3 General Relativity

The general theory of relativity (Einstein 1915) is the theory of gravity. General relativity ("Einstein's theory") replaced the previous theory of gravity, Newton's theory. The fundamental idea in (both special and general) relativity is that space and time form together a 4-dimensional spacetime. The fundamental idea in general relativity is that gravity is manifested as *curvature* of this spacetime. While in Newton's theory gravity acts directly as a force between two bodies, in Einstein's theory the gravitational interaction is mediated by the spacetime. A massive body curves the surrounding spacetime. This curvature then affects the motion of other bodies. "Matter tells spacetime how to curve, spacetime tells matter how to move" [1]. From the viewpoint of general relativity, gravity is not a force at all; if there are no other forces than gravity acting on a body, we say the body is in *free fall.* A freely falling body is moving as straight as possible in the curved spacetime, along a *geodesic line.* If there are other forces, they cause the body to deviate from the geodesic line. It is important to remember that the viewpoint is that of *spacetime*, not just space. For example, the orbit of the earth around the sun is curved in space, but as straight as possible in spacetime.

If a spacetime is not curved, we say it is *flat*, which just means that it has the geometry of Minkowski space. In the case of 2- or 3-dimensional space, "flat" means that the geometry is Euclidean.

3.1 Curved 2-d and 3-d space

If you are familiar with the concept of curved space and how its geometry is given by the metric, you can skip the following discussion of 2- and 3-dimensional spaces and jump to Sec. 3.3.

Ordinary human brains cannot visualize a curved 3-dimensional space, let alone a curved 4-dimensional spacetime. However, we can visualize *some* curved 2-dimensional spaces by considering them embedded in flat 3-dimensional space.[1] So let us consider first a 2-d space. Imagine there are 2-d beings living in this 2-d space. They have no access to a third dimension. How can they determine whether the space they live in is curved? By examining whether the laws of Euclidean geometry hold. If the space is flat, then the sum of the angles of any triangle is 180°, and the circumference of any circle with radius χ is $2\pi\chi$. If by measurement they find that this does not hold for some triangles or circles, then they can conclude that the space is curved.

A simple example of a curved 2-d space is the sphere. The sum of angles of any triangle on a sphere is greater than 180°, and the circumference of any circle is less than $2\pi\chi$. (Straight, i.e. geodesic, lines, e.g. sides of a triangle, on the sphere are sections of *great circles*, which divide the sphere into two equal hemispheres. The radius of a circle has to be measured along the sphere surface).

[1] This embedding is only an aid in visualization. A curved 2-d space is defined completely in terms of its 2 independent coordinates, without any reference to a higher dimension, the geometry being given by the metric (a part of the definition of the 2-d space), an expression in terms of these coordinates. Some such curved 2-d spaces have the same geometry as some 2-d surface in flat 3-d space. We then say that the 2-d space can be embedded in flat 3-d space. But other curved 2-d spaces have no such corresponding surface, i.e., they can not be embedded in flat 3-d space.

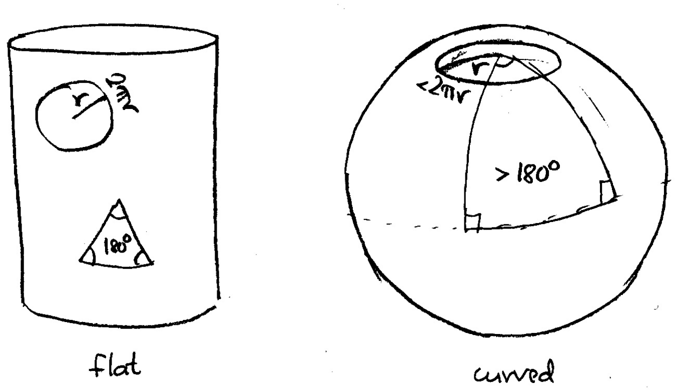

Figure 1: Cylinder and sphere.

Note that the surface of a cylinder has Euclidean geometry, i.e., there is no way that 2-d beings living on it could conclude that it differs from a flat surface, and thus by our definition it is a flat 2-d space. (Except that by traveling around the cylinder they could conclude that their space has a strange *topology*).

In a similar manner we could try to determine whether the 3-d space around us is curved, by measuring whether the sum of angles of a triangle is 180° or whether a sphere with radius r has surface area $4\pi r^2$. In fact, the space around Earth is curved due to Earth's gravity, but the curvature is so small, that more sophisticated measurements than the ones described above are needed to detect it.

3.2 The metric of 2-d and 3-d space

The tool to describe the geometry of space is the *metric*. The metric is given in terms of a set of coordinates. The coordinate system can be an arbitrary curved coordinate system. The coordinates are numbers which identify locations, but do not, by themselves, yet say anything about physical distances. The distance information is in the metric.

To introduce the concept of a metric, let us first consider Euclidean 2-dimensional space with Cartesian coordinates x,y. A parameterized curve $x(\eta)$, $y(\eta)$, begins at η_1 and ends at η_2. The length of the curve is given by

$$s = \int ds = \int \sqrt{dx^2 + dy^2} = \int_{\eta_1}^{\eta_2} \sqrt{x'^2 + y'^2} d\eta \,, \tag{1}$$

where $x' \equiv dx/d\eta$, $y' \equiv dy/d\eta$. Here $ds = \sqrt{dx^2 + dy^2}$ is the *line element*. The square of the line element, the *metric*, is

$$ds^2 = dx^2 + dy^2 \,. \tag{2}$$

The line element has the dimension of distance. If our coordinates are dimensionless, we need to include the distance scale in the metric. If the separation of neighboring coordinate lines, e.g., $x = 1$ and $x = 2$ is a (say, $a = 1$cm), then we have

$$ds^2 = a^2 \left(dx^2 + dy^2\right) \tag{3}$$

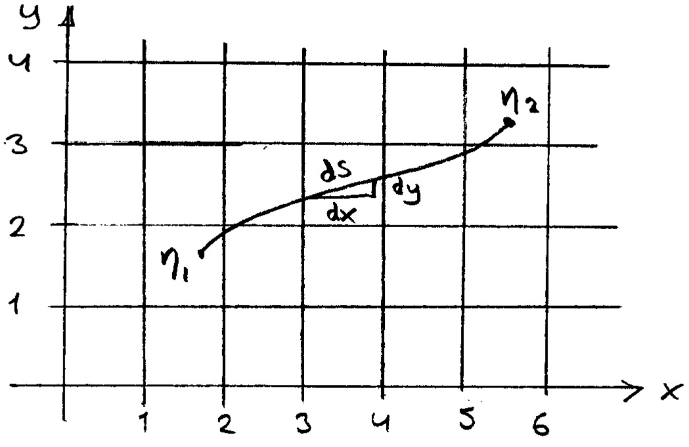

Figure 2: A parameterized curve in Euclidean 2-d space with Cartesian coordinates.

where a could be called the *scale factor.* As a working definition for the *metric*, we can use that *the metric is an expression which gives the square of the line element in terms of the coordinate differentials.*

We could use another coordinate system on the same 2-dimensional Euclidean space, e.g., polar coordinates. Then the metric is

$$ds^2 = a^2\left(dr^2 + r^2 d\varphi^2\right) , \tag{4}$$

giving the length of a curve as

$$s = \int ds = \int a\sqrt{dr^2 + r^2 d\varphi^2} = \int_{\eta_1}^{\eta_2} a\sqrt{r'^2 + r^2\varphi'^2}d\eta\,. \tag{5}$$

In a similar manner, in 3-dimensional Euclidean space, the metric is

$$ds^2 = dx^2 + dy^2 + dz^2 \tag{6}$$

in (dimensionful) Cartesian coordinates, and

$$ds^2 = dr^2 + r^2 d\vartheta^2 + r^2\sin^2\vartheta d\varphi^2 \tag{7}$$

in spherical coordinates (where the r coordinate has the dimension of distance, but the angular coordinates ϑ and φ are of course dimensionless).

Now we can go to our first example of a curved (2-dimensional) space, the sphere. Let the radius of the sphere be a. For the two coordinates on this 2-d space we can take the angles ϑ and φ. We get the metric from the Euclidean 3-d metric in spherical coordinates by setting $r \equiv a$,

$$ds^2 = a^2\left(d\vartheta^2 + \sin^2\vartheta d\varphi^2\right) . \tag{8}$$

The length of a curve $\vartheta(\eta), \varphi(\eta)$ on this sphere is given by

$$s = \int ds = \int_{\eta_1}^{\eta_2} a\sqrt{\vartheta'^2 + \sin^2\vartheta\varphi'^2}d\eta\,. \tag{9}$$

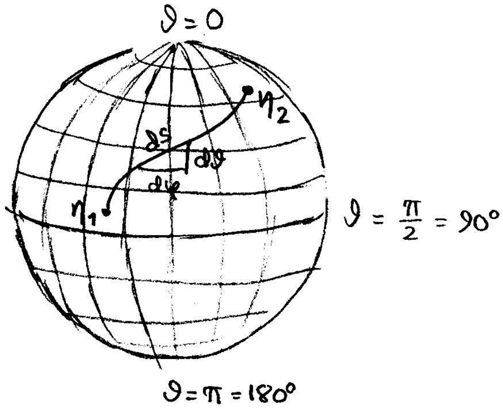

Figure 3: A parameterized curve on a 2-d sphere with spherical coordinates.

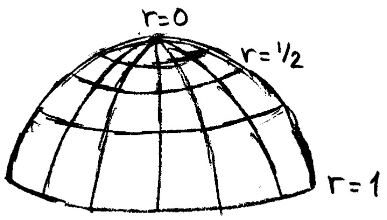

Figure 4: The part of the sphere covered by the coordinates in Eq. (10).

For later application in cosmology, it is instructive to now consider a coordinate transformation $r = \sin\vartheta$ (this new coordinate r has nothing to do with the earlier r of 3-d space, it is a coordinate on the sphere growing in the same direction as ϑ, starting at $r = 0$ from the North Pole ($\vartheta = 0$)). Since now $dr = \cos\vartheta d\vartheta = \sqrt{1-r^2}d\vartheta$, the metric becomes

$$ds^2 = a^2\left(\frac{dr^2}{1-r^2} + r^2 d\varphi^2\right) . \tag{10}$$

For $r \ll 1$ (in the vicinity of the North Pole), this metric is approximately the same as Eq. (4), i.e., it becomes polar coordinates on the "Arctic plain", with scale factor a. Only as r gets bigger we begin to notice the deviation from flat geometry. Note that we run into a problem when $r = 1$. This corresponds to $\vartheta = 90°$, i.e. the "equator". After this $r = \sin\vartheta$ begins to decrease again, repeating the same values. Also, at $r = 1$, the $1/(1-r^2)$ factor in the metric becomes infinite. We say we have a *coordinate singularity* at the equator. There is nothing wrong with the space itself, but our chosen coordinate system applies only for a part of this space, the region "north" of the equator.

3.3 4-d flat spacetime

Let us now return to the 4-dimensional spacetime. The coordinates of the 4-dimensional spacetime are (x^0, x^1, x^2, x^3), where $x^0 = t$ is a time coordinate. Some examples are "Cartesian" (t, x, y, z) and spherical $(t, r, \vartheta, \varphi)$ coordinates. The coordinate system can be an arbitrary curved coordinate system. The coordinates do

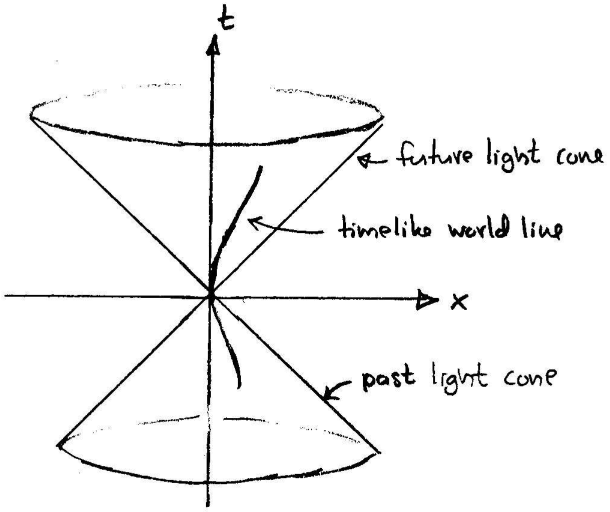

Figure 5: The light cone.

not, by themselves, yet say anything about physical distances. The distance information is in the metric. We shall often use a *Greek index* to denote an arbitrary spacetime coordinate, x^μ, where it is understood that μ can have any of the values 0, 1, 2, 3. *Latin* indices are used to denote space coordinates, x^i, where it is understood that i can have any of the values 1, 2, 3.

The metric of the Minkowski space of *special relativity* is

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2, \tag{11}$$

in Cartesian coordinates. In spherical coordinates it is

$$ds^2 = -dt^2 + dr^2 + r^2 d\vartheta^2 + r^2 \sin^2\vartheta\, d\varphi^2, \tag{12}$$

The fact that time appears in the metric with a different sign, is responsible for the special geometric features of Minkowski space. (I am assuming you already have some familiarity with special relativity.) There are three kinds of directions,

- timelike, $ds^2 < 0$
- lightlike, $ds^2 = 0$
- spacelike, $ds^2 > 0$.

The lightlike directions form the observer's future and past *light cones.* Light moves along the light cone, so that everything we see lies on our past light cone. To see us as we are now, the observer has to lie on our future light cone. As we move in time along our world line, we drag our light cones with us so that they sweep over the spacetime. The motion of any massive body is always timelike.

3.4 Curved spacetime

These features of the Minkowski space are inherited by the spacetime of general relativity. However, spacetime is now *curved*, whereas in Minkowski space it is

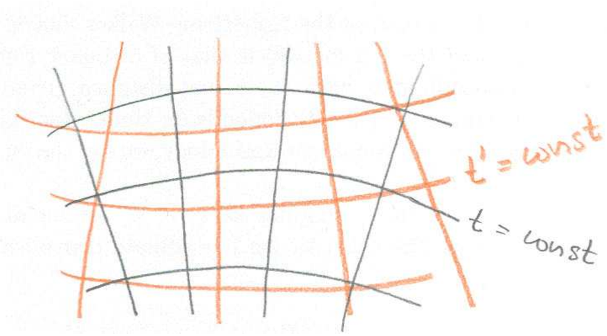

Figure 6: Two coordinate systems with different time slicings.

flat (i.e., not curved). (When we say space, or the universe, is flat, we mean it has Euclidean geometry, when we say spacetime is flat, we mean it has Minkowski geometry.) The (proper) length of a spacelike curve is $\Delta s \equiv \int ds$. Light moves along lightlike world lines, $ds^2 = 0$, massive objects along timelike world lines $ds^2 < 0$. The time measured by a clock carried by the object, the *proper time*, is $\Delta\tau = \int d\tau$, where $d\tau \equiv \sqrt{-ds^2}$, so that $d\tau^2 = -ds^2 > 0$. The proper time τ is a natural parameter for the world line, $x^\mu(\tau)$. The *four-velocity* of an object is defined as

$$u^\mu = \frac{dx^\mu}{d\tau}. \tag{13}$$

The zeroth component of the 4-velocity, $u^0 = dx^0/d\tau = dt/d\tau$ relates the proper time τ to the *coordinate time* t, and the other components of the 4-velocity, $u^i = dx^i/d\tau$, to *coordinate velocity* $v^i \equiv dx^i/dt = u^i/u^0$. To convert this coordinate velocity into a "physical" velocity (with respect to the coordinate system), we still need to use the metric, see Eq. (38).

In an *orthogonal* coordinate system the coordinate lines are everywhere orthogonal to each other. The metric is then diagonal, meaning that it contains no cross-terms like $dxdy$. We shall only use orthogonal coordinate systems in this course.

The three-dimensional subspace ("hypersurface") $t = const$ of spacetime is called the space (or the *universe*) at time t, or a *time slice* of the spacetime. It is possible to slice the same spacetime in many different ways, i.e., to use coordinate systems with different $t = const$ hypersurfaces.

3.5 Robertson–Walker metric

In cosmology we often assume that the spacetime is *homogeneous* (in space, but not in time). This means that there exists a coordinate system whose $t = const$ hypersurfaces are homogeneous. The time coordinate is then called the *cosmic time*.

There is good evidence, that the universe is indeed rather homogeneous (all places look the same) and isotropic (all directions look the same) at sufficiently large scales (i.e., ignoring smaller scale features), > 100 Mpc. Therefore we shall now consider homogeneous and isotropic spacetimes only. This means that the curvature of spacetime must be the same everywhere and into every direction, but it may change in time. It can be shown, that the metric can then be given (by a suitable choice of the coordinates) in the form

$$ds^2 = -dt^2 + a^2(t)\left[\frac{dr^2}{1-Kr^2} + r^2 d\vartheta^2 + r^2 \sin^2\vartheta\, d\varphi^2\right], \tag{14}$$

the *Robertson–Walker* (RW) metric. This is thus the metric of our universe, to first approximation, and we shall work with this metric for a large part of this course.[2] The time coordinate t is the *cosmic time.* Here K is a constant, related to curvature of *space* and $a(t)$ is a function of time, related to expansion (or possible contraction) of the universe. We call

$$r_{\text{curv}} \equiv a(t)/\sqrt{|K|} \tag{15}$$

the *curvature radius* of space (at time t). The above metric is given in spherical coordinates. We notice immediately that the 2-dimensional surfaces $t = r = const$ have the metric of a sphere with radius ar. The time-dependent factor $a(t)$ is called the *scale factor.* We will need the Einstein equation to solve $a(t)$. (This will be done in the next chapter.) For now, it is an arbitrary function of the time coordinate t. Another common notation for $a(t)$ is $R(t)$.

We have the freedom to rescale the radial coordinate r. For example, we can multiply all values of r by a factor of 2, if we also divide a by a factor of 2 and K by a factor of 4. The geometry of the spacetime stays the same, the meaning of the coordinate r has just changed: the point that had a given value of r has now twice that value in the rescaled coordinate system. There are two common ways to rescale. If $K \neq 0$, we can rescale r to make K equal to ± 1. In this case K is usually denoted k, and it has three possible values, $k = -1, 0, +1$. (In this case r is dimensionless, and $a(t)$ has the dimension of distance.) The other way is to rescale a to be one at present[3], $a(t_0) \equiv a_0 = 1$. (In this case $a(t)$ is dimensionless, whereas r and $K^{-1/2}$ have the dimension of distance.) To choose one of these two scalings would simplify some of our equations, but we resist the temptation, and keep the general form (14). This way we avoid the possible confusion resulting from comparing equations using different scaling conventions.

If $K = 0$, the space part ($t = const$) of the Robertson–Walker metric is flat. The 3-metric (the space part of the full metric) is that of ordinary Euclidean space written in spherical coordinates, with the radial distance given by ar. The *spacetime*, however, is curved, since $a(t)$ depends on time, describing the expansion or contraction of space. In common terminology, we say the "universe is flat" in this case.

If $K > 0$, the coordinate system is singular at $r = 1/\sqrt{K}$. (Remember our discussion of the 2-sphere!) With the substitution (coordinate transformation) $r = K^{-1/2}\sin\chi$ the metric becomes

$$ds^2 = -dt^2 + a^2(t)K^{-1}\left[d\chi^2 + \sin^2\chi\, d\vartheta^2 + \sin^2\chi \sin^2\vartheta\, d\varphi^2\right]. \tag{16}$$

The space part has the metric of a *hypersphere*, a sphere with one extra dimension. χ is a new angular coordinate, whose values range over 0–π, just like ϑ. The singularity at $r = 1/\sqrt{K}$ disappears in this coordinate transformation, showing that it was just

[2] That is, for the whole of Cosmology I. In Cosmology II we shall consider deviations from this homogeneity.

[3] In some discussions of the early universe, it may also be convenient to rescale a to be one at some particular early time.

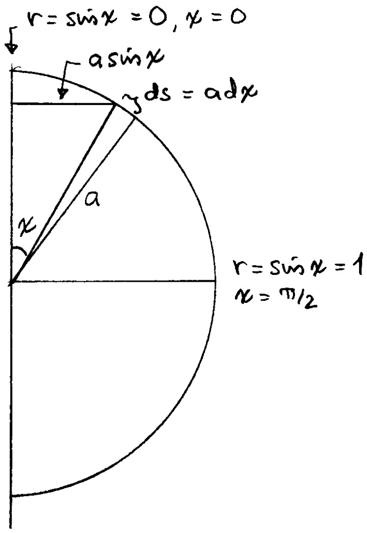

Figure 7: The hypersphere. This figure is for $K = k = 1$. Consider the semicircle in the figure. It corresponds to χ ranging from 0 to π. You get the (2-dimensional) sphere by rotating this semicircle off the paper around the vertical axis by an angle $\Delta\varphi = 2\pi$. You get the (3-dimensional) hypersphere by rotating it twice, in two extra dimensions, by $\Delta\vartheta = \pi$ and by $\Delta\varphi = 2\pi$, so that each point makes a sphere. Thus each point in the semicircle corresponds to a full sphere with coordinates ϑ and φ, and radius $(a/\sqrt{K})\sin\chi$.

a coordinate singularity, not a singularity of the spacetime. The original coordinates covered only half of the hypersphere, as the coordinate singularity $r = 1/\sqrt{K}$ divides the hypersphere into two halves. The case $K > 0$ corresponds to a *closed* universe, whose (spatial) curvature is *positive.*[4] This is a finite universe, with circumference $2\pi a/\sqrt{K} = 2\pi r_{\text{curv}}$ and volume $2\pi^2 K^{-3/2} a^3 = 2\pi^2 r_{\text{curv}}^3$, and we can think of r_{curv} as the radius of the hypersphere.

If $K < 0$, we do not have a coordinate singularity, and r can range from 0 to ∞. The substitution $r = |K|^{-1/2}\sinh\chi$ is, however, often useful in calculations. The case $K < 0$ corresponds to an *open* universe, whose (spatial) curvature is *negative.* The metric is then

$$ds^2 = -dt^2 + a^2(t)|K|^{-1}\left[d\chi^2 + \sinh^2\chi\left(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2\right)\right]. \tag{17}$$

This universe is infinite, just like the case $K = 0$.

The Robertson–Walker metric (at a given time) has two associated length scales. The first is the curvature radius, $r_{\text{curv}} \equiv a|K|^{-1/2}$. The second is given by the time scale of the expansion, the *Hubble time*, $t_H \equiv H^{-1}$, where $H \equiv \dot{a}/a$ is the *Hubble parameter.* The Hubble time multiplied by the speed of light, $c = 1$, gives the *Hubble length*, $\ell_H \equiv ct_H \equiv H^{-1}$. In the case $K = 0$ the universe is flat, so the only length scale is the Hubble length.

The coordinates $(t, r, \vartheta, \varphi)$ of the Robertson–Walker metric are called *comoving* coordinates. This means that the coordinate system follows the expansion of space, so that the space coordinates of objects which *do not move* remain the same. The

[4] Positive (negative) curvature means that the sum of angles of any triangle is greater than (less than) 180° and that the area of a sphere with radius r is less than (greater than) $4\pi r^2$.

homogeneity of the universe fixes a special frame of reference, the *cosmic rest frame* given by the above coordinate system, so that, unlike in special relativity, the concept "does not move" has a specific meaning. The coordinate distance between two such objects stays the same, but the physical, or *proper* distance between them grows with time as space expands. The time coordinate t, the *cosmic time*, gives the time measured by such an observer at rest, at $(r, \vartheta, \varphi) = \mathit{const.}$

It can be shown that the expansion causes the motion of an object in free fall to slow down with respect to the comoving coordinate system. For nonrelativistic velocities,

$$v(t_2) = \frac{a(t_1)}{a(t_2)}\, v(t_1). \tag{18}$$

(This expression refers to the "physical" velocity of Eq. (38).) The *peculiar velocity* of a galaxy is its velocity with respect to the comoving coordinate system.

3.6 Distance

Let us now ignore the peculiar velocities of galaxies (i.e., we assume they are $= 0$), so that they will stay at fixed coordinate values (r, ϑ, φ), and consider the distances between them. We set the origin of our coordinate system at galaxy O (observer). Let the r-coordinate of galaxy A be r_A. Since we assumed the peculiar velocity of galaxy A to be 0, the coordinate r_A stays constant with time. To calculate the proper distance between the galaxies at time t, we need the metric, $s(t) = \int_0^{r_\mathrm{A}} ds$. We integrate along the path $t, \vartheta, \varphi = \mathit{const}$, or $dt = d\vartheta = d\varphi = 0$, so $ds^2 = a^2(t)\,\frac{dr^2}{1 - Kr^2}$, and get

$$\begin{aligned} s(t) &= \int_0^{r_\mathrm{A}} a(t) \frac{dr}{\sqrt{1 - Kr^2}} \\ &= \begin{cases} K^{-1/2} a(t) \arcsin(K^{1/2} r_A) = K^{-1/2} a(t) \chi\,, & (K > 0) \\ a(t)\, r_A\,, & (K = 0) \\ |K|^{-1/2} a(t)\, \mathrm{arsinh}(|K|^{1/2} r_A) = |K|^{-1/2} a(t) \chi\,. & (K < 0) \end{cases} \end{aligned} \tag{19}$$

There are also other distance concepts in cosmology, which will be discussed in the next chapter.

3.7 Redshift

Let us now find the redshift of galaxy A. Light leaves the galaxy at time t_1 with wavelength λ_1 and arrives at galaxy O at time t_2 with wavelength λ_2. It takes a time $\delta t_1 = \lambda_1/c = 1/\nu_1$ to send one wavelength and a time $\delta t_2 = \lambda_2/c = 1/\nu_2$ to receive one wavelength. Follow now the two light rays sent at times t_1 and $t_1 + \delta t_1$ (see figure). t and r change, ϑ and φ stay constant (this is clear from the symmetry of the problem). Light obeys the *lightlike* condition

$$ds^2 = 0\,. \tag{20}$$

We have thus

$$ds^2 = -dt^2 + a^2(t) \frac{dr^2}{1 - Kr^2} = 0 \tag{21}$$

$$\Rightarrow \quad \frac{dt}{a(t)} = \frac{-dr}{\sqrt{1 - Kr^2}}\,. \tag{22}$$

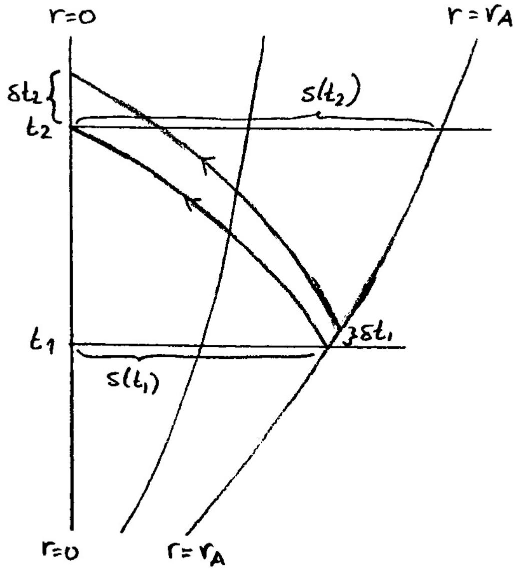

Figure 8: The two light rays to establish the redshift.

Integrating this, we get for the first light ray,

$$\int_{t_1}^{t_2} \frac{dt}{a(t)} = \int_0^{r_{\mathrm{A}}} \frac{dr}{\sqrt{1-Kr^2}}\,, \tag{23}$$

and for the second,

$$\int_{t_1+\delta t_1}^{t_2+\delta t_2} \frac{dt}{a(t)} = \int_0^{r_{\mathrm{A}}} \frac{dr}{\sqrt{1-Kr^2}}\,. \tag{24}$$

The right hand sides of the two equations are the same, since the sender and the receiver have not moved (they have stayed at $r = r_A$ and $r = 0$). Thus

$$0 = \int_{t_1+\delta t_1}^{t_2+\delta t_2} \frac{dt}{a(t)} - \int_{t_1}^{t_2} \frac{dt}{a(t)} = \int_{t_2}^{t_2+\delta t_2} \frac{dt}{a(t)} - \int_{t_1}^{t_1+\delta t_1} \frac{dt}{a(t)} = \frac{\delta t_2}{a(t_2)} - \frac{\delta t_1}{a(t_1)}, \tag{25}$$

and the time to receive one wavelength is

$$\delta t_2 = \frac{a(t_2)}{a(t_1)}\,\delta t_1. \tag{26}$$

As is clear from the derivation, this *cosmological time dilation* effect applies to observing any event taking place in galaxy A. As we observe galaxy A, we see everything happening in "slow motion", slowed down by the factor $a(t_2)/a(t_1)$, which is the factor by which the universe has expanded since the light (or any electromagnetic signal) left the galaxy. This effect can be observed, e.g., in the light curves of supernovae (their luminosity as a function of time).

For the redshift we get

$$1 + z \equiv \frac{\lambda_2}{\lambda_1} = \frac{\delta t_2}{\delta t_1} = \frac{a(t_2)}{a(t_1)}. \tag{27}$$

Thus the redshift directly tells us how much smaller the universe was when the light left the galaxy. The result is easy to remember; the wavelength expands with the universe.

3.8 Conformal time

In the comoving coordinates of Eqs.(14), (16), and (17), the space part of the coordinate system is expanding with the expansion of the universe. It is often practical to make a corresponding change in the time coordinate, so that the "unit of time" (i.e., separation of time coordinate surfaces) also expands with the universe. The *conformal time* η is defined by

$$d\eta \equiv \frac{a_0}{a(t)}dt, \qquad \text{or} \qquad \eta = a_0 \int_0^t \frac{dt'}{a(t')}. \tag{28}$$

Exercise: Write the RW metric in the coordinates $(\eta, r, \vartheta, \varphi)$ and $(\eta, \chi, \vartheta, \varphi)$. The latter form is especially nice for studying radial ($d\vartheta = d\varphi = 0$) light propagation, where $ds^2 = 0$.

3.9 Vectors, tensors, and the volume element

The *metric* of spacetime can always be written as

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu \equiv \sum_{\mu=0}^{3}\sum_{\nu=0}^{3} g_{\mu\nu}dx^\mu dx^\nu. \tag{29}$$

We introduce Einstein's *summation rule*: there is a sum over repeated indices (that is, we don't bother to write down the summation sign $\sum$ in this case). Greek (spacetime) indices go over the values 0–3, Latin (space) indices over the values 1–3, i.e., $g_{ij}dx^i dx^j \equiv \sum_{i=1}^3 \sum_{j=1}^3 g_{ij}dx^i dx^j$. The objects $g_{\mu\nu}$ are the components of the *metric tensor*. They have, in principle, the dimension of distance squared. In practice one often assigns the dimension of distance (or time) to some coordinates, and then the corresponding components of the metric tensor are dimensionless. These *coordinate distances* are then converted to *proper* ("real" or "physical") distances with the metric tensor. The components of the metric tensor form a symmetric 4×4 matrix.

Example 1. The metric tensor for a sphere (discussed above as an example of a curved 2-d space) has the components

$$[g_{ij}] = \begin{bmatrix} a^2 & 0 \\ 0 & a^2 \sin^2\vartheta \end{bmatrix}. \tag{30}$$

Example 2. The metric tensor for Minkowski space has the components

$$g_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{31}$$

in Cartesian coordinates, and

$$g_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2\vartheta \end{bmatrix} \tag{32}$$

in spherical coordinates.

Example 3. The Robertson-Walker metric of Eq.(14) has components

$$g_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & \frac{a^2}{1-Kr^2} & 0 & 0 \\ 0 & 0 & a^2r^2 & 0 \\ 0 & 0 & 0 & a^2r^2\sin^2\vartheta \end{bmatrix}. \tag{33}$$

Note that the metric tensor components in the above examples always formed a diagonal matrix. This is the case when the coordinate system is orthogonal.

The vectors which occur naturally in relativity are *four-vectors*, with four components, e.g., the four-velocity. The values of the components depend on the basis $\{\mathbf{e}_\alpha\}$ used. Note that the index of the basis vector does not refer to a component, but specifies which one of the four basis vectors is in question. The components of the basis vectors in the basis they define are, of course,

$$(\mathbf{e}_\alpha)^\beta = \delta^\beta_\alpha. \tag{34}$$

Given a coordinate system, we have two bases (also called *frames*) naturally associated with it, the *coordinate basis* and the corresponding normalized basis. If the coordinate system is orthogonal, the latter is an *orthonormal basis.* When we use the coordinates to define the components of a vector, like the 4-velocity in Eq. (13), the components naturally come out in the coordinate basis. The basis vectors of a coordinate basis are parallel to coordinate lines, and their length represents the distance from changing the value of the coordinate by one unit. For example, if we move along the coordinate x^1 so that it changes by dx^1, the distance traveled is $ds = \sqrt{g_{11}dx^1dx^1} = \sqrt{g_{11}}dx^1$. The length of the basis vector $\mathbf{e}_1$ is thus $\sqrt{g_{11}}$. Since in the coordinate basis the basis vectors usually are not unit vectors, the numerical values of the components give the wrong impression of the magnitude of the vector. Therefore we may want to convert them to the normalized basis

$$\mathbf{e}_{\hat\alpha} \equiv (\frac{1}{\sqrt{|g_{\alpha\alpha}|}})\mathbf{e}_\alpha. \tag{35}$$

(It is customary to denote the normalized basis with a hat over the index, when both bases are used. In the above equation there is no sum over the index α, since it appears only once on the left hand side.) For a four-vector $\mathbf{w}$ we have

$$\mathbf{w} = w^\alpha\mathbf{e}_\alpha = w^{\hat\alpha}\mathbf{e}_{\hat\alpha}, \tag{36}$$

where

$$w^{\hat\alpha} \equiv \sqrt{|g_{\alpha\alpha}|}w^\alpha. \tag{37}$$

For example, the components of the coordinate velocity of a massive body, $v^i = dx^i/dt$ could be greater than one; the “physical velocity”, i.e., the velocity measured by an observer who is at rest in the comoving coordinate system, is [5]

$$v^{\hat i} = \sqrt{g_{ii}}dx^i/\sqrt{|g_{00}|}dx^0\,, \tag{38}$$

with components always smaller than one.

[5] When $g_{00} = -1$, this simplifies to $\sqrt{g_{ii}}dx^i/dt$.

The volume of a region of space (given by some range in the spatial coordinates x^1, x^2, x^3) is given by

$$V = \int_V dV = \int_V \sqrt{\det[g_{ij}]} dx^1 dx^2 dx^3 \tag{39}$$

where $dV \equiv \sqrt{\det[g_{ij}]} dx^1 dx^2 dx^3$ is the *volume element.* Here $\det[g_{ij}]$ is the determinant of the 3×3 submatrix of the metric tensor components corresponding to the spatial coordinates. For an orthogonal coordinate system, the volume element is

$$dV = \sqrt{g_{11}} dx^1 \sqrt{g_{22}} dx^2 \sqrt{g_{33}} dx^3 . \tag{40}$$

The metric tensor is used for taking scalar (dot) products of four-vectors,

$$\mathbf{w} \cdot \mathbf{u} \equiv g_{\alpha\beta} u^\alpha w^\beta \, . \tag{41}$$

The (squared) *norm* of a four-vector $\mathbf{w}$ is

$$\mathbf{w} \cdot \mathbf{w} \equiv g_{\alpha\beta} w^\alpha w^\beta . \tag{42}$$

Exercise: Show that the norm of the four-velocity is always -1.

For an *orthonormal* basis we have

$$\begin{aligned} \mathbf{e}_{\hat{0}} \cdot \mathbf{e}_{\hat{0}} &= -1 \\ \mathbf{e}_{\hat{0}} \cdot \mathbf{e}_{\hat{\jmath}} &= 0 \\ \mathbf{e}_{\hat{\imath}} \cdot \mathbf{e}_{\hat{\jmath}} &= \delta_{ij} . \end{aligned} \tag{43}$$

We shall use the short-hand notation

$$\mathbf{e}_{\hat{\alpha}} \cdot \mathbf{e}_{\hat{\beta}} = \eta_{\alpha\beta} , \tag{44}$$

where the symbol $\eta_{\alpha\beta}$ is like the Kronecker symbol $\delta_{\alpha\beta}$, except that $\eta_{00} = -1$.

3.10 Contravariant and covariant components

We sometimes write the index as a subscript, sometimes as a superscript. This has a precise meaning in general relativity. This is explained in this small-print (indicating stuff not really needed in this course) subsection. The component w^α of a four-vector is called a *contravariant* component. We define the corresponding *covariant* component as

$$w_\alpha \equiv g_{\alpha\beta} w^\beta . \tag{45}$$

The norm is now simply

$$\mathbf{w} \cdot \mathbf{w} = w_\alpha w^\alpha . \tag{46}$$

In particular, for the 4-velocity we always have

$$u_\mu u^\mu = g_{\mu\nu} u^\mu u^\nu = \frac{ds^2}{d\tau^2} = -1 . \tag{47}$$

We defined the metric tensor through its covariant components (Eq. 29). We now define the corresponding covariant components $g^{\alpha\beta}$ as the inverse matrix of the matrix $[g_{\alpha\beta}]$,

$$g_{\alpha\beta} g^{\beta\gamma} = \delta^\gamma_\alpha . \tag{48}$$

Now

$$g^{\alpha\beta}w_\beta = g^{\alpha\beta}g_{\beta\gamma}w^\gamma = \delta^\alpha_\gamma w^\gamma = w^\alpha. \tag{49}$$

The metric tensor can be used to lower and raise indices. For tensors,

$$\begin{aligned} A_\alpha^{\ \beta} &= g_{\alpha\gamma}A^{\gamma\beta} \\ A_{\alpha\beta} &= g_{\alpha\gamma}g_{\beta\delta}A^{\gamma\delta} \\ A^{\alpha\beta} &= g^{\alpha\gamma}g^{\beta\delta}A_{\gamma\delta}. \end{aligned} \tag{50}$$

Note that the *mixed components* $A_\alpha^{\ \beta} \neq A^\beta_{\ \alpha}$, unless the tensor is symmetric, in which case we can write A^β_α.

For an orthonormal basis,

$$g_{\hat\alpha\hat\beta} = g^{\hat\alpha\hat\beta} = \eta_{\alpha\beta}, \tag{51}$$

and the covariant and contravariant components of vectors and tensors have the same values, except that the raising or lowering of the time index 0 changes the sign. These orthonormal components are also called "physical" components, since they have the "right" magnitude.

Note that the symbols $\delta_{\alpha\beta}$ and $\eta_{\alpha\beta}$ are not tensors, and the location of their index carries no meaning.

3.11 Einstein equation

From the first and second partial derivatives of the metric tensor,

$$\partial g_{\mu\nu}/\partial x^\sigma, \qquad \partial^2 g_{\mu\nu}/(\partial x^\sigma \partial x^\tau), \tag{52}$$

one can form various *curvature tensors*. These are the Riemann tensor $R^\mu_{\ \nu\rho\sigma}$, the Ricci tensor $R_{\mu\nu} \equiv R^\alpha_{\ \mu\alpha\nu}$, and the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$, where R is the Ricci scalar $g^{\alpha\beta}R_{\beta\alpha}$, also called the "scalar curvature" (not to be confused with the scale factor $R(t)$ of the Robertson–Walker metric). We shall not discuss these curvature tensors in this course. The only purpose of mentioning them here is to be able to show the general form of the Einstein equation, before we go to the much simpler specific case of the Friedmann models.

In Newton's theory the source of gravity is mass, in the case of continuous matter, the mass density ρ. According to Newton, the gravitational field $\vec{g}_N$ is given by the equation

$$\nabla^2\Phi = -\nabla\cdot\vec{g}_N = 4\pi G\rho \tag{53}$$

Here Φ is the gravitational potential.

In Einstein's theory, the source of spacetime curvature is the *energy-momentum tensor*, also called the *stress-energy tensor*, or, for short, the "energy tensor" $T^{\mu\nu}$. The energy tensor carries the information on energy density, momentum density, pressure, and stress. The energy tensor of frictionless continuous matter (a *perfect fluid*) is

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + pg^{\mu\nu}, \tag{54}$$

where ρ is the energy density and p is the pressure in the *rest frame* of the fluid. In cosmology we can usually assume that the energy tensor has the perfect fluid form. T^{00} is the energy density in the coordinate frame. (T^{i0} gives the momentum density, which is equal to the energy flux T^{0i}. T^{ij} gives the flux of momentum i-component in j-direction.)

We can now give the general form of the Einstein equation,

$$G^{\mu\nu} = 8\pi G T^{\mu\nu}. \tag{55}$$

This is the *law of gravity* according to Einstein. Comparing to Newton (Eq. 53) we see that the mass density ρ has been replaced by $T^{\mu\nu}$, and $\nabla^2\Phi$ has been replaced by the Einstein tensor $G^{\mu\nu}$, which is a short way of writing a complicated expression containing first and second derivatives of $g_{\mu\nu}$. Thus the gravitational potential is replaced by the 10 components of $g_{\mu\nu}$ in Einstein's theory.

In the case of a weak gravitational field, the metric is close to the Minkowski metric, and we can write, e.g.,

$$g_{00} = -1 - 2\Phi \tag{56}$$

(in suitable coordinates), where Φ is small. The Einstein equation for g_{00} becomes then

$$\nabla^2\Phi = 4\pi G(\rho + 3p). \tag{57}$$

Comparing this to Eq. (53) we see that the density ρ has been replaced by $\rho + 3p$. For relativistic matter, where p can be of the same order of magnitude than ρ this is an important modification to the law of gravity. For nonrelativistic matter, where the particle velocities are $v \ll 1$, we have $p \ll \rho$, and we get Newton's equation.

When applied to a homogeneous and isotropic universe filled with ordinary matter, the Einstein equation tells us that the universe cannot be static, it must either expand or contract.[6] When Einstein was developing his theory, he did not believe this was happening in reality. He believed the universe was static. Therefore he modified his equation by adding an extra term,

$$G^{\mu\nu} + \Lambda g^{\mu\nu} = 8\pi G T^{\mu\nu}. \tag{58}$$

The constant Λ is called the *cosmological constant*. Without Λ, a universe which was momentarily static, would begin to collapse under its own weight. A positive Λ acts as repulsive gravity. In Einstein's model for the universe (the *Einstein universe*), Λ had precisely the value needed to perfectly balance the pull of ordinary gravity. This value is so small that we would not notice its effect in small scales, e.g., in the solar system. The Einstein universe is, in fact, unstable to small perturbations.[7] When Einstein heard that the Universe was expanding, he threw away the cosmological constant, calling it "the biggest blunder in my life".

In more recent times the cosmological constant has made a comeback in the form of *vacuum energy*. Considerations in quantum field theory suggest that, due to vacuum fluctuations, the energy density of the vacuum should not be zero, but some constant ρ_{vac}.[8] The energy tensor of the vacuum would then have the form $T_{\mu\nu} = -\rho_{\text{vac}} g_{\mu\nu}$. Thus vacuum energy has exactly the same effect as a cosmological constant with the value

$$\Lambda = 8\pi G\rho_{\text{vac}}. \tag{59}$$

[6] It leads to $\ddot{a} < 0$, which does not allow $a(t) = \text{const}$. If we momentarily had $\dot{a} = 0$, a would immediately begin to decrease.

[7] "If you sneeze, the universe will collapse."

[8] In field theory, the fundamental physical objects are fields, and particles are just quanta of the field oscillations. *Vacuum* means the ground state of the system, i.e., fields have those values which correspond to minimum energy. This minimum energy is usually assumed to be zero. However, in quantum field theory, the fields cannot stay at fixed values, because of quantum fluctuations. Thus even in the ground state the fields fluctuate around their zero-energy value, contributing a positive energy density. This is analogous to the zero-point energy of a harmonic oscillator in quantum mechanics.

Vacuum energy is observationally indistinguishable from a cosmological constant. This is because in physics, we can usually measure only energy differences. Only gravity responds to absolute energy density, and there a constant energy density has the same effect as the cosmological constant. In principle, however, they represent different ideas. The cosmological constant is an "addition to the left-hand side of the Einstein equation", a *modification of the law of gravity*, whereas vacuum energy is an "addition to the right-hand side", a contribution to the energy tensor, i.e., a form of energy.

A problem with the idea of vacuum energy is that its value from vacuum fluctuations should naturally come out huge (because all frequencies of oscillations should contribute), but observations restrict its value to be much smaller (by a factor of about 10^{-120}, the biggest fine-tuning problem in theoretical physics) compared to what one would expect from vacuum fluctuations. Probably there is some deep symmetry reason, not yet understood, why the vacuum energy is exactly zero.[9]

In fact, as we shall discuss later, observational evidence agrees with a positive cosmological constant / vacuum energy with such a small magnitude. It would explain the observed acceleration in the expansion of the universe. Because of the above fine-tuning problem, it is generally thought, that instead of vacuum energy we have something which is not really vacuum energy, but just something resembling it, called *dark energy*. The other possibility is that the law of gravity needs to be *modified* at cosmological scales in such a way that the effect is close to that of a cosmological constant. This is an active research area at the moment, but it seems that we are nowhere close to a solution. The hope is to find a model where the fine-tuning would have a natural explanation.

References

[1] C.W. Misner, K.S. Thorne, J.A. Wheeler, Gravitation (Freeman 1973).

[9] Or perfectly balanced with a cosmological constant of the same magnitude but opposite sign!