Lecture I: FLAT SPACE–TIMES

Einstein formulated the theory of special relativity in 1905. I am not going to attempt to give a comprehensive description of it, but only mention briefly a few aspects that will help us most in our approach to curved space–times.

1. SPACE–TIME DIAGRAMS; CAUSAL REGIONS

One of the most revolutionary concepts of special relativity is the idea of unifying space and time into a single entity called space–time. Throughout these lectures, the emphasis will be on the geometry of space–time, so it will often prove helpful to draw space–time diagrams of what we are trying to study. It is not that easy to visualize four-dimensional space–time, but many problems have sufficient symmetry so that they reduce effectively to three or even two dimensions of space–time. For example, consider two situations that are essentially three-dimensional:

i) a planet in circular motion around the Sun: the planet describes a helix in space–time (Fig. I.1a);

ii) a circular wave produced by dropping a stone in a pond: the spreading ripples will produce a cone in space–time (Fig. I.1b).

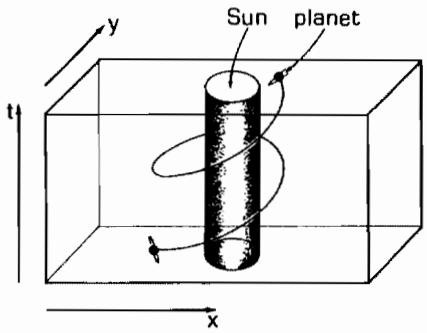


Fig. I.1a A planet in circular motion around the Sun

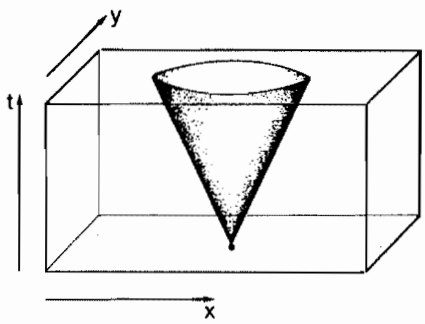


Fig. I.1b Circular ripples produced by a stone thrown into a pond

Points in space–time are called *events*, and a *world-line* is the path traced out in space–time by the events representing the history of a particular particle or light-ray.

Special relativity has two very important things to say about the speed of light c: firstly, it is a constant *in vacuo*, and secondly it is a limiting speed for all communication and for all motion of massive bodies; indeed it is the limiting speed for propagation of all causal influences. In terms of space–time diagrams, the first of these statements means that light-rays are straight lines. For convenience, we plot ct against the spatial coordinates; then the light-rays are at 45° to the ct-axis. The *light-cone* of an event O is the set of all light-rays through the event (Figs. I.2a,b). The second statement means that all world-lines of massive particles must lie within the light-cone since their speeds are less than c, so the regions outside the light-cone are inaccessible to causal influence by O and cannot influence O. We see that by knowing the positions of the light-cones, we know a great deal about possible motion in the space and about its causal properties.

INTRODUCTION

In these lectures, I want to give an introduction to general relativity, and in particular to the aspects that are involved in cosmological models. The approach I am going to take is slightly unusual in that I am not going to use explicitly any tensor calculus—not because I think it is unimportant or not useful, but because there is a danger of getting lost in the details of tensor manipulation and losing sight of what I feel is really important, the geometric structure of space–time. Therefore the lectures will include many diagrams and only fairly simple equations.

Mainly for pedagogical reasons, I shall begin by discussing flat space–times. Many of the techniques we use in studying curved space–times are used also for flat space–times where it is usually easier to see what is going on. I shall then introduce the idea of curved space–times, emphasizing the importance of the metric interval and the resulting light-cone structure, which determines the causal properties of the space–time. As examples of these ideas, I shall first describe the space–time around an isolated spherically symmetric object and the possibility of collapse to a black hole, with the associated formation of an event horizon, and secondly the space–time produced by the matter in the universe as a whole, and the causal structure of simple model universes. Throughout these lectures, I will be talking about the classical theory, with only brief mention of some possible differences when quantum effects are taken into account.

These lectures follow closely the book *Flat and Curved Space–Times* written jointly with George Ellis and with diagrams by Mauro Carforo. I am very grateful to both of them for this collaboration. I am also grateful to John Bell for inviting me to give these lectures at CERN, and for the help, encouragement and inspiration he provided. I dedicate these notes to his memory.

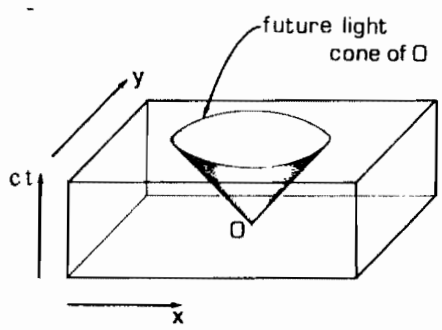


Fig. I.2a The future light-cone of O (a projection for fixed z)

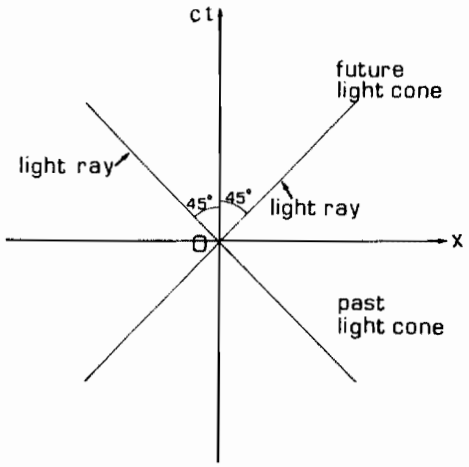


Fig. I.2b The light-cone of O (a projection for fixed y and z)

2. MEASUREMENTS; THE DOPPLER SHIFT

To make measurements in space–time, we first of all assume the existence of ideal clocks that measure time accurately along their world-lines. This time, measured along the world-line, is called *proper-time*. (The usual t-coordinate is therefore the proper-time for an observer at rest at the spatial origin O.) We then make use of the invariance of the speed of light: we use radar to measure distances. We also use radar or light to establish *simultaneity,* which is an important concept in special relativity, and one about which observers in relative motion will disagree, as we shall see. Here is an operational definition of simultaneity. Suppose an observer A sends a light signal at event E. The signal is reflected back at point P and received again by A at R (Fig. I.3). Since the light travel-time must be the same for both the outward and the

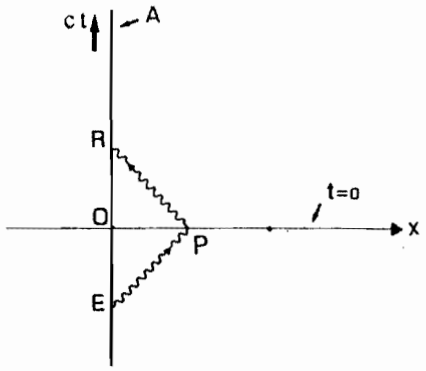


Fig. I.3 The light-rays used by observer A to determine that P is simultaneous with O

return journey, A concludes that the event O on his world-line, half-way between E and R, is simultaneous with P. Similarly, all other points on the x-axis are simultaneous with O, and so in general, for an observer at rest, the surfaces of simultaneity are just the surfaces of constant t.

Now consider a second observer, B, moving with constant speed v relative to A. He goes through exactly the same process, determining that P' is simultaneous with

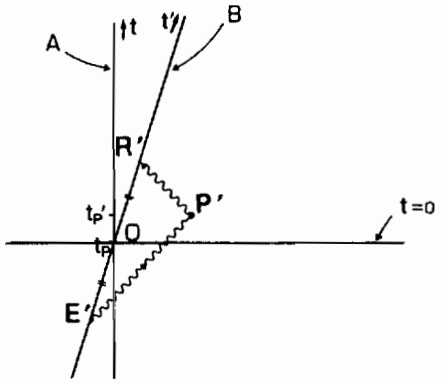


Fig. I.4 The light-rays used by observer B to determine that P′ is simultaneous with O

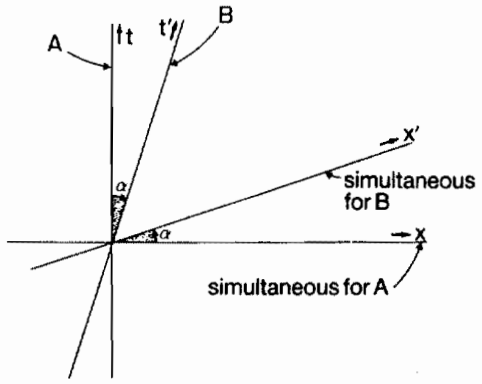


Fig. I.5 The angle between the surfaces of simultaneity for A and B is the same as the angle between their world-lines

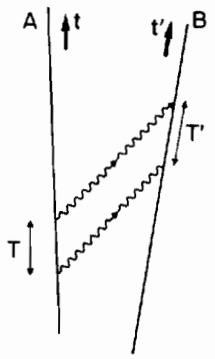


Fig. I.6 Light signals between observers A and B in relative motion

O, which is half-way between E' and R'. From the geometry of Fig. I.4 we see that P' must be above the x-axis and so does not coincide with any of the events that were simultaneous with O for A. In fact the events that are simultaneous with O for B are on a straight line through O, and it can be proved that the angle of this line to the x-axis is the same as the angle of B's world-line to the t-axis. Now B's world-line is his time axis, t' say, and his surface of simultaneity through O, $t' = 0$, is the x'-axis (Fig. I.5). However, we cannot read off from a space–time diagram drawn from A's point of view what measurements B will make (that will come later!). What we can see is that space–time is split differently into space (surfaces of simultaneity) and time (measured along world-lines) for observers in relative motion.

Now let us look at one of the simplest measurements we can make, that of the *Doppler shift* (and hence the red-shift). Suppose that light signals are sent by A at an interval of T and received at an interval of T' by B, moving with speed v relative to A (the times being measured by their respective clocks) (Fig. I.6). The Doppler shift K is defined by

$$K = T'/T \ .$$

A standard way of measuring K is from the observed wavelength of electromagnetic radiation, provided the intrinsic wavelength is known. Since the periods $\Delta\tau_e$ and $\Delta\tau_o$ of the emitted and observed radiation, for observers in relative motion, are related by

$$\Delta\tau_o = K\Delta\tau_e \ ,$$

then the emitted and observed wavelengths are related by

$$\lambda_o = K\lambda_e \ .$$

The red-shift parameter z is defined by

$$z = \frac{\text{change in wavelength}}{\text{emitted wavelength}} = \frac{\lambda_o - \lambda_e}{\lambda_e} = K - 1 \ .$$

We see that $z > 0$ and $K > 1$ correspond to a red-shift of the radiation and that $-1 < z < 0$ and $0 < K < 1$ correspond to a blue-shift. By measuring red-shifts from distant galaxies, astronomers can determine their speed of recession by using the formula which we will now derive.

We must first assume uniformity of K (i.e. that it is independent of T and constant in time) and reciprocity, which is a consequence of the principle of relativity. (If B is moving away from A with speed v, then A is moving away from B with speed v, so $K_{AB} = K_{BA}$). Suppose, now, that B is moving with speed v relative to A and that they coincide at the origin O. At time T, A sends a signal to B which is received at time T' according to B, and then reflected back to A who receives it at time T'' (Fig. I.7). Then we have

$$T' = KT \ ,$$
$$T'' = KT' = K^2T \ .$$

According to A, the travel time for the light is

$$T'' - T = (K^2 - 1)T \ ,$$

so the distance measured to B is

$$D = \frac{1}{2}\, c\, (K^2 - 1)T \ .$$

This is the distance to B at a time

$$t = \frac{1}{2}\,(T + T'') = \frac{1}{2}\,(K^2 + 1)T \ .$$

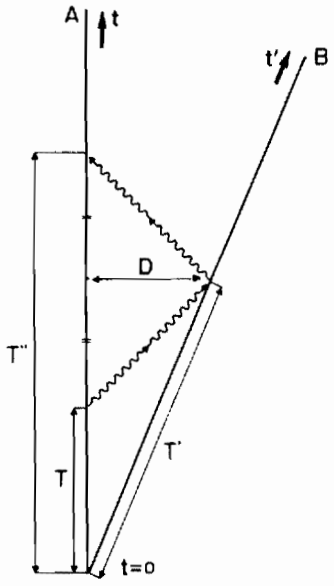


Fig. I.7 Light signals by which A can determine the relative velocity of B

Thus B's velocity relative to A is

$$v = \frac{D}{t} = \frac{c(K^2-1)}{K^2+1} .$$

Solving this formula for K, we obtain

$$K = \left(\frac{1+v/c}{1-v/c}\right)^{1/2}$$

(a formula usually derived from the Lorentz transformation).

By similar arguments, using what is known as Bondi's K-calculus, one can derive all the kinematic effects of special relativity.

3. THE LORENTZ TRANSFORMATION; INVARIANTS AND THE METRIC FORM

3.1 Active and passive transformations

In the conventional treatment of special relativity, one starts from the Lorentz transformation relating the coordinates of relatively moving observers and derives the kinematic effects from it. However, it is possible—and perhaps more appealing intuitively—to derive the Lorentz transformation from arguments about surfaces of simultaneity, etc., in space–time diagrams, using the K-calculus. I will not go through the derivation here, but merely state the result.

Consider an observer B moving with speed v in the x- and x'-directions relative to A. The Lorentz transformation between the coordinates of a point P in A's frame and in B's frame (Fig. I.8) is given by

$$t' = \gamma(t - vx/c^2) \ , \quad x' = \gamma(x - vt) \ , \quad y' = y \ , \quad z' = z \ ,$$

with $\gamma = (1 - v^2/c^2)^{-1/2}$.

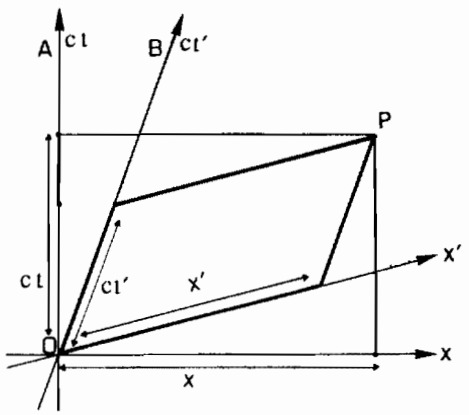


Fig. I.8 The event P has coordinates (t, x, y, z) in A's frame and $(t', \mathrm{x}', \mathrm{y}', z')$ in B's frame

In the sense in which I have described it, relating the coordinates of a single event according to different observers, the Lorentz transformation is *passive*. We can also regard it in an active sense. Let us first see what this distinction means for rotations in a plane.

Consider a rotation through an angle θ from a frame F with coordinates (x, y) to a frame F' with coordinates (x', y') (Fig. I.9a). The coordinates of a single point P are related by

$$x' = x\cos\theta + y\sin\theta$$
$$y' = -x\sin\theta + y\cos\theta \ .$$

This is a passive transformation. In an active transformation (Fig. I.9b), the space as a whole rotates relative to the fixed frame F, the rotation of F' dragging the points with it. Hence a point P with coordinates (x', y') relative to F moves to P' with the same coordinates (x', y') relative to F' [and (x, y) relative to F, with (x, y) and (x', y') related by the rotation formula above]. The movement of points as a whole is shown in Fig. I.9c.

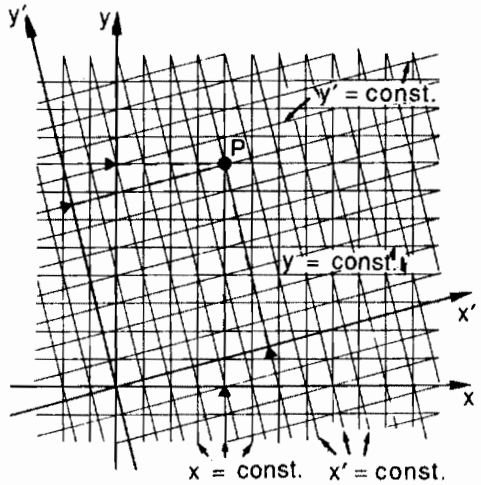


Fig. I.9a A passive rotation relates the coordinates (x, y) and (x', y') of a single point P under rotation of axes

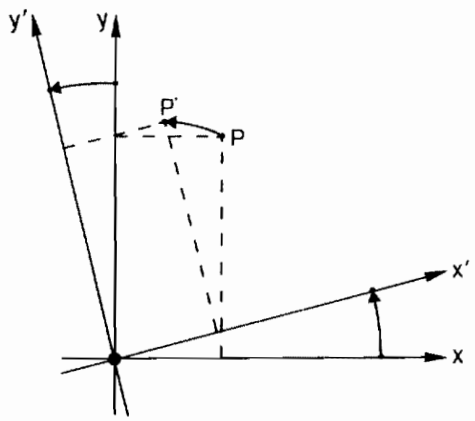


Fig. I.9b In an active rotation, the point P moves with the axes and coordinates to a new point P'

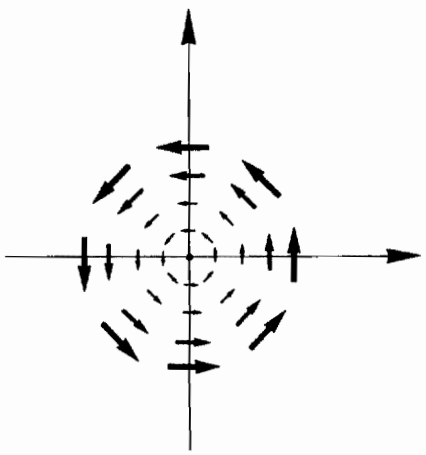


Fig. I.9c The movement of points in the plane generated by an active rotation

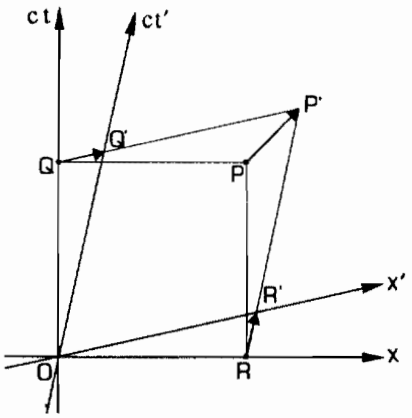


Fig. I.10a The effect of a boost on the points P, Q, and R, moving them to P', Q', and R'

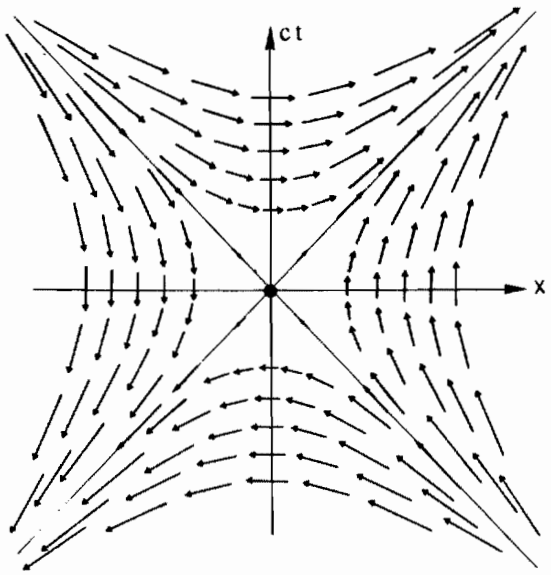


Fig. I.10b The pattern of motion generated by a boost

To understand the *active* sense of a Lorentz transformation, consider two Lorentz frames, F and F', which coincide initially. We then give frame F' a velocity v in the x-direction: we say that we give F' a *boost* through v. Space–time points are dragged along with it, so an event P with coordinates (x', t') in both F and F', goes to an event with coordinates (x', t') in the boosted F' [and (x, t) in F] (Fig. I.10a). Thus if we look only at what happens relative to F, a point (x', t') goes under the boost to (x, t), related to (x', y') by the Lorentz transformation. (In this sense the boost corresponds to an inverse Lorentz transformation.) The movement of points under a boost is shown in Fig. I.10b. It is clear from the definition that length and time measurements are preserved under an active Lorentz transformation.

If we keep on repeating the boost for a particular relative velocity v, we get an infinite series of frames, each related to the previous one by a Lorentz transformation. We can see the effect on a unit time vector in A's frame (Fig. I.11): these arrows represent unit clock measurements made by observers moving at different velocities

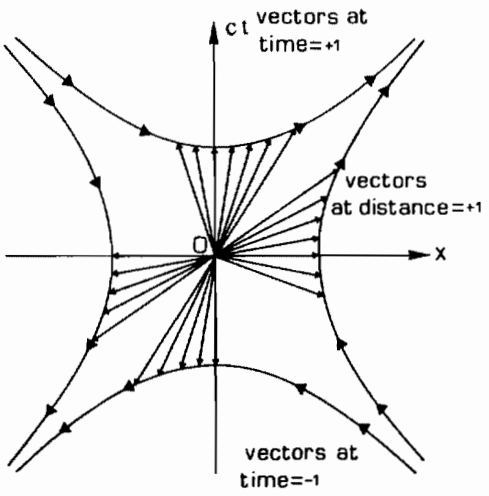


Fig. I.11 The effect of a repeated series of boosts on the unit time-like and space-like vectors along the axes of the reference frame of a fixed observer A

relative to A, and the surface they define is at unit proper-time from O. Thus this surface enables us to compare the units of time on different lines through the origin, representing the uniform motion of particles at different speeds. Similarly, repeated boosting of a spatial vector will give a series of vectors representing unit spatial measurements in the surfaces of simultaneity of the family of observers in the boosted frames, defining a unit spatial distance from O. This surface enables us to compare the units of spatial distance along different space-like lines, all passing through the origin.

3.2 Invariants and the metric form

We have seen how observers in relative motion disagree about simultaneity and about the coordinates they assign to events. Let us now define some quantities about which they agree.

As we have already seen, and as can be easily checked, the distance S from the origin O to an event (t, x, y, z) defined by

$$S^2 = -c^2t^2 + x^2 + y^2 + z^2$$

is invariant under Lorentz transformations. The surfaces of constant S^2 can be plotted (Fig. I.12), and we see how different regions correspond to events whose distances from the origin are time-like ($S^2 < 0$), null ($S^2 = 0$), or space-like ($S^2 > 0$).

By a shift of origin, we can show that the distance between any two points is also invariant, as is the infinitesimal distance defined by

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \ .$$

This expression is known as the *metric form*, and the proper-time τ is defined in terms of it by

$$\mathrm{d}s^2 = -c^2\mathrm{d}\tau^2 \ .$$

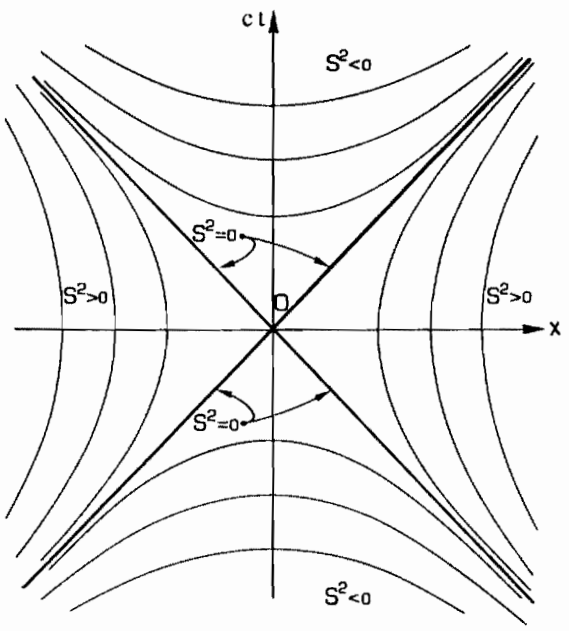


Fig. I.12 The surfaces $\{S^2 = \text{const}\}$

Of course, ds^2 can be expressed in terms of other coordinates; e.g. in spherical polars,

$$ds^2 = -c^2dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \ .$$

Given ds^2, we can find the light-cone of a point P by using the fact that light-rays are characterized by null intervals, $ds^2 = 0$. This leads to

$$c^2dt^2 = dx^2 + dy^2 + dz^2 \ ,$$

which shows that in flat space–time, the light-cones are parallel to each other (Fig. I.13).

The significance of the metric is that it enables us to calculate distances in space–time along any given path or world-line, which is clearly very important. (For example, the source of the time difference in the 'twin paradox' is the fact that the proper-time elapsed along different space–time paths is different.)

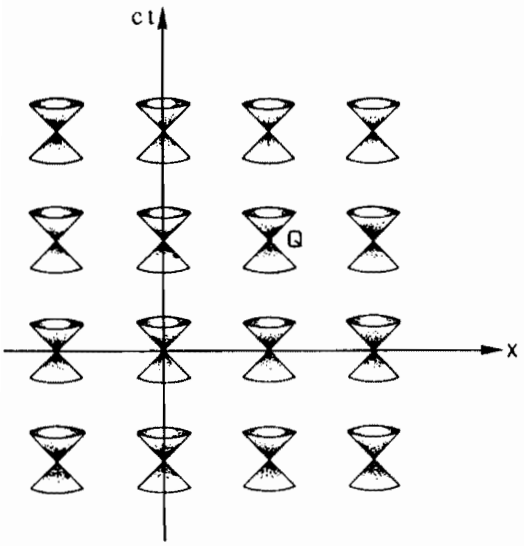


Fig. I.13 In a flat space–time, the light-cones at each point are parallel to each other

4. SOME FLAT-SPACE COSMOLOGICAL MODELS

I now want to illustrate these ideas about special relativity by describing three cosmological models in flat space–time (i.e. neglecting gravitational effects). Similar features will arise in some cosmologies in curved space–times, where we take gravity properly into account. For pedagogical reasons, I will concentrate on two-dimensional versions, but these can easily be extended to four dimensions. From now on, I will use units in which $c = 1$.

In the real universe, we observe matter clustered into galaxies that are measured to have systematically increasing red-shifts as their distances from us increase. This suggests that there is a well-defined average motion of matter in each region of the universe, and a model of the universe must therefore specify both the space–time itself and this average motion of matter. A space–time is a *model universe* when a family of preferred world-lines is specified in it, representing the average motion of matter at each point in space–time. These world-lines, called *fundamental world-lines,* represent the history of galaxies moving with the average motion of matter at each point. Observers moving with them are called *fundamental observers.* We can calculate the results of their observations and compare them with real observations in order to test how realistic the model is.

The models we consider here are based on symmetries of space–time which pick out certain world-lines to be 'naturally preferred'. We consider the Minkowski universe, the Rindler universe, which has some properties similar to those of a black hole, and the Milne universe, which is a simple expanding universe.

4.1 Minkowski universe

Let us consider the two-dimensional version first. This is just flat space–time with metric

$$ds^2 = -dt^2 + dx^2 \ ,$$

The world-lines of the fundamental observers are lines $\{x = \text{const}\}$ and the density of galaxies is uniform in the $\{t = \text{const}\}$ surfaces (Fig. I.14).

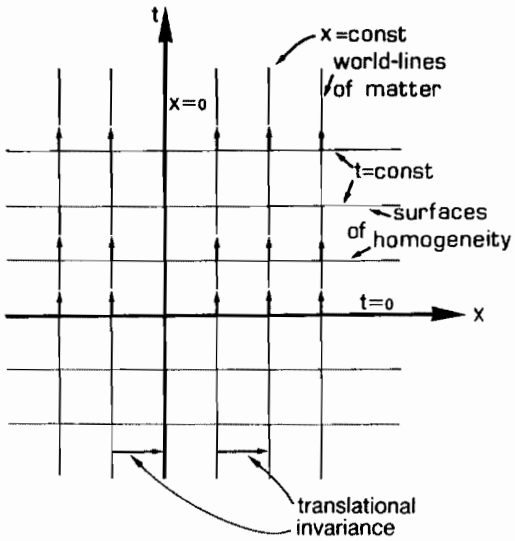


Fig. I.14 The Minkowski universe. The fundamental world-lines are $\{x = \text{const}\}$, and the surfaces of simultaneity and homogeneity are $\{t = \text{const}\}$.

This universe is based on the *translational invariance* of the space–time. The world-lines are moved into themselves by the time translation

$$t' = t + t_0 \ , \quad x' = x \ , \quad \text{for all } t_0 \ .$$

Thus the world-lines stay a constant distance from each other. Also they are moved into each other by the spatial translation

$$x' = x + x_0 \ , \quad t' = t \ , \quad \text{for some } x_0 \ ,$$

which implies spatial homogeneity. Notice that the translations are symmetries because they leave the metric invariant.

Whether we think of this universe in the continuum case where there is a world-line through every space–time point, or in the discrete case where there is an infinite set of uniformly distributed world-lines, we can construct the whole universe by starting with a single world-line and generating the others by translation. We clearly have a static uniform distribution of matter.

A four-dimensional version of this universe with metric

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$$

is the simplest kind of model universe—a static uniform distribution of matter in a flat space–time, without beginning or end or spatial limit. It is rather uninteresting because there are no red-shifts and it clearly does not correspond to the real universe.

4.2 Rindler universe

Although this model is again based on flat space–time, it exhibits some of the features of a black hole. It is based on the *boost invariance* or Lorentz invariance of flat space–time. For the two-dimensional version, we again start with flat space–time with metric

$$ds^2 = -dt^2 + dx^2 \ .$$

We use the spatial translation

$$x' = x + x_0 \ , \ t' = t \ , \quad \text{for some } x_0 \ ,$$

to determine the initial positions of a family of world-lines in the surface $\{t = 0\}$, resulting in an initially uniform distribution of matter. We then use the boosts about O to determine the world-lines L elsewhere from their initial positions (recall Fig. I.10b). The interval is invariant under boosts, so the distances between them remain constant in their surfaces of simultaneity, which are straight lines through the origin (see Fig. I.15). Thus at all times the fundamental observers measure the density of matter to be constant.

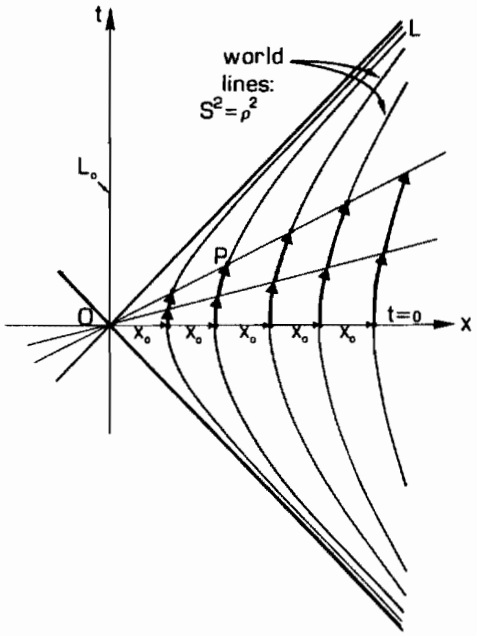


Fig. I.15 The Rindler universe. The fundamental world-lines L are obtained by boosts applied to their initial positions at equal distances along the x-axis

Consider a world-line through the point $(\rho, 0)$. A general point on the world-line is obtained from this by applying a boost through some velocity v to give

$$x = \gamma(v)\rho \ , \quad t = \gamma(v)v\rho \ .$$

Thus the equation of the world-line is

$$-t^2 + x^2 = \rho^2 \ ,$$

and v serves as a parameter along the world-line labelled by ρ.

Note that O is a fixed point of the boosts, so they do not generate a world-line L_0 through O. Instead, we define L_0 to be the line $\{x = 0\}$.

Let us now look at some of the other properties of this model universe.

4.2.1 Uniform acceleration

The world-lines L are not straight lines, so the observers must be accelerating. To determine the acceleration, consider two events $Q = (x, t)$ and $Q' = (x', t')$ on L $\{\rho = \rho_0\}$, related to each other by a boost through Δv (Fig. I.16). The interval of proper-time $\Delta\tau$ between Q and Q' is given by

$$\Delta\tau^2 = (t' - t)^2 - (x' - x)^2 \ .$$

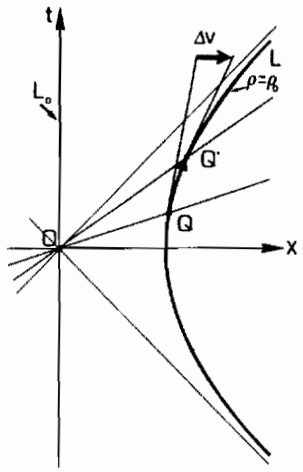


Fig. I.16 Two neighbouring points Q and Q' on the world-line $\{\rho = \rho_0\}$ have velocities differing by Δv

We substitute

$$x' = \gamma(\Delta v)[x + (\Delta v)t] \ ,$$
$$t' = \gamma(\Delta v)[t + (\Delta v)x] \ ,$$

[recall that boosting (x, t) to (x', t') in this sense involves what is effectively the inverse transformation] and we use

$$-t^2 + x^2 = \rho_0^2$$

to obtain

$$\Delta\tau^2 = 2[\gamma(\Delta v) - 1]\rho_0^2 \ .$$

Now for small Δv, which we need to consider so that the proper-time on the straight line between Q and Q' converges to the proper-time along the world-line between them, we have

$$\gamma(\Delta v) - 1 = (1 - \Delta v^2)^{-1/2} - 1 \simeq \frac{1}{2}\Delta v^2 \ ,$$

which gives

$$\Delta\tau = \Delta v \rho_0 \ .$$

Thus each observer will measure his acceleration to be

$$\frac{\Delta v}{\Delta \tau} = \frac{1}{\rho_0} \ ,$$

which is constant on each world-line, and is smaller the further the world-line is from O.

4.2.2 Red-shifts measured by fundamental observers

Let us now calculate the red-shift for observers O_1 and O_2 on world-lines $\{\rho = \rho_1\}$ and $\{\rho = \rho_2\}$ (Fig. I.17). Suppose that light emitted by O_1 at event r_1 is received by O_2 at event r_2. If event r_1 is boosted through Δv to an event r_1' a proper-time $\Delta\tau_1$ later, the light-ray is boosted to another light-ray which is received by O_2 at event r_2',

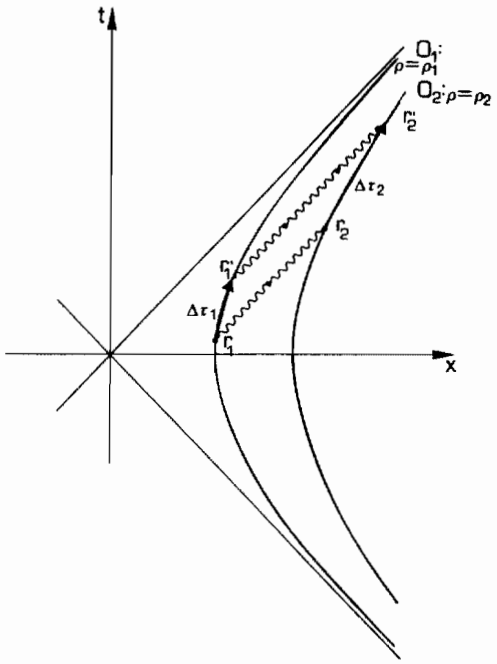


Fig. I.17 Light-rays used to determine the red-shift between observers O_1 and O_2

a proper-time $\Delta\tau_2$ later than r_2. Now we have already seen in the previous section that $\Delta\tau$ is proportional to ρ for fixed Δv. Hence the K-factor observed by O_2 is given by

$$K = \frac{\Delta\tau_2}{\Delta\tau_1} = \frac{\rho_2}{\rho_1} \ ,$$

which is independent of Δv and of t. We see that the red shift increases the further apart the world-lines of the two observers are.

4.2.3 The event horizon

We now consider light signals between a fundamental observer A on the world-line L and an observer A_0 on the world-line through the origin L_0. By trying to draw light-rays in Fig. I.18, we can see that an observer on L_0 can receive signals from L only when $t > 0$, and can send signals, which will be received on L, only when $t < 0$. Thus it is impossible for A to send a signal to A_0 and receive an answer! In fact, all events with $x < t$ cannot send signals to A_0 and events with $x < -t$ cannot receive signals from A. The surfaces $x = \pm t$ are called *event horizons,* and the observer A on

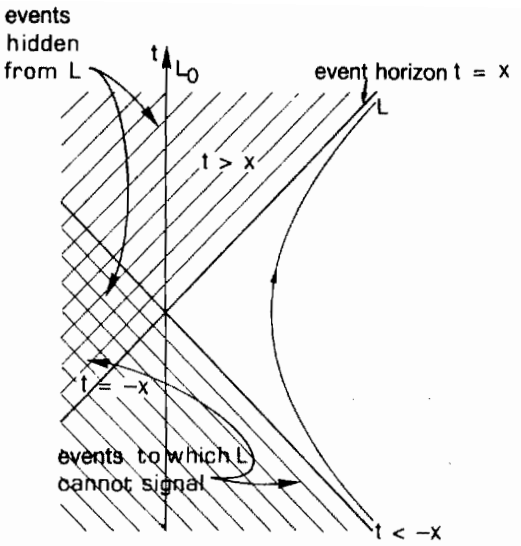


Fig. I.18 The event horizons $x = \pm t$ in a Rindler universe

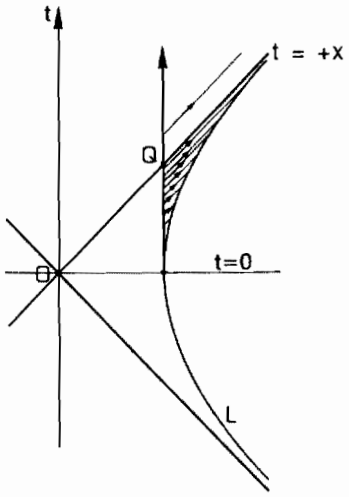


Fig. I.19 A massive object leaves the world-line L and crosses $x = t$ at the event Q

L cannot know what lies behind the future event horizon $x = t$ and cannot influence what happens behind the past event horizon $x = -t$.

Suppose a massive body leaves a world-line L and crosses the future event horizon at event Q (Fig. I.19). It can never re-cross the horizon and return to L because it would have to move faster than light. It is trapped by the event horizon, a surface in space–time which it can cross in only one direction. If it sends out signals before crossing the horizon, these will take longer and longer times to reach L. The red-shift will diverge, and the image intensity, which depends on K^{-4}, will tend to zero. Thus all activity on the body will appear to slow down and the image will fade away. However, as far as the body goes, nothing special happens to it locally at event Q as it crosses the horizon. This behaviour is exactly analogous to what happens as a body crosses the horizon of a black hole.

4.3 Milne universe

We again start with flat two-dimensional space with the usual metric. Let the world-line L_0 be $\{x = 0\}$, as in the Rindler universe. We repeatedly apply a boost to this through $\pm\Delta v$ to generate an infinite family of world-lines all passing through O (Fig. I.20). This represents an expanding universe with an infinite number of galaxies. Let us look at some of its properties.

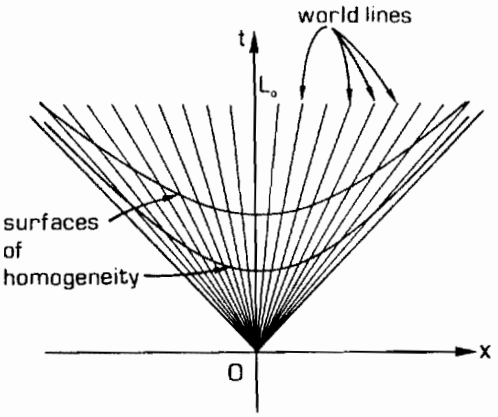


Fig. I.20 The Milne universe. The world-lines are generated by repeatedly applying a boost to L_0

4.3.1 Equivalent world-lines

All world-lines are constructed in the same way, and each fundamental observer will determine the same history for the universe as every other one. Thus this model obeys the *cosmological principle*: all fundamental observers are equivalent to each other. Note that the world-lines are all straight lines, representing inertial motion.

4.3.2 Homogeneous spatial sections

Consider the surfaces

$$t^2 - x^2 = \tau^2 \ .$$

These are at constant space–time distance from O, and in fact τ is just the proper-time along the world-lines. Let us look at the intersections Q and Q' of two world-lines with this surface (Fig. I.21). The space–time distance between them satisfies

$$\Delta\rho^2 = 2[\gamma(\Delta v) - 1]\tau^2$$

(cf. the calculation of $\Delta\tau^2$ in the Rindler universe), and in the limit of small Δv we obtain

$$\Delta\rho = \tau\Delta v \ .$$

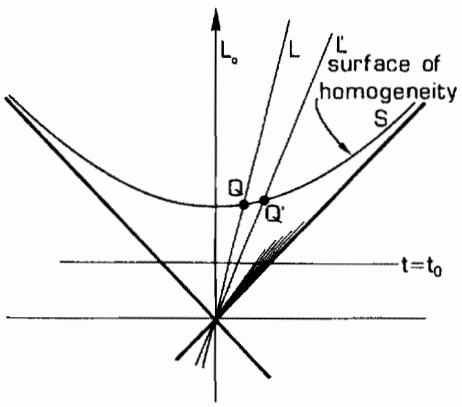


Fig. I.21 A boost through Δv applied to the event Q on L moves it to the event Q' where L' intersects the surface of homogeneity

Thus on a surface of constant τ, with boosts through fixed Δv separating the world-lines, the distance between world-lines is constant. Thus there is a uniform density of matter on each surface of constant τ.

4.3.3 Linear expansion and observed red-shifts

As we have seen, the distance between galaxies scales linearly with τ, so the matter in this universe is expanding uniformly. The observers move inertially, so the red-shifts are given by the formula we derived for constant relative velocity, and they increase systematically as one goes to further and further galaxies. A fundamental observer on L will measure all other galaxies to be receding linearly from him, but this will also be the experience of all the other observers. (Our diagram suggests that L_0 is privileged, but that is just because we have drawn it in terms of the coordinates of an observer on L_0.)

4.3.4 Initial singularity

If the expansion is followed back in time to O, there is a 'Big Bang' where all the world-lines intersect and the density of matter is infinite (see Fig. I.22).

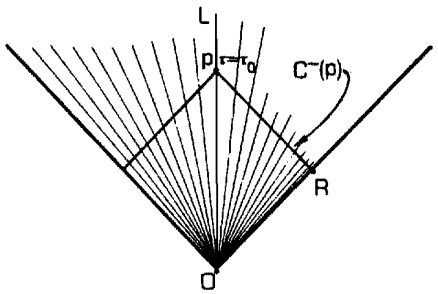


Fig. I.22 The past light-cone $C^-(P)$ of an event P on a world-line L intersects all the other fundamental world-lines

The past light-cone of any point P at time τ_0 on a world-line L intersects all the other world-lines, so in principle each observer can at all times see and communicate with every other galaxy (even though there are an infinite number of them). This means that there are no event horizons. However, the red-shifts will diverge as one looks to earlier and earlier times, and the intensity of the received light will fade away to zero.

Although each observer can receive signals from every galaxy in the universe, the distance measured by radar to the limiting observable event (e.g. R in Fig. I.22) is $\tau_0/2$ in each direction. Thus at time τ_0, the size of the observable is τ_0.

By extending these ideas to four dimensions we can construct a Milne universe with many of the properties of the expanding universes described in Lecture IV.

Lecture II: CURVED SPACE–TIMES

In this lecture, we shall look at some general features of curved space–times. In flat space–times, we can choose coordinates so that all the light-cones are parallel to each other, but in curved space–times the situation is very different. According to general relativity the gravitational fields of massive objects not only curve the paths of other massive objects but also cause light-rays to bend. This feature affects the causal and observational properties of curved space–times in intriguing ways.

1. GENERAL CONCEPT; PRINCIPLE OF EQUIVALENCE; GEODESICS

First of all, how do we distinguish a curved space–time from a flat one? Let us start with a two-dimensional space. One easy test for curvature is to attempt to flatten it out onto a plane (Fig. II.1); if distortion, gaps, or overlap occur anywhere, then it is curved. We see by this test that the surface of a cylinder is not curved, but that of a sphere is.

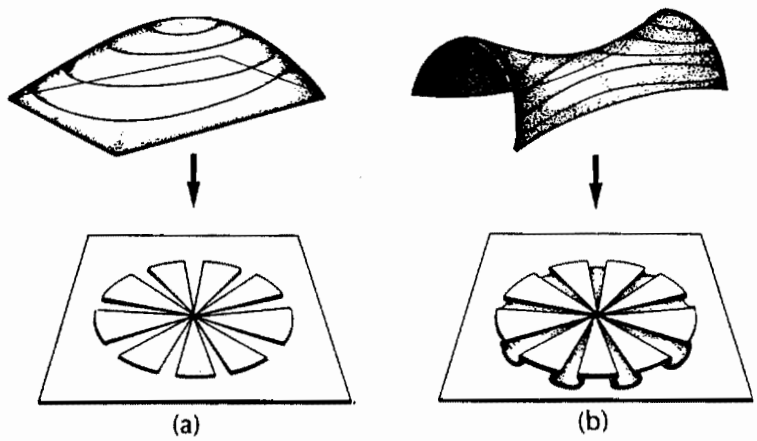


Fig. II.1 a) For a surface with positive curvature, there are gaps if it is flattened onto a plane. b) For a surface with negative curvature, there is overlap if it is flattened onto a plane

Geometric relations also differ in curved spaces. In flat spaces, the laws of Euclidean geometry hold: e.g. parallel lines never meet, the angles of a triangle add up to π. But this is not so in curved spaces. The analogues of straight lines are *geodesics* (the shortest lines between points, or, equivalently, curves that always move straight ahead without deviation). If we consider the surface of a sphere (Figs. II.2a,b),

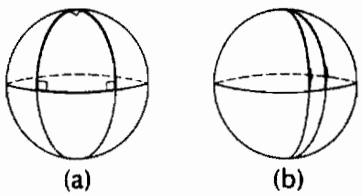


Fig. II.2 a) A spherical triangle formed by three great circles on the surface of the Earth (the equator and two lines of longitude meeting at a right angle at the North Pole). Each of the angles of the triangle is a right angle. b) Two great circles (lines of longitude), parallel to each other at the equator, meet at the North Pole.

geodesics parallel to each other at one stage can meet, and the angles of a triangle can add up to more than π!

Higher-dimensional curved spaces and curved space–times are more difficult to visualize, but similar effects occur.

To see why we actually need curved space–times, let us consider the relationship between acceleration and gravity. We look at four observers in different physical situations. Observer A is in a lift at rest relative to the Earth. If he releases an object it will fall to the ground (Fig. II.3a). Observer B (Fig. II.3b) is in a rocket moving with constant acceleration g far from any massive body. The results of experiments will be the same for him as for observer A. Observer C is in a lift which is falling freely under gravity because the cable has broken (Fig. II.3c). He will be falling at the same rate as any object he releases, so his measurements will be the same as for an observer D (Fig. II.3d) in a stationary rocket far from any gravitational field. The experiences of these observers can be summarized in the *principle of equivalence:* there is no way of distinguishing between the effects, on an observer, of a uniform gravitational field and of constant acceleration.

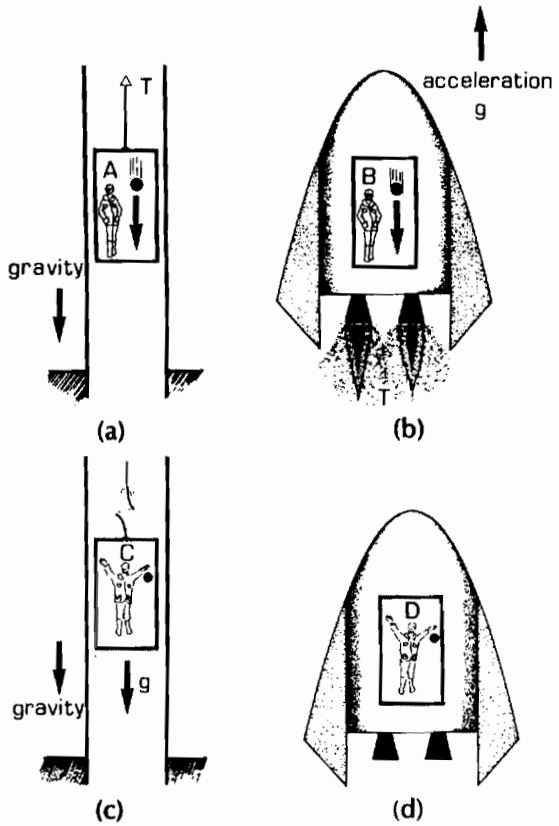


Fig. II.3 a) Observer A is in a lift at rest relative to the Earth. b) Observer B is in a rocket moving with constant acceleration g far from any massive body. c) Observer C is in a lift falling freely under gravity. d) Observer D is in a stationary rocket far from any gravitational field.

If we can mimic a gravitational field by going to an accelerating frame, why do we need curved space–times? Let me give you two reasons.

i) Observer D will see light-rays travelling in straight lines across the cabin of his rocket, so they will also be straight for the freely falling observer C in his lift

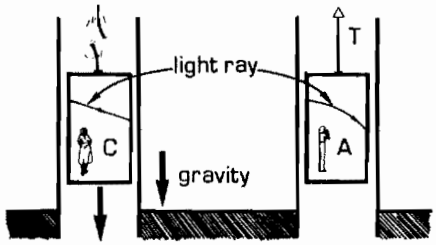


Fig. II.4 Observer C will measure a light-ray travelling across the lift to move in a straight line. The same light-ray will appear curved to observer A.

(Fig. II.4). However, the stationary observer A will regard C's light-ray as being bent down because C is accelerating relative to A, and will conclude that space–time is not flat. Thus to be able to describe the experiences of all observers, we need curved space–times.

ii) The second argument (Figs. II.5a–c) is that real gravitational fields are non-uniform, so whilst we can transform away the effect at one point, it will not go at other points. For example, to transform away the gravitational field of the Earth (Fig. II.5a), we would need infinitely many accelerated frames. However, if we allow curved space–times, then we can represent any gravitational field in a single frame.

It follows that in general relativity where we use accelerated frames, it is no longer possible to make a clear-cut distinction between motion under gravity and inertial motion (motion under no forces), because what is inertial motion in one frame

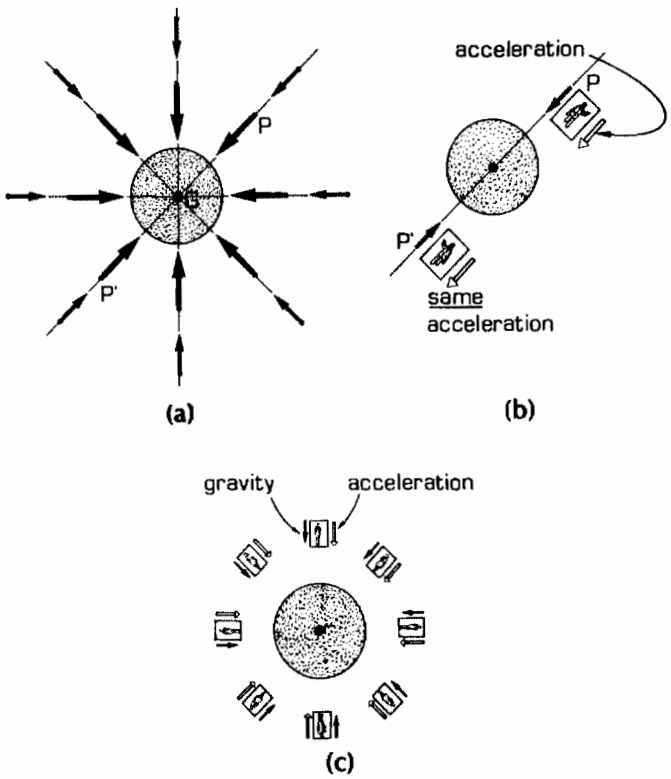


Fig. II.5 a) The direction of the gravitational field at various points around the Earth. b) An acceleration which transforms away the gravitational field at P will double it at P'. c) In a flat space–time, a separate accelerated frame is needed at each point to transform away the gravitational field.

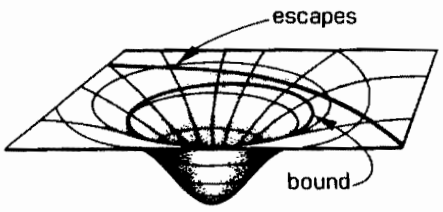


Fig. II.6 A ball moving on a curved surface is held in its path by the shape of the surface

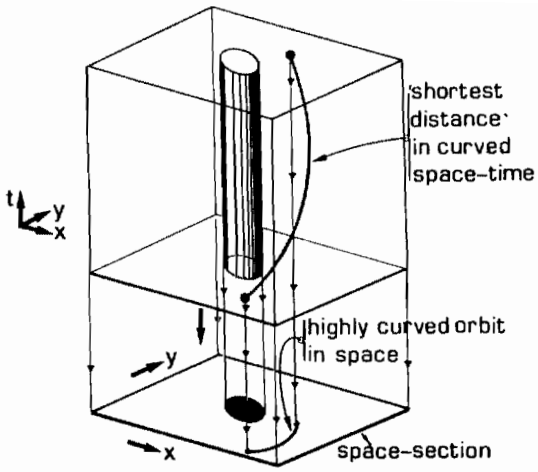


Fig. II.7 The orbit of a planet around the Sun

may not be inertial in another. However, we can give a clear meaning to *free fall,* which is motion under gravity and inertia alone. We then postulate that the paths of freely falling objects in space–time are time-like geodesics. This prescription gives a satisfactory description of all observed motion.

Consider the planets for example; gravity curves their paths through space–time. (Note the inherent non-linearity of the theory: massive bodies produce space–time curvature which then affects their motion! This non-linearity is the reason why some calculations in curved space–time are so difficult. However, for the moment we shall consider the motion of test particles, neglecting their effect on the curvature of space–time.) An analogue of planetary motion is that of a ball on a curved surface; its motion is determined by the structure of the space (Fig. II.6). Similarly, the planets are held in their orbits by the curvature of space–time caused by the gravitational field of the Sun (Fig. II.7). They follow paths corresponding to the shortest distance in space–time, which actually corresponds to the longest proper-time. The spatial projection of such a path can be highly curved.

We cannot measure the strength of a gravitational field by the amount it bends a single light-ray or particle path, because this depends on the reference frame used. (For example, in the case of a particle we could take a reference frame moving with it.) However, we can look at the relative motion of two particles or light-rays in order to investigate the space–time curvature and the related gravitational field. For example, if we consider two particles falling freely towards a star, we can see that the geodesics become closer to each other in time and eventually meet (Figs. II.8a,b). The shorter the time before this happens, the stronger the gravitational field.

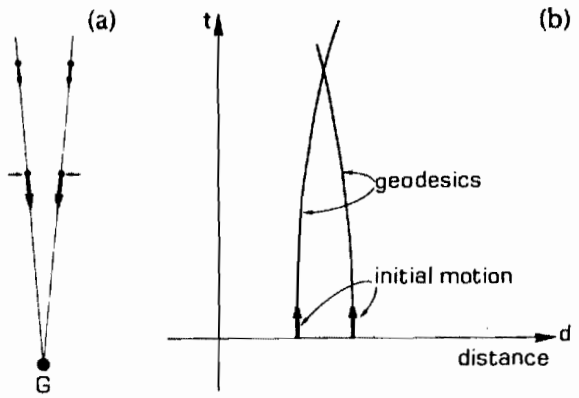


Fig. II.8 a) Two particles falling freely from rest towards a star G. b) The space–time paths of the particles.

2. THE METRIC FORM

Let us now consider how to describe curved space–times mathematically. As we saw for flat space–times, the metric form determines all distance measurements and angles, so we can describe a curved space–time by giving ds^2 in some suitable coordinate system.

We shall first look at the case of two dimensions. Consider, for example, the metric form for the surface of a sphere of radius a:

$$ds^2 = a^2(d\theta^2 + \sin^2\theta\, d\phi^2) \ ,$$

where θ and ϕ are standard polar coordinates (Fig. II.9a). Increments in θ and ϕ produce displacements of $a\ d\theta$ and $a\sin\theta\ d\phi$ along the lines of longitude (ϕ = constant) and latitude (θ = constant), as shown in Fig. II.9b. For small increments, the displacements form two sides of an approximately flat right-angled triangle, and Pythagoras' theorem gives ds^2, the square of the hypotenuse, as in the metric form.

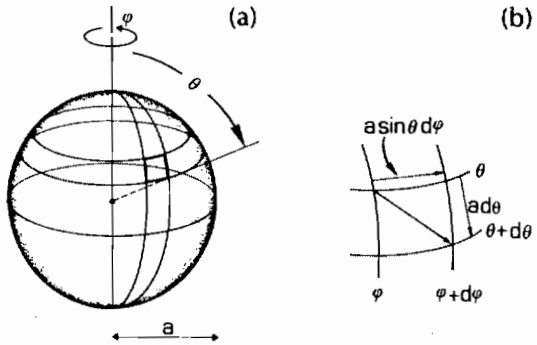


Fig. II.9 a) The angles θ and ϕ used to describe position on the surface of a sphere. b) The displacements on the surface of the sphere produced by small increments in θ and ϕ.

We see that the geometry of the curved space agrees with the flat-space result in the limit of very small displacements. Thus locally a curved space is just like a flat space. The distinction is that for a flat space we can find coordinates where the metric takes the form

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

everywhere, whereas this is not possible for a curved space. Similarly, for a flat space–time there are coordinates for which the metric is everywhere of the form

$$ds^2 = -dt^2 + dx^2 \ ,$$

but again this is not possible in the curved case. The extension to four dimensions follows immediately by adding extra spatial increments to the metric form.

Now let us suppose that we are given the metric form ds^2 for a four-dimensional space–time. This then determines all time measurements by ideal clocks (moving on time-like curves for which $ds^2 < 0$) through .

$$\tau = \int(-ds^2)^{1/2} \ .$$

It also determines the motion of light (paths for which $ds^2 = 0$). Thus it defines the light-cone at each point and therefore also the causal structure of the space–time. As an example, suppose that we are given a metric

$$ds^2 = -dt^2 + t^{4/3}(dx^2 + dy^2 + dz^2) \ .$$

How do we interpret this? Firstly, along a world-line $\{x, y, z \text{ const}\}$, we have $dx = dy = dz = 0$, giving $ds^2 = -dt^2$, and so t measures proper-time along these world-lines, the fundamental world-lines of the universe described by this metric. Secondly, along a curve $\{t, y, z \text{ const}\}$, we obtain $ds^2 = t^{4/3}\, dx^2$, so the proper distance along the curve is measured by $t^{2/3}x$ rather than x, which implies (as we shall see later) that this is an expanding universe. Thirdly, the light-cones are determined by $ds^2 = 0$, so a displacement along a light-cone must satisfy

$$dt^2 = t^{4/3}(dx^2 + dy^2 + dz^2) \ .$$

Projection onto a plane $\{y, z \text{ const}\}$ gives

$$dt = \pm t^{2/3}\, dx \ .$$

Thus for small t, a displacement dx gives small dt, whereas for large t, dt is much larger. This means that the light-cones flatten out as the $t = 0$ surface is approached, as shown in Fig. II.10.

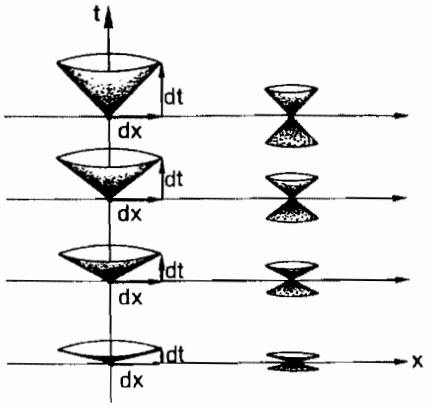


Fig. II.10 The light cones for the metric form $ds^2 = -dt^2 + t^{4/3}(dx^2 + dy^2 + dz^2)$

We see that the geometry of space–time is determined or described by the metric ds^2. What then determines the form of the coefficients in ds^2? They are actually determined by the distribution of matter and energy in the universe through Einstein's equations, which are non-linear tensor equations relating the geometry of the space–time to its matter content. We shall not look at these equations in these lectures but just examine the properties of various solutions.

3. LIGHT-RAYS AND CAUSALITY

We have already identified the paths of freely falling objects with time-like geodesics, obtained by finding the extremal value of $\tau = \int(-ds^2)^{1/2}$, corresponding to maximum proper-time. We now identify light-rays in a curved space–time with *null geodesics*. Let us look at some of their properties.

3.1 Bending

We have already seen that the stationary observer A observes the light-rays in the freely falling lift (C's lift) to be bent (see Fig. II.4). It is true, in general, that relative to an observer at rest on a massive body, light-rays will be bent by the gravitational field of that body. Experimental evidence for this was found from the apparent positions of stars in a solar eclipse in 1919, and led to widespread acceptance of the theory of general relativity.

3.2 Gravitational red-shifts

For a rocket in free fall, observer D should measure no change in frequency for light emitted from the floor and detected at the roof. However, for observer B in an accelerating rocket (Fig. II.11a), the roof accelerates away from the position of the floor when the light was emitted. Thus in every time interval measured by B, the light has to travel further before reaching the roof, so observer B will detect a red-shift (cf. calculation of the red-shift in the Rindler universe). By the principle of equivalence, the same will be true for an observer A stationary on the surface of the Earth (Fig. II.11b). This is gravitational red-shift and has been verified experimentally by observations of distant stars and also by experiments at Harvard Tower.

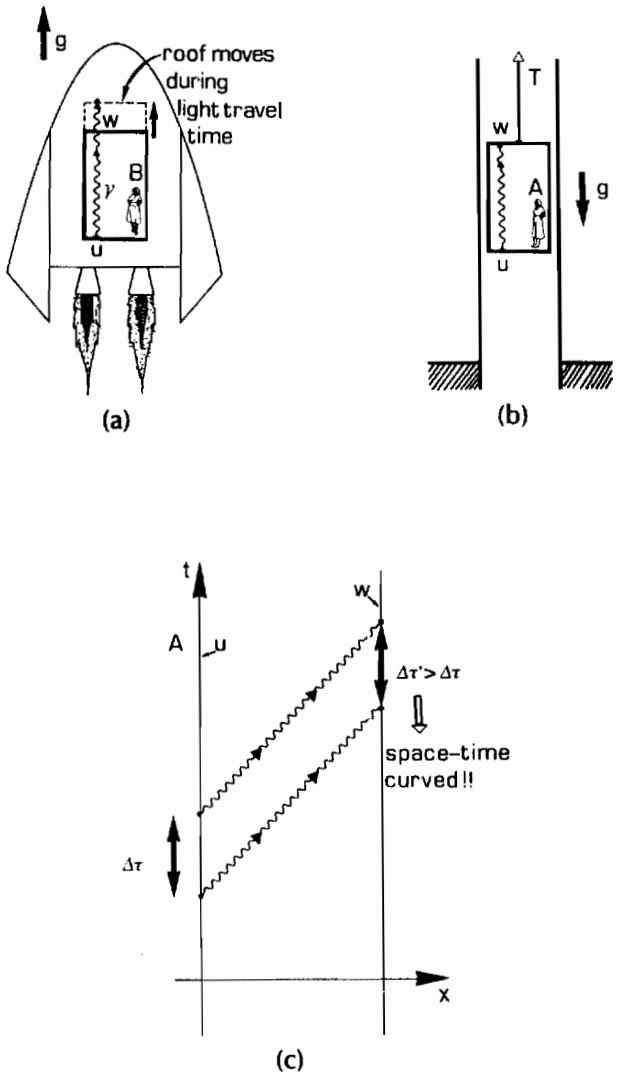


Fig. II.11 a) Light is observed to be red-shifted in an accelerated rocket. b) Light must also be red-shifted in a stationary lift in the Earth's gravitational field. c) The time interval $\Delta\tau'$ between reception of signals is greater than the time-interval $\Delta\tau$ between their emission, not because of relative motion of the emission point u and the reception point w, but because of space–time curvature.

3.3 Geodesic deviation of light-rays

The bending of light-rays means that the relationship between observed angles and sizes is changed from the flat-space case (Figs. II.12a–c). In particular, if light-rays are bent towards each other, as we expect for an attractive gravitational field, they will be closer at the object than would appear from their angular separation and the object will appear to be larger than its real size (Fig. II.12c).

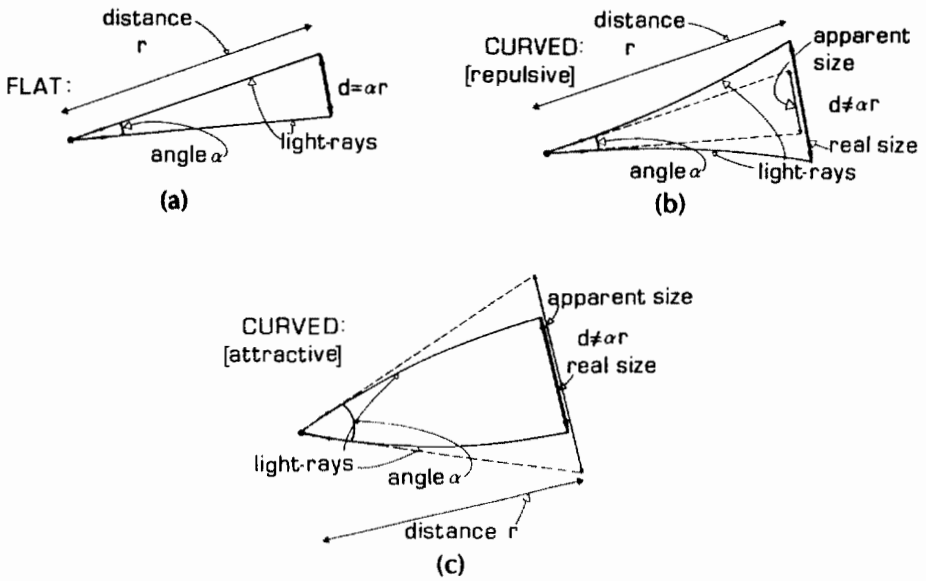


Fig. II.12 a) In a flat space, the size d of an object viewed with angular width α at a distance r must be αr. b) In a space of negative curvature, the apparent size αr will be smaller than the true size d. c) In a space of positive curvature, αr will be larger than d.

3.4 Gravitational lensing

In extreme cases, the presence of matter can cause sufficient bending to produce refocusing of the light-rays. This lensing effect can occur locally or over the whole past light-cone.

3.4.1 Local lensing

In a cosmological model, a massive object can refocus light-rays from more distant objects to produce multiple images. This effect has been observed in double images of quasars, where the spectra of the light from the two images show that it originates at the same source (Fig. II.13).

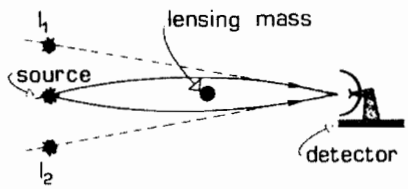


Fig. II.13 A massive object refocuses light from a more distant source, producing multiple images I_1 and I_2 of the source

3.4.2 Large-scale refocusing

In this case, the light-cone as a whole can be bent back on itself. In flat space–time, the area of a wavefront necessarily increases with the distance from the observer (Fig. II.14a). In a curved space–time this is not so, because neighbouring light-rays are focused towards each other. Our past light-cone will reach a maximum distance from our world-line C and then start refocusing towards it (Fig. II.14b). We believe that the density of matter in the real universe is sufficient to cause this kind of refocusing.

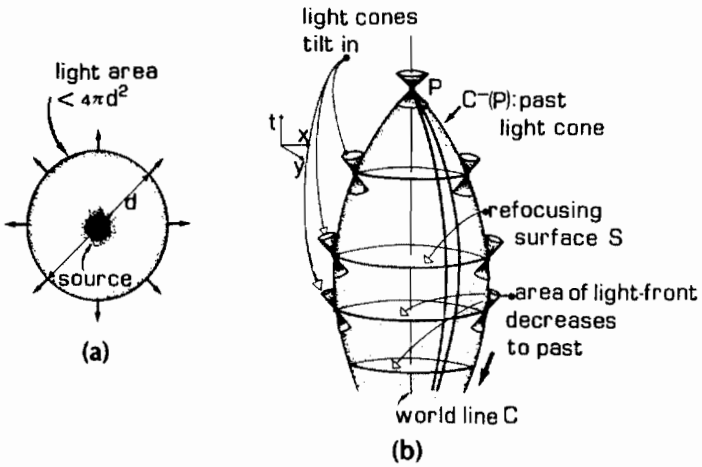


Fig. II.14 a) Light spreading out from a source at distance d has area less than $4\pi d^2$. b) In a space–time view, the light cone reaches a surface S of maximum area and then bends back on itself. (The local light-cones tip over, remaining tangent to the light-cone of P.)

We see that in a curved space–time the local behaviour of the light-cones can be very different from that in a flat space–time, which means that the causal properties can also be very different. One feature is the existence of certain types of *horizons* which limit observation in various ways. The simplest example is our past light-cone; we shall go on to discuss the event horizon around a black hole (cf. the Rindler universe) and a particle horizon in cosmology. Another possible feature is the violation of our ideas of causality. To see how this can happen, notice that the local light-cones can tip over relative to each other, as we might expect, for example, in a rotating system. The speed of light determined by them is still a limiting speed, so the light-cones and associated particle paths and light-rays still determine what parts of the space–time can influence the other parts. An example where this happens is Gödel's stationary rotating universe. On the axis the light cones are vertical, but away from the axis the rotation causes them to tilt over (Fig. II.15). This tilting increases with

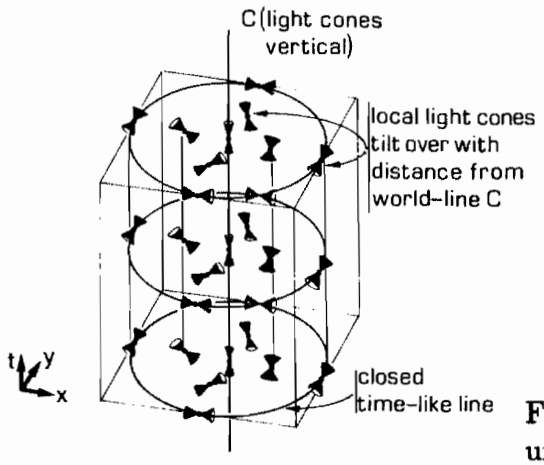


Fig. II.15 The light-cones in Gödel's stationary universe

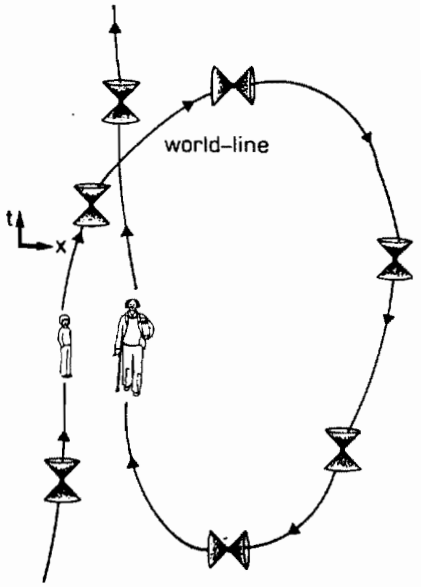


Fig. II.16 In a universe with closed time-like lines, world-lines can come back to themselves so it would be possible for an old man to stand next to himself as a child

distance from the axis, so that eventually they are horizontal and then there are closed time-like lines. As shown in Fig. II.16, this gives rise to the possibility of bizarre causal violations, such as an adult returning to stand next to himself as a child! We have no evidence that this happens in the real universe, but it is a theoretical possibility in some curved-space–time model universes.

Lecture III: SPHERICAL STARS AND STELLAR COLLAPSE

Having discussed general properties of curved space–times, we shall spend the remaining lectures looking at two of the most basic examples: the space–time produced by an isolated body such as a star, and the space–time produced by the matter in the universe as a whole. In this lecture, we study the gravitational field of a spherically symmetric star and consider how stellar collapse can take place, leading to the formation of a black hole.

1. THE SCHWARZSCHILD SOLUTION AND ITS PROPERTIES

A single massive object such as the Earth or the Sun produces curvature in the space–time around it. If we assume that it is spherically symmetric and isolated from other massive objects (as an approximation), then it can be shown from Einstein's equations that the space–time around it is described by the exterior Schwarzschild solution, with metric interval

$$\mathrm{d}s^2 = -\left(1-\frac{2m}{r}\right)\,\mathrm{d}t^2 + \frac{1}{1-(2m/r)}\,\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\ ,$$

where m is its mass in geometric units. This is valid for $r > R_s$ (the value of r at the surface of the object), and, as we shall see, we require $R_s > 2m$. For simplicity, we shall refer to the object as a star. Let us look at the properties of the space–time.

1.1 Symmetries

The space–time is static (unchanging in time) because the metric coefficients are independent of time. We shall refer to observers at constant r, θ, and ϕ as static observers, since for them all properties are constant in time. The space–time is also spherically symmetric about the central body. To see this, note that $r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$ is just the metric for a two-sphere of radius r, which is clearly spherically symmetric.

1.2 Distances and times

The $(1-2m/r)$ factors in the metric mean that distances in the r-direction and times are not the same as they would be in flat space–time.

1.2.1 Radial distances

The significance of the coordinate r lies in its association with area; it is chosen so that the area of the two-sphere $\{t, r \text{ const}\}$ is $4\pi r^2$ (this follows from putting $\mathrm{d}t = \mathrm{d}r = 0$ in the metric from). However, r does not directly measure distance between the two-spheres, as it would in flat space–time. To find the distance D between two-spheres at $r = r_1$ and $r = r_2$ (Fig. III.1), we must integrate $\mathrm{d}s$ with $\mathrm{d}t = \mathrm{d}\theta = \mathrm{d}\phi = 0$:

$$D = \int_{r_1}^{r_2} \left(1 - \frac{2m}{r}\right)^{-1/2} \mathrm{d}r$$
$$= r_2 \left(1 - \frac{2m}{r_2}\right)^{1/2} - r_1 \left(1 - \frac{2m}{r_1}\right)^{1/2} + 2m \ln \left[\frac{(r_2 - 2m)^{1/2} + r_2^{1/2}}{(r_1 - 2m)^{1/2} + r_1^{1/2}}\right] .$$

It can be shown that this is larger than $r_2 - r_1$, the flat-space value, which indicates that the space–time has positive curvature.

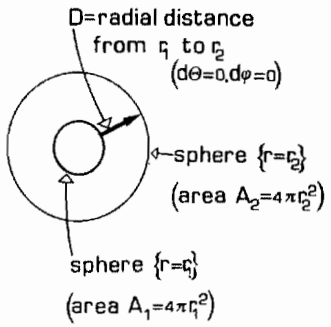


Fig. III.1 A spatial section $\{t = \text{const}\}$ of the Schwarzschild solution

1.2.2 The time coordinate

We must see how t is related to the proper-time τ measured by a static observer (they would coincide in flat space–time). The proper-time interval $\Delta\tau$ corresponding to a coordinate time interval $\Delta t = t_2 - t_1$ for a static observer with $\mathrm{d}r = \mathrm{d}\theta = \mathrm{d}\phi = 0$ is given by

$$\Delta\tau = \int (-\mathrm{d}s^2)^{1/2} = \int_{t_1}^{t_2} \left(1 - \frac{2m}{r}\right)^{1/2} \mathrm{d}t = \left(1 - \frac{2m}{r}\right)^{1/2} \Delta t \ ,$$

which is less than Δt, since $r > R_s > 2m$.

1.3 Asymptotic behaviour

For very large r, $2m/r$ becomes negligible, and we obtain

$$\mathrm{d}s^2 \simeq -\mathrm{d}t^2 + \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta \ \mathrm{d}\phi^2) \ ,$$

the flat-space metric in spherical polars. Thus the solution represents an asymptotically flat space–time, as we would expect.

Even when r is not very large, it turns out that for many stars we can obtain a good approximation that is valid everywhere outside the surface, by just retaining the first-order terms in m/r. This is because $r > R_s$ implies that $m/r < m/R_s$, and m/R_s is very small for many bodies (e.g. for the Earth, $m/R_s = 6.9 \times 10^{-10}$, and for the Sun, $m/R_s = 2.1 \times 10^{-6}$). In this case the metric form is given by

$$\mathrm{d}s^2 \simeq -\left(1 - \frac{2m}{r}\right) \mathrm{d}t^2 + \left(1 + \frac{2m}{r}\right) \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta \ \mathrm{d}\phi^2) \ ,$$

which results in

$$D = r_2 - r_1 + m \ln (r_2/r_1)$$

and

$$\Delta\tau = \left(1 - \frac{m}{r}\right) \Delta t \ .$$

1.4 Red-shifts

There are observable gravitational red-shifts in Schwarzschild space–times. We shall derive the red-shift formula for two static observers situated radially relative to each other, at the same values of θ and ϕ but at different values r_1 and r_2 of r, as shown in Fig. III.2. The path of a light-ray travelling between them will satisfy $\mathrm{d}\theta = \mathrm{d}\phi = 0$ and $\mathrm{d}s^2 = 0$, and so $\mathrm{d}r$ and $\mathrm{d}t$ will satisfy

$$\frac{\mathrm{d}r}{\mathrm{d}t} = 1 - \frac{2m}{r} \ .$$

If the signal is emitted at t_1 and received at t_2 (Fig. III.3), then integration of this expression gives

$$t_2 - t_1 = r_2 - r_1 + 2m \ \ln\left(\frac{r_2 - 2m}{r_1 - 2m}\right) \ .$$

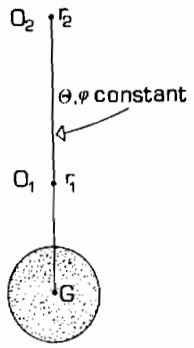


Fig. III.2 Two static observers O_1 and O_2 on the same radial line

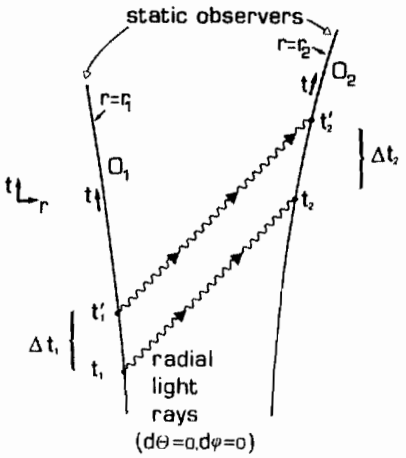


Fig. III.3 Radial light-signals are emitted by O_1 at t_1 and t'_1 (at coordinate interval Δt_1) and received by O_2 at t_2 and t'_2 (at coordinate interval Δt_2)

Similarly, for the second signal emitted at t'_1 and received at t'_2

$$t'_2 - t'_1 = r_2 - r_1 + 2m \ \ln\left(\frac{r_2 - 2m}{r_1 - 2m}\right) \ .$$

Upon subtraction, we see that

$$\Delta t_1 = t'_1 - t_1 = t'_2 - t_2 = \Delta t_2 \ .$$

The K-factor is the ratio of the corresponding intervals of proper-time:

$$K_{12} = \frac{\Delta\tau_2}{\Delta\tau_1} = \left(\frac{1-2m/r_2}{1-2m/r_1}\right)^{1/2} \quad .$$

We see that, as in the flat-space case, the red-shift is independent of t and of Δt (which is essential since the space–time is static). However, it is no longer true that $K_{12} = K_{21}$. In fact

$$K_{21} = 1/K_{12} \ ,$$

and light travelling towards the star will be blue-shifted. These results have been observed experimentally (e.g. in observations of white dwarf stars, and in experiments at the Harvard Tower).

One can derive other properties of the Schwarzschild space–time from the metric form: for example, perihelion shifts in particle orbits and the bending of light, the predictions of which form the basis of the classic tests of general relativity. The results of these tests strongly support the validity of the Schwarzschild solution as a description of the local gravitational field in many parts of the universe.

2. SPHERICAL COLLAPSE; BLACK HOLES; THE EVENT HORIZON

We now consider what happens when an isolated spherical body such as a massive star collapses to form a 'black hole'. Two of the most important effects are the creation of a singularity at the end of the collapse, and the formation of an event horizon which restricts communication between the star and the outside world, hiding the singularity. These causal features follow from the structure of the light-cones in the Schwarzschild space–time.

2.1 Applicability of the Schwarzschild solution

According to Birkhoff's theorem, the Schwarzschild solution represents the gravitational field of any spherically symmetric star, not only if it is static but also if it is collapsing, expanding, or pulsating. Thus the metric form is of very wide applicability.

Since the coefficient of dr^2 diverges at $r = 2m$, the metric is clearly singular there. However, it can be shown that it is the coordinates that are badly behaved, not the space–time itself. It is not a curvature singularity but just a coordinate singularity.

2.2 Use of a null coordinate to describe collapse

Consider a star in which the density is so high that the gravitational forces dominate and it shrinks in upon itself, collapsing to form a black hole (Fig. III.4). The surface radius R_s decreases to zero and a physical singularity occurs because the star has zero volume and infinite density. During the collapse, the interior geometry will be described by some dynamic metric, which we shall not investigate here; we are interested only in the exterior solution.

As it collapses, the star's surface will fall through the critical value $R_s = 2m$, so we need a new coordinate system that will cover $r = 2m$ in a non-singular manner.

There are various possibilities; we shall use the *Eddington–Finkelstein coordinates*, which are particularly appropriate for gravitational collapse. The resulting metric for the exterior space–time (i.e. for $r > R_s$) is

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dv^2 + 2\,dv\,dr + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \ ,$$

where $v \equiv t + r + 2m \ln(r/2m - 1)$. Note that this is just the Schwarzschild solution in new coordinates that are well-behaved at $r = 2m$, and so allows us to investigate what happens when objects cross the surface $r = 2m$.

Let us consider radial light-rays. Using $d\theta = d\phi = 0$ and $ds^2 = 0$ we obtain

$$dv\left[2dr - \left(1 - \frac{2m}{r}\right) dv\right] = 0 \ .$$

Hence the light-rays are given by $dv = 0$, i.e. $v =$ constant (so v is called a null coordinate) and by $dr = \frac{1}{2}(1 - 2m/r)\ dv$, which we shall interpret in a moment. It is convenient to draw the space–time in the following way. The vertical axis represents time (but surfaces {t = const} are not perpendicular to it); r and θ are polar coordinates in planes perpendicular to the t-axis, so surfaces of constant r are cylinders around the t-axis; the coordinate ϕ has been suppressed. Lines of constant v, which correspond to one generator of the light-cones, are drawn at 45° to the t-axis.

Figure III.4 represents not only the exterior solution but also the interior one, inside the surface $r = R_s$. (The interior solution is not represented in any detail because it depends on the equation of state of the star.) The surface radius R_s decreases steadily with time until it reaches zero, and the remains of the star form a singularity which is inside the surface $r = 2m$. Outside $r = R_s$, we have the exterior solution in the new coordinates. Let us look in more detail at the structure of the light-rays. The ingoing ones, corresponding to constant v, are at 45° to the t-axis. The outgoing ones, which satisfy

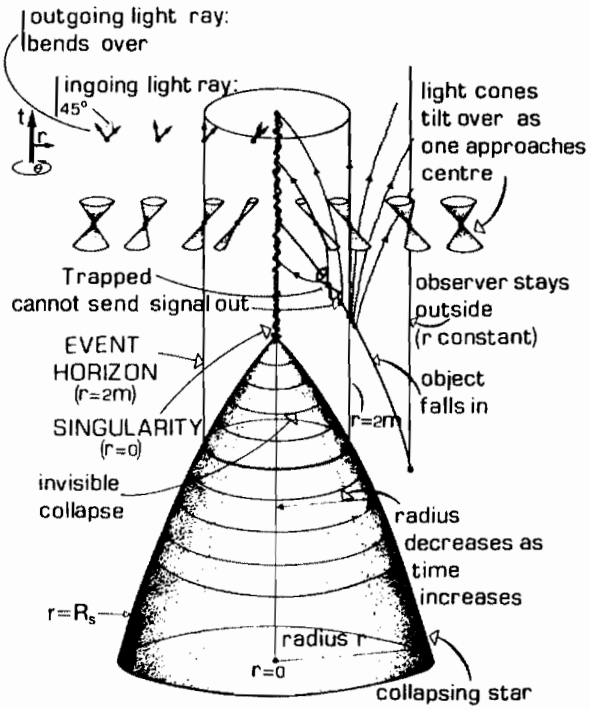


Fig. III.4 A space–time diagram of the collapse of a star to form a black hole

$$\frac{\mathrm{d}r}{\mathrm{d}v} = \frac{1}{2}\left(1 - \frac{2m}{r}\right) \ ,$$

clearly depend on radial distance. For large r, they behave as though space were flat, but as r decreases, they tilt over towards the spatial origin; they are parallel to the t-axis for $r = 2m$ and approach $v = \text{const}$ as r decreases to zero. Thus, far outside the surface they are tilted outwards, whilst inside they are tilted inwards; the light-cones actually 'tip over' under the influence of the gravitational field. Notice that the surface $r = 2m$ is in fact a null surface, and this is why it has physical significance. We shall see how the causal properties of the space–time follow from the structure of the local light-cones.

2.3 The event horizon

The surface $r = 2m$ is a one-way trapping surface which allows matter and radiation to fall inwards but not escape outwards. Why should this be? The point is that at $r = 2m$, an outgoing light-ray is just held back by the star's gravitational field. Light emitted just outside the event horizon can escape to infinity, but light emitted just inside cannot escape because the light-rays are dragged back by the gravitational field and eventually fall into the singularity. Clearly, any massive object inside the horizon cannot escape because it cannot exceed the speed of light; its possible future histories are bounded by the light-rays. No matter how hard he accelerates, an observer who has crossed the horizon cannot go back through it (and indeed cannot escape falling into the singularity). Hence the name 'black hole': in the classical theory, no radiation or signal of any kind can reach the outside from the inside. There is no way in which external observers can know what is happening inside the horizon.

This trapping effect happens at very small radii: e.g. the Sun would have to be compressed from a radius of nearly 700,000 km to less than 3 km, and the Earth to less than 0.9 cm.

2.4 Collapse seen from outside

Let us consider how the collapse appears to an external observer. A stationary observer O_1 outside the event horizon sees the star shrinking towards $r = 2m$ but never actually reaching this radius, because light from there would take infinitely long to reach him. An inward-moving observer O_2 will see the collapse, but he himself will be drawn into the singularity and will not be able to send signals back to O_1 once he has crossed the horizon (Fig. III.5).

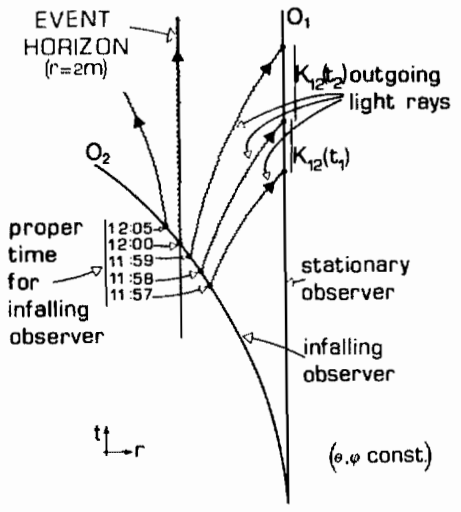


Fig. III.5 An in-falling observer O_2 emits radial light-signals each minute; a stationary observer O_1 receives them at longer and longer intervals, and the final minute before O_2 crosses the horizon appears to O_1 to last for ever

How will the inward-moving observer O_2 look to the stationary observer O_1? When they are at the same radial distance there will be a Doppler shift due to their relative velocities, as in special relativity. Then as O_2 moves away from O_1, there will also be a gravitational red-shift,

$$K_{21} = \left(\frac{1-2m/r_1}{1-2m/r_2}\right)^{1/2} ,$$

which gets larger and larger as O_2 approaches $r_2 = 2m$, where it becomes infinite. Thus the event horizon is a surface of infinite red-shift (as in the Rindler universe).

It is clear, then, that the surface of the star will also be infinitely red-shifted as it crosses the horizon and the observed intensity will decrease to zero, so the star will just fade away. Note that the speed at which the surface of the star crosses the horizon is not necessarily large (and is certainly less than c) so the infinite red-shift is purely a gravitational effect.

2.5 The central singularity

What happens at the central singularity? The gravitational field is unbounded so any object there will be torn apart by infinite tidal forces. In fact our model of space–time breaks down there; general relativity predicts an end to space–time at that point. In order to say anything more about what happens there, we need a theory that takes quantum effects into account. In fact we know that moving from the purely classical theory to quantum gravity produces a rather different picture of black holes. Hawking has shown that, owing to quantum effects, one would expect a black hole to emit black-body radiation with a temperature depending on its mass (see, for example, 'The quantum mechanics of black holes', by S.W. Hawking, *Scientific American,* January 1977). Thus black holes are no longer completely opaque!

2.6 Existence of black holes

Do black holes really exist? We believe on theoretical grounds that many should occur at the end of the life of massive stars which cannot be prevented from collapsing by any known physical force. Because of their nature, it is difficult to detect black holes, but there is reasonably good evidence that we have seen the effects of their gravitational fields in several compact star-like objects observed to emit X-rays. This radiation is emitted by matter which accelerates as it falls into a massive object, and by analysing the radiation we can deduce that the object is sufficiently compact to be a black hole. There is reasonable evidence of this type for the existence of several stellar-mass black holes in our galaxy. Clearly the evidence is not entirely conclusive, but the existence of such black holes seems to be the best available explanation of the data.

Many astronomers also believe that much larger black holes exist at the centres of quasars and provide some explanation of their behaviour. It is also prossible that there are black holes at the centres of galaxies and it is hoped that the Hubble space telescope will provide some evidence for this.

Lecture IV: SOME SIMPLE COSMOLOGICAL MODELS

We shall now look at some models of the large-scale structure of the universe, which explain our observations of the red-shifts of distant galaxies in terms of expansion from a 'Big Bang'. We again use the idea of a cosmological model as a space–time with a set of preferred world-lines. We shall show that, although the space–time is very different from a black hole, there are still some interesting causal limits resulting from the light-cone structure and leading to the existence of particle horizons.

1. SPACE–TIME GEOMETRY OF FLRW MODELS

The basic assumptions that we make for a cosmological model are that, on the large scale, the universe is spatially *homogeneous* (i.e. the same at every point) and *isotropic* (i.e. the same in all directions). This is clearly not true locally, but gives a good lowest-order approximation into which one can introduce perturbations in order to study the more detailed structure. The two assumptions imply that the universe is isotropic about every point (i.e. every observer will see the large-scale properties of the universe to be the same in all directions around him). Such universe models are referred to as *Friedmann–Lemaître–Robertson–Walker universes*. Let us look at their geometry.

1.1 The metric

Coordinates can be chosen so that the metric takes the form

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + R^2(t)\{dr^2 + f^2(r)(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2)\}\ ,$$

with $f(r) = \sin r$, r, or $\sinh r$, depending on the nature of the universe. The fundamental world-lines (Fig. IV.1), representing the average flow of matter in the universe, are the curves $\{r, \theta, \phi\ \text{const}\}$. Because of the spatial homogeneity, there must be a uniform distribution of matter, and so quantities such as density and pressure are functions only of time. The fact that there is no spatial gradient means that there is nothing to make the fundamental galaxies and observers move non-inertially, so they are in free fall.

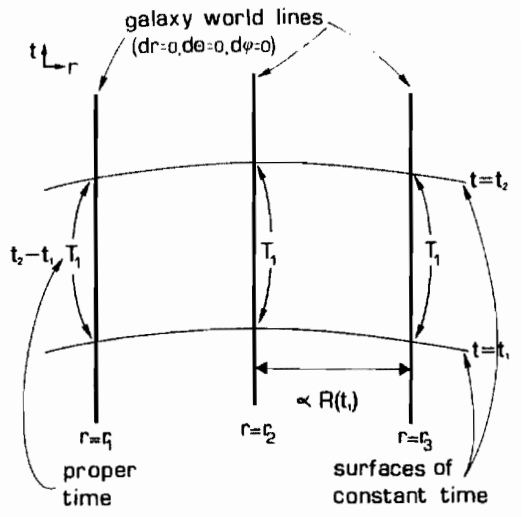


Fig. IV.1 The fundamental world-lines in a FLRW universe

We see from the metric that the coordinate t measures proper-time along the world-lines. It also follows from the metric that the universe is spatially homogeneous and isotropic. For an observer at $r = 0$, the space around him is isotropic, since for fixed r and t the metric reduces to that of a two-sphere that is spherically symmetric. Since all observers are equivalent, this is true for all of them.

1.2 The space-sections

The surfaces $\{t = \text{const}\}$ are surfaces of simultaneity for the fundamental observers since they are orthogonal to the matter world-lines. Let us look in detail at the geometry of these space-sections.

For the surface $t = t_0$, the metric form reduces to

$$\mathrm{d}s^2 = R^2(t_0)\{\mathrm{d}r^2 + f^2(r)(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2)\}\ .$$

The coordinates are centred on the arbitrary point $r = 0$ (Fig. IV.2). Moving out radially from this to $r = r_s$, we obtain a two-sphere with metric

$$\mathrm{d}s^2 = R^2(t_0)f^2(r_s)(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2)\ .$$

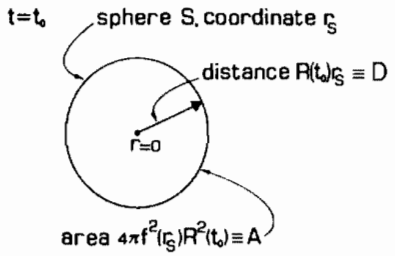


Fig. IV.2 A sphere with $r = r_s$. In this case, r is neither the radial distance nor an 'area coordinate'.

The area of this two-sphere is therefore

$$A = 4\pi R^2(t_0)f^2(r_s)\ ,$$

and the distance D from the origin to it is

$$D = R(t_0)r_s\ .$$

Let us look at the implications of these formulae for the different forms of f.

1.2.1 Flat space

If $f(r) = r$, then

$$A = 4\pi R^2(t_0)r_s^2 = 4\pi D^2\ ,$$

the usual relation in flat space. Thus the spatial sections here are flat (i.e. surfaces of zero curvature) and they also continue indefinitely. We have a spatially infinite universe with an infinite number of galaxies.

1.2.2 Hyperbolic space

If $f(r) = \sinh r$, then

$$A = 4\pi R^2 \sinh^2 r_s \geq 4\pi R^2 r_s^2 = 4\pi D^2\ ,$$

with equality only when $r_s = 0$. This is a hyperbolic three-space of constant negative curvature (cf. Fig. II.1b). Again the space-sections continue indefinitely and there is an infinite number of galaxies.

1.2.3 Elliptic space

If $f(r) = \sin r$, then

$$A = 4\pi R^2 \sin^2 r_s \leq 4\pi R^2 r_s^2 = 4\pi D^2 \ ,$$

again with equality only when $r_s = 0$. This is an elliptic three-space of constant positive curvature (cf. Fig. II.1a).

It is convenient to label these three cases by a parameter k, so that the curvature is given by

$$\kappa = \frac{k}{R^2(t_0)} \ ,$$

where $k = 1, 0, -1$ for the elliptic, flat, and hyperbolic space-sections, respectively (Fig. IV.3).

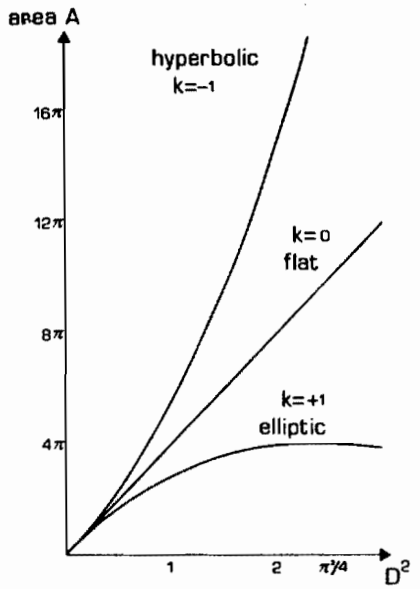
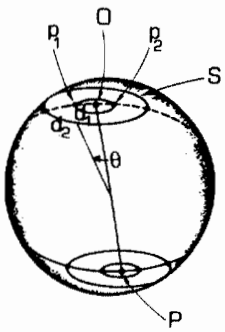


Fig. IV.3 The relation between the area A of a two-sphere and the square of its radius D, in the $\{t = \text{const}\}$ surfaces for the three classes of FLRW universe.

In the elliptic case, a new feature arises. As D increases, A increases, reaches a maximum at $r_s = \pi/2$, and then decreases to zero at $r_s = \pi$, the 'antipodal' point to $r_s = 0$. This cyclic behaviour is repeated as r_s continues to increase.

As a model of this situation, let us look at the analogy one dimension down. Consider a two-sphere of radius a (Fig. IV.4). Start at any point O and move a distance $d = a\theta$ in both directions along a great circle to reach opposite points p_1 and p_2 on a circle with circumference

$$C = 2\pi a \sin\theta = 2\pi a \sin(d/a) \ .$$



A 2-sphere model

Fig. IV.4 Geodesics in opposite directions at O on a two-sphere meet again at the antipodal point P. En route, they cut each circle centred at O in opposite points p_1 and p_2.

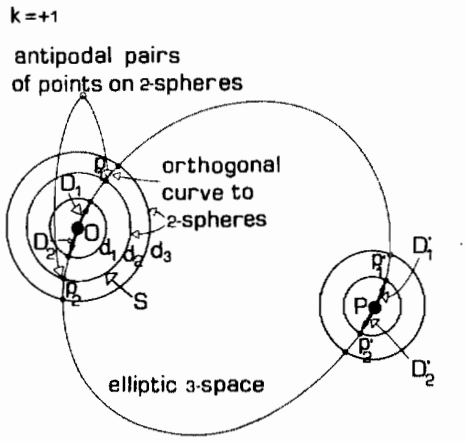


Fig. IV.5 The global geometry of an elliptic three-space. A geodesic starting at O will pass through the antipodal point P and return to O.

As d increases, C increases to a maximum where $\theta = \pi/2$, then decreases again to zero at P, the antipodal point to O. Any geodesic at O goes through P and arrives back at O from the opposite direction.

In the full three-dimensional case (Fig. IV.5), we have to imagine a succession of two-spheres rather than circles. Geodesics D_1 and D_2 in opposite directions from O cut a series of two-spheres in antipodal points p_1 and p_2. Because the area of the two-spheres eventually goes to zero, the geodesics eventually meet again at P, antipodal to O.

Note that in this elliptic case, the space-sections are necessarily finite, as is the number of galaxies.

1.3 The scale factor

From the form of the metric, we see that distances in the surfaces $\{t = \text{const}\}$ scale with $R(t)$, which is therefore called the *scale factor*. The distances between fundamental observers (at constant spatial coordinate values) will also scale with $R(t)$, as shown in Fig. IV.6. This means that since in general R is a varying function of time, the universe must be evolving in time. Not only does the density of matter vary, but also the curvature of space–time. The way $R(t)$ varies is determined, by Einstein's

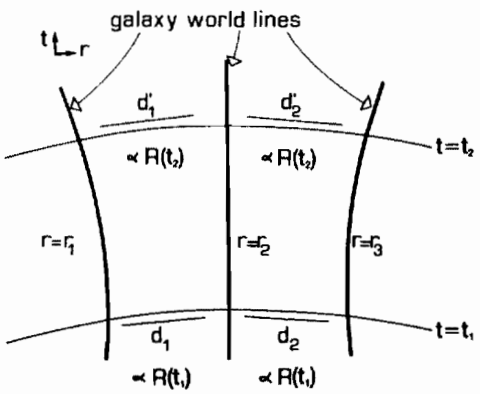


Fig. IV.6 The distances between galaxies scale with $R(t)$

equations, from the matter and radiation in the universe. Experimental evidence suggests that we live in an expanding universe; this is an expansion of the universe as a whole, not an expansion into anything. In the $k = 0$ and $k = -1$ cases, the spatial sections are infinite anyway, and the expansion means that distances between every pair of galaxies is increasing. In the $k = +1$ case, the spatial sections are finite but without edge, so there is not a boundary moving out into anything. The expansion is isotropic and without a centre; we can equally well take any galaxy to be at the origin of the spatial coordinates. There are various types of expansion depending on the values of k, and we shall now look briefly at the possibilities.

2. THE EVOLUTION OF THE UNIVERSE

Einstein's field equations give equations for the time evolution of $R(t)$; we now consider their solutions and ways of measuring $R(t)$.

2.1 The Einstein static universe

This is an exceptional solution when the 'cosmological constant' Λ is non-zero, and there is an extra term in the equations representing a universal repulsive force that balances the gravitational forces and so allows a static solution. The solution corresponds to a FLRW metric with $R(t) = R_0$ and $k = 1$ (i.e. closed spatial sections and therefore only a finite number of galaxies). In other respects, it is similar to the Minkowski universe: it is unchanging in time and has no systematic red-shifts. It is unstable under density perturbations, which destroy the fine tuning between Λ and gravity. For these reasons, it is not considered a good model of the real universe.

We shall now take $\Lambda = 0$, which excludes the possibility of static solutions.

2.2 Evolving universes

We consider the evolution first at early times and then at later times when it is qualitatively different.

2.2.1 The early universe

For all types of FLRW universes, the behaviour is the same at very early times. At that stage, the universe is filled with radiation, and $R(t)$ behaves like $t^{1/2}$, so the universe begins by expanding from a state of infinite compression, the 'hot Big Bang'. It is widely accepted that the real universe began in this way, and although theories about what happened at very early times are speculative (since, for example, a purely classical theory cannot describe a singularity like the Big Bang), the physics involved in the expansion of the universe at times later than about one second is fairly well understood. The universe was filled with a very hot mixture of particles and radiation in equilibrium with each other, and this mixture cooled as it expanded. As the temperature dropped, element formation took place at about 10^8 K, and then matter and radiation decoupled at about 3000 K. [What this means is that the universe was opaque to electromagnetic radiation at earlier times when electrons, moving freely between nuclei, scattered light strongly, but was transparent afterwards when the electrons were bound together with nuclei to form atoms (Fig. IV.7)]. The remnant radiation from this time is observed by us today as black-body radiation at a temperature of about 3 K. The discovery of this in 1965 was very important, as it

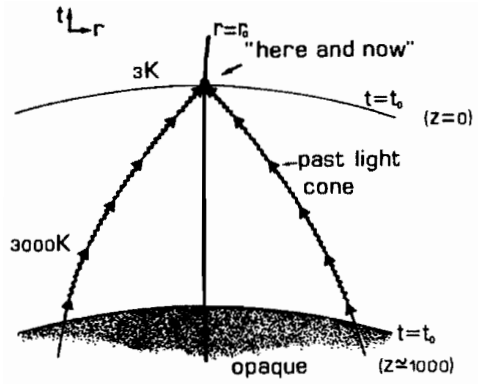


Fig. IV.7 Black-body radiation arriving along the past light-cone. Before the decoupling time $t = t_d$, the universe was opaque.

gives evidence that the universe was once much hotter, with $R(t)$ much less, and it indicates strongly by its isotropy that the universe was originally very uniform.

2.2.2 The late universe

The later behaviour of the universe differs according to whether the spatial curvature is negative, zero, or positive. As shown in Fig. IV.8, if $k = -1$, the universe is a low-density one which expands for ever. If $k = 0$, it is a high-density universe that just manages to expand for ever. If $k = +1$, it is a high-density universe that expands to a maximum value of $R(t)$ and then recollapses to a second singularity.

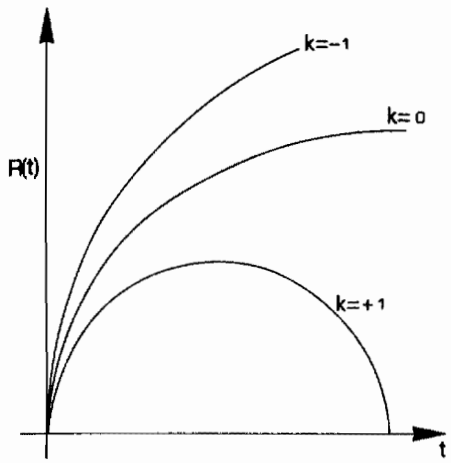


Fig. IV.8 The scale factor $R(t)$ plotted against t for the different values of k

Clearly it is very interesting to find out the value of k, since this not only determines whether the spatial sections are infinite but also whether the universe will expand for ever. In principle, the value of k can be inferred from the behaviour of $R(t)$, determined by observations of distant galaxies. Current evidence suggests that $k = -1$, but there is the possibility of a large amount of undetected dark matter which could lead instead to a high-density universe.

2.2.3 Red-shifts

As we shall see, observations of red-shifts in a FLRW universe give a way of measuring its expansion. Consider radial light-rays, which must satisfy $\mathrm{d}\theta = \mathrm{d}\phi = 0$ and $\mathrm{d}s^2 = 0$, leading to

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{1}{R(t)} \ .$$

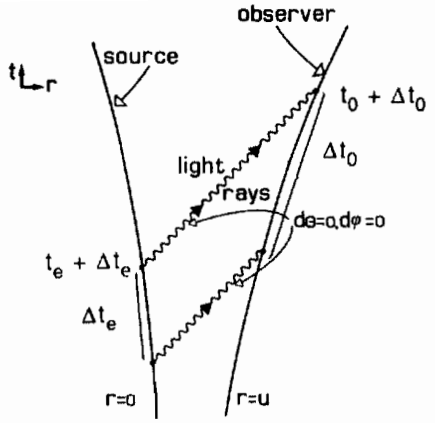


Fig. IV.9 Radial light-rays emitted at an interval Δt_e at $r = 0$ and received at an interval Δt_0 at $r = u$

Suppose that light is emitted by an observer O_1 at $r = 0$ at time t_e and received by O_2 at $r = u$ at time t_o (Fig. IV.9). Integration of the expression for dr/dt gives

$$u = \int_{t_e}^{t_o} \frac{dt}{R(t)} \ .$$

Similarly, for a second light-signal emitted at $r = 0$ at $t_e + \Delta t_e$ and received at $r = u$ at $t_o + \Delta t_o$,

$$u = \int_{t_e+\Delta t_e}^{t_o+\Delta t_o} \frac{dt}{R(t)} \ .$$

Equating the integrals and rearranging the limits, we obtain

$$\int_{t_e}^{t_e+\Delta t_e} \frac{dt}{R(t)} = \int_{t_o}^{t_o+\Delta t_o} \frac{dt}{R(t)} \ .$$

For small Δt_e and Δt_o during which intervals $R(t)$ will be approximately constant and may be taken outside the integrals, we obtain

$$\frac{\Delta t_e}{R(t_e)} = \frac{\Delta t_o}{R(t_o)} \ .$$

The red-shift z and K-factor are then given by

$$1 + z = K \equiv \frac{\Delta t_o}{\Delta t_e} = \frac{R(t_o)}{R(t_e)} \ .$$

Thus, observed red-shifts directly measure the expansion which has taken place. Note that the effect is symmetric but of course not independent of time (cf. Milne universe).

Red-shifts have been observed up to $z = 3.2$ for distant galaxies, $z = 3.8$ for quasars, and $z = 1000$ for the cosmic microwave background. Since $R(t)$ tends to zero as t tends to zero, the red-shifts from the earliest times would diverge, but it is not possible to receive electromagnetic radiation from such early times because of the opaqueness of the universe before the decoupling time.

3. CAUSAL STRUCTURE AND HORIZONS

We now look at some of the implications of the light-cone structure of FLRW universes.

3.1 Refocusing and the initial singularity

We have already mentioned the possibility of refocusing of the past light-cone (see, for example, Fig. II.14b). From the form of the metric, we can show that the area of the past light-cone of the event $t = t_0, r = 0$, for light emitted at $t = t_e$, $r = u$ and distance corresponding to red-shift z, is

$$A = 4\pi R^2(t_e) f^2(u) = \frac{4\pi R^2(t_0) f^2(u)}{(1+z)^2} \ .$$

Whatever the form of f, this tends to zero as z tends to infinity. Before this happens, A must reach a maximum on the surface of reconvergence, lying between t_0 and the initial singularity.

We see (Fig. IV.10) that our past is trapped inside a light-cone that goes back to the initial singularity. This claim is based on the FLRW model, which is highly idealized, and the question arises as to whether inhomogeneities in the real universe could lead to an avoidance of that singularity. We still do expect a surface of refocusing to occur in more realistic models, and then the singularity theorems of Hawking and Penrose show that in the classical theory, once refocusing has taken place, an initial singularity is inevitable no matter how irregular the early universe might be. Clearly we would need to look at quantum gravity to investigate this question further.

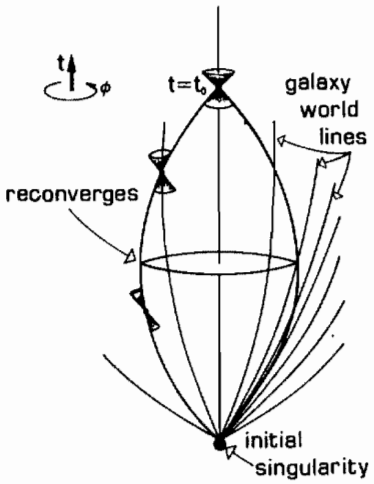


Fig. IV.10 The reconvergence of the past light-cone of an event at $t = t_0$

3.2 Particle horizons

As we saw in our investigation of red-shifts an observer at $r = 0$ at time t_0 can see to radial distance u at time t_e (Fig. IV.11), where

$$u = \int_{t_e}^{t_0} \frac{\mathrm{d}t}{R(t)} \ .$$

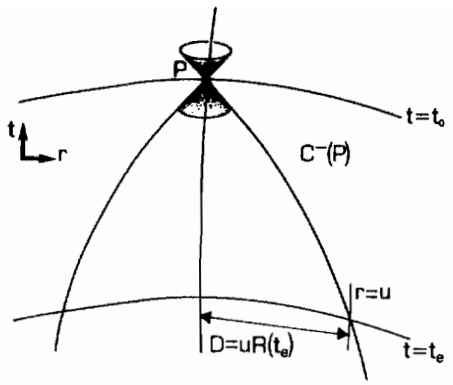


Fig. IV.11 An observer at $t = t_0$ can see matter up to a radial coordinate value u by light emitted at $t = t_e$

The maximum value of u, $u_{\max}$, is obtained by letting t_e tend to zero, and it can be shown that all the possible forms for $R(t)$, satisfying Einstein's equations, give a finite value for $u_{\max}$. Thus only a fraction of the galaxies in the universe can be observed. (Strictly speaking, this fraction is zero if $k = 0$ or -1, since the number of galaxies is infinite.) Thus a fundamental feature of the universe is a limitation on the regions with which any observer can have had causal communication; there are many galaxies which we could never hope to observe, no matter how long we wait.

This feature is difficult to understand from an ordinary space–time diagram because everything gets squashed together as $R(t)$ tends to zero. We can choose new coordinates where this does not happen; in particular, we choose coordinates in which the light-cones are at 45° (Fig. IV.12). For example, for $k = 0$, the spatial coordinate is r, and the time coordinate w, called the conformal time, is defined by

$$w = \int_0^t \frac{\mathrm{d}\tau}{R(\tau)} \ .$$

The resulting 'conformal diagram' is Fig. IV.12. Galaxy world-lines are vertical, and the initial singularity is a whole line, not just a point. The penalty is that spatial distances are badly distorted near the initial singularity [and must be multiplied by $R(t)$ to scale like measured distances].

Let us now look at the interior past light-cone of a typical galaxy G (Fig. IV.13). At $t = t_0$, the galaxies E, F, and F' will be visible to G, but not J, M, and N. In

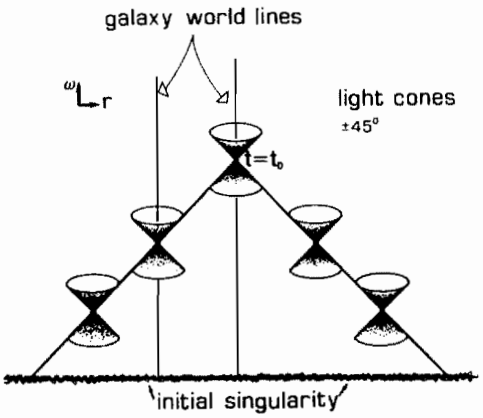


Fig. IV.12 Figure IV.10 drawn in new coordinates with the light-cones at 45° to the time-axis

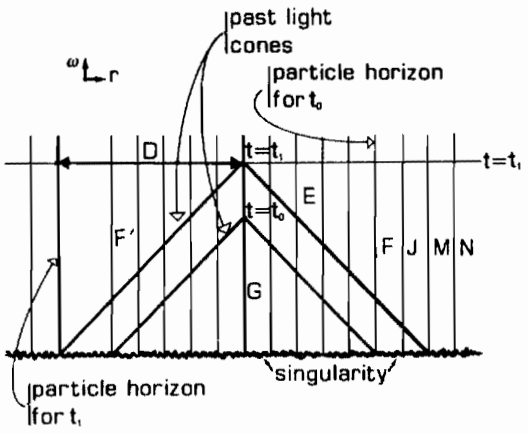


Fig. IV.13 The particle horizon of an observer on galaxy G is formed at $t = t_0$ by the galaxies F and F'

fact, F and F' are the limiting case: they are 'particles' that separate what G can see from what is invisible, and they constitute G's *particle horizon* at $t = t_0$. At a later time t_1, G's particle horizon has moved out and M lies on it; it is just visible. The physical size of the particle horizon at $t = t_0$ is

$$D = R(t_0)u_{\max} \ .$$

Does this mean that galaxies flash out of nothing? In fact the limit of the particle horizon is at $R = 0$, a surface of infinite red-shift, and this means that new galaxies emerge gradually into view, rather than appear suddenly.

The existence of particle horizons leads to an important puzzle related to the isotropy of the observed microwave background radiation, which indicates that conditions were very similar in regions that can have had no causal communication with each other. For example, Q and Q' in Fig. IV.14 are on the past light-cone of P, but their causal pasts do not intersect.

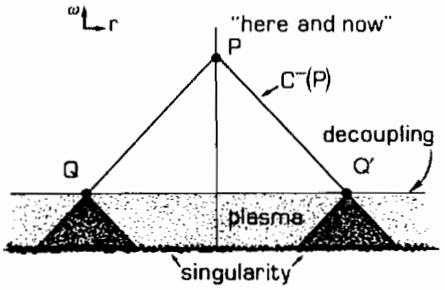


Fig. IV.14 Events Q and Q' at the decoupling time are on the past light-cone of P, but there can be no causal connection between them since their causal pasts do not intersect

There are several possible solutions to this problem. The two which we shall discuss in some detail in the next section are (a) dropping some of the usual field equations or equations of state (as in inflationary models), and (b) assuming a different topology (i.e. global connectivity) of the universe, resulting in a 'small' universe which we have already seen around many times.

There are also philosophical implications of particle horizons, since our predictive powers must surely be limited by their existence.

4. OTHER UNIVERSE MODELS

In this final section, we consider some other models of the universe, which are related to the FLRW models already described.

4.1 The de Sitter universe

This model has the FLRW metric with flat spatial sections and exponential expansion, i.e. $k = 0$, $R(t) = \mathrm{e}^{Ht}$, H = constant. It was originally discovered in 1917 as a vacuum solution of Einstein's equations with $\Lambda \neq 0$. Because it contains no matter, it was abandoned as a model of the real universe in 1930, when the expanding universe models of Friedmann and Lemaître became widely known.

4.2 The steady-state universe

The de Sitter universe was rediscovered in 1948 by Bondi, Gold and Hoyle as the steady-state universe. It is the only expanding universe model for which every

point in space–time (not just every point in space) is equivalent. It has no hot Big Bang at the beginning but just exists unchanging for ever. In general, in an expanding universe model the density of matter decreases with time, so to avoid this one has to modify Einstein's field equations to include an effective Λ-term, which gives continuous creation of matter throughout the universe. However, evidence from radio-source number counts was against this, and the discovery of the microwave background radiation, interpreted as a relic from a hot early state, led to a widespread rejection of the theory.

4.3 The inflationary universe

While Big Bang models have proved satisfactory in describing most of the history of the universe, there have been problems in understanding the initial conditions in the standard models, in particular the causal problems associated with the existence of particle horizons. This led Guth and others to propose a new model of the early universe, the inflationary universe, which in some senses is the steady-state universe in a new disguise! The basic idea is that quantum field effects in the very early universe, associated with symmetry breaking, led to an effective Λ-term in the field equations and so to a period of exponential expansion. During this time, the radius $R(t)$ increased rapidly, with the pressure and density remaining effectively constant, and so the size of a region lying within the particle horizon at the time of decoupling is enormously greater than in the standard model. This enables causal communication between regions visible to us in different directions in the sky.

To see how this works, let us look at the conformal diagram (Fig. IV.15). Because of the period of inflation, the initial singularity is now much further back in the past than it is in the standard model. Therefore the causal pasts of Q and Q' overlap to a considerable extent, and there can be a common physical cause for the similarity between conditions at Q and Q'. However, there are still some parts of their pasts that are not shared, so they are still susceptible to independent influences that could in principle interfere with the observed isotropy. There are also other problems associated with the theory (e.g. issues related to galaxy formation).

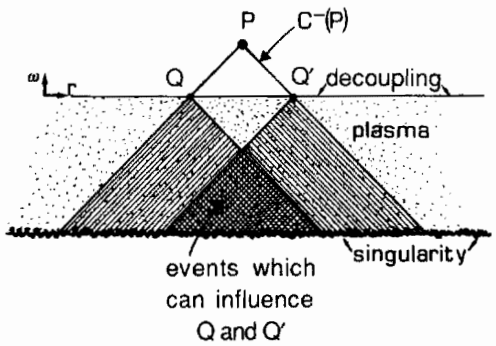


Fig. IV.15 A conformal diagram of the inflationary universe situation

4.4 Small universes

Let me finish by describing a universe model with a new element that provides a neat explanation of a number of problems. This is a small universe where the curvature properties and expansion history are the same as for a FLRW model but its topology or global connectivity is different.

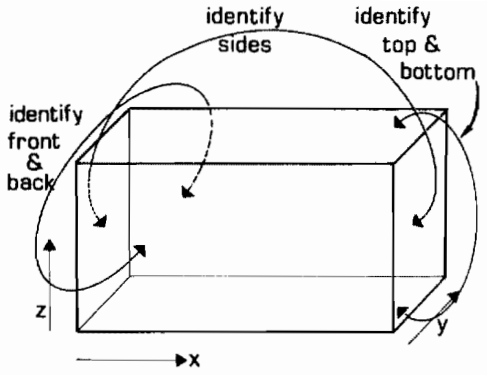


Fig. IV.16 A 'small universe' formed from a section of a $k = 0$ FLRW model universe by identifying opposite faces of a rectangular block in a space-section

The simplest model is the $k = 0$ flat space with a torus topology. Consider a rectangular block in a space-section $\{t = \text{const}\}$. We identify the opposite faces in pairs, as shown in Fig. IV.16. This means, for example, that when an observer travelling in the z-direction reaches the top face he continues his journey up from the corresponding position in the bottom face.

The volume of the space sections is finite and there are a finite number of galaxies in the universe. However, the universe will look infinite! To see how this works, let us look at a representation with one spatial dimension only. We unwrap the space-time diagram to get space-sections that are apparently infinite (Fig. IV.17), but all objects at intervals of some distance d are actually the same object. We see how our past light-cone can intersect the world-line of the same galaxy many times, producing many images of the same object. Thus this finite universe will look to an observer like an infinite FLRW universe, with images of galaxies fading away into the distance.

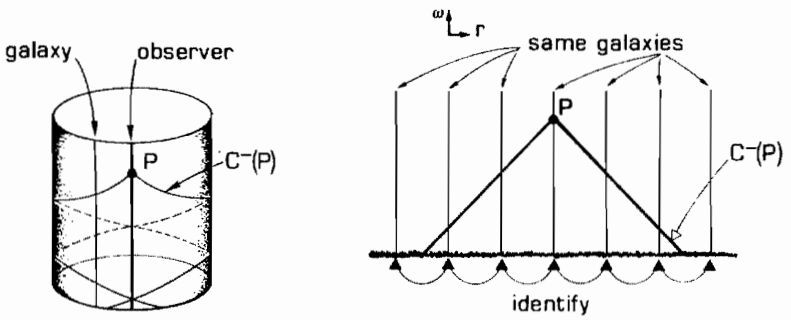


Fig. IV.17 The representation of one spatial dimension and the time dimension in a small universe. All the vertical lines in the diagram on the right represent the same galaxy and can be identified with each other to give the diagram on the left.

We can see the same features in a diagram of two spatial dimensions (Fig. IV.18). Opposite sides of the rectangle are identified, so unwrapping the space we obtain a repetition of the basic cell and its contents in all directions. Looking out to a distance r, characterized by red-shift z, we see the same material many times over. On a large enough scale, this model will look spatially homogeneous even if the basic cell is not.

It is actually very difficult to prove that the real universe is not like this because it would not be easy to show that all images of an object have a common origin, since they would be seen at different red-shifts, at different stages in their history, in different directions.

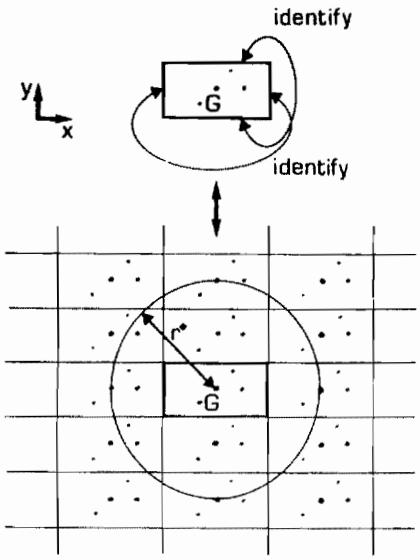


Fig. IV.18 A two-dimensional representation of a small universe at a time $t = t_0$. Looking out to a distance r, one will see the same basic cell over and over again.

This type of model has a number of very attractive features. There are no particle horizons since we can see all the matter. There are no boundaries and hence there is no problem of finding appropriate boundary conditions. Finally, the apparent homogeneity and isotropy of the universe are explained in the simplest possible way.

In this brief survey we have looked at some very idealized models, which nevertheless describe some of the overall features of the universe. There are some problems with all of the models; the small-universe idea is particularly interesting but, as we have seen, it is hard to prove or disprove its correctness. In all the models we have studied, the use of space–time diagrams has helped us towards a clearer understanding of many possible features of the physical universe that are otherwise rather difficult to comprehend.

Suggestions for further reading

Anyone who wants to follow up the subject in more detail should read some of the books that follow a similar line of approach. For obvious reasons, the one to which these lectures are closest is *Flat and curved space-times*, by G.F.R. Ellis and myself (Oxford University Press, 1988). Other similar books are *Discovering relativity for yourself*, by S. Lilley (Cambridge University Press, 1981) and *General relativity from A to B*, by R. Geroch (University of Chicago Press, 1978). The application of general relativity to cosmology is described in *Cosmology*, by E.R. Harrison (Cambridge University Press, 1981), which is at about the same level as the books already mentioned. More advanced books are *The first three minutes*, by S. Weinberg (Basic Books, 1977), and *Cosmology*, by M. Rowan-Robinson (Oxford University Press, 1977).