1.2 The Geometry of the Universe

In this section and the next we will investigate the theory of the expanding Universe. There are two elements to this:

- What is the large-scale structure of space? ("Geometry")
- How and why does that structure change with time? ("Dynamics")

The first may seem an unlikely problem -- how can empty space have any kind of structure? But it was realised in the 19th century that the traditional geometry of Euclid does indeed define a particular structure for space, and moreover, that there are many other possibilities. Once this was recognised, it became natural to ask whether the geometry was fixed or whether it could change with time, and if so, what were the forces that controlled geometry? In 1915, Albert Einstein published his **General Theory of Relativity** (GR, for short), which provided answers to just these questions. It is a theory of the geometry, not of space alone but of space-time. According to GR, the force of gravity is just a manifestation of the distorted geometry of space-time. We know that gravity is produced by mass, so you might guess that GR will tell us how mass affects geometry. In fact, Einstein had already shown that mass and energy are equivalent ($E = mc^2$); now he found that to make GR consistent, geometry is controlled not by mass/energy alone but by a combination of energy and momentum.

Over the last century, GR has passed many tests against observations, giving us confidence that it gives an accurate description of how the dynamics of the Universe are controlled by the energy-momentum of its contents. But the best evidence for GR is on the relatively small scales of individual stars, as we saw in the section of this course on pulsars, and also on smaller scales still in Earth-bound laboratories. There are a number of speculative theories according to which GR breaks down on cosmological scales. So it is worth making a distinction between cosmological results that rely on the dynamics supplied by GR (the subject of the next section), and those results that rely purely on geometry.

1.2.1 Non-Euclidean Geometry

Euclid's *Elements* was the most successful textbook in history. Written around 300 BC, it was still the basic school text on geometry in the 19th century. Even today, school geometry follows Euclid in content, if not in style. Geometry ("Earth measuring") had a huge impact on the Greek way of thinking. They were enormously impressed that, from five seemingly-obvious postulates (example: "a circle can be drawn with any centre and radius") one could logically derive true statements about the real world, that were practically useful in fields such as surveying and architecture. This attitude persisted into the 18th century, and served as the basis of Immanuel Kant's idea that it is possible to intuit truths about the real world independently of experience.

But in the 19th century mathematicians including C. F. Gauss and G. B. Riemann showed that alternative postulates could lead to equally self-consistent geometries. Gauss realised that the geometry of the real world could only be determined by observation. Because Euclidean geometry does work for surveying with good accuracy, the real world must be close to Euclidean on this scale; but most non-Euclidean geometries become approximately Euclidean when figures (triangles, circles etc.) are small compared to some characteristic length. As Gauss recognised, our experience on Earth is no guide to the geometry of the Universe as a whole.

1.2.2 Positive Curvature

As the Earth is a sphere, one of the earliest applications of geometry was to the properties of figures drawn on a sphere. This is called **spherical geometry**. For more than two thousand years spherical geometry was studied as a set of results in 3-D Euclidean geometry; but in 1854, Riemann realised that we can also describe it as analogous to the 2-D Euclidean geometry of a plane, but with the points related to each other in a different (non-Euclidean) way. From Riemann's point of view, there is no need for a 3-D space in which the sphere is embedded; attention is confined solely to the points on the surface and the way they are "connected".

Let us look at the geometry of the sphere more closely (from now we will follow the convention in geometry according to which "sphere" means just the surface, not the solid object). We know that the shortest distance between two points on a sphere is along a great circle. But a straight line is defined as the shortest distance between two points, so great circles become the "straight lines" of the non-Euclidean geometry of the sphere. Since any two great circles will always intersect, in this geometry there are no parallel lines, violating Euclid's fifth postulate. A circle is defined as the locus of points a given distance from a fixed point. This corresponds to a small circle on a sphere (Fig. 1.2a). From the figure we see that the radius (measured along a great circle of course) is $x = \theta R_0$, where R_0 is the radius of the sphere and θ is measured in radians. But the circumference of the circle is not $2\pi x$ as in Euclidean geometry but the smaller value $2\pi R_0 \sin\theta = 2\pi R_0 \sin(x/R_0)$. By replacing Euclid's postulates with appropriate equivalents, all of spherical geometry can be deduced. For instance, in Euclidean geometry the internal angles of a triangle add up to π radians; while in spherical geometry the angles always add to $> \pi$ (Fig 1.2b).

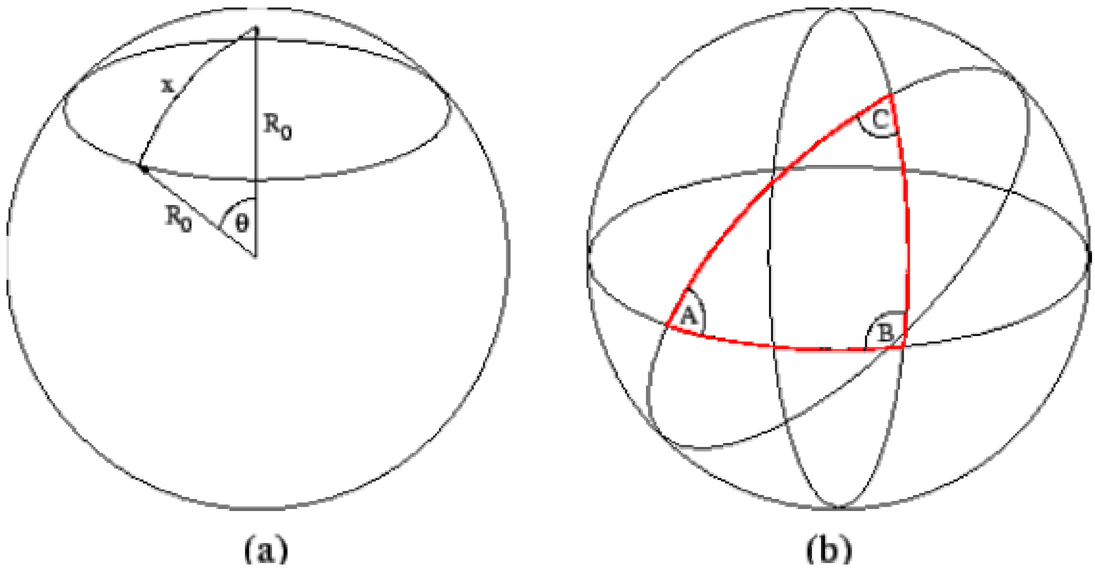

Figure 1.2: *Spherical geometry: (a) The radius of a small circle subtends an angle θ from the centre of the sphere. (b) A spherical triangle has sides which are sections of great circles. The sum of the angles $A + B + C > \pi$. Note that all great circles eventually intersect.*

All points on a sphere are equivalent, and there is no preferred direction, so the geometry of this 2-D "space" is homogeneous and isotropic. It is common to say that this space is curved, in contrast to "flat" Euclidean 2-D space. This is a rather misleading description, but we are stuck with it. By convention the curvature is "positive"; we will meet the alternative, negatively curved space, in the next section.

Thinking of a sphere as a 2-D space is a useful analogy, but in reality space is three dimensional. What would positive curvature mean for a 3-D space? Simply, that the properties of triangles and circles are exactly the same as in 2-D curved space (they are still plane figures, after all). In 3-D there are infinitely many planes, separated vertically and at different angles (Fig. 1.3); but if space is homogeneous and isotropic, the properties of figures cannot depend on their position or orientation.

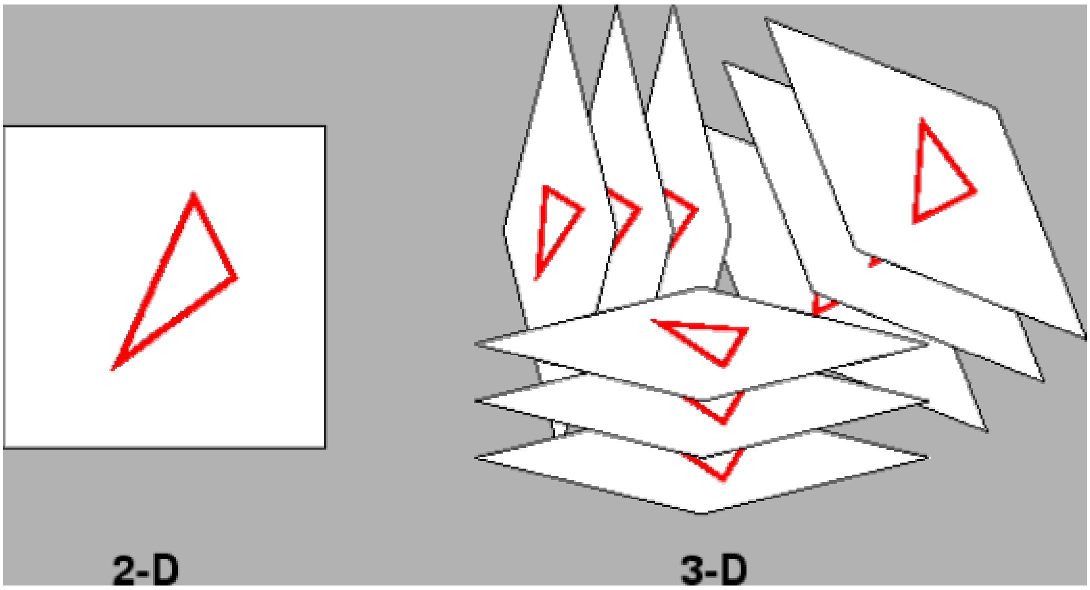

Figure 1.3: *The difference between 2-D and 3-D space is that we no longer have just one plane!*

In 3-D the locus of points a given distance (say x) from a fixed point is a sphere. We can think of this as being built up of circles with every possible orientation, all centred on the fixed point and with radius x. By isotropy, all these circles have the same circumference, so the geometry on the sphere in our curved 3-D space is exactly the same as on a sphere embedded in Euclidean 3-D space, with one exception: the radius of curvature is not x but

$$r_A = R_0 \sin(x / R_0) .$$

That is, a great circle on the sphere has circumference $2\pi r_A$, the area of the sphere is $4\pi r_A^2$, and the distance between two points on the sphere separated by ψ radians (as seen from the centre) is $r_A \psi$.

1.2.3 Negative Curvature

Fig. 1.4 shows a kind of curved surface that is fundamentally different from the surface of a sphere. Using the straightest possible lines on the surface (shortest distance between two points, as ever), we can draw a triangle, for instance, as shown in the figure. It is clear that the three angles will add to less than π; similarly a circle on this surface will have a circumference $> 2\pi x$. The surface is distorted from a flat plane in the opposite way to the surface of a sphere, so we say it has **negative curvature**. On the surface shown in the diagram, the curvature changes from point to point. We would have preferred to show the negatively curved equivalent of a sphere, a "pseudosphere" which has constant negative curvature at each point. Unfortunately it is impossible to construct a 2-D pseudosphere in 3-D flat space, unlike the case for a sphere (it is possible in 4-D flat space, but we can't visualise that!). But this does not prevent us from working out its geometry; in fact this was the first non-Euclidean geometry to be discovered.

On a pseudosphere, the circumference of a circle with radius x is

$$2\pi r_A = 2\pi R_0 \sinh(x / R_0)$$

where sinh is the hyperbolic sine. We call R_0 the radius of curvature of the pseudosphere. Whereas $\sin\theta$ is always smaller than θ (measured, as always, in radians), $\sinh\theta$ is always greater, so $r_A > x$, as we found in Fig. 1.4. Unlike a sphere, a pseudosphere extends to infinity in all directions. In this case there are an infinite number of "straight lines" passing through a given point that will never intersect a given line, rather than just one parallel.

Just as for positive curvature, there is a possibility that our 3-D space is really negatively curved, and the details of the geometry carry over from the 2-D case in the same way.

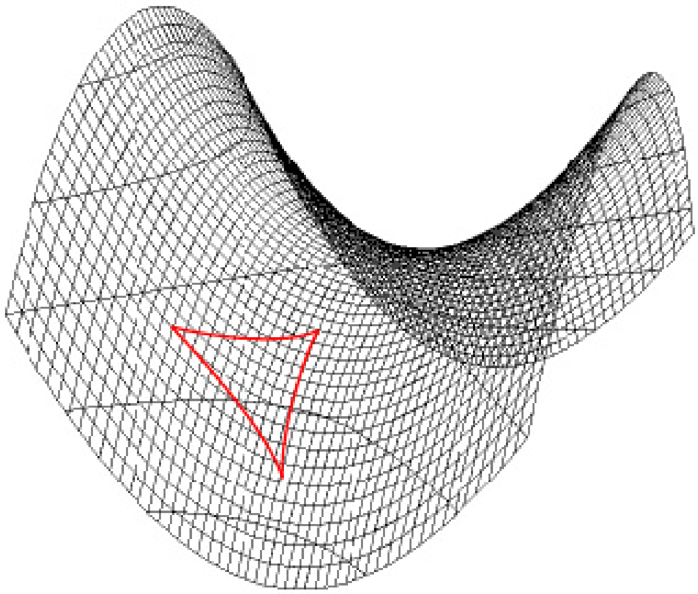

Figure 1.4: *A saddle - an example of a negatively curved surface.*

1.2.4 Flat space

If R_0 becomes very large, then $\chi = x/R_0$ becomes very small, in which case both $\sin\chi$ and $\sinh\chi$ are well approximated by χ. In the limit of infinite R_0, the radius of curvature of a sphere tends to the actual radius: $r_A = x$. This means that Euclidean geometry is a limiting case of the other two. By the same token, when we are dealing with lengths x small compared to R_0, geometry will be effectively Euclidean as we claimed at the beginning. Because a space with Euclidean geometry shows neither positive nor negative curvature is often called **flat space**.

It is often useful to talk about all three geometries simultaneously. To do this we use the symbol $S_k(\chi)$ to stand for all three cases, according to the value of the **curvature constant** k, as follows:

k	Curvature	$S_k(\chi)$	
-1	negative	$S_{-1}(\chi)$	$= \sinh\chi$
0	zero	$S_0(\chi)$	$= \chi$
+1	positive	$S_{+1}(\chi)$	$= \sin\chi$

It is awkward to use a formally infinite R_0 to describe flat space, and we can avoid doing so if we use the S_k notation, because then we can pick *any* value of R_0, and it will cancel out:

$$r_A R_0 S_0(x/R_0) = R_0 \frac{x}{R_0} = x \quad .$$

So we can leave R_0 in the equations without any ill effects.

1.2.5 Topologies

We have carefully avoided mentioning the most obvious difference between a sphere and a flat plane: on a sphere, if you travel far enough in a "straight line" you will arrive back where you started. This tells us that, taken as a whole, points on the sphere are linked in a fundamentally different way from points on a plane. These large-scale connections would remain if the sphere was distorted, e.g. sat on, stretched into a rugby ball, or just heavily dented; on the other hand the lengths of circles drawn on the surface would certainly be changed. The large-scale connections define the **topology** of the surface (sometimes called "rubber-sheet" geometry because topology is unaffected by stretching or squashing the surface). In contrast, geometry, strictly speaking, is concerned with the actual lengths and angles, and not with large-scale connections.

We say that the topology of a sphere is **closed**, meaning that it has a finite surface area, but no edges. The topology of a pseudosphere (or a flat plane) is **open**, meaning that it extends infinitely in all directions. This suggests that there is a necessary connection between geometry and topology, but this is not correct. For instance, Euclidean geometry also applies on the surface of a cylinder, in the sense that circles have radius exactly $2\pi r$ etc.; but the topology is different because straight lines in one particular direction return to their starting point. This illustrates an important point: if we compare the properties of any two small regions of a cylinder they are the same and independent of direction, so we say that **locally** the cylinder is homogeneous and isotropic. But topologically, the direction along the cylinder behaves very differently from the direction around it, so we say that **globally** the cylinder is anisotropic (though still homogeneous).

Figure 1.5: *A torus embedded in 3-D space.*

Many other topologies are consistent with Euclidean geometry, including completely closed ones. The simplest closed Euclidean geometry is a **torus**, generated by connecting together the two ends of a segment of a cylinder. Unlike a cylinder, a 2-D torus embedded in 3-D space (e.g. Fig. 1.5) cannot have a strictly Euclidean geometry, but, as for a pseudosphere, this is an accident which does not affect the self-consistency of the geometry. In this case it is quite easy to visualise what is going on. A classic example of a 2-D Euclidean torus is the space of video games such as *Asteroids* (Applet 1.6), where objects leaving the screen on one edge return through the opposite edge. Notice that the "edges" only appear because we have to cut the torus to unroll it onto a flat plane; they are not special places as far as inhabitants of the game are concerned.

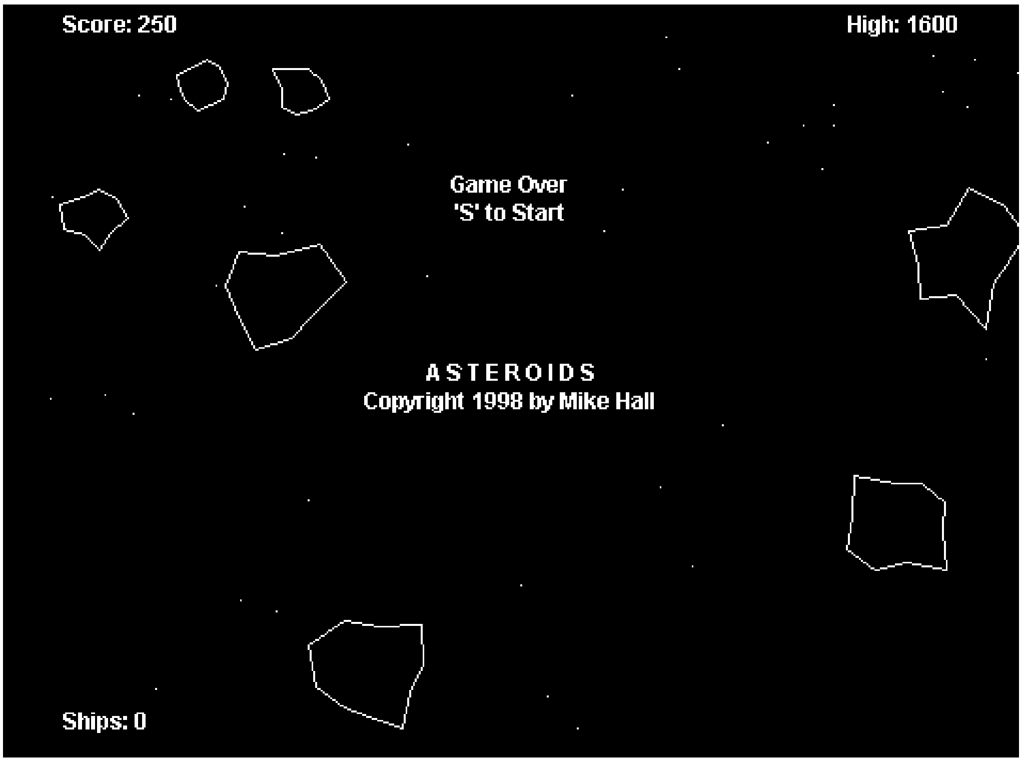

Figure 1.6: *The classic video game "Asteroids" takes place in a torus projected onto flat 2-D space. Try playing this Java version of "Asteroids" written by Mike Hall (www.brainjar.com).*

Another way of visualising a 2-D torus is to divide an infinite plane up into a repeating pattern, as shown in Fig. 1.7. There are only two people in this universe, but the connections of the torus make it seem to be populated by a large crowd. We have outlined the basic repeating pattern with two (equally valid) centres. This gives a clue as to how to construct more complicated closed topologies with Euclidean geometry: simply divide up the plane into other repeating patterns, such as diamonds.

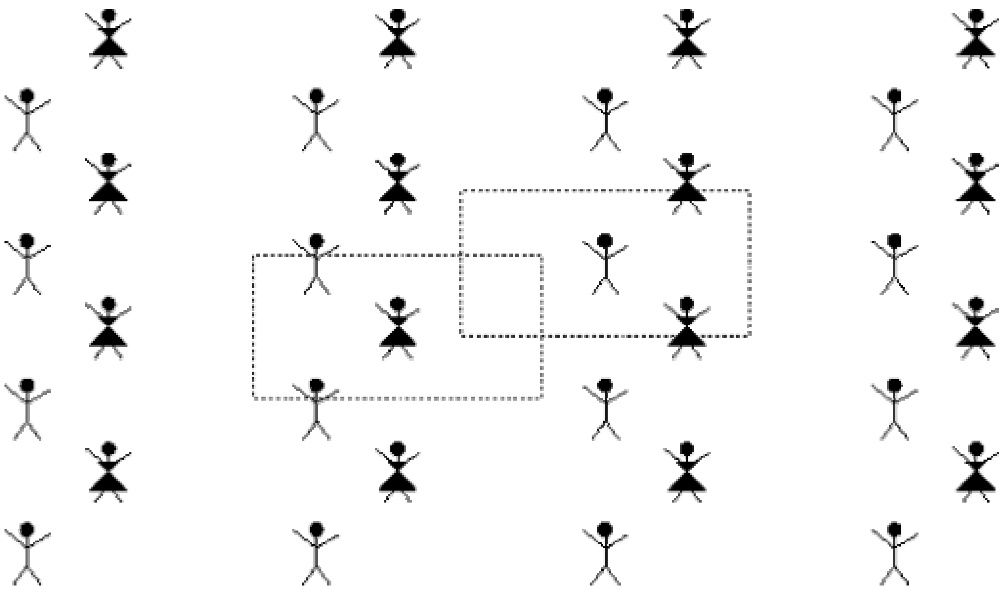

Figure 1.7: *A 2-D torus seen as an infinitely-repeated pattern.*

1. Convince yourself that a repeating pattern like that in Fig. 1.7, except with a diamond instead of a rectangular basic unit, would be caused if a cylinder was twisted by 180° around its axis before joining the two ends.
Answers to question are at end of this document.

Three dimensional spaces can have closed topologies in just the same way as 2-D spaces. The first person to realise this was G. B. Riemann, who suggested in 1854 that our Universe might be a **spherical space** (also known as a hypersphere), i.e. a 3-D space with constant positive curvature and the topology of a sphere (return to starting point after travelling a distance of $2\pi R_0$ in any direction). This avoided the classical difficulties of imagining a space which was either infinite, or had some sort of edge.

In positively curved space we saw that the area of a sphere of radius x around any point is $4\pi (R_0 \sin(x/R_0))^2$. In a spherical space we can travel to $x = \pi R_0$, where the sphere shrinks to a single point, the **antipode**, the most distant point from our starting-place. In other words, travelling a distance πR_0 in any direction takes us to the same place. An alternate topology is so-called **elliptical space**. Here, when we travel to $x = \pi R_0/2$, we find that we have reached the same point as if we had set out in the opposite direction, and so we return to the start after only travelling πR_0. An elliptical space has only half the volume of a spherical space with the same radius.

There are only four combinations of topology and geometry which are globally isotropic. These are: infinite Euclidean space and infinite negatively curved space, both open; and spherical and elliptical space, both positively curved and closed. But there are an infinity of possibilities, analogous to the 2-D torus, if we include spaces which are only locally isotropic. For instance the 3-torus is a closed Euclidean space constructed by joining opposing faces of a cube. Closed spaces with negative and positive curvature can be constructed by connecting opposite faces of other polyhedra; these are sometimes called **compact spaces**. If our Universe is a closed or compact space, light may be able to travel around it several times, so that distant galaxies are actually repeat images of nearer ones, just as the man and woman in Fig. 1.7 see themselves surrounded by a large crowd who are in fact themselves.

Except for flat space, the size of each compact space must have a definite relation to the radius of curvature R_0. The smallest compact space with negative curvature, Weeks space, is made by identifying the faces of an 18-sided polyhedron. It has a volume of $0.94R_0^3$. More complex connections give spaces with larger volumes. In contrast, compact positively curved spaces get smaller as the complexity increases, so that the largest possible space with constant positive curvature is a spherical space, which has a volume of $2\pi^2 R_0^3$. There is no way to make an infinite space with constant positive curvature.

1.2.6 From space to space-time

So far we have talked only about 3-D space. What about time, the fourth dimension? Einstein's Special theory of relativity (SR) shows that space and time are exchangeable to a certain extent: two events which happen at the same time but in different places according to one observer may happen at different times according to another observer moving past at high speed. This concept is incorporated into GR.

Now we have an apparent problem. The universe is homogeneous and also expanding; homogeneity means, among other things, that the density ρ is independent of position, while of course expansion means that it changes with time. But relativity theory allows us to change our coordinate system so that events which were previously at different times are now regarded as simultaneous; in which case we have ρ varying with position (and also direction, breaking isotropy as well).

The solution is that homogeneity and isotropy are only visible to a special set of observers, the fundamental observers (FOs) who "go with the flow". When we talk about time in cosmology, we mean time as measured by the FOs. Despite the propaganda of relativity theory that all observers are equal, the fundamental observers are more equal than the rest: anyone who presumes to move relative to them will get a distorted view of the universe in which the observer's direction of motion relative to the fundamental observers will pick out a special direction in space.

Let's look at this in more detail. Lost in the wastes of four dimensional space-time, we orient ourselves by labelling each space-time point (**event**) by the local density of matter. If density changes smoothly, we have divided space-time into a sheaf of "slices", each a 3-D space with constant density. To say that space is homogeneous and isotropic is to say that all these slices are homogeneous and isotropic. Since every event is on one or another slice, the slices must pack together perfectly. This is only possible[1.1] if they all share the same geometry -- for instance it's obvious that you can't smoothly stack a closed space with an open one. We now set up an array of local (x, y, z, t) coordinate systems centered on each event, in which the three space axes lie in the 3-D slice of constant density (Fig. 1.8). Consider the paths through space-time (**world lines**) of particles in a small region of one slice. There will be be random motions, but we can take an average to define a flow of matter through space-time. This flow must run along the time axis we have set up -- matter must be on average stationary in space -- otherwise the flow pattern in each slice would make the slices inhomogeneous (unless the flow had the same speed and direction at every point, but that would still be anisotropic). By definition the world lines of FOs follow this average flow, and so are always perpendicular to the constant-density surfaces; in other words a FO will always find herself surrounded by a constant-density universe, and as Fig. 1.8 shows, will see her colleagues moving away from her according to Hubble's law. Finally, all FOs must measure the same amount of time between different slices, otherwise the rate of change of density would be different at different points in the slice, contrary to homogeneity (it would also be impossible to stack the slices perfectly if the gaps between them changed with position). If they can agree on a standard starting point, such as the Big Bang, the FOs can then define a **cosmic time** which is constant on each slice.

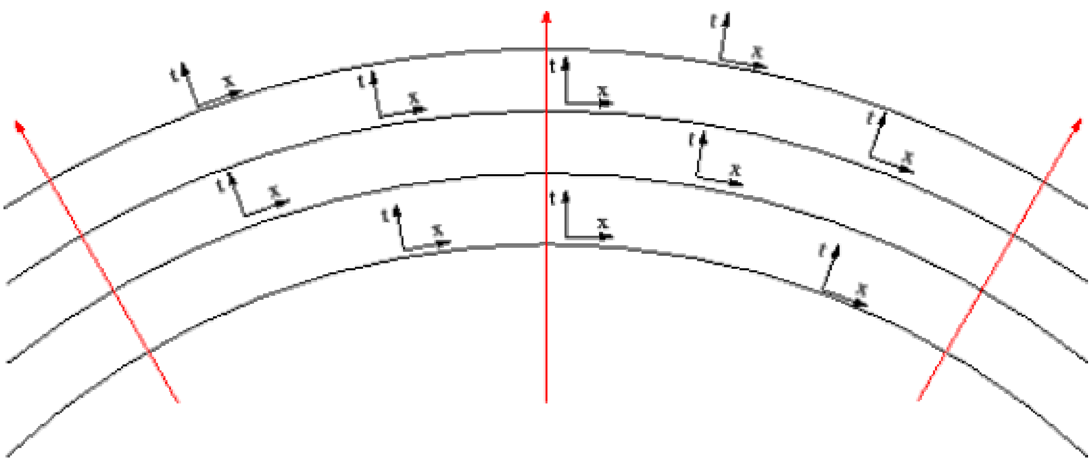

Figure 1.8: *Space-time diagram of homogeneous expansion. The curved lines represent the constant-density slices of space-time. Two spatial axes are suppressed for simplicity. Coordinate systems chosen so that the space (x) axis lies in the slices are shown at selected points. The red arrows show the world lines of fundamental observers, which are parallel to our local time axes if, as here, the slices are homogeneous and self-similar.*

This argument may seem subtle, but the bottom line is simple: in cosmology it makes sense to separate "space" from "time", despite the best efforts of relativity theory to blend them together. In a very real sense, the absolute space and time of Isaac Newton are restored (the very opposite of what Einstein was trying to achieve!). Of course, we have relied on homogeneity and isotropy, which are only true on the large scale. When we come to look at irregularities in the Universe, the problem of separating space from time will recur.

We now have enough information to sort out the mathematics of expanding, curved space, At the present cosmic time t_0, space has curvature k (= - 1, 0 or +1) and radius R_0. As the expansion is homogeneous, we know that at some other time t, all distances have changed by a scale factor $a(t)$. For a particular circle (e.g. with us at the centre), the proper radius x becomes ax, and the circumference becomes

$$a \times 2\pi R_0 S_k\left(\frac{x}{R_0}\right) = 2\pi a R_0 S_k\left(\frac{ax}{aR_0}\right) ,$$

Let's define $R(t) \equiv a(t)R0$, and replace the co-moving distance x by the co-moving dimensionless quantity

$$\chi = \frac{x}{R_0} = \frac{ax}{R} \; .$$

This is analogous to the polar angle θ in spherical polar coordinates. At any time t, the proper radius is $R(t)\chi$ and the circumference is $2\pi r$A = $2\pi R(t)$ Sk(χ). The rôle of the dimensionless scale factor $a(t)$ has now been taken over by the length $R(t)$ which is often called the **radius of curvature of the universe**. Notice that the sign of the curvature is fixed once and for all: homogeneous expansion can't convert negative to positive curvature, just as we found by thinking about stacking the slices in space-time.

1.2.7 The light cone

A central concept in relativity theory is the **light cone** (Fig 1.9): the past light cone of an event is the surface (actually 3-D) in space-time swept out by all the photons arriving at that event, while the future light-cone is the surface swept out by all the photons starting out at that event. Our world-line must be enclosed within the light cone because we cannot overtake a photon. Notice that two observers moving rapidly past each other would share the same light cone: the light cone for each event is a fixed object in space-time, independent of coordinate systems or frames of reference. Directions in space-time enclosed by the light cone are called **time-like** and can be taken as the time axis for a coordinate system, corresponding to the viewpoint of an observer whose world-line was in that direction. Directions outside the cone are **space-like**.

The light-cone nicely illustrates the way that observations in cosmology probe the history of the Universe, because photons arriving now from objects far away must have started travelling towards us a long time ago. These photons have travelled along our past light cone: only this slice through space-time is accessible to observation, no matter how good our telescopes become.

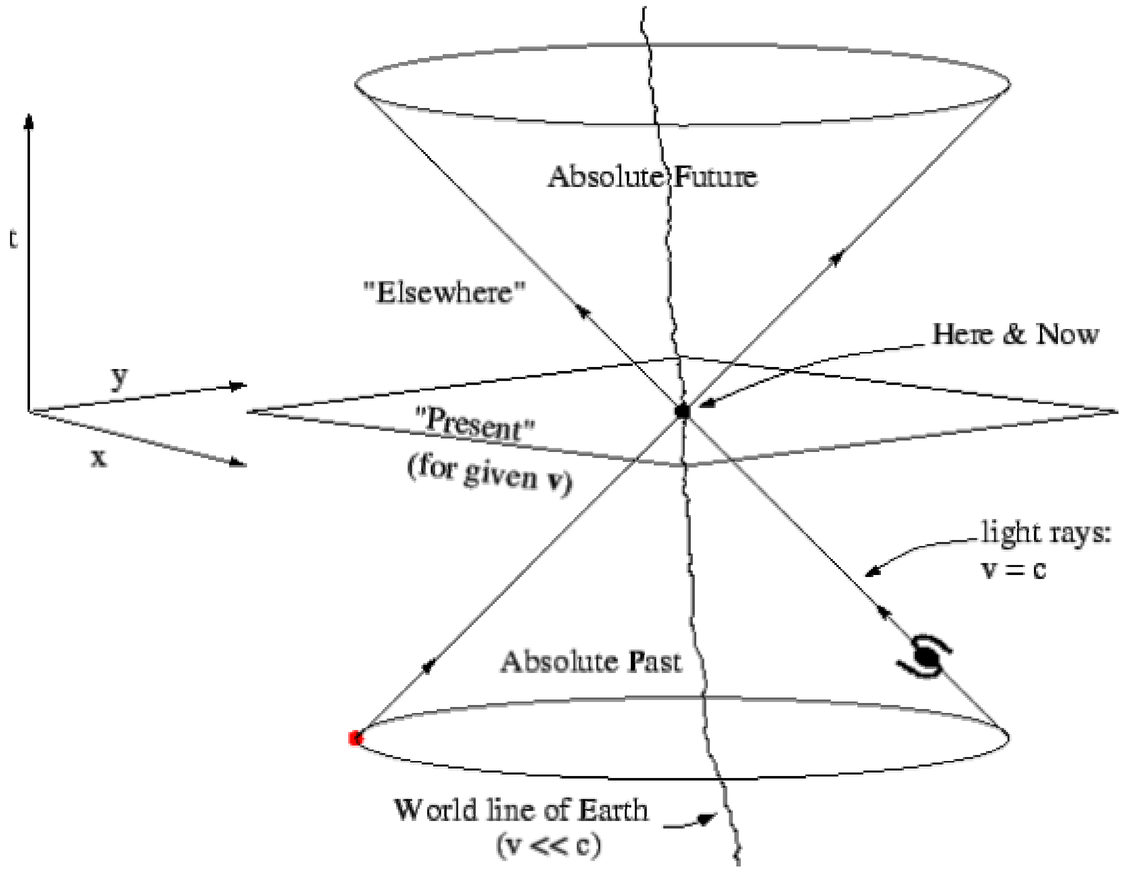

Figure 1.9: *Space-time diagram (with the z axis suppressed, i.e. held constant), showing the light cone. With two spatial dimensions (x, y), the set of points at a given time t on the light cone is a circle surrounding the observer; in 3-D it will of course be a sphere.*

1.2.8 The many distances of cosmology

Ask a cosmologist the distance to some galaxy and he is liable to reply "what kind of distance?". We have already met

- proper distance $r(t) = R(t)\chi$ (do you want the value now, at t_0, or when the light was emitted, at t_{em}?),
- co-moving distance $x = R_0\chi$.
- dimensionless co-moving distance χ.

Two other kinds of distance are often used. The radius of curvature of the sphere at co-moving distance χ is better known as

- **Angular size distance**, $r_A = R(t_{em})S_k(\chi) = R_0 S_k(\chi)/(1+z)$.

As the name suggests, this is the distance needed to convert from a measured angle ψ across some distant object to the physical size D perpendicular to the line of sight:

$$D = r_A \psi$$

Note that D is measured along a circle that had radius $r(t_{em}) = R(t_{em})\ \chi$ at the time the light was emitted.

The last kind of distance comes from astronomers' habit of using the inverse-square law for brightness to infer distances. In a non-expanding Euclidean space, the apparent brightness or **flux density** S of an object at distance r is related to the intrinsic brightness or **luminosity** L by

$$S = \frac{L}{4\pi r^2}$$

because the light from the object is now spread over a sphere of area $4\pi r^2$. In an expanding universe, photons are redshifted, $\lambda \propto (1 + z)$. A photon has energy $E = hc/\lambda$, where h is Planck's constant, so the redshift reduces the energy by a factor $(1 + z)$. Also, the rate of arrival is reduced: consider two photons heading in the same direction emitted a time dt apart. At first they will be separated by $dl = cdt$, but the expansion of the universe increases this to $(1 + z)dl$ on arrival; so the arrival rate is reduced by $(1 + z)$. As we have seen, the (present) area of a sphere with a co-moving radius χ is $4\pi(R_0 S_k(\chi))^2$. Combining all these, the formula for flux density is

$$S = \frac{L}{4\pi(R_0 S_k(\chi))^2(1+z)^2}.$$

For convenience, we write this as

$$S = \frac{L}{4\pi r_L^2}$$

where

- **Luminosity distance** $r_L = R(t_0)S_k(\chi)(1+z)$.

We can use the formulae for angular size and luminosity distance to see how **surface brightness** depends on distance. Surface brightness (or intensity) is the flux density per unit **solid angle**, i.e. area on the sky. A small square region with angular size in ψ each direction will have solid angle ψ^2, corresponding to a cross-sectional area $A = D^2$ at the source. The observed surface brightness I_{obs} is related to the emitted brightness $I_{em} = L/(4\pi A)$ by

$$I_{obs} = \frac{S}{\psi^2} = \frac{L}{4\pi r_L^2}\frac{r_A^2}{D^2} = I_{em}\left(\frac{r_A}{r_L}\right)^2 = \frac{I_{em}}{(1+z)^4}.$$

This is called **cosmological surface brightness dimming**. Notice that it depends only on the redshift, not on the geometry of the universe or the present scale size. Going as the fourth power of $(1 + z)$, it is a strong effect; it makes galaxies at large redshift much harder to see than might be expected.

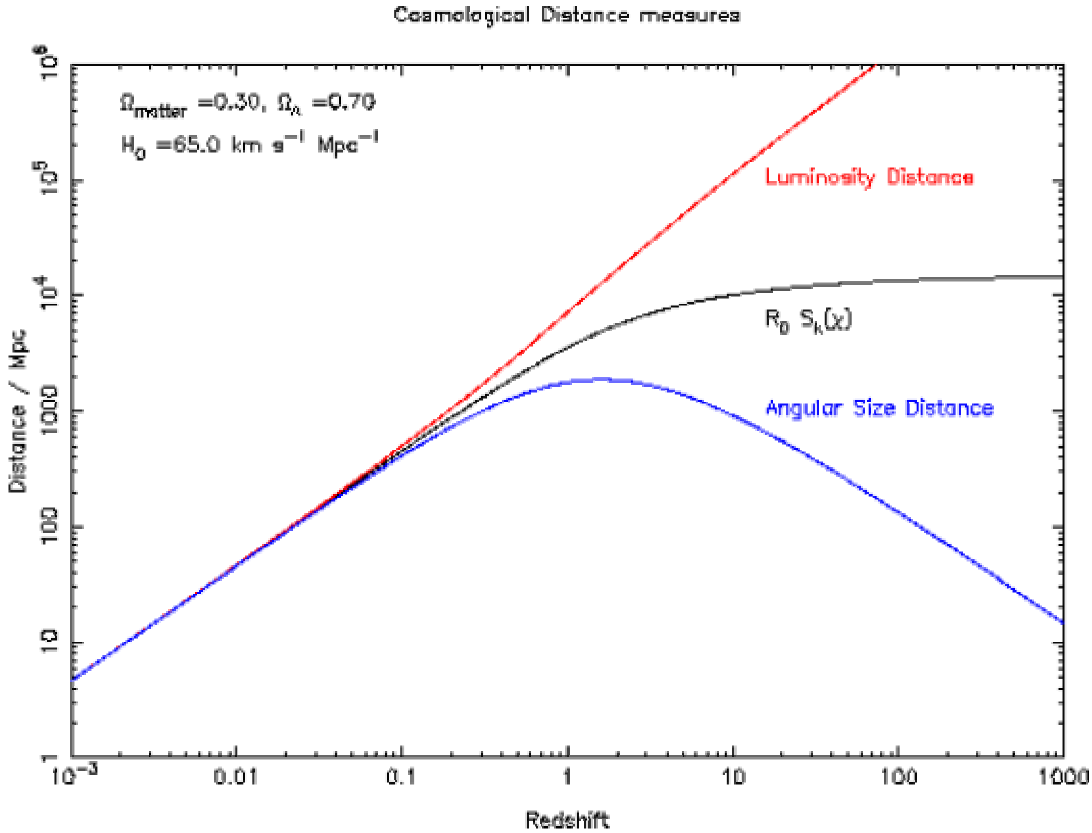

Figure 1.10: *Plot of angular size distance, luminosity distance and* $R_0 S_k(\chi)$ *against redshift. In this case k = 0, so that* $R_0 S_k(\chi) = R_0 \chi$*, the proper distance. The detailed shape of these curves depends on the assumed R(t); the model used here is consistent with observations of our own Universe.*

Answers to questions

1. Convince yourself that a repeating pattern like that in Fig. 1.7, except with a diamond instead of a rectangular basic unit, would be caused if a cylinder was twisted by 180° around its axis before joining the two ends.

Answer to question

A diamond repeating pattern is shown below, as before outlined with a dashed line. I have also drawn in a square cut (dot-dashed rectangle) equivalent to our cylinder, as we will see.

As before, there are only two people here. Suppose the woman goes for a walk, as shown. From the point of view of the repeating pattern, the central man sees copy A of the woman leave to his lower right, but shortly after, copy B passes by from his upper left. An equal time interval later, copy C arrives from his lower left.

Now consider our cylinder, which consists of the dash-dot square. The scale bar shows the angle around the cylinder, which I'll call longitude: as the top and bottom edges are joined, 0° = 360°. The woman is standing at a longitude of about 90°. If we had joined the cylinder up as a torus, if the woman walks along the cylinder her longitude will stay the same, so she will return from the left also at longitude 90°, as in *Asteroids*. But if we make a 180° twist before joining the ends, when she leaves the right side at longitude 90°, she will return at longitude 90° + 180° = 270°, i.e. at position B. (If you try to visualize this, I suggest you imagine a long thin cylinder, rather than a square one!). In turn, when she crosses the cylinder again and leaves the right-hand edge at longitude 270°, she will return on the left edge at longitude 270° + 180° = 450° ≡90°, i.e. at position C.

In other words, the connectivity of the twisted cylinder is the same as the diamond repeat pattern, as I claimed.