1

Introduction

From the earliest days of history mankind has shown an avid interest in the heavenly phenomena, and astronomers have good reasons to claim that theirs is the oldest profession of the world but one. This interest arose largely from practical needs. In a differentiated society where rituals play an important role it is useful to know the direction of the North and to be able to predict the turn of the seasons, days of festivities, and so on. Astronomy was still tightly interwoven with religion and astrology. The Babylonians had an extensive knowledge of practical mathematics and astronomy. The two important issues were the *calendar* (i.e. the question of the relative length of the year, months, days and the time of important feast-days), and the *ephemeris* (the positions of the Sun, Moon and the planets, lunar and solar eclipses, etc., as a function of time). In parallel to this practical knowledge, a whole variety of mythological ideas developed about the origin of the world around us. It is a peculiar coincidence that the Hindus arrived at time scales close to what we now think to be the age of the universe. The Hindus believed in a cyclic universe. It was created by Brahma, and exists in an orderly state for a period of one Brahma day (4.32×10^9 year).[1] At the end of the day Brahma will go to rest, and the universe will turn into chaos. Light, orderly motion and life only return when Brahma wakes up again. Ultimately Brahma himself will die, and the universe and the Hindu pantheon will perish with him. A new Brahma will then be born, and the endless cycle of creation and destruction will repeat itself.

The Greek were the first to develop rational concepts about the world. According to Pythagoras and his followers (ca. 500 B.C.) the Earth is spherical. The Sun, Moon and planets reside on concentric spheres revolving around the central fire Hestia. The stars are located on the outermost sphere. The idea that the Earth is not at the centre of the universe is therefore very old. Eudoxus (about 408-355) and Aristotle (384 - 322) developed a spherical world model consisting of a great number of concentric spheres with the

[1] Thomas, P.: 1975, *Hindu religion, customs and manners*, Taraporevala Sons & Co, Bombay.

Earth located at the centre. Each celestial body (Sun, Moon, and the five known planets) has a set of spheres associated with it, and is located on the innermost sphere of its own set. Each sphere of a set revolves around an axis attached to the sphere directly within. Because the axes of the spheres are not aligned, the apparent motions of the planets could be reproduced approximately. To the Greek, esthetic considerations played an important role, and this trend has persisted in physics to this day because it is often productive ('a theory is plausible *because* it is elegant'). Religious aspects played a role as well, and this has also lingered on for a very long time (cf. for example Newton). And haven't we all at times been overwhelmed by the beauty of the night sky – a strong emotional experience bordering to a religious experience? In a letter to his brother Theo, Vincent van Gogh wrote '.. It does not prevent me from having a terrible need of – shall I say the word – of religion, then I go outside in the night to paint the stars ..' [2]

Based on Babylonian observations Hipparchus (ca. 190 - 125) catalogued some 850 stars and their positions. He also invented the concept of epicycles to explain the brightness variations associated with the apparent motion of the planets. It should be kept in mind that in those days stars and planets were regarded as independent light sources of a divine nature, and that only the Earth and the Moon were thought to be lit by the Sun. The insight that the Earth and the planets are actually comparable objects came much later. Geocentric world models with epicycles were gradually refined. Ptolemy $(87 - 150)$ recorded his version in the Almagest[3], a summary of ancient astronomy and one of the most influential texts in the development of Western thinking. Much earlier, Aristarchus (ca. $310 - 230$) had proposed a simpler, truly heliocentric model with the Earth rotating around its axis and around the Sun. He was therefore 1800 years ahead of his time, but his ideas did not prevail. The history of astronomy would arguably have been quite different if they had, and this example may serve as a consolation for those who feel that the world does not hear their voice. The heliocentric theory became gradually accepted only after the publication of the work of Copernicus in 1543. For more information on these matters see Koestler (1959), Dijsterhuis (1969), Pannekoek (1989), Evans (1998), and Bless (1995).

The transition from a geocentric to a heliocentric world model meant that mankind had to give up its privileged position at the centre of the universe. This development continued well into the last century, one might say, until Hubble proved in 1924 that the spiral nebulae are actually galaxies located far outside our own galaxy, as Kant had already postulated in 1755. As a result, our galaxy became one among many. This led to the formulation of the *cosmological principle*, which says that our position in the universe is in no way special – the complete antithese of the geocentric view.

---

[2] J. van Gogh-Bonger (ed.), *Verzamelde brieven van Vincent van Gogh*, Wereldbibliotheek, Amsterdam (1973), Vol III, letter 543, p. 321.

[3] From the Arabic-Greek word Kitab al-megiste, the Great Book.

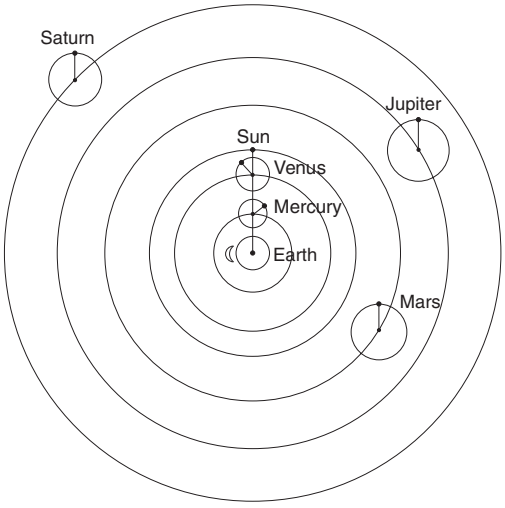

Fig. 1.1. Ptolemy's world model, very much simplified and not to scale. The centres of the epicycles of the inner planets are on the Sun-Earth line, while the radii of the epicycles of the outer planets run parallel to this line. The innermost sphere around the Earth (the 'sublunary') belongs to the Moon. The stars are located on an outermost sphere (not shown). The whole system operates like a clockwork as the Sun moves around the Earth. To the modern eye, a strange aspect of the model is that the motion of the other planets is connected with the motion of the Sun around the Earth. This coincidence is removed in Copernicus's heliocentric model. After Dijksterhuis (1969).

1.1 Special relativity (SR)

Modern cosmology is based on the theory of general relativity (GR), which is a natural generalisation of the theory of special relativity (SR). This section recapitulates the main ideas of special relativity, that is, physics in the absence of gravity. For a more thorough discussion we refer to Schutz (1985). We consider space and time to be a 4-dimensional continuum, called *Minkowski spacetime*. A (global) co-ordinate system in Minkowski spacetime is usually called a *reference frame* or just a *frame*. A point P with co-ordinates $\{x^\alpha\}$ is called an *event*. The motion of a particle can be represented by its *worldline*, Fig. 1.2. SR is based on two postulates:

- The *principle of relativity*, which states that the laws of physics must have the same form in every inertial frame.

- The *speed of light has a constant value* c in all inertial frames.

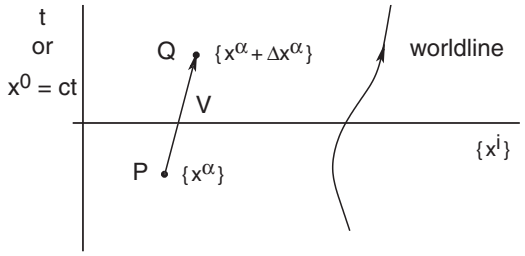

Fig. 1.2. The Minkowski spacetime, with events P and Q, a vector $\boldsymbol{V}$ connecting these events, and the worldline of a particle.

An inertial frame is a rigid system of spatial co-ordinates with synchronised clocks to measure t, in which test particles on which no forces are exerted move uniformly with respect to each other. An example of an inertial frame is a frame that does not move (no rotation, no translation) with respect to the distant galaxies. Inertial frames in SR are global, and they all move uniformly with respect to each other. In this section we admit only inertial frames. The principle of relativity is very old and goes back to Galilei. The second postulate is Einstein's innovative step, which he based, among other things, on Michelson and Morley's experiment which demonstrated the impossibility of measuring the velocity of the Earth with respect to the ether. The consequence is that invariance for Galilean transformations, as e.g. Newton's laws possess, no longer applies.

Simultaneity exit

SR often evokes major conceptual problems due to the fact that some very deeply rooted (Newtonian) ideas about space and time are not consistent with observations. Paramount among these is the fact that simultaneity has no longer an invariant meaning. Consider an inertial observer W, who tries to locate the events in his co-ordinate system (x, t) that are simultaneous with the origin $x = t = 0$, see Fig. 1.3, left. W argues: all events P that reflect light such that the moments of emission and detection are symmetrical with respect to $t = 0$ (emission at $t = -t_0$, detection at $t = t_0$ for all t_0). W's conclusion is: all events on the x-axis. Now consider observer $\overline{W}$ who moves uniformly to the right in W's frame, Fig. 1.3, right. At $t = 0$, W and $\overline{W}$ are both at the origin. $\overline{W}$'s worldline serves as the $\overline{t}$-axis of his frame, and $\overline{t} = 0$ is chosen at the common origin. $\overline{W}$ repeats W's experiment, but since the value of c is frame-independent, $\overline{W}$ identifies a different set of events, effectively his $\overline{x}$-axis, as being simultaneous with the origin. The $\overline{x}$-axis lies tilted in W's frame, and the tilt angle depends on $\overline{W}$'s velocity. Different observers $\overline{W}$ will

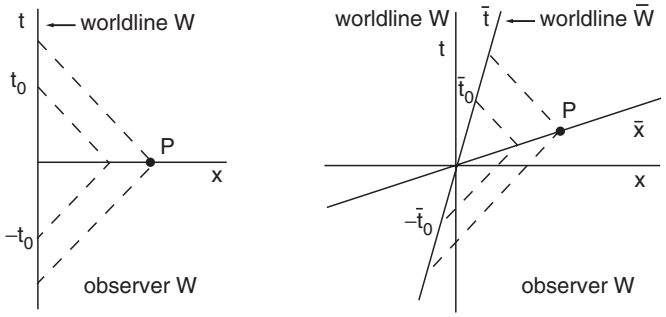

Fig. 1.3. As explained in the text, an invariant definition of simultaneity is impossible in SR.

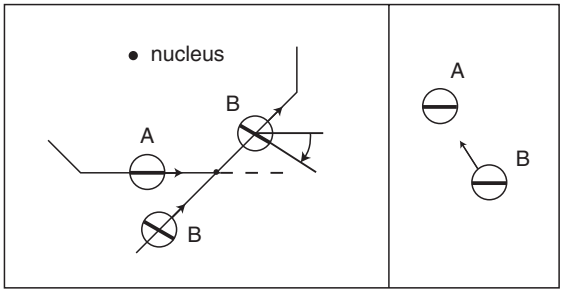

Fig. 1.4. Thomas precession of an electron orbiting a nucleus explained in the spirit of Fig. 1.3. After Taylor and Wheeler (1966).

therefore disagree as to which events are simultaneous with the origin.

Inaccurate reasoning in SR has led to many paradoxes (clock paradox, car-in-garage paradox). A vivid illustration of how drastically SR turns our perception of space and time upside down is the Thomas precession of the spin of an electron in an atom, a purely special-relativistic effect. Fig. 1.4 shows the classical orbit, approximated by a polygon. The heavy line is the projection of the spin axis on the plane of the orbit. After the electron has rounded a corner, its spin axis has turned. An analysis of what happens during the acceleration at the corner can be avoided by replacing electron A there by electron B, demanding that the spin vectors are aligned in a frame moving with A (A's rest-frame; right figure). But in the laboratory frame these orientations are different – this is a consequence of the relative meaning of simultaneity as explained in Fig. 1.3. Note that the electron is subject to

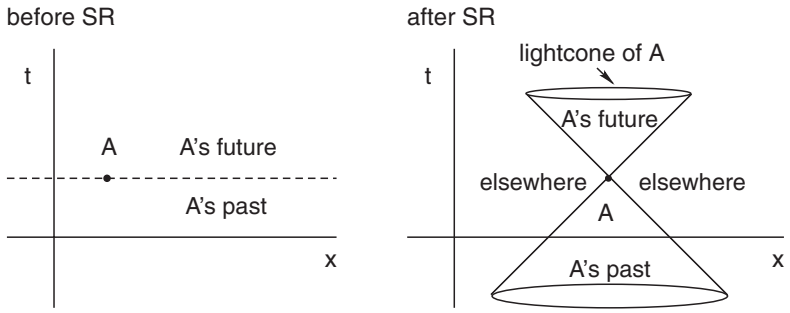

Fig. 1.5. The causal structure of Minkowski space. In SR every event A has its own invariant light-cone that divides Minkowski space into a past, a future and an elsewhere.

an additional precession due to electromagnetic interaction with the nucleus. The question arises whether a gyroscope in orbit around the Earth will also exhibit a precession. At the time of writing, the Gravity Probe B mission is performing the experiment, see further Ch. 8.

Lorentz metric

An important concept in SR is the *interval* Δs^2 between two events P and Q with co-ordinates x^α and $x^\alpha + \Delta x^\alpha$:

$$\Delta s^2 \,=\, c^2 \Delta t^2 \,-\, \Delta x^i \Delta x^i \,=\, \eta_{\alpha\beta}\, \Delta x^\alpha \Delta x^\beta \;; \tag{1.1}$$

$$\eta_{\alpha\beta} \,=\, \begin{pmatrix} 1 & & & \emptyset \\ & -1 & & \\ & & -1 & \\ \emptyset & & & -1 \end{pmatrix} \;. \tag{1.2}$$

Notation:

$$x^0 \,=\, ct\,, \qquad \Delta t^2 \,\equiv\, (\Delta t)^2\,. \tag{1.3}$$

Relation (1.1) defines the *metric*, i.e. the distance between two events in Minkowski space, and is called the *Lorentz metric*. Here and everywhere else: summation convention ; Roman indices run from 1 to 3 and Greek indices from 0 to 3. Note that we adopt the signature : $1, -1, -1, -1$.[4]

[4] The sign convention is important as it leads to sign differences everywhere, but it has of course no influence on the physics. The advantage of the present choice is that for timelike geodesics the curve parameter p, the interval length s and proper time τ are proportional, see § 2.5.

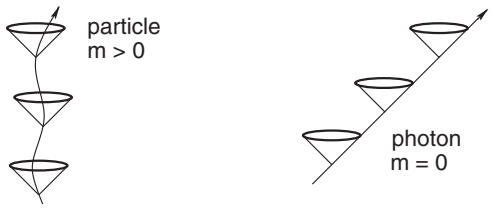

Fig. 1.6. The worldline of a particle with nonzero mass is located inside the light-cone, that of a photon is tangent to it.

Two events connected by a light ray have $\Delta s^2 = 0$, irrespective of their spatial distance. No matter how enormous the distance of some objects in the universe may be, the interval Δs^2 between them and the telescope is zero. The value of Δs^2 is also invariant: if some other observer $\overline{W}$ computes $\Delta \overline{s}^2 \equiv \eta_{\alpha\beta}\, \Delta \overline{x}^\alpha \Delta \overline{x}^\beta$ in his rest-frame (i.e. in a comoving inertial frame), the value he finds is equal to Δs^2 (proof: e.g. Schutz (1985), p. 11). This leads to an important relation between events, see Fig. 1.5. Prior to the advent of SR, all events were located either in the future, in the past, or they were simultaneous with a given event A. In SR there is A's *light-cone* $\Delta s^2 = 0$ that divides Minskowski space into a past and a future (with which A can have causal relations), and an 'elsewhere' (with which A cannot have any interaction). This division is independent of the reference frame because Δs^2 is invariant. Hence we can speak of *the* light-cone. The worldline of a particle with non-zero mass is always located inside the light-cone, see Fig. 1.6. Depending on the value of Δs^2, the vector connecting events P and Q in Fig. 1.2 is called a

$$\left.\begin{array}{ll} \text{timelike vector :} & \text{when } \Delta s^2 > 0 \ ; \\ \text{null vector :} & \text{when } \Delta s^2 = 0 \ ; \\ \text{spacelike vector :} & \text{when } \Delta s^2 < 0 \ . \end{array}\right\} \tag{1.4}$$

The *proper time* interval $\Delta\tau$ between two (timelike connected) events on the worldline of a particle is defined as:

$$c^2 \Delta\tau^2 \equiv \Delta s^2 = c^2 \Delta t^2 - \Delta x^i \Delta x^i \ . \tag{1.5}$$

For positive Δs^2 we may define $\Delta s \equiv (\Delta s^2)^{1/2}$ and proper time intervals as $\Delta\tau = \Delta s/c$. Proper time intervals are invariant because Δs^2 is. By transforming to the rest-frame of the observer $\overline{W}$, so that $\Delta \overline{x}^i = 0$, we find that $\Delta\tau^2 = \Delta \overline{t}^2$, which shows that the proper time is just the time of a clock moving with the observer (his own wristwatch). Now substitute $\Delta x^i = (\Delta x^i/\Delta t)\, \Delta t = v^i \Delta t$ in (1.5) and compute the limit:

$$\mathrm{d}\tau = \sqrt{1 - (v/c)^2}\ \mathrm{d}t \ , \tag{1.6}$$

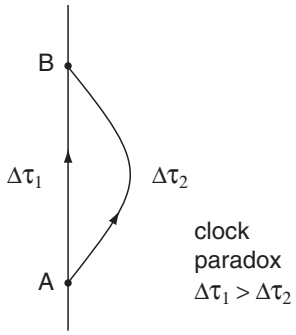

Fig. 1.7. The clock paradox. Two clocks moving from event A to B along different worldlines indicate different readings $\Delta\tau$ for the duration of the trip.

where v is the speed of the particle (the ordinary 3-velocity). The proper time $\Delta\tau$ elapsed between two events can be found by integrating (1.6) along the worldline connecting the events. The answer will depend on the shape of the worldline, which leads to the famous clock paradox, Fig. 1.7, explained in detail in Schutz (1985), § 1.13.

Lorentz transformations

The co-ordinates x^α and $\overline{x}^\alpha$ of an event with respect to two different inertial frames can be expressed into each other by means of a *Lorentz transformation*:

$$\overline{x}^\alpha \;=\; L^\alpha{}_\nu \, x^\nu \;. \tag{1.7}$$

The $L^\alpha{}_\nu$ are constants that depend only on the relative velocity $\boldsymbol{v}$ of the two frames. Relation (1.7) is a linear transformation that leaves Δs^2 invariant. If the co-ordinate axes (x, t) and $(\overline{x}, \overline{t})$ are defined as in Fig. 1.3 the transformation is

$$L^\alpha{}_\nu \;=\; \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.8}$$

with $\beta = v/c$ and $\gamma = (1 - \beta^2)^{-1/2}$. The mathematical formulation of SR proceeds in terms of 4-vectors and tensors, that transform according to a Lorentz transformation. The trick is to try and write the laws of physics as relations between scalars, vectors and tensors *only*, because in that case they are automatically invariant for Lorentz transformations.

Lorentz transformations are global. In GR we allow arbitrary curvilinear

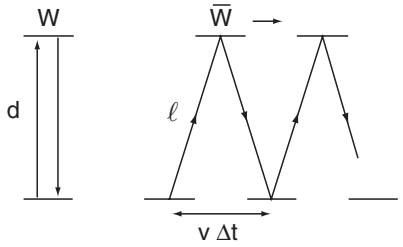

Fig. 1.8. An Einstein clock consists of photons traveling between two parallel mirrors at a distance d; the time for a round trip $\Delta t = 2d/c$ serves as the time unit. This clock will run slower if it moves with respect to the observer because the photons traverse a distance $\ell > d$ while c is constant. The merit of this example is that the time dilation is immediately obvious, but it is not so evident that it is impossible to eliminate the effect by using another clockwork. However, it can be shown that the effect is quite general and independent of the way the clock is constructed.

reference frames. As we shall see in § 2.3, the effect is that the global Lorentz transformation is replaced by a mesh of local Lorentz transformations that are different at each position in spacetime.

Exercise 1.1: Explain the time dilation with the help of Einstein's clock, Fig. 1.8:

$$(\Delta t)_{\text{measured by W}} = \frac{(\Delta \bar{t})_{\text{measured by } \overline{\text{W}}}}{\sqrt{1 - v^2/c^2}} \ . \tag{1.9}$$

Hint: W observes $\overline{W}$'s clock as it travels to the right at velocity v. W measures $\Delta t = 2\ell/c$, and $\ell^2 = d^2 + (v\Delta t/2)^2 = d^2 + (v\ell/c)^2$, from which $\ell = d/\sqrt{1-(v/c)^2}$, and $(\Delta \bar{t})_{\text{measured by } \overline{\text{W}}} = 2d/c$.

Exercise 1.2: Below relation (1.6) it was said that the proper time elapsed between events depends on the worldline connecting the two events. Doesn't that contradict the fact that $\mathrm{d}\tau$ is invariant?

Hint: In a given set of $\mathrm{d}\tau$, each $\mathrm{d}\tau$ is invariant under co-ordinate transformations, but another integration path simply implies a different set of $\mathrm{d}\tau$.

1.2 General relativity (GR)

If we extend SR to arbitrarily moving reference frames, we would be able to do physics from the point of view of an accelerated observer. There is, however, another important motivation. Since in doing so apparent forces appear that are closely related to gravity, we may perhaps also be able to address gravity. And this turns out to be true. But if we are only after gravity, it would seem more straightforward to try and incorporate gravity in the framework of SR. Unfortunately, that doesn't work. Newtonian gravity may be summarised by $\nabla^2\Phi = 4\pi G\rho$ and $K = -m\nabla\Phi$. It follows that gravity operates instantaneously – a change in ρ alters Φ everywhere at the same moment. This is inconsistent with SR because what is instantaneous in one frame is no longer so in another. This theory holds therefore only in one preferred frame. The problem might be overcome by replacing the equation for the potential by $\Box\Phi = (c^{-2}\partial^2/\partial t^2 - \nabla^2)\Phi = -4\pi G\rho$, for example, but then other difficulties appear, see e.g. Robertson and Noonan (1969) and Price (1982). Special relativistic theories of gravity using a flat spacetime and a single global reference frame don't work because they cannot accommodate the gravitational redshift and the weak equivalence principle. A different approach is needed.

Weak equivalence

At this point we need to be more precise about the concept of mass. A force $\boldsymbol{K}$ acting on a particle with *inertial mass* m_i causes an acceleration $\boldsymbol{a}$ given by Newton's law $\boldsymbol{K} = m_\mathrm{i}\boldsymbol{a}$. The inertial mass expresses the fact that objects resist being accelerated. To compute the force $\boldsymbol{K}$ we need the field(s) in which the particle moves, and the charge(s) that couple to those field(s). For example, $\boldsymbol{K} = q(\boldsymbol{E} + \boldsymbol{v}\times\boldsymbol{B}/c)$ for a particle with electric charge q moving with speed $\boldsymbol{v}$ in an electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{B}$. For a particle with a gravitational charge m_g, usually called the *gravitational mass*, we have $\boldsymbol{K} = -m_\mathrm{g}\nabla\Phi$. It follows that $\boldsymbol{a} = -(m_\mathrm{g}/m_\mathrm{i})\nabla\Phi$.

It is an experimental fact that materials of different composition and mass experience exactly the same acceleration in a gravitational field. Eötvös verified that with an accuracy of 10^{-8} in 1896, and Dicke attained 2×10^{-11} in 1962. Both experiments used a torsion balance. Presently, torsion balance and free-fall experiments achieve an accuracy of $\sim 10^{-12}$.[5] [6] Hence $m_\mathrm{g}/m_\mathrm{i}$ is

[5] Chen and Cook (1993) § 4.8; Will (1993) Ch. 14. With the help of lunar laser-ranging an accuracy of 7×10^{-13} has been achieved (Dickey, J.O. et al., *Science 265* (1994) 482). The idea is that the lunar orbit as a whole must be displaced along the Earth-Sun line in case the Moon and the Earth experience a slightly different acceleration with respect to the Sun.

[6] The gravitational constant G, however, is only known with a precision of a few times 10^{-4}.

a universal constant, taken to be unity in classical mechanics. This is called the *weak principle of equivalence*. It follows that the concept of gravity loses its meaning, as the field can be made to vanish by transforming to a freely falling reference frame. For an electromagnetic field this is impossible as q/m_{i} is most certainly not a universal constant. From this Einstein (and others before him) concluded that light must be deflected by a gravitational field, because it moves along a straight line in a freely falling frame where there is no gravity. This trick of transforming gravity away works only locally. In a frame that moves with a freely falling particle, neighbouring particles will initially move uniformly with respect to each other, but not after some time. In the famous elevator thought experiment it is impossible to distinguish *locally* gravity from an externally imposed acceleration. But a distinction is possible by observing two test particles at some distance from each other, because the latter is homogenous while the former is not. The so-called *tidal forces* cannot be transformed away, because a 'real' gravitational field is inhomogenous.

The fact that inertial and gravitational mass are identical is an unexplained coincidence, in some sense comparable to the unexplained coincidence in Ptolemy's world model, Fig. 1.1. Einstein took that as a basis for a new theory. The fact that motion in a gravitational field depends neither on the composition nor on the mass of the particles suggests that the particle orbits might perhaps be determined by the structure of spacetime. In SR the worldlines of free particles are straight, independent of the nature of the particles. If we now switch on gravity, maybe a more general formulation is possible, in which the worldlines remain 'straight' (i.e. geodesics) in a *curved* spacetime.[7] In that case gravity would no longer be a force, but rather a consequence of the curvature of spacetime. The elaboration of this idea is what we now know as the theory of General Relativity (GR). Global inertial frames no longer exist, only local inertial frames do. For according to GR there are no forces working on freely falling particles, while it is at the same time not possible to define a reference frame in which two freely falling particles move uniformly with respect to each other.

Curvature

That curvature is the way to go ahead may be gleaned, for instance, from the experiments of Pound, Rebka and Snider.[8] Photons moving vertically in the Earth's gravity field turn out to be slightly redshifted, see Fig. 1.9. The

[7] A space is said to be *flat* when Euclides's 5th postulate on the existence of a single parallel holds (in metric terms: the Riemann tensor is zero). A space is said to have Euclidean geometry if the metric can be cast in the form $\mathrm{d}s^2 = \mathrm{d}x^\alpha \mathrm{d}x^\alpha$. The Minkowski spacetime of SR is flat but not Euclidean.

[8] Pound, R.V. and Rebka, G.A. *Phys. Rev. Lett. 4* (1960) 337; Pound, R.V. and Snider, J.L. *Phys. Rev. B 140* (1965) 788.

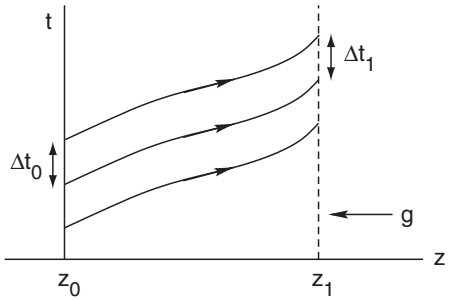

Fig. 1.9. The Pound-Rebka-Snider experiment. Photons move vertically upwards over a distance of $z_1 - z_0 = 22.5$ meters and get redshifted.

required precision could be attained with the help of the Mössbauer effect. The worldlines of subsequent wave crests in the Minkowski diagram must be congruent because the gravity field does not depend on time. Therefore Δt_0 should be equal to Δt_1, regardless of the shape of the worldlines, but the experiment shows that $\Delta t_1 > \Delta t_0$ (a redshift). This suggests (but does not prove) that one can no longer assume that the Minkowski spacetime is globally flat in the presence of gravity. A curved spacetime is descibed by a *local* metric:

$$c^2 \mathrm{d}\tau^2 \;=\; \mathrm{d}s^2 \;=\; g_{\alpha\beta}\, \mathrm{d}x^\alpha \mathrm{d}x^\beta \;, \tag{1.10}$$

and $\mathrm{d}s^2$ is the interval ('distance') between two events at x^α and $x^\alpha + \mathrm{d}x^\alpha$; $g_{\alpha\beta}$ is the *metric tensor*. The relation $\mathrm{d}s^2 = c^2\mathrm{d}\tau^2$ between interval and proper time is taken to remain valid (for particles with mass), but the relation between $\mathrm{d}t$ and $\mathrm{d}\tau$ is no longer as simple as in (1.6) because $g_{\alpha\beta} \neq \eta_{\alpha\beta}$. The possibility of transforming gravity away locally amounts to the following requirement: at any point $\{x^\mu\}$ of spacetime there should exist a transformation that casts (1.10) into the SR form $\mathrm{d}s^2 = \eta_{\alpha\beta}\, \mathrm{d}x^\alpha \mathrm{d}x^\beta$. In doing so we have constructed a local inertial (i.e. freely falling) frame in $\{x^\mu\}$ where gravity does not exist[9] – provided the frame is not too big, otherwise we will notice the effect of curvature in the form of tidal forces. Sufficiently small sections of spacetime are flat, 'small' meaning small compared to the typical dimension of the system (the Schwarzschild radius, the scale factor S of the universe, etc.). Spacetime curvature and tidal forces will be the hallmark of a real gravitational field. Weight is merely a pseudo-force caused by being in the wrong (not freely falling) frame, just as centrifugal and Coriolis forces are pseudo-forces caused by being in a wrong (rotating) frame.

[9] The terms 'local inertial frame' and 'local freely falling frame' will be used interchangeably. A local rest-frame is a local inertial frame in which a particle or an observer is instantaneously at rest.

Strong equivalence and general covariance

In order to generalise existing physical laws to GR we broaden the scope of the weak equivalence principle, and assume that it is impossible to detect locally any effect of gravity in a freely falling frame, whatever other forces may be acting. In other words, in a freely falling frame *all* laws of physics have the form they have in SR in the absence of gravity. This is called the *strong principle of equivalence*. These laws / equations are then generalised by replacing the tensors that appear in them by tensors that are invariant for *arbitrary* co-ordinate transformations instead of only for Lorentz transformations. This is called the *principle of general covariance*. The application of this principle is somewhat arbitrary, as we shall see, but the obvious way out of adopting the simplest possible generalisation has sofar proven to be effective. The term 'principle of general covariance', incidentally, is misleading in that it has nothing to do with the covariant form of tensors. Principle of general *in*variance (for arbitrary co-ordinate transformations) would have been a much better name. Note also that general covariance has no deeper significance of its own (Friedman, 1983). It is a self-imposed regime of great heuristic value in finding physically correct equations, in some way comparable to checking the correct dimension of an expression.

Mach's principle

A number of ideas, collectively known today under the name *Mach's principle*, have strongly influenced Einstein in his formulation of GR. Mach rejected the Newtonian concept of absolute space, as Leibniz had done earlier. Mach was struck by the fact that the frame defined by the distant matter in the universe happens to be an inertial frame, and that inertia manifests itself only if masses are accelerated with respect to this frame. He argued that this cannot be just a coincidence, and that the inertial mass may somehow be 'induced' by the gravitational mass of all matter in the universe. This led Einstein to seek a theory in which the geometry of spacetime, i.e. $g_{\alpha\beta}$, is determined by the mass distribution. The frame-dragging effect near rotating massive objects, for example (Ch. 6), may be seen as a manifestation of Mach's principle. However, Gödel's solution[10] of the field equations indicates that Mach's principle is only partially contained in GR as it is presently formulated, see Friedman (1983) for more information.

Exercise 1.3: GR and cosmology are fields of many principles. Formulate in your own words the meaning of these principles: relativity, strong and weak equivalence principle, Mach, general covariance, cosmological and the anthropic principle (§ 13.4).

---

[10] Gödel, K., *Rev. Mod. Phys. 21* (1949) 447.

1.3 The need for GR in astrophysics

While SR was born out of the need to resolve a major conflict, namely the failure to measure the velocity (of the Earth) with respect to the ether, GR was created rather for esthetic reasons: the wish to have a relativistic theory of gravity. But there was no compelling conflict with observations that called for a solution. The problem of the perihelium precession of Mercury was known at the time, but was considered to be a nut for the astronomers to crack – not as a stumble block to progress in physics. Consequently, after its conception, GR remained for a long time what is was: an elegant but rather inconsequential theory that was accepted by the physics community precisely because of its elegance. After the correct prediction of the perihelium shift and the spectacular confirmation of the deflection of starlight in 1919, there weren't many other things that could be measured. The technology of the day, for example, was inadequate to detect the gravitational redshift in the solar spectrum. SR on the other hand, led to many observable consequences and was soon completely integrated in the framework of physics as an indispensable basic element. It was recognised that GR was relevant for cosmology,[11] but in the first half of 20th century cosmology was very much a slightly esoteric field that a decent physicist did not touch, because there were very few observations that could show the way. Notions such as a hot big bang, light element synthesis and structure formation were as yet unheard of. And so GR remained outside the mainstream of physics. That state of affairs began to change only in the second half of the 20th century. In particular the 60ies saw a rapid succession of novel developments and discoveries. Technological advances led to a demonstration of the gravitational redshift in the laboratory (1960), soon followed by a measurement in the solar spectrum (1962). Radar reflections from Venus (1964) showed that the travel time of light increases when it moves closely past the Sun. This effect had been predicted by GR as a consequence of the warping of spacetime near a massive object, causing distances to be generally longer.

Astrophysics, too, began to profit from several new developments. Most important were the emergence of radio astronomy, and the possibility to deploy instruments in space which opened up the field of X-ray astronomy. Non-solar X-rays were first detected in 1962 and led to the discovery of X-ray binaries. The X-ray emission is believed to be due to accretion of matter onto a neutron star or black hole, two objects whose existence is predicted by GR. The energy released per unit mass by accretion on such a compact object depends on various parameters, and is of the order of 10% of the infalling rest mass energy – a factor 10-20 more than hydrogen fusion. As the matter falls into the deep potential well, it is heated to X-ray temperatures and serves as

[11] In particular the work of Lemaître was influential in this regard (Lemaître, G., *Ann. Soc. Sci. Bruxelles 47A* (1927) 49 and *M.N.R.A.S. 91* (1931) 483).

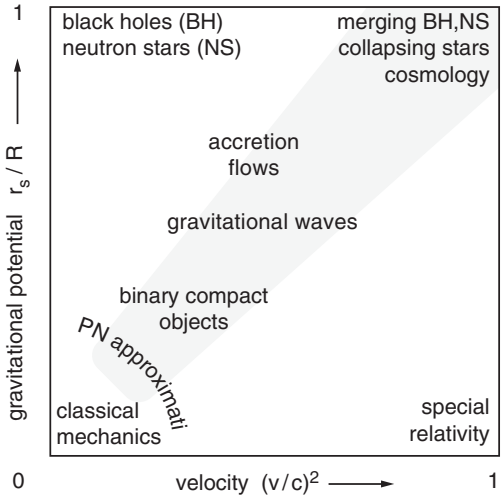

Fig. 1.10. A classification of some applications of General Relativity. For weak fields there is only the horizontal axis. The world of GR unfolds as we move upward to stronger gravitational potential, measured by $r_{\rm s}/R = 2GM/Rc^2 \sim |\Phi|/c^2$, where $\Phi \sim -GM/R$ and $R =$ typical size of the object ($r_{\rm s} =$ its Schwarzschild radius). The post-Newtonian approximation gives first order corrections to classical mechanics. Neutron stars and black holes are in the strong field corner. Binary objects have $\Phi \sim -v^2$ and are approximately on the grey diagonal, slowly moving up to their eventual merger and generating gravitational waves as they do so. The latter may also be generated to the left of the diagonal (oscillating / rotating neutron stars) or to the right (close encounters). To position cosmology the universe is considered to be a compact object with expansion velocities approaching c near the horizon (though fields and velocities are locally small).

a bright probe of conditions very close to the compact object. In some cases the mass of the object could be shown to be larger than $3M_\odot$. Since this is larger than the theoretical maximum mass of a neutron star, the object is, in all likelihood, a black hole. Accretion flows thus provide an important diagnostic tool of these compact systems, but it is not the only one. Direct proof of the existence of neutron stars came in 1967 with the discovery of pulsars. It was soon realised (1968) that pulsars are spinning neutron stars equipped with a radio beacon, a feat no one had ever dreamt of. Neutron stars had been hypothesized by Baade and Zwicky (1934) following the discovery of the neutron (1932). They suggested that neutron stars are formed in a supernova explosion, a gravitational collapse of a heavy, evolved star that has run out of nuclear fuel. In 1939 Oppenheimer and Volkoff calculated the structure of a neutron star and showed that it is completely determined by

GR. Now, after 33 years, it turned out that these objects actually did exist. And if stellar evolution, that great creator, can make neutron stars, it may very well produce black holes too. These and other developments led to a revival of theoretical studies in GR which had been stagnant for years. The properties of these mysterious black holes and the generation of gravitational waves, for example, drew much attention. Experimental gravity received a boost as well, leading to the development of detectors for gravitational waves and Gravity Probe B, a space mission for detecting relativistic precession effects – to name only two.

The first binary pulsar was discovered by Hulse and Taylor in 1975. This system turned out to be a perfect cosmic experiment featuring two neutron stars in a tight orbit, one of which is a precision clock. Since the system is clean, application of GR permitted determination of all system parameters. In 1979 it was shown that the system loses energy at a rate that is consistent with energy loss by gravitational waves. This is a strong if indirect argument for the existence of gravitational waves. Several of these binaries have now been found, and there should be many more out there that we cannot see because they contain no pulsar. However, the gravitational waves they emit should be detectable. As the binary loses energy it shrinks and moves slowly along the diagonal in Fig. 1.10 until the components merge in a gigantic explosion, unleashing a final burst of gravitational radiation and γ-rays into space which should be visible throughout the universe. Perhaps this is the explanation of the so-called short-duration γ-ray bursts, whose nature is still not understood. And the hunt for gravitational waves is on: detectors for gravitational waves are in an advanced state of development and several are operating in science mode.

The discovery of quasi-stellar objects or quasars (1963) showed that there are distant objects that are typically 100 times brighter than ordinary galaxies in our neighbourhood. It was gradually understood that these and other objects (Seyferts, BL Lac objects,..) are different visual manifestations of active galactic nuclei (AGNs) with a huge power release, up to 10^{48} erg s^{-1}. Rapid variability pointed to a small gravitational powerhouse casting as the main actors a black hole of $10^6 - 10^9\, M_\odot$, a surrounding disc swallowing matter (in some cases as much as $10 - 100\, M_\odot$ per year), and collimated bipolar outflows. Another line of evidence for the existence of massive black holes comes from galactic rotation curves which demonstrate that many galaxies contain heavy objects ($10^6 - 10^9\, M_\odot$) within a small radius at the centre, very likely a black hole. And there is very strong evidence that a $\sim 3.6 \times 10^6\, M_\odot$ black hole is lurking at the centre of our own galaxy, which is currently not accreting any appreciable amount of mass.

The gravitational deflection of light by the Sun discovered in 1919 received a spectacular follow-up in 1979 when the quasars Q0957+561 A and

B were identified as two images of the same object whose light is deflected by an intervening galaxy. Many gravitational lenses have been found since then. In principle this opens the possibility to weigh the lens including the dark matter it contains, and to study magnified images of very distant objects. There have been many other advances in cosmology, but there are two that outshone all others. The first is the cosmic microwave background (CMB), discovered in 1965, with suggestions as to its existence dating back to 1946. The CMB was a monumental discovery that marked the beginning of cosmology as a quantitative science. It put an end to the so-called steady state model and permitted for example a quantitative prediction of the synthesis of the light elements in the universe (1967), which has been confirmed by observations. The latest highlight is the WMAP mission which has measured the tiny fluctuations in the temperature of the CMB across the sky. This has resulted in a determination of the basic parameters that fix the structure and evolution of our universe. The second very important development was of a theoretical nature and took place in 1981: the discovery of the possibility of an inflation phase right after the Big Bang. The inflation concept repairs some basic defects of the classical Friedmann-Robertson-Walker cosmology that had to do with causality. The inflation paradigm is very powerful but speculative. Pending some unsettled 'fine-tuning' it seems to explain why the universe expands, why it is homogeneous and flat, as well as the origin of the density fluctuations out of which galaxies evolve later.

This overview illustrates that GR is nowadays being studied in all corners of the diagram of Fig. 1.10. The field has really opened up and there is a great sense of anticipation and promise of new results every day. Particle physicists turn to cosmology in the hope to find answers to questions that particle accelerators seem unable to address. This symbiosis of cosmology and particle physics has sparked off the new field of astroparticle physics. And although it may take years before the detectors for gravitational waves currently in operation actually observe a wave, it may also be tomorrow! This element of suspense and impending surprise renders GR and its application to astrophysics and cosmology a highly attractive field, and some of the thrill, it is hoped, will transpire in the following chapters.