1

How the theory of relativity came into being (a brief historical sketch)

1.1 Special versus general relativity

The name 'relativity' covers two physical theories. The older one, called special relativity, published in 1905, is a theory of electromagnetic and mechanical phenomena taking place in reference systems that move with large velocities relative to an observer, but are not influenced by gravitation. It is considered to be a closed theory. Its parts had entered the basic courses of classical mechanics, quantum mechanics and electrodynamics. Students of physics study these subjects before they begin to learn general relativity. Therefore, we shall not deal with special relativity here. Familiarity with it is, however, necessary for understanding the general theory. The latter was published in 1915. It describes the properties of time and space, and mechanical and electromagnetic phenomena in the presence of gravitational field.

1.2 Space and inertia in Newtonian physics

In the Newtonian mechanics and gravitation theory the space was just a background – a room to be filled with matter. It was considered obvious that the space is Euclidean. The masses of matter particles were considered their internal properties independent of any interactions with the remaining matter. However, from time to time it was suggested that not all of the phenomena in the Universe can be explained using such an approach. The best known among those concepts was the so-called Mach principle. This approach was made known by Ernst Mach in the second half of the nineteenth century, but had been originated by the English philosopher Bishop George Berkeley, in 1710, while Newton was still alive. Mach began with the following observation: in the Newtonian mechanics a seemingly obvious assumption is tacitly made, namely that all the space points can be labelled, for example by assigning Cartesian coordinates to them. One can then observe the motion of matter by finding in which point of space a given particle is located at a given instant. However, this is not actually possible. If we accept another basic assumption of Newton, namely that the space is Euclidean, then its points do not differ from one another in any way. They can be labelled only by matter being present in the space. In truth, we thus can observe only the motion of one portion of matter relative to another portion of matter. Hence, a correctly formulated theory should speak only about relative motion

(of matter relative to matter), not about absolute motion (of matter relative to space). If this is so, then the motion of a single particle in a totally empty Universe would not be detectable. Without any other matter we could not establish whether the lone particle is at rest, or is moving or experiencing acceleration. But the reaction of matter to acceleration is the only way to measure its inertia. Hence, that lone particle would have zero inertia. It follows then that inertia is, likewise, not an absolute property of matter, but is induced by the remaining matter in the Universe, supposedly via the gravitational interaction.

One can question this principle in several ways. No-one will ever be able to find him/herself in an empty Universe, so any theorems on such an example cannot be verified. It is possible that the inertia of matter is a 'stronger' property than the homogeneity of space, and would still exist in an empty Universe, thus making it possible to measure absolute acceleration. Criticism of Mach's principle is made easier by the fact that it has never been formulated as a precise physical theory. It is just a collection of critical remarks and suggestions, partly based on calculations. It happens sometimes, though, that a new way of looking at an old theory, even if not sufficiently justified, becomes a starting point for meaningful discoveries. This was the case with Mach's principle that inspired Einstein at the starting point of his work.

1.3 Newton's theory and the orbits of planets

In addition to the above-mentioned theoretical problem, Newton's theory had a serious empirical problem. It was known already in the first half of the nineteenth century that the planets revolve around the Sun in orbits that are not exactly elliptic. The real orbits are rosettes – curves that can be imagined as follows: let a point go around an ellipse, but at the same time let the ellipse rotate slowly around its focus in the same direction (see Fig. 1.1). Newton's theory explained this as follows: an orbit of a planet is an exact ellipse only if we assume that the Sun has just one planet.[1] Since the Sun has several planets, they interact gravitationally and mutually perturb their orbits. When these perturbations are taken into account, the effect is *qualitatively* the same as observed.

However, in 1859, Urbain J. LeVerrier (the same person who, a few years earlier, had predicted the existence of Neptune on the basis of similar calculations) verified whether the calculated and observed motions of Mercury's perihelion agree. It turned out that they do not – and the discrepancy was much larger than the observation error. The calculated rate of perihelion shift was smaller than the one observed by 43 arc seconds per century (the modern value is $43.11 \pm 0.45''$ per century (Will, 2018)). Astronomers and physicists tried to explain this effect in various simple ways, e.g. by assuming that yet another planet, called Vulcan, revolves around the Sun inside Mercury's orbit and perturbs it; by allowing for gravitational interaction of Mercury with the interplanetary dust; or by assuming that the Sun is flattened in consequence of its rotation. In the last case, the gravitational field of the Sun would not be spherically symmetric, and a sufficiently large flattening would

[1] More assumptions were actually made, but the other ones seemed so obvious at that time that they were not even mentioned: that the Sun is exactly spherical, and that the space around the Sun is exactly empty. None of these is strictly correct, but the departures of observations from theory caused by the nonsphericity of the Sun and by the interplanetary matter are insignificant.

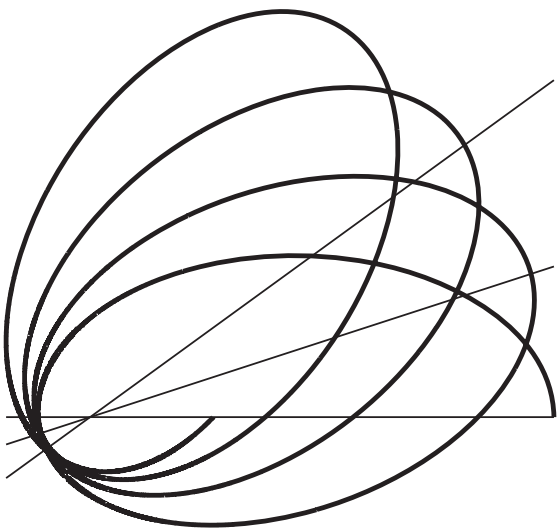

Fig. 1.1 Real planetary orbits, in consequence of various perturbations, are not ellipses but non-closed curves. The perihelion shift in this figure is greatly exaggerated. In reality, the greatest angle of perihelion shift observed in the Solar System, for Mercury, equals approximately 1.5° per 100 years.

explain the additional rotation of Mercury's perihelion. All these hypotheses did not pass the observational tests. The hypothetical planet Vulcan would have to be so massive that it would be visible in telescopes, but wasn't. There was not enough interplanetary dust to cause the observed effect. The Sun, if it were sufficiently flattened to explain Mercury's motion, would cause yet another effect: the planes of the planetary orbits would swing periodically around their mean positions with an amplitude of about 43″ per century, and that motion would have been observed, but wasn't (Dicke, 1964).

In spite of these difficulties, nobody doubted the correctness of Newton's theory. The general opinion was that Mach's critique would be answered by formal corrections in the theory, and the anomalous perihelion motion of Mercury would be explained by new observational discoveries. Nobody expected that any other gravitation theory could replace Newton's, which had been going from one success to another for over 200 years. General relativity was not created in response to experimental or observational needs. It resulted from speculation, it preceded all but one of the experiments and observations that confirmed it, and it became broadly testable only about 50 years after it had been created, in the 1960s. So much time did technology need to catch up and go beyond the opportunities provided by astronomical phenomena.

1.4 The basic assumptions of general relativity

It is interesting to follow the development of relativistic ideas in the same order as that in which they actually appeared in the literature. However, this was not a straight and smooth road. Einstein made a few mistakes and put forward a few hypotheses that he had

to revoke later. He had been constructing the theory gradually, while at the same time learning the Riemannian geometry – the mathematical basis of relativity. If we followed that gradual progress, we would have to take into account not only some blind paths but also competitors of Einstein, some of whom questioned the need for the (then) new theory, while some others tried to get ahead of Einstein, but without success (Mehra, 1974). Learning relativity in this way would not be efficient, so we will take a shortcut. We shall begin by justifying the need for this theory, then we shall present the basic elements of Riemann's geometry and then we will present Einstein's theory in its final shape. The history of relativity's taking shape is presented in Mehra's book (Mehra, 1974), and its original presentation is to be found in the collection of classic papers (Einstein et al., 1923).

Einstein's starting point was a critique of Newton's theory based on Mach's ideas. Newtonian physics said that in a space free of any interactions, material bodies would either remain at rest or would move by uniform rectilinear motion. Since, however, the real Universe is permeated by gravitational fields that cannot be shielded, all bodies in the Universe move on curved trajectories in consequence of gravitational interactions.

There is a problem here. When we say that a trajectory is curved, we assume that we can define a straight line. But how can we do this when no real body follows a straight line? The terrestrial standards of straight lines are useful only because no distances on the Earth are truly great, and at short distances the deformation of 'rigid' bodies due to gravitation is unmeasurably small. Maybe then the trajectory of a light ray would be a good model of a straight line?

To see whether this could be the case, consider two Cartesian reference systems K and K', whose axes (x, y, z) and (x', y', z') are, respectively, parallel. Let K be inertial, and let K' move with respect to K along the z-axis with acceleration $g(t, x, y, z)$. Let the origins of both systems coincide at $t = 0$. Then

$$x' = x, \qquad y' = y, \qquad z' = z - \int_0^t \mathrm{d}\tau \int_0^\tau \mathrm{d}s g(s, x, y, z).$$

Hence, the equations of motion of a free particle, that in K are

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = \frac{\mathrm{d}^2 y}{\mathrm{d}t^2} = \frac{\mathrm{d}^2 z}{\mathrm{d}t^2} = 0,$$

in K' assume the form

$$\frac{\mathrm{d}^2 x'}{\mathrm{d}t^2} = \frac{\mathrm{d}^2 y'}{\mathrm{d}t^2} = 0, \qquad \frac{\mathrm{d}^2 z'}{\mathrm{d}t^2} = -g(t, x, y, z).$$

The quantity that we interpreted in K as acceleration would be interpreted in K' as the intensity of a gravitational field (with opposite sign). The gravitational field can thus be simulated by accelerated motion, or, more exactly, the gravitational force is simulated by the force of inertia. If so, then light in a gravitational field should behave similarly to when it is observed from an accelerated reference system.

How would we see a light ray in such a system? Imagine a space vehicle that flies across a light ray. Let the light ray enter through the window W and fall on a screen on the other side of the vehicle (see Fig. 1.2). If the vehicle were at rest, then the light ray entering at W would hit the screen at the point A. Since the vehicle keeps flying, it will move a bit

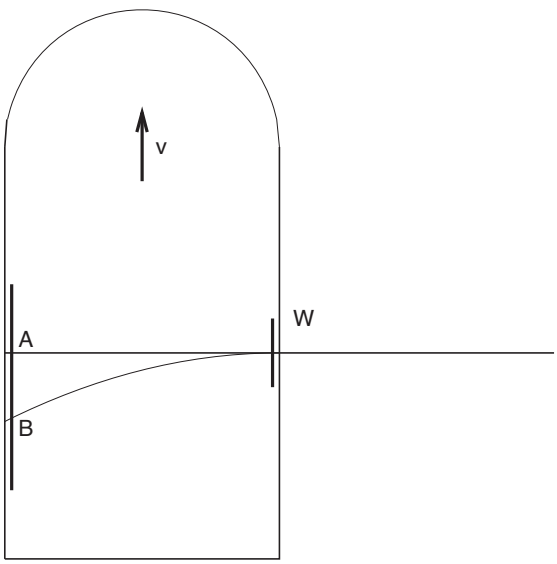

Fig. 1.2 A space vehicle flying across a light ray. See the explanation in the text.

before the ray hits the screen, and the bright spot will appear at the point B. Now assume that the light ray indeed moves in a straight line when observed by an observer who is at rest. Then it is easy to see that the path WB will be straight when the vehicle moves with a constant velocity, whereas it will be curved when the vehicle moves with acceleration. Hence, if the gravitational field behaves analogously to the field of inertial forces, then the light ray should be deflected also by gravitation. Consequently, it cannot be the standard of a straight line.

If we are unable to provide a physical model of a fundamental notion of Newton's physics, let us try to do without it. Let us assume that no such thing exists as 'gravitational forces' that curve the trajectories of celestial bodies, but that the geometry of space is modified by gravitation in such a way that the observed trajectories are paths of free motion. Such a theory might be more complicated than Newton's in practical instances, but it will use only such notions as are related to actual observations, without an unobservable background of the Euclidean space.

A modified geometry means non-Euclidean geometry. A theory created in order to deal with broad classes of non-Euclidean geometries is differential geometry. It is the mathematical basis of general relativity, and we will begin by studying it.

Part I

Elements of differential geometry

2

A short sketch of 2-dimensional differential geometry

2.1 Constructing parallel straight lines in a flat space

The classical Greek geometric constructions, with the help of rulers and compasses, fail over large distances. For example, if we wish to construct a straight line parallel to the momentary velocity of the Earth that passes through a given point on the Moon, compasses and rulers do not help. What method might work in such a situation? For the beginning, let us assume that great distance is our only problem – that we live in a space without gravitation, so we can use a light ray or the trajectory of a stone shot from a sling as a model of a straight line.

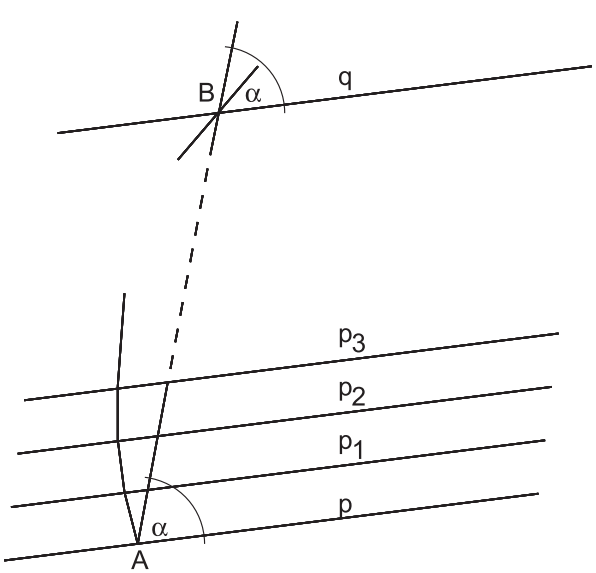

Fig. 2.1 Constructing parallel straight lines at a distance in a flat space. See the explanation in the text.

Assume that an observer is at the point A (see Fig. 2.1) on the straight line p, and wants to construct a straight line parallel to p through the point B. The following programme is 'technically realistic': we first determine the straight line passing through both A and B (for example, by directing a telescope towards B), then we measure the angle α between

the lines p and AB, then, from B, we construct a straight line q that is inclined to AB at the same angle α *and lies in the same plane as* p *and* AB. The second condition requires that we can control points of q other than B, and it can pose some problems. However, if our observer is able to construct parallel straight lines that are not too distant from the given one, he/she can carry out the following operation: the observer moves from A to A_1, constructs a straight line $p_1 \parallel p$, then moves on to A_2, constructs a straight line $p_2 \parallel p_1$, etc., until, in the n-th step, he/she reaches B and constructs $q = p_n \parallel p_{n-1}$ there.

This construction can be generalised. The observer does not have to move from A to B on a straight line. He/she can start from A in an arbitrary direction and, at a point A_1, construct a straight line parallel to p; it has to lie in the plane pAA_1 and be inclined to AA_1 at the same angle as p. Then, from A_1 the observer can continue in still another direction and at a point A_2 repeat the construction: a straight line p_2 has to lie in the plane $p_1A_1A_2$ and be inclined to A_1A_2 at the same angle as p_1 was. When the broken line he/she is following reaches B, the last straight line will be the desired one.

We can imagine broken lines whose straight segments are becoming still shorter. In the limit, we conclude that we would be able to carry out this construction along an arbitrary differentiable curve. The plane needed in the construction will be in each step determined by the tangent vector of the curve and the last straight line we had constructed.

In this way, we arrived at the idea of constructing parallel straight lines by parallely transporting directions. Note that a straight line is privileged in this construction: it is the only one to which the parallely transported direction is inclined always at the same angle. In particular, a parallely transported vector tangent to a straight line remains tangent to it at every point. A straight line can be defined by this property, provided we are able to independently define what it means to be parallel. One possible definition is this: a vector field $\mathbf{v}(x)$ along a curve $C \subset \mathbb{R}^n$ consists of parallel vectors when there exists a coordinate system such that $\partial v^i/\partial x^j \equiv 0$.

2.2 Generalisation of the notion of parallelism to curved surfaces

On a curved surface, the analogue of a straight line is a geodesic line. This is a curve whose arc PQ (see Fig. 2.2) is the shortest among all curved arcs connecting P and Q. Note that the vector tangent to a curve on a curved surface S is not a subset of this surface. The collection of all vectors tangent to S at a point $P \in S$ spans a plane tangent to S at P.

On a curved surface S, parallel transport is defined as follows. Suppose that we are given the pair of points P and Q, an arc of a curve C connecting P and Q and a vector tangent to S at P that we plan to parallely transport to Q. If C is a geodesic, then we transport the vector v along it in such a way that it is everywhere inclined to the tangent vector of C at the same angle. If C is not a geodesic, then we proceed as follows:

1. We divide the arc PQ into n segments.
2. We connect the ends of each arc by a geodesic.
3. We transport $\mathbf{v}$ parallely along each geodesic arc.
4. We calculate the result of this operation as $n \to \infty$.

It is easy to note that the parallel transport thus defined depends on the curve along which the transport was carried out. For example, consider a sphere, its pole C and two

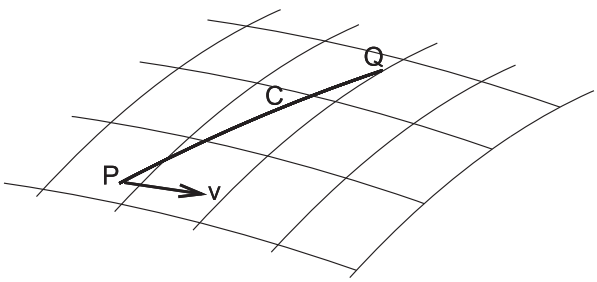

Fig. 2.2 Parallel transport of vectors on a curved surface. See the explanation in the text.

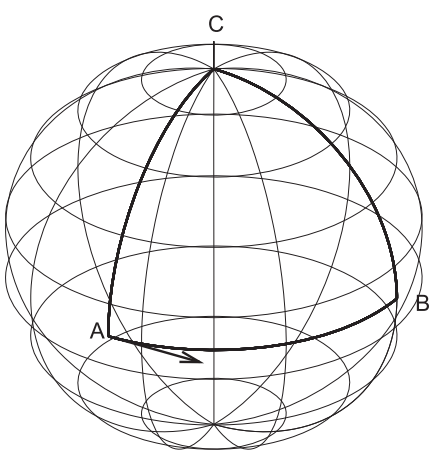

Fig. 2.3 Parallel transport of vectors on a sphere. See the explanation in the text.

points A and B lying on the equator, 90° away from each other (Fig. 2.3). Let $\mathbf{v}$ be the vector tangent to the equator at A. Transport $\mathbf{v}$ parallely to C along the arc AC, and then again along the arcs AB and BC. All three arcs are parts of great circles, which are geodesics, so $\mathbf{v}$ makes always the same angle with the tangent vectors of the arcs. The first transport will yield a vector at C that is tangent to BC, while the second one will yield a vector at C perpendicular to BC. In consequence, if we transport (in differential geometry one says 'drag') a vector along a closed loop, we will not obtain the same vector that we started with. The curvature of the surface is responsible for this. The connection between the initial vector, the final vector and the curvature is rather complicated; we will come to it further on.

We have discussed 2-dimensional surfaces in this chapter in order to visualise things more easily. However, this gave us an unfair advantage: on a 2-dimensional surface, the direction inclined to a given tangent vector at a given angle is uniquely determined. In spaces of higher dimension we will need a definition of 'parallelism at a distance' that will be analogous to $\partial v^i/\partial x^j = 0$ that we used in a flat space.

3
Tensors, tensor densities

3.1 What are tensors good for?

In Newtonian physics, a preferred class of reference systems is used. They are the inertial systems – those in which the three Newtonian principles of dynamics hold true. However, it may be difficult to identify them in practice. As we have seen in Chapter 1, it may not be easy to decide whether a given object moves with acceleration or remains at rest in a gravitational field. Hence, laws of physics should be formulated in such a way that no reference system is privileged. The choice of a reference system, even when it is evidently convenient (like, e.g., the centre of mass system), is an act of human will, while the laws of physics should not depend on our decisions.

Tensors are objects defined so that no reference system is privileged. For the beginning, we will settle for a vague definition that we will make precise later. Suppose we change the coordinate system in an n-dimensional space from $\{x^\alpha\}$, $\alpha = 1, 2, \ldots, n$, to $\{x^{\alpha'}\}$, $\alpha' = 1, 2, \ldots, n$. A tensor is a collection of functions on that space that changes in a specific way under such a coordinate transformation. The appropriate class of spaces and the 'specific way' in which the functions change will be defined in subsequent sections.

3.2 Differentiable manifolds

As already stated, in relativity we will be using non-Euclidean spaces. The most general class of spaces that we will consider are **differentiable manifolds**. This is a generalisation of the notion of a curved surface for which a tangent plane exists at every point of it. An n-dimensional differentiable manifold of class p is a space M_n in which every point x has a neighbourhood $\mathcal{O}_x$ such that the following conditions hold:

1. There exists a one-to-one mapping κ_x of $\mathcal{O}_x$ onto a subset of $\mathbb{R}^n$, called a **map** of $\mathcal{O}_x$. The coordinates of the image $\kappa_x(x)$ are called the coordinates of $x \in M_n$.
2. If the neighbourhoods $\mathcal{O}_x$ and $\mathcal{O}_y$ of $x, y \in M_n$ have a nonempty intersection ($\mathcal{O}_x \cap \mathcal{O}_y \neq \emptyset$), κ_x is a map of $\mathcal{O}_x$ and κ_y is a map of $\mathcal{O}_y$, then the mappings $(\kappa_y \circ {\kappa_x}^{-1})$ and $(\kappa_y \circ {\kappa_x}^{-1})^{-1}$ are mappings of class p of $\mathbb{R}^n$ into itself.

A **tangent space** to the manifold M_n at the point x is a vector space spanned by vectors tangent at x to curves in M_n that pass through x.