PHY 475/375
Lecture 5
(April 9, 2012)

Describing Curvature (contd.)

So far, we have studied homogenous and isotropic surfaces in 2-dimensions. The results can be extended easily to three dimensions.

As in 2-dimensions, if we want a homogenous and isotropic space in 3 dimensions, we have only three possibilities: that the space can be uniformly flat, that it can have uniform positive curvature, or it can have uniform negative curvature.

Again, as in 2-dimensions we can specify the geometry of these 3-dimensional surfaces by two quantities, κ and R, with $\kappa = 0$ for a flat space, $\kappa = +1$ for a positively curved space, and $\kappa = -1$ for a negatively curved space.

If a 3-dimensional surface is flat ($\kappa = 0$), we can write its metric in cartesian coordinates as

$$ds^2 = dx^2 + dy^2 + dz^2 \tag{3.12}$$

The counterpart of polar coordinates in 2-dimensions is now the standard form of the spherical coordinates (r, θ, ϕ) in 3-dimensions, and the metric of a flat 3-dimensional space in spherical coordinates can be written as

$$ds^2 = dr^2 + r^2\left[d\theta^2 + \sin^2\theta\, d\phi^2\right] \tag{3.13}$$

If a 3-dimensional space has uniform positive curvature ($\kappa = +1$), its metric can be written as

$$ds^2 = dr^2 + R^2 \sin^2\left(\frac{r}{R}\right)\left[d\theta^2 + \sin^2\theta\, d\phi^2\right] \tag{3.14}$$

A positively curved 3-dimensional space has finite volume, just as a positively curved 2-dimensional space has finite area.

If a 3-dimensional space has uniform negative curvature ($\kappa = -1$), its metric can be written as

$$ds^2 = dr^2 + R^2 \sinh^2\left(\frac{r}{R}\right)\left[d\theta^2 + \sin^2\theta\, d\phi^2\right] \tag{3.15}$$

Like flat space, a negatively curved 3-dimensional space has infinite volume.

Looking at equations (3.13), (3.14), and (3.15), we see that they have a similar basic structure, with the differences due to the curvature captured within one of the terms. Whenever this is the case, the standard procedure is to write one compact equation, with new symbols defined to cover all exhaustive cases. We shall now proceed to do this.

We can express the three possible metrics for a homogenous, isotropic, and 3-dimensional space more compactly by writing them in the form:

$$ds^2 = dr^2 + S_\kappa(r)^2\, d\Omega^2 \tag{3.16}$$

where

$$d\Omega^2 \equiv d\theta^2 + \sin^2\theta\, d\phi^2 \tag{3.17}$$

and

$$S_\kappa(r) = \begin{cases} R\sin\left(\dfrac{r}{R}\right), & \kappa = +1 \\ r, & \kappa = 0 \\ R\sinh\left(\dfrac{r}{R}\right), & \kappa = -1 \end{cases} \tag{3.18}$$

Note that in the limit $r \ll R$, $S_\kappa \approx r$, regardless of the value of κ; this makes sense, since between two points separated by a distance that is much smaller than the radius of curvature of the surface, the surface will seem flat, no matter what the curvature.

Note also that for flat or negatively curved spaces, S_κ increases monotonically with r, so that eventually $S_\kappa \to \infty$ as $r \to \infty$. In contrast, when space is positively curved, S_k increases to a maximum of $S_{\rm max} = R$ at $r/R = \pi/2$, then decreases to 0 at $r/R = \pi$, the antipodal point to the origin.

Note that we can also choose to work in different coordinate systems. For example, if we switch the radial coordinate from r to $x \equiv S_\kappa(r)$, the metric for a homogenous, isotropic, 3-dimensional space takes the form:

$$ds^2 = \frac{dx^2}{1 - \kappa\, x^2/R^2} + x^2\, d\Omega^2 \tag{3.19}$$

Even though equations (3.16) and (3.19) look different, they represent the same homogenous, isotropic spaces. They merely have a different functional form due to the different choice of radial coordinates.

The Metric in $(3+1)$ Dimensions

So far, we have only considered the metrics for 2-dimensional and 3-dimensional spaces. In relativity, though, we must work with four dimensional space-time. It turns out that in $(3+1)$ dimensions, we have the same set of homogenous and isotropic manifolds, but promoted to higher dimensionality, meaning that it is effectively impossible to picture them with our 3-dimensional brains. Fortunately, however, we can describe these manifolds in terms of mathematics. Just as we can compute the distance between two points in space, for example, using the appropriate metric for that space, so too we can compute the 4-dimensional distance between two events in space-time.

Consider two events, one occurring at the space-time location (ct, r, θ, ϕ), and another at the space-time location $(ct + cdt, r + dr, \theta + d\theta, \phi + d\phi)$. Using the laws of special relativity, we can write the space-time separation between these two events as

$$ds^2 = -c^2dt^2 + dr^2 + r^2d\Omega^2 \tag{3.20}$$

Such a metric is called the *Minkowski metric*, and the space-time which it describes is called the *Minkowski space-time*. If we compare equation (3.20) with equation (3.16), we see immediately that the spatial component of Minkowski space-time is flat (i.e., Euclidean).

Another complication that is introduced when we get into space-time is that, unlike 3-dimensional space where ds^2 has to be positive, there is no such restriction in space-time. You can see this clearly in equation (3.20), and so we must separate our geodesics based on the sign of ds^2:

- Geodesics with $ds^2 > 0$ are called *time-like geodesics*; that is, information can be conveyed from one event to the other at less than the speed of light.
- Geodesics with $ds^2 = 0$ are called *null geodesics*; that is, information can be conveyed from one point to the other at exactly the speed of light. It follows that two events separated by a null geodesic can be connected by a light ray. In fact, it is a fundamental postulate of relativity that $ds^2 = 0$ for photons, that is, the path of a photon through 4-dimensional space-time is a null geodesic.
- Geodesics with $ds^2 < 0$ are called *space-like geodesics*; therefore, no information can be conveyed from one point to the other, because to do so would involve going faster than the speed of light.

In Minkowski space-time, therefore, a photon's trajectory obeys the relation

$$ds^2 = 0 = -c^2dt^2 + dr^2 + r^2d\Omega^2 \tag{3.21}$$

The Notation of General Relativity

While, e.g., equation (3.20) is written out in full in the above treatment, one usually tends to find more compact notation in advanced texts. We will continue to write the more explicit forms for now, but it is worth getting used to the more compact notation, if for nothing else than to give us the ability to follow the published literature on the subject.

In more compact notation, the metric is usually written by using the invariant interval

$$ds^2 = \sum_{\mu,\nu} g_{\mu\nu}\, dx^\mu\, dx^\nu \tag{3.a}$$

where the indices μ and ν both run over the values $(0, 1, 2, 3)$ in order to correspond to the time and the three spatial coordinates, and dx^μ is a differential space-time coordinate interval. The metric is specified by $g_{\mu\nu}$. As we are free to use different coordinate systems, the actual form of the metric depends on the coordinates used, and below we will look at some examples. As stated previously, however, the physics must not depend on the form of the metric used.

In equation (3.21), we wrote the Minkowski metric in spherical polar coordinates, but let us write it in Cartesian coordinates here first to illustrate the different forms of the metric $g_{\mu\nu}$.

With Cartesian coordinates specified by $(x_0, x_1, x_2, x_3) \equiv (ct, x, y, z)$, where as usual, we take the time coordinate corresponding to $\mu = 0$ to be $dx^0 = c\,dt$ so that it has the units of length, we have

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 \tag{3.b}$$

so that the metric tensor for the Minkowski metric in Cartesian coordinates is given by

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix} \quad \text{or} \quad g_{\mu\nu} = \begin{pmatrix} -1 & & & \\ & +1 & & \\ & & +1 & \\ & & & +1 \end{pmatrix} \tag{3.c}$$

where we've written the form on the right with the zeros suppressed to increase visibility.

In spherical polar coordinates, writing out equation (3.20) in full form gives:

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

so that the metric tensor for the Minkowski metric in spherical polar coordinates is given by

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix} \quad \text{or} \quad g_{\mu\nu} = \begin{pmatrix} -1 & & & \\ & +1 & & \\ & & +r^2 & \\ & & & +r^2\sin^2\theta \end{pmatrix} \tag{3.d}$$

The Robertson-Walker Metric

The Minkowski metric of equation (3.20) has a spatial component that is flat, so it applies only within the context of the Special Theory of Relativity. The "special" here means that the theory deals only with the special case of the absence of gravity, that is, the special case in which space-time is not curved by the presence of mass/energy. In the 1930's, Howard Robertson and Arthur Walker independently wrote down a space-time metric for a homogenous and isotropic universe in which distances are allowed to expand (or contract) as a function of time. It is called the *Robertson-Walker metric* (the full form is the *Friedmann-Lemaitre-Robertson-Walker* or *FLRW metric*), and may be written in the form

$$ds^2 = -c^2dt^2 + a(t)^2\left[\frac{dx^2}{1-\kappa\,x^2/R_0^2} + x^2\,d\Omega^2\right] \tag{3.24}$$

From this equation, we see that the spatial component of the Robertson-Walker metric consists of the spatial metric for a uniformly curved space of radius of curvature R_0 (e.g., see equation 3.19), scaled by the square of the scale factor $a(t)$ that we introduced in an earlier lecture to describe how distances in a homogenous, isotropic universe expand or contract with time. We do this because the cosmological principle tells us that the Universe is spatially homogenous and isotropic in time, but we also need to change the overall scale of the Universe as a function of time.

The Robertson-Walker metric may be written in the form whose spatial part matches equation (3.16):

$$ds^2 = -c^2dt^2 + a(t)^2\Big[dr^2 + S_\kappa(r)^2 d\Omega^2\Big] \tag{3.25}$$

where $S_\kappa(r)$ for the three different types of curvature is given by equation (3.18).

The time variable t in the Robertson-Walker metric is the *cosmological proper time*, called the *cosmic time* for short, and is the time measured by an observer who sees the Universe expanding uniformly around their location.

The spatial variables (x, θ, ϕ) or (r, θ, ϕ) are called the *comoving coordinates* of a point in space. If the expansion of the Universe is perfectly homogenous and isotropic, the comoving coordinates of any point remain constant with time.

In summary, the assumption of homogeneity and isotropy is a very powerful assumption, so powerful that Robertson and Walker made it in the 1930's, long before the available observational evidence gave firm support for such a conclusion. If the Universe is truly homogenous and isotropic, then everything we need to know about its geometry is contained within $a(t), \kappa$, and R_0.

- The scale factor $a(t)$ is a dimensionless function of time which describes how distances increase (or decrease) with time; it is normalized so that $a(t_0) = 1$ at the present moment.
- The curvature constant κ is a dimensionless number which can take on one of three discrete values: $\kappa = +1$ if the universe has positive spatial curvature, $\kappa = 0$ if the universe is spatially flat, $\kappa = -1$ if the universe has negative spatial curvature.
- The radius of curvature R_0 has dimensions of length, and gives the radius of curvature of the universe at the present moment.

Much of modern cosmology, as we shall see in future lectures, is directed toward finding the values of $a(t)$, κ, and R_0.

It is worth noting at this stage that the Universe is not homogenous and isotropic on small scales.

- Small dense lumps like humans and interstellar dust grains are held together by electromagnetic forces and do not expand.
- Larger lumps like planetary systems, galaxies, and clusters of galaxies are held together by their own gravity and do not expand.

It is only on scales larger than $\sim$100 Mpc that the expansion of the Universe can be treated as the ideal homogenous and isotropic expansion to which the Robertson-Walker metric can be strictly applied.

Proper Distance

The expansion of the Universe requires that we specify a procedure to determine the spatial distance between two objects. That is, since in an expanding universe, the distance between two objects increases with time, we must specify the time t at which the measured distance is the correct one. Typically, there are several measures of distance based on different considerations. We will look at one of them here, the so-called *proper distance* $d_p(t)$. This proper distance between two points is equal to the length of the spatial geodesic between them when the scale factor is fixed at the value $a(t)$.

Suppose we are at the origin, and that the galaxy we are observing is at a comoving coordinate position (r, θ, ϕ). The proper distance between us, the observer, and the galaxy can be found using the Robertson-Walker metric at a fixed time t:

$$ds^2 = a(t)^2 \Big[dr^2 + S_\kappa(r)^2 d\Omega^2 \Big] \tag{3.26}$$

Along the spatial geodesic between the observer and galaxy, the angle (θ, ϕ) is constant, so that

$$ds = a(t)\, dr \tag{3.27}$$

The proper distance $d_p(t)$ can then be found by integrating over the radial comoving coordinate r:

$$d_p(t) = a(t) \int_0^r dr = a(t) r \tag{3.28}$$

That is, if you could freeze time and set down a ruler between yourself and an object, you would get the proper distance of that object from your location.

Since the comoving coordinate r is constant with time, the rate of change for the proper distance between us and a distant galaxy is

$$\dot{d}_p(t) = \dot{a}\, r = \left(\frac{\dot{a}}{a} \right) d_p(t) \tag{3.30}$$

Recall that we are using the convention of writing d/dt with a dot above the quantity: $\dfrac{d(d_p)}{dt} = \dot{d}_p$.

If we measure $\dot{d}_p(t)$ at the current time t_0, it is just the recession speed $v_p(t_0)$ of the galaxy. Therefore, equation (3.30) tells us that at the current time ($t = t_0$) there is a linear relation between the recession speed of a galaxy and the proper distance to the galaxy:

$$v_p(t_0) = H_0\, d_p(t_0) \tag{3.31}$$

where

$$H_0 = \left(\frac{\dot{a}}{a} \right)_{t=t_0} \tag{3.33}$$

While this was demonstrated earlier from observations, we are now explicitly interpreting the change in distance between galaxies in our mathematical framework as being associated with the expansion of space.

The relation given in equation (3.31) does have the unfortunate consequence that for points separated by a proper distance greater than a critical value

$$\Big[d_p(t_0)\Big]_H \equiv \frac{c}{H_0} \tag{3.34}$$

where $[d_p(t_0)]_H$ is called the *Hubble distance*, we will have $v_p > c$!!! That is, galaxies beyond the Hubble distance are currently moving away from us faster than the speed of light!!!

There is much in the literature on this subject. Ever since high redshifts have been measured, observational astronomers have tended to use the special relativity correction to make sure the conversion to velocities doesn't come out with superluminal speeds (as I've asked PHY 375 students to do in HW 1). However, this makes theoreticians, especially general relativists, bristle. In the purest interpretation of general relativity, there is no unique way to compare vectors at widely separated space-time points in a curved space-time, and so the notion of the relative velocity of a distant galaxy is almost meaningless. Put another way, general relativity does not have a problem in having two points move away from each other at speeds faster than that of light. It is only cropping up as a problem because you're trying to do something that is not well defined in a curved space-time, namely, you're trying to compare vectors at different points in a curved space-time. Many suggest the problem arises with interpreting the observed redshift as a Doppler shift, hence tying it to a velocity. In fact, we can leave aside an interpretation for the redshift, and consider it as a number in its own right, especially since it reveals information on the scale factor of the universe at the time of emission, as we will show below.

Redshift and Scale Factor

The redshift z of a galaxy is a measured quantity; recall that it is measured as

$$z \equiv \frac{\lambda_{\rm obs} - \lambda_{\rm rest}}{\lambda_{\rm rest}}$$

where $\lambda_{\rm obs}$ is the observed wavelength of a particular line emitted by an object in the Universe (e.g., a galaxy), and $\lambda_{\rm rest}$ is the rest wavelength of that line measured in the laboratory. Below, we will show how z is linked to the scale factor a of the Universe at the time of emission of the line whose wavelength is being measured.

Consider now the light that was emitted by a galaxy at time t_e and observed by us at time t_0. During its travel from the distant galaxy to us, the light traveled along a null geodesic, with $ds = 0$. Along the null geodesic, we consider θ and ϕ constant, so that

$$c^2 dt^2 = a(t)^2 dr^2 \tag{3.37}$$

which, upon rearranging, becomes

$$c\,\frac{dt}{a(t)} = dr \tag{3.38}$$

The left hand side is a function of t only, whereas the right hand side is independent of t.

Suppose the galaxy emits light with a wavelength λ_e, as measured by an observer in the emitting galaxy. Let us fix our attention on a single wave crest of the emitted light. The wave crest is emitted at a time t_e and observed at a time t_0, such that

$$c\int_{t_e}^{t_0}\frac{dt}{a(t)} = \int_0^r dr = r \tag{3.39}$$

The next wave crest of light is emitted at a time $t_e + \lambda_e/c$, and is observed at a time $t_0 + \lambda_0/c$, where in general, $\lambda_e \neq \lambda_0$. So, for the second wave crest

$$c\int_{t_e+\lambda_e/c}^{t_0+\lambda_0/c}\frac{dt}{a(t)} = \int_0^r dr = r \tag{3.40}$$

Comparing equations (3.39) and (3.40), we obtain

$$\int_{t_e}^{t_0}\frac{dt}{a(t)} = \int_{t_e+\lambda_e/c}^{t_0+\lambda_0/c}\frac{dt}{a(t)} \tag{3.41}$$

That is, the integral of $dt/a(t)$ between the time of emission and the time of observation is the same for every wave crest in the emitted light.

We can now split up the range of integration on both sides to eliminate an overlap:

$$\int_{t_e}^{t_e+\lambda_e/c}\frac{dt}{a(t)} + \int_{t_e+\lambda_e/c}^{t_0}\frac{dt}{a(t)} = \int_{t_e+\lambda_e/c}^{t_o}\frac{dt}{a(t)} + \int_{t_0}^{t_0+\lambda_0/c}\frac{dt}{a(t)}$$

The two inner terms are equal and cancel, so we are just left with

$$\int_{t_e}^{t_e+\lambda_e/c}\frac{dt}{a(t)} = \int_{t_0}^{t_0+\lambda_0/c}\frac{dt}{a(t)} \tag{3.43}$$

Equation (3.43) is telling us that the integral of $dt/a(t)$ between the emission of successive wave crests is equal to the integral of $dt/a(t)$ between the observation of successive wave crests.

Now, the Universe doesn't have the time to expand by a significant amount during the time between the emission or observation of two successive wave crests. For example, while the time scale for expansion of the Universe is the Hubble time, $H_0^{-1} \approx 14$ Gyr, the time between wave crests for visible light is $\lambda/c \approx 2\times10^{-15}$ s $\approx 10^{-32}\, H_0^{-1}$. This means that $a(t)$ is effectively constant in each of the integrals of equation (3.43), and so we may write

$$\frac{1}{a(t_e)}\int_{t_e}^{t_e+\lambda_e/c} dt = \frac{1}{a(t_0)}\int_{t_0}^{t_0+\lambda_0/c} dt \tag{3.44}$$

Integrating equation (3.44), we get

$$\frac{1}{a(t_e)}\Big[t\Big]_{t_e}^{t_e+\lambda_e/c} = \frac{1}{a(t_0)}\Big[t\Big]_{t_0}^{t_0+\lambda_0/c}$$

so that

$$\frac{1}{a(t_e)}\left[t_e + \frac{\lambda_e}{c} - t_e\right] = \frac{1}{a(t_0)}\left[t_0 + \frac{\lambda_0}{c} - t_0\right]$$

and finally

$$\frac{\lambda_e}{a(t_e)} = \frac{\lambda_0}{a(t_0)} \tag{3.45}$$

Now, let us rewrite the definition of redshift z using the symbols employed in equation (3.45). If we replace $\lambda_{\rm obs}$ with λ_0, and $\lambda_{\rm rest}$ with λ_e, the definition of redshift is

$$z = \frac{\lambda_0 - \lambda_e}{\lambda_e}$$

Add 1 to both sides:

$$1 + z = 1 + \frac{\lambda_0 - \lambda_e}{\lambda_e} = \frac{\lambda_e + \lambda_0 - \lambda_e}{\lambda_e}$$

so

$$1 + z = \frac{\lambda_0}{\lambda_e}$$

and, since from equation (3.45), we have $\lambda_0/\lambda_e = a(t_0)/a(t_e)$, we obtain finally

$$1 + z = \frac{a(t_0)}{a(t_e)}$$

With the usual convention that $a(t_0) = 1$, this becomes

$$1 + z = \frac{1}{a(t_e)} \tag{3.46}$$

Thus, if we observe a galaxy with a redshift of 2, we are observing it when the Universe had a scale factor

$$a(t_e) = \frac{1}{1+z} = \frac{1}{1+2} = \frac{1}{3}$$

In summary, the redshift we observe depends only on the relative scale factors at the time of emission and the time of observation. It does not depend on how the transition between $a(t_e)$ and $a(t_0)$ was made. It doesn't matter if the expansion was gradual or abrupt. All that matters are the scale factors at the time of emission and the time of observation.

The section on redshift and scale factor was done on W (4/11), but is included here for continuity.