 268
STERN and GERLACH, 246
SWANN, W. F. G., 29, 41, 187, 191

THALES, 119
THIRRING, H., 206, 268
THOMSON, G. P., 260
THOMSON, J. J., 188, 242
TOLMAN, R. C., 38, 49, 177

VEBLEN, O., 74

WEYL, H., 231, 268, 277, 284, 295
WHITEHEAD, A. N., 49, 65, 80
WHITTAKER, E. T., 118
WILSON, W., 247

YOUNG, J. W., 74
YOUNG, THOMAS, 158

ZEEMAN, P., 181
ZERMELO, E., 168

INDEX OF SUBJECTS

(The numbers refer to the pages)

Acceleration, 88
Action, 239
  principle of least, 260
A priori, 11, 21, 43
Aspect, 3, 15
Atomistic theory of electricity, 179 *ff.*

Balmer series, 244
Body, 15
Bohr atom, 242

Carnot cycle, 143
Carnot's theorem, 148
Cartesian coordinates, 30
Causality, 10, 285 *ff.*
  functional concept of, 289
  primitive concept of, 285
  principle of, 288
Causal law, 175, 289
Coincidence, 18
Combination principle, 245
Concepts, origin of, 42
Conductivity, electrical, 126
Conservation of energy, 113, 279
Constants, 277
Corpuscular theory of light, 159, 242
Correspondence principle, 249
Coulomb's law, 122
Curl of a vector, 39
Curvature, 211

d'Alembert's principle, 115
Definition, 12, 46, 50, 57
Derivative, 33
Determinism, 9, 175
Dielectric constant, 124
Differential equation, 34
Displacement current, 128
Divergence of a vector, 38
Dynamics, 96 *ff.*
  fundamental principle of, 115

Elasticity, 156
Electric charge, 121
Electric current, 125
Electric displacement, 123
Electric field, 121
Electrification, 119
Electrodynamic conception of nature, 178 *ff.*
Electrodynamic theory of mechanical force, 190
Electrolysis, 179
Electromagnetic field, First equation of the, 127
  Second equation of the, 129
Electromagnetic mass, 188
Electron, 181
Electron theory, 182 *ff.*
Electrostatics, 119 *ff.*
Energy, 112, 279
  internal, 145
Entropy, 149
Equivalence, principle of, 224
Euclidean, geometry, 50 *ff.*, 208
Event, 18, 75
Experimental physics, 40

Force, kinetical concept of, 105
  statical concept of, 97

Gaussian coordinates, 212
Gauss' theorem, 122
Generalized coordinates, 115
Geodesic line, 211
Geometry, abstract, 65
  analytic, 58
  Euclidean, 50 *ff.*
  ideal, 64
Gravitation, 154
  Einstein's law of, 229
Group, 62

Hamilton-Jacobi equation, 261
Hamilton's canonical equations, 116
Hamilton's principle, 117
Hooke's law, 100

Indeterminacy, principle of, 266
Indeterminism, 9, 266, 295
Induction, magnetic, 129
Integral, 34
Invariant representation of physical laws, 31
Invariants, 281
Irreversible process, 143, 147

Kinematics, 85 *ff.*
Kinetic theory, of gases, 162
of matter, 160
Knowledge, theory of, 10

Lagrange's equations, 116
Length, 28
Lever, principle of the, 99
Light, 72, 158
Light quantum, 241
Lorentz-transformation, 199

Magnet, 124
Magnetism, 124
Magnetization, 186
Magnitude, 22
Mass, 28, 110
Matrix, 253
Matrix mechanics, 253 *ff.*
Measurement, 19, 55
Mechanical conception of nature, 153 *ff.*
Mechanical theory of heat, 140
Metaphysics, 4
Metrical tensor, 215
Michelson-Morley experiment, 194
Microscopic electrodynamics, 178 *ff.*
Microscopic entities, 6, 161, 179
Momentum, 105

Newton's laws of motion, 105

Ohm's law, 126

Parallel displacement of a vector, 220
Physical reality, character of, 283
criterion of, 3
Polarization, 185
Positivism, 5
Potential, 121
Proton, 181

Quantity, 22, 29
of heat, 137
Quantum hypothesis, 237
Quantum theory, 9, 235 *ff.*

Relativity, principle of, 113, 193, 197
special theory of, 9, 192 *ff.*
Resistance, electrical, 126
Reversibility, 117
Riemannian geometry, 207 *ff.*

Scalar, 30
Schrödinger's equation, 262
Simultaneity, 9, 18, 76
Sound, 157
Space, 18, 50 *ff.*
structure of, 63
Space-time, 203
Specific heat, 137
Speed, 86
Stationary state, 246
Statistical interpretation of quantum mechanics, 265
Statistical law, 9, 175
Statistical mechanics, 169
Statistical method, 162
Substance, 120, 138, 274 *ff.*
Successive approximation, 46
Successive definition, 47
Surfaces, theory of, 210
System of reference, transformation of, 91

Temperature, 133, 148
Tensor, 30
Theoretical physics, 40
Thermodynamics, 142
First law of, 144
Second law of, 146, 174
Thermometer, 134
Thing, 3, 17
Time, 9, 17, 75, 195
Truth, criterion of, 271

Universal, 16

Vector, 30
Vector analysis, 35
Velocity, 89
  of light, principle of the constancy of, 197
Virtual displacements, principle of, 103

Wave mechanics, 258
Wave motion, 93
Wave theory of light, 159
Weight, 99
Work, 37, 101