PROF. KASEN, SPRING 2023

C161 RELATIVISTIC ASTROPHYSICS AND COSMOLOGY

Contents

1
Relativity and Coordinate Systems

1.1 Coordinates and the Metric

1.1.1 Charting out Spacetime

Philosophically, the fundamental nature of space and time is a subject of debate. Mathematically, we have a functional description that is widely used in physics. Spacetime is a set of points (called a *manifold*) that are labeled by real numbers (called *coordinates*), and for which we have a rule for measuring distance (called a *metric*).

The dimensionality of a manifold indicates the number of real numbers needed to uniquely describe each point; for example, a 4D manifold requires four numbers (call them t, x, y, z) to uniquely identify each point (i.e., event). Think about the coordinates as just a sticker that we post at each spacetime point to identify it. Such labels do not (in themselves) fully define the geometry of the space; to discuss distances and angles on the manifold, we need to add additional mathematical structure – the *metric* – which provides a rule to construct invariant quantities from our coordinate labels[1].

The metric plays the role of the "scale bar" on a map which tells you "1 inch corresponds to 1 mile". Many maps have a single scale bar that applies everywhere and in all directions. This works fine for describing an approximately flat 2D region (e.g., the Cal campus). But maps may have a scale bar that varies from place to place; indeed, there is no way to display a curved 2D surface (e.g., of the earth) on a flat 2D plane without having a varying scale bar.

Figure 1.1 shows an example of a Mercator projection map, the north and south poles look "distorted" and too big. The apparent distortion is because the coordinate spacing of the map (indicated by the dotted grid lines) does not correspond directly to physical distance. If we use the local scale bar to translate coordinate spacing into physical distances, however, we will get the right measurements.

Figure 1.2 shows a different way of mapping the earth (called a

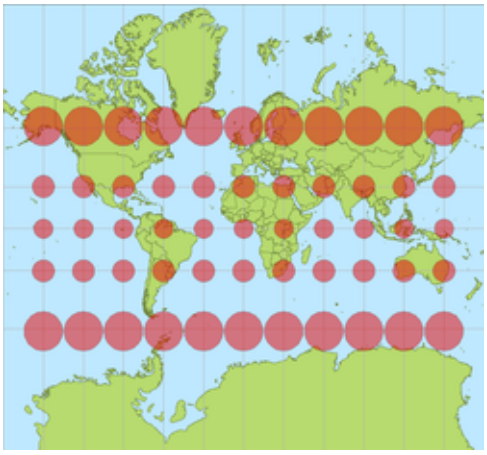


Figure 1.1: A Mercator projection map of the earth. The red circles (called *Tissot's Indicratices*) indicate the scale bar – e.g., one circle diameter equals 1000 miles. This scale bar is different at different places on the map.

[1] For example, if we go to a concert in an arena, knowing you are in seat 5A and I am in seat 9D, uniquely identifies our positions, but does not tell us the distance (or angle) between us. We need additional information to turn the seat labels (i.e., coordinates) into physical measurements.

Behrmann map) in which the scale bar not only varies from place to place, but is also different along the vertical and horizontal directions. Compared to the Mercator map, the north and south poles are squeezed vertically and therefore don't look as extended. While the coordinate (x, y) positions of places on earth differ in the Behrmann and Mercator map, all distance measurements will be the same if we use the appropriate scale bars. The combination of coordinates *and* metric give us the full geometrical picture of the space.

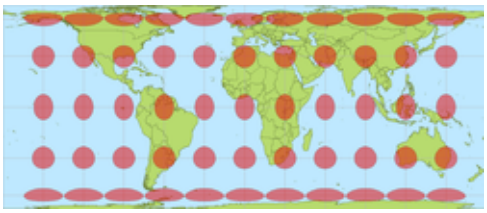


Figure 1.2: A Behrmann map of the earth. The scale bar (shown as the red elliposidal Tissot Indicratices) is different along the vertical direction than the horizontal, and varies from place to place on the earth.

1.1.2 *The Metric*

We now make the discussion of maps and scale bars mathematical, by defining a procedure which takes the coordinates of two points on a manifold and returns a physical distance. The simplest example applies to Euclidean (i.e., uncurved or flat space) labeled using Cartesian coordinates. For two points (x_1, y_1) and (x_2, y_2) on a 2D manifold, the Euclidean metric gives the distance between them

$$\Delta l^2 = \Delta x^2 + \Delta y^2 \tag{1.1}$$

where $\Delta x = x_2 - x_1$ and $\Delta y = y_2 - y_1$. Since we will be considering metrics that may vary with position, we will usually write metrics replacing Δ with a d to indicate that the displacements are all infinitesimally small

$$dl^2 = dx^2 + dy^2 \tag{1.2}$$

The Euclidean metric Eq. 1.2 is just a statement of the Pythogorean theorem. Indeed, once we supply this metric, all of the results of Euclidean geometry are derivable.

We are free to use a different coordinate system to label points, in which case the metric rule may look different. For example, polar coordinates maps the 2D plane using labels (ρ, ϕ) related to Cartesian coordinates by

$$x = \rho \cos\phi \qquad y = \rho \sin\phi \tag{1.3}$$

If we calculate the differentials dx, dy of the above equations[2] , and plug them into the metric Eq. 1.2 we get the metric expressed in polar coordinates

$$dl^2 = d\rho^2 + \rho^2 d\phi^2 \tag{1.4}$$

While this equation looks superficially different than Eq. 1.2, it describes the exact same metric (geometry) just using a different coordinate labeling convention.

It is possible, however, for a metric to encode the *non-Euclidean* geometry of a curved space. For example, we can label points on the surface of a sphere by coordinates (θ, ϕ) (corresponding essentially to

[2] We won't be too rigorous in dealing with differentials; for our purposes we can calculate them like we would derivatives, using the chain rule

$$dx = d\rho\ \cos\phi - \rho \sin\phi\ d\phi$$

$$dy = d\rho\ \sin\phi + \rho \cos\phi\ d\phi$$

Then plugging these into $dx^2 + dy^2$ gives Eq. 1.4.

latitude and longitude). The metric for the distance between points is[3]

$$dl^2 = a^2 d\theta^2 + a^2 \sin^2\theta d\phi^2 \tag{1.5}$$

where a is the radius of the sphere. Although it is not immediately obvious, the metric Eq. 1.5 is fundamentally different than the Pythagorean theorem Eq. 1.2 (no coordinate transform can turn one into the other)[4]. The geometry of curved spherical surface is fundamentally different than that of a flat plane.

In general, we can write a metric in 2D space labeled with arbitrary coordinates (x_1, x_2) as[5]

$$dl^2 = g_{11}\, dx_1^2 + g_{22}\, dx_2^2 + 2g_{12}\, dx_1\, dx_2 \tag{1.6}$$

where the metric components g_{11}, g_{22}, g_{12}, can be functions of the coordinates. For example, for a spherical surface (Eq. 1.5), the metric components are $g_{11} = a^2, g_{22} = a^2 \sin^2\theta, g_{12} = 0$. We can think of g_{11} as providing the scale bar in the horizontal direction and g_{22} the scale bar in the vertical direction[6], and as we saw in the maps of the earth, these scale bars may vary with the coordinates (x_1, x_2).

Since ancient times, many assumed that geometry was a form of *a-priori* knowledge; i.e., a set of mathematical propositions that could be known to be true independent of experience. However, once it was realized that non-Euclidean (i.e., curved) metrics are possible, the geometry of our universe became seen as *a-posteriori knowledge* that could only determined by going out and doing experiments. The core idea of General Relativity is that the metric of spacetime is determined by the distribution of mass/energy within it, and the Einstein equations of GR allow us to calculate the metric given a mass/energy distribution.

1.2 *Symmetry and Coordinate Transformations*

The notion of *Symmetry* is one of the most important concepts in physics. By definition, a symmetry is the quality of remaining unchanged under a transformation. We will be particular interested in *spacetime symmetries*, where the transformation is a changing of the spacetime coordinates (by, e.g., rotating the coordinates). We must define what exactly is being left unchanged by the transformation; two important possibilities are

1. The laws of physics

2. The contents of the universe

[3] We will return to this later and derive this metric, and after that generalize it to a 3D curved space which may describe our entire universe.

[4] It takes the more advanced mathematics of differential geometry to be able to determine whether any given metric corresponds to a curved or uncurved spacetime (one needs to calculate a quantity called the Reimann curvature tensor). We won't go so far in this class, but the interested student should consult the recommended books on GR.

[5] The metric is properly speaking a tensor, and Eq. 1.6 is one way of representing how it acts on differential displacement. Another way of representing the metric tensor is as a symmetric matrix

$$g^{\mu\nu} = \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{33} \end{pmatrix}$$

The metric tensor is necessarily symmetric, so $g_{21} = g_{12}$.

A proper discussion of the metric would introduce the ideas of differential geometry, include distinguishing between covariant and contravariant vectors, inner products and so on. We will sidestep this mathematics in this class, but the interested student is advised to look at the discussion given in most GR textbooks.

[6] Having non-zero off-diagonal metric component g_{12} indicates that the coordinate system at a given location is not orthogonal. We will usually be avoid these terms by using orthogonal coordinate systems.

We will at first be interested in 1), where it is the equations of physics (and in particular the spacetime metric equation) that have the symmetry. We discuss first translations and rotations, and after that the Lorentz transformation of special relativity.

1.2.1 *Space Translational Symmetry*

A *space translation* simply shifts the origin of the coordinate system by a constant amount. For example, in 2D Cartesian coordinates we can define a new labeling that shifts the x coordinate by a constant amount D

$$x' = x - D \qquad y' = y \tag{1.7}$$

This transformation leaves the x-distance between points unchanged, since

$$\Delta x' = x_2' - x_1' = (x_2 - D) - (x_1 - D) = x_2 - x_1 = \Delta x \tag{1.8}$$

And similarly $dx' = dx$ The Euclidean metric Eq. 1.2 remains unchanged under space translations

$$dl^2 = dx^2 + dy^2 = dx'^2 + dy'^2 \tag{1.9}$$

So the form of the metric equation remains unchanged under spatial translation. We expect that other equations of physics should also obey this symmetry (e.g., the Lagrangian used in Lagrangian mechanics, or the Hamiltonian used in quantum mechanics[7]).

We make use of translational symmetry all the time in physics without thinking much about it (e.g., we get to arbitrarily choose the origin of our coordinate system when solving a problem). But this symmetry actually says something deep about our universe – that there is *no preferred point in space*. This is a non-trivial realization. Indeed, in the Aristotelean physics that dominated for centuries after Aristotle, the center of the earth was thought to be special point, and gravity explained as the tendency of objects to seek the center. This physics lacked translational symmetry, and it took many years before cosmology dislodged the earth from its special place at took a "Copernican" view that all places in space are equivalent.

[7] In some cases it may look like an equation does not obey translational symmetry. For example, the Hamiltonian (i.e., energy) of a non-relativistic harmonic oscillator (mass on a spring with spring constant k) is

$$H = \frac{1}{2}mv^2 + \frac{1}{2}kx^2$$

If we perform a transformation $x' = x + D$ we find that this equation does not stay the same. But of course what matters in this problem is not the "absolute" position of the mass, but its displacement from the equilibrium (unstretched) point x_0. Above we implicitly took $x_0 = 0$ but we could more explicitly write this equation as

$$H = \frac{1}{2}mv^2 + \frac{1}{2}k(x - x_0)^2$$

which has translational symmetry. A consequence of translational symmetry is that a only relative difference in spatial coordinates will appear in equations.

1.2.2 *Rotational Symmetry*

A rotation in the 2D plane by an angle θ "mixes up" the two coordinates by the transformation

$$x' = x\cos\theta + y\sin\theta \qquad y' = -x\sin\theta + y\cos\theta \tag{1.10}$$

Written in vector form, this transformation is

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{1.11}$$

It is straight-forward to show[8] that in the metric in the primed coordinate system is

$$dl'^2 = dx'^2 + dy'^2 \tag{1.12}$$

So the metric is the same in both the primed and unprimed from, and hence has *rotational symmetry*.

Modern physics asserts that equations should be invariant under rotations. Again, this is a deep statement about our universe – that there is no preferred *direction* in space; the laws of physics are not different in one direction than another. Again, this need not be the case. One could imagine a universe where the laws of physics (e.g., the strength of gravity) behaved differently along one direction than another.

Due to rotational symmetry, there is no fundamental way to distinguish between the "x-direction" and the "y-direction"; the difference is merely a matter of convention. Similarly, the coordinate distances Δx and Δy between two points will be relative to the particular coordinate system we are using. The distance supplied by the metric, however, is an invariant that remains unchanged under rotation, and so has an absolute (not relative) meaning.

[8] Taking differentials of the primed coordinates

$$dx' = dx\cos\theta + dy\sin\theta$$

$$dy' = -dx\sin\theta + dy\cos\theta$$

Here θ is a constant parameter describing the rotation amount, not a coordinate and so we don't take a differential in the chain rule.

Multiplying out dx'^2 and dy'^2 (and using $\cos^2\theta + \sin^2\theta = 1$) you can show that

$$dx'^2 + dy'^2 = dx^2 + dy^2 = dl^2$$

1.2.3 *The Homogeneous, Isotropic Universe*

Modern theories assume the laws of physics are translationally and rotationally symmetry – i.e., the same at all places and in all directions. The *contents* of the universe, however, obviously do not possess the same symmetries, at least on small scales. The stuff in my room is different than the stuff in my office, which is different than the stuff on the moon or in another galaxy.

How is it that the contents of the universe evolved to vary from place to place, when the laws of physics that govern that evolution are the same everywhere? This is an important question[9] that modern cosmology at least partly tries to address. As we will see later, the leading theory is that symmetry was broken spontaneously by quantum fluctuations, which randomly set the density of the universe to be slightly different from place to place. Over time, the gravity of these "seeds" pulled in additional material, and the inhomogeneities grew until they formed cosmic structures we see today.

However, if we look on large enough scales – scales bigger than galaxy clusters – the universe does look, on average, pretty much the same at all places and in all directions. When studying the universe on these scales, we will often assume that the contents are homogeneous (the same everywhere) and isotropic (the same in all directions), an assumption known as the *cosmological principle*.

[9] In some sense this is *the* question of why we are here – a universe that is identical at all points would have no differentiation, no way to distinguish here from there, or this from that, or you from me. Such a universe would be, in practice, indistinguishable from nothingness. We see why the creation stories of so many different cultures often begin with a breaking of symmetry, e.g., the separation of "light from darkness" or "heaven from earth".

1.3 *Spacetime and Spacetime Diagrams*

In relativity, time becomes another coordinate that is combined with space into a 4D *spacetime* manifold. Points on the manifold are labeled by (in Cartesian coordinates) (t, x, y, z) and correspond to *events* which happen at a certain time and a certain place.

The path of a moving object is described by a curve through the 4D manifold. We illustrate such a path on a *spacetime diagram*, where time is plotted on the vertical axis and one of the space dimensions (say x) on the horizontal axis. Typically we plot ct (where c is the speed of light) so that both axes have units of length[10]. A light ray is represented by the line $x = ct$, and so has a slope $dx/d(ct) = 1$. Light moves along the diagonal (i.e., 45^o) in a (flat) spacetime diagram.

[10] Commonly one uses units where the speed of light $c = 1$, so we do not have to bother with the factors of c. But we will usually retain the factor to clarify units

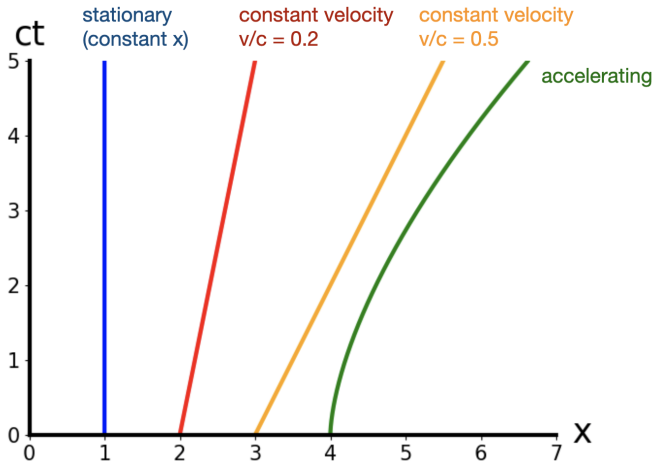


Figure 1.3: Example spacetime diagram of paths of moving objects. Because we choose to plot time on the vertical axis, the slope of the lines, dt/dx, gives the inverse of the coordinate velocity $v = dx/dt$ of the spacetime curve. The steeper the slope, the *slower* the motion.

Because coordinates are merely labels, we will need a metric to define physical distances on a spacetime diagram. What may seem at first natural would be to define *two* metrics, one for spatial distances dl and one for time "distances" (i.e., durations), $d\tau$. This was what was essentially used in Newtonian, non-relativistic physics

$$dl^2 = dx^2 + dy^2 + dy^2 \qquad d\tau = dt \qquad \text{(Newtonian)} \qquad (1.13)$$

In this approach, lengths and time durations would each be *invariant* quantities which could each be measured independently of the other.

The insight of special relativity is that time and space are in fact part of a 4D manifold with a single metric. In flat (uncurved) 4D spacetime, in Cartesian coordinates, that metric is

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 \qquad \text{(Minkowski metric)} \qquad (1.14)$$

We call ds a "spacetime distance" or "spacetime interval"[11]. Eq. 1.14

[11] The idea of the spacetime interval ds may seem unnatural. We experience x, y, and z as different dimensions of the same kind of spatial "stuff", so it seems natural to combine them into a single measure of "length". But time *feels* fundamentally different than space, so what exactly does it mean to combine them to form a single quantity ds? By analogy, it makes sense to mix together a red color and a blue one, or to mix together a musical C note with a G note, but to mix together a red color with a C note seems nonsensical!

This is a quite reasonable intuition that we must counter (and explore). Consider this: when we look out at the world, we take in a 2D image landing on our retinas from which our brain constructs a 3D rendering that visualizes objects with depth (using stereo vision and other visual cues). We take this rendering to be a single snapshot in time, but of course it is not – light has a finite travel time, so the objects further away come from longer ago than those nearby. If the speed of light weren't so fast, our brain would presumably have to account for this, and map out not just *where* objects were, but *when* they were. Perhaps then our entire visual field would be rendered in 4D with a time component stamped on everything, and the spacetime distance ds would seem like a natural measure for objects in this field.

Alas, the speed of light is so fast that any time delay between objects in our view is negligible. Our brain did not evolve to "see" spacetime, instead taking our entire visual field to be "now". We'll have to retrain our brain using mathematics.

is known as the Minkowski metric and is the simplest and most symmetric way to add time into the metric. The time coordinate is treated the same as all of the spatial coordinates *except* for a minus sign. While a mere minus sign may not seem to capture all of the experiential factors that make time *feel* different than space for us, it does have important physical implications. In particular, we will see that the sign difference for the time coordinate makes it natural to introduce a *casual structure* to the spacetime manifold.

1.4 Spacetime Distances in the Minkowski Metric

Unlike spatial lengths, the spacetime distance ds^2 is *not* necessarily a positive quantity. This introduces a new, interesting mathematical structure where we can make a distinction between different distances[12]

$$ds^2 < 0 \quad \text{timelike distance} \tag{1.15}$$
$$ds^2 > 0 \quad \text{spacelike distance} \tag{1.16}$$
$$ds^2 = 0 \quad \text{lightlike (or null) distance} \tag{1.17}$$

The naming convention is such that in a timelike interval, the time difference c^2dt is greater than the space difference $dx^2 + dy^2 + dz^2$, and vice versa for a spacelike interval.

To understand the physical meaning of these distinctions, consider an object moving on a wordline as shown in the spacetime diagram Figure 1.4. If at some point **O** the object emits light, these light rays move diagonally in the spacetime figure. The cone defined by the light rays is known as the *future light cone* at point **O**. Starting from **O**, any point inside the light cone can potentially be reached by traveling at a speed $v < c$, while to reach a point outside the cone would require faster than light travel.

The sign of the spacetime interval indicates just this distinction. A point **T** inside the light cone will have $dx < cdt$ and so the interval between **O** and **T** is a *timelike* ($ds^2 < 0$) interval. A point **S** outside the light cone has $dx > cdt$ and corresponds to a *spacelike* ($ds^2 > 0$) interval from **O**. Any point **L** along one of the diagonal light rays has $dx = cdt$ and represents a *lightlike* ($ds^2 = 0$) interval.

We make the following postulate: **Massive objects are required to travel on timelike paths** i.e., move such that $ds^2 < 0$ at all segments on the path. This is equivalent to requiring objects move at speeds $v < c$ when using ordinary Minkowski coordinates. We make this postulate so that we can preserve a meaningful notion of *causality* in our spacetime. We will see later that the time ordering of timelike events (e.g., event O occurs before event T) is invariant (the same in

[12] We use a metric convention where the minus sign is in front of the time coordinate, but other books use the opposite convention (or "signature") where the Minkowski metric is

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

In this signature, timelike distances correspond to $ds^2 > 0$.

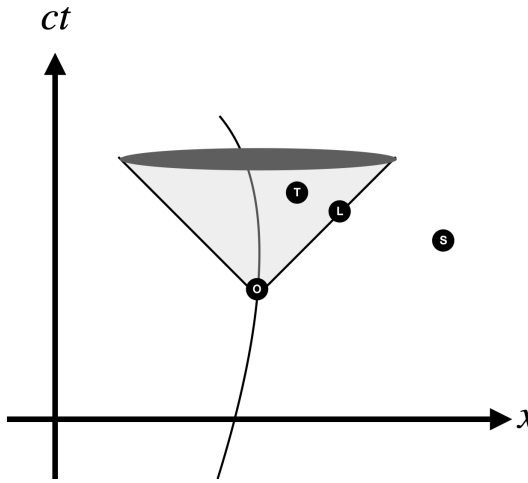


Figure 1.4: Illustration of spacetime intervals. Consider event **O** along the world line of an object (curved line). Rays of light emitted from **O** move along diagonals and define the future light cone. Points inside the light cone are timelike separated ($ds^2 < 0$) from **O**, while points outside the light cone are spacelike separated ($ds^2 > 0$) from **O**. Points along the cone itself are lightlike separated ($ds^2 = 0$).

all frames) and so a notion of causality (O caused T) makes sense. For events with spacelike separation, the time ordering the events is not invariant (event **O** may occur before **S** in one frame, put after **S** in another). Thus we assure that spacelike separated events are not casually related by postulating that no physical signal could have traveled them.

1.5 Proper Time of a Spacetime Path

The metric only tells us how to calculate infinitesimal distances. If we want to determine a finite distance, we must integrate the metric. For example consider the spacetime path shown in Figure 1.5. This is a timelike path where each segment of the path has $ds^2 < 0$ (i.e., at each point on the path the speed is $v < c$). IF we were to take a square root to get ds, the result would be imaginary, so instead we define a new variable, $d\tau^2$ that is opposite sign of ds^2

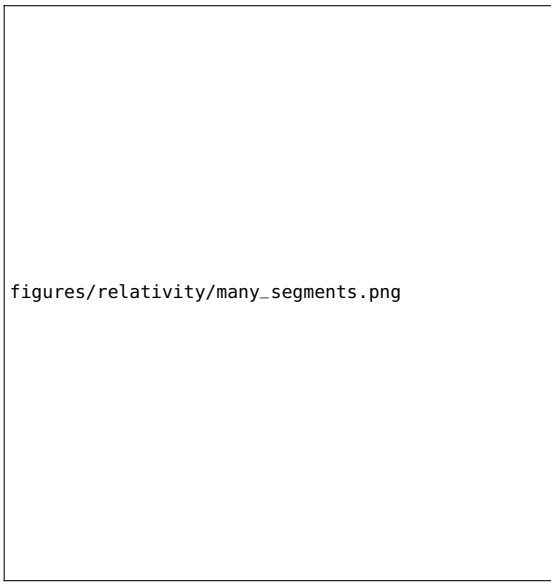


Figure 1.5:

$$d\tau^2 = -\frac{ds^2}{c^2} \qquad \text{(proper time interval)} \tag{1.18}$$

where the factor of $1/c^2$ is there to make τ have units of time, and τ is called the *proper time*. Since $d\tau^2$ is by definition positive for a timelike path, we can take a square root and integrate over the path

$$\tau = \int d\tau = \int \sqrt{-ds^2/c^2} \tag{1.19}$$

The value of τ gives the spacetime "length" of the entire path.

The proper time is an invariant quantity, with the same value no matter what coordinated system or frame you use to caclaulte it. To get a better sense of why it is called a "proper time", consider for example the Minkowski metric, where we have

$$d\tau^2 = -\frac{ds^2}{c^2} = dt^2 - \frac{dx^2}{c^2} \tag{1.20}$$

from which we have

$$d\tau = \sqrt{dt^2 - \frac{dx^2}{c^2}} = dt\sqrt{1 - \frac{dx^2}{dt^2}\frac{1}{c^2}} = dt\sqrt{1 - \frac{v^2}{c^2}} \tag{1.21}$$

where we used the coordinate velocity $v = dx/dt$. Integrating along the path gives[13]

$$\tau = \int d\tau = \int dt\sqrt{1 - \frac{v^2}{c^2}} \tag{1.22}$$

[13] Note that Eq. 1.22 applies only for the Minkowski metric, and will look different for other metrics of curved space time.

Now imagine your friend Alice moves along a path like the one shown in Figure 1.5. Since the path is timelike ($v < c$ at all points) we can always define a coordinate frame that moves along with the path, such that in this frame Alice does not move. We call such a

frame the "rest frame" of the path (it need not be an inertial frame). In this rest frame $dx' = 0$ (and $v'_x = 0$) for each segment of the path and so the integral Eq. 1.22 is trivial, giving $\tau = \Delta t'$, where $\Delta t'$ is the total time elapsed according to Alice's wristwatch. Since τ is an invariant spacetime interval, we find the same value if we calculate it in the "rest frame" or the original "lab frame". But we also now have the physical interpretation that the "proper time" τ of a path is the **time that will have elapsed on a clock that moves along that path**.

We can also consider the length of spacelike paths, even though no physical object can traverse such a path. Since $ds^2 > 0$ for spacelike intervals, we do not need to multiply by a minus sign. Such intervals are sometimes called "proper lengths".

1.6 *Lorentz Symmetry*

The Minkowski metric (like the Euclidean one) has rotational symmetry – we can perform a rotation to "mix up" any two spatial coordinates (e.g., a rotation about the z axis "mixes up" x and y, a rotation about the x axis "mixes up" y and z). We suspect there may be some analogous transformation that would "mix up" a space dimension and the time dimension. Given the minus sign in front of the dt^2 term in the Minkowski metric, this cannot be a rotation, but it will be something very analogous to a rotation, namely a *Lorentz transformation*. The content of the theory of special relativity is encoded in the Minkowski metric and its Lorentz symmetry.

Historically, the interconnection between space and time was explored by Galileo when he was trying to displace the earth from the center of the universe. Physicists of Galileo's era quite reasonably argued that we would notice if the earth were in motion around the sun. Galileo argued that, in fact, the laws of physics would remain unchanged (i.e., obey a symmetry) under a transformation to a coordinate system moving at speed v in the x direction[14]

$$x' = x - vt \qquad t' = t \qquad \text{(Galilean transformation)} \tag{1.23}$$

Writen in vector form, Galileo's transformation is

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} \tag{1.24}$$

where we defined $\beta = v/c$.

While Galileo had touched upon a key physical concept (that physics remains the same if we transform to a moving frame) he got the mathematical form of the transformation incorrect. The Galilean transformation Eq. 1.24 treats space and time distinctly, and so is not as mathematically "beautiful" as can be imagined. A more symmetric

[14] A person moving at constant velocity in the unprimed frame follows the line $x = vt$. The Gallilean transformation then gives their location in the unprimed frame as

$$x'(t) = x - vt = vt - vt = 0$$

and so they are stationary in the primed frame, as desired. So this serves as a transformation to a moving frame.

transformation would transform time and space in a similar fashion, that is

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \mathbf{\Lambda}(\beta) \begin{pmatrix} ct \\ x \end{pmatrix} \qquad \text{where } \mathbf{\Lambda}(\beta) = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \tag{1.25}$$

where γ is some overall scaling factor that we will see in a moment is necessary. This matrix equation corresponds to the set of equations

$$x' = \gamma\,(x - \beta ct) \qquad ct' = \gamma\,(ct - \beta x) \tag{1.26}$$

which is known as a *Lorentz transformation*.

Like an ordinary rotation, the Lorentz transformation mixes up two coordinates. However, instead of rotating both axes in the same direction, the Lorentz transformation "squeezes" (or "stretches") them in opposite directions. A consequence of this (visible in Fig 1.6) is that the diagonal remains unchanged – i.e., in both the primed and unprimed coordinate systems, the blue line passes through one unit of x in one unit of ct. Since the diagonal represents the trajectory of light, we see that the speed of light remains unchanged under a Lorentz transformation (which is the behavior that motivated the development of special relativity.)

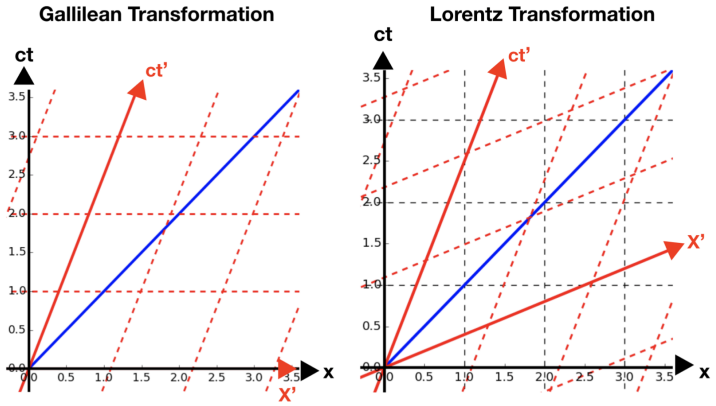


Figure 1.6: Illustration of spacetime transformations. A Galilean transformation treats space and time distinctly, shifting only the vertical axis. A Lorentz transformation treats space and time similarly, shifting them an equal way in opposite directions. The speed of light (shown as the blue line) remains the same under a Lorentz transformation (light covers one unit of space in one unit of time) but not under a Gallilean transformation.

We can show why the γ scaling factor is needed in the Lorentz transformation. If we transform to a frame moving at speed v, then transform in reverse by a speed $-v$, we should get back to the frame we started from. In other words, $\mathbf{\Lambda}(\beta)$ times $\mathbf{\Lambda}(-\beta)$ should equal the identity matrix

$$\mathbf{\Lambda}(\beta)\Lambda(-\beta) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{1.27}$$

Carrying out the matrix multiplication we find

$$\Lambda(\beta)\Lambda(-\beta) = \begin{pmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & \gamma \end{pmatrix} \begin{pmatrix} \gamma & \gamma\beta \\ \gamma\beta & \gamma \end{pmatrix} = \begin{pmatrix} \gamma^2(1-\beta^2) & 0 \\ 0 & \gamma^2(1-\beta^2) \end{pmatrix}$$

For this to equal the identity matrix we must have

$$\gamma = \frac{1}{\sqrt{1-\beta^2}} \tag{1.28}$$

which is the definition of the *Lorentz factor*, γ. The scaling factor γ effectively normalizes the Lorentz transformation so that the volume of the boxes in the spacetime diagram Fig 1.6 doesn't change as they get squeezed.

The connection between Lorentz transformations and rotations can be made more explicit by using hyperbolic functions [15] cosh and sinh. The range of $\cosh\eta$ is from 1 to infinity, which is the same range as the Lorentz factor γ. We therefore can make a change of variables by defining $\gamma = \cosh\eta$. It can then be shown that $\beta\gamma = \sinh\eta$ and $\beta = \tanh\eta$. The Lorentz transformation can then be written

$$\mathbf{\Lambda} = \begin{pmatrix} \cosh\eta & -\sinh\eta \\ -\sinh\eta & \cosh\eta \end{pmatrix} \tag{1.30}$$

which shows strong similarity to the matrix of an ordinary rotation

$$\mathbf{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{1.31}$$

Thus a Lorentz transformation can be understood to be a *hyperbolic rotation* by a hyperbolic angle, η.

1.7 *Implications of the Lorentz Transformation*

From the mathematical point of view, the Lorentz transformation is a natural construct and the obvious counterpoint to a rotation transformation. Nature is not obligated to follow the most beautiful mathematics, but in this case it seems to have. This structure, however, leads to several physical results that may be counter-intuitive.

1.7.1 *Relativity of Simultaneity:*

In the spacetime diagram of Figure 1.7, the lines of constant t are horizontal lines. All events along such a line happen *simultaneously* in the unprimed frame. However, we see that the lines of constant t' are not parallel to the lines of constant t. Events that are simultaneous in the primed frame are *not* simultaneous in the unprimed frame, and vice versa. There is no absolute notion of "at the same time" – simultaneity depends on your frame of reference.

While this may seem unintuitive, note that the equivalent feature of space is not unusual at all. If I am driving in a car and clapping my hands to the beat of a song, all of the claps happen at the same

[15] The hyperbolic functions are defined by

$$\cosh\eta = \frac{e^{\eta} + e^{-\eta}}{2}$$

$$\sinh\eta = \frac{e^{\eta} - e^{-\eta}}{2}$$

With these definitions we have the identity

$$e^{\eta} = \cosh\eta + \sinh\eta$$

which can be compared to the famous Euler formula relating cosine and sine to the complex exponential

$$e^{i\theta} = \cos\theta + i\sin\theta$$

It follows that

$$\cosh\eta = \cos(i\eta)$$

$$\sinh\eta = -i\sin(i\eta)$$

So the hyperbolic functions are analogues to the ordinary trig functions, that differ in whether the exponent is real or imaginary. For hyperbolic functions we find the identity

$$\cosh^2\eta - \sinh^2\eta = 1$$

which can be compared to the trigonmetric relation

$$\cos^2\theta + \sin^2\theta = 1 \tag{1.29}$$

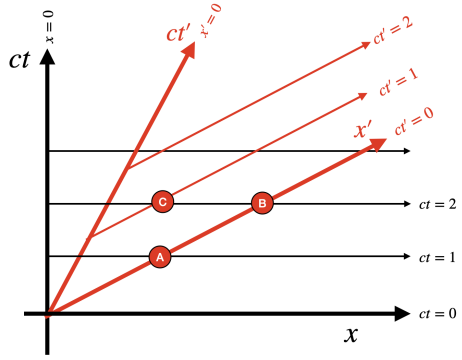


Figure 1.7: Under a Lorentz transformation, the lines of constant time get skewed and demonstrate relativity of simultaneity. In the figure, the two events A and B happen at two different times (ct = 1 and ct = 2) in the black (unprimed) frame, but occur at the same time (ct' = 0) in the red primed frame. Then events B and C occur at the same time (ct = 2) in the black unprimed frame, but at different times (ct' = 0 and ct' = 1) in the red primed frame.

point in space in the car frame. But my claps will happen at different points in space according to somebody standing beside the road watching the car drive by. The notion of *cospatiality* is obviously relative to the frame of reference. The relativity of simultaneity (or co-temporality) is the same idea applied to time.

1.7.2 *Time Dilation:*

Because the Lorentz transformation scales the coordinates by a factor of γ, intervals of time will be relative to the reference frame. Imagine I am driving in a car and clap twice, at times t'_1 and t'_2 second. In my reference frame, both events happen at the same place $x'_1 = x'_2 = 0$. For somebody watching the car go by on the street, the Lorentz transformation gives

$$\begin{aligned} ct_1 &= \gamma(ct'_1 + \beta x'_1) = \gamma ct'_1 \\ ct_2 &= \gamma(ct'_2 + \beta x'_2) = \gamma ct'_2 \end{aligned} \tag{1.32}$$

So the relationship between the time intervals is

$$(t_2 - t_1) = \gamma(t'_2 - t'_1) \tag{1.33}$$

or

$$\Delta t = \gamma \Delta \tau \tag{1.34}$$

where we use the label τ to describe the *proper time*, i.e., the time ticked off by a clock that is at rest in the frame (in this case at $x' = 0$) If the clock in the car ticks off a proper time interval of $\Delta\tau = 1$ second, somebody on the road will experience a longer time interval of $\Delta t = \gamma\Delta\tau$. We say that **moving clocks run slowly** by a factor of γ. This effect is called *time dilation*.

1.7.3 *Lorentz Contraction:*

Intervals of length are also relative to the reference frame. Say that in the frame of a car, the back of the car is at $x'_1 = 0$ and the front of the car at $x'_2 = L_0$. The ends of the car remain fixed in the car frame, and the car's length is $x'_2 - x'_1 = L_0$. But if somebody on the road watching the car drive by measures the position of each end of the car *at the same time in the road frame*, they find (taking the time of the measurements to be $t_1 = t_2 = 0$)

$$\begin{aligned} x'_1 &= \gamma(x_1 - \beta c t_1) = \gamma x_1 \\ x'_2 &= \gamma(x_2 - \beta c t_2) = \gamma x_2 \end{aligned} \tag{1.35}$$

from which we have

$$x'_2 - x'_1 = \gamma(x_2 - x_1) \tag{1.36}$$

The length measured in the road frame $L = x_2 - x_1$ is thus given by

$$L = L_0/\gamma \tag{1.37}$$

We call L_0 the *proper length*, as it is the length measured in a frame where the car is at rest. In the road frame where the car is moving, it's length is shorter by a factor $1/\gamma$. This is known as *Lorentz contraction*. It only applies along the direction of motion.

1.8 The Invariant Spacetime Interval

Consider two pointsin space, A and B. We have discussed how coordinate distances between A and , such as Δx and Δy, are relative quantities, as they depend upon the orientation of our coordinate system. This is illustrated on the left panel of Figure 1.8, which shows that as you perform a rotation, a point B is moved along a circle. The values of Δx and Δy will depend on exactly which rotated coordinate system, we choose to use. However, the distance between A and B given by the Euclidean metric, $\Delta l^2 = \Delta x^2 + \Delta y^2$, is invariant under rotations.

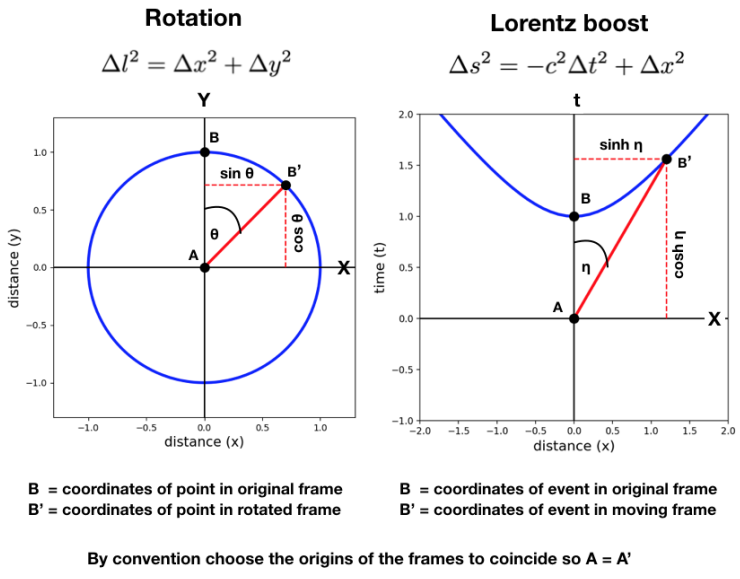


Figure 1.8: Comparison of ordinary rotations, and Lorentz transformations. The latter can be considered to be *hyperbolic rotations*

Similarly we have seen that coordinate distances like Δt and Δx are relative to our coordinates (or reference frame). However, the *spacetime interval* defined by the Minkowski metric

$$\Delta s^2 = -c^2\Delta t^2 + \Delta x^2 \tag{1.38}$$

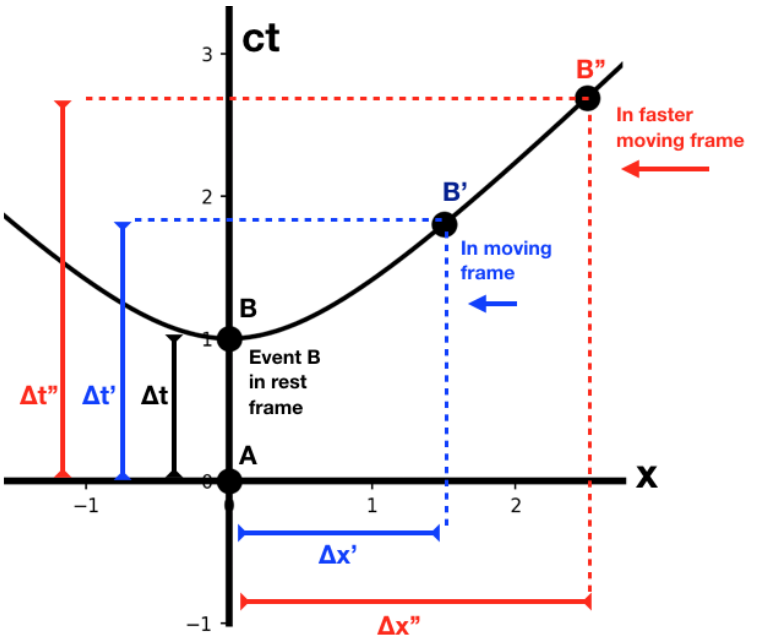
can be shown to be invariant. That is, if we apply a Lorentz transfor-

mation we will find in the primed frame that

$$-c^2\Delta t'^2 + \Delta x'^2 = -c^2\Delta t^2 + \Delta x^2 \tag{1.39}$$

and so $\Delta s'^2 = \Delta s^2$. Unlike Δt or Δx, the spacetime interval Δs has the same value in every inertial frame[16].

While the equation for Euclidan metric $\Delta l^2 = \Delta x^2 + \Delta y^2$ was that of a circle, the equation for the spacetime interval Δs^2 is that of a *hyperbola*. Figure 1.9 illustrates the effect of a Lorentz transform[17]. While Δt and Δx differ depending on the frame of reference, the value $\Delta s^2 = c^2\Delta t^2 - \Delta x^2$ remains fixed; thus a Lorentz transformation can only move events in spacetime along a hyperbola. This can be compared with ordinary rotations, which move points along a circle (Figure 1.8).



It can be hard to retrain our brain to visualize distances in a Minkowski metric. When we look at Figure 1.9, our brain immediately concludes that point **B'** is further from point **A** than is point **B**. This would be true if we were using a Euclidean metric, but **under the Minkowski metric, all points along a hyperbola have the same spacetime interval from the origin**. Figure 1.10 illustrates this metrical structure by showing lines of constant spacetime distance from the origin. Timelike distances correspond to vertically oriented hyperbolas while spacelike distances correspond to horizontally oriented hyperbolas. Lightlike (or null) distances are diagonals lines. The simple the minus sign in front of c^2dt^2 in the Minkowski metric has given us a more interesting metrically structure to spacetime that

time interval, take diferentials of the Lorentz transformation to get

$$dx' = \gamma(dx - \beta c\, dt)$$

$$cdt' = \gamma(dt - \beta dx)$$

Squaring these we have

$$(dx')^2 = \gamma^2(dx^2 + \beta^2 c^2\, dt^2 - 2\beta dx\, cdt$$

$$(c\, dt')^2 = \gamma^2(c^2dt^2 + \beta^2 dx^2 - 2\beta dx\, cdt$$

Adding these two squares together won't get us anywhere, but if we subtract them then the 3rd term goes cancels and we find

$$-(c\, dt')^2 + (dx')^2 = -c^2dt^2\gamma^2(1-\beta^2) + dx^2\gamma^2(1-\beta^2)$$

Then using $\gamma^2 = 1/(1-\beta^2)$ we have

$$-(c\, dt')^2 + (dx')^2 = -c^2dt^2 + dx^2$$

So the spacetime interval is the same in both frames.

Figure 1.9: Spacetime diagram illustrating the invariant spacetime interval. Event B occurs at $(t_B = 1, x_B = 0)$ in the unprimed frame. In frames moving with respect to the unprimed frame, event B happens at a different time and position, but remains on the hyperbola defined by $c^2\Delta t^2 - \Delta x^2$. We can see the effects of time dilation (i.e., $\Delta t' > \Delta t$).

[17] In previous spacetime diagrams we demonstrated Lorentz transformation by redrawing the axes but keeping events fixed. In Figure 1.9 we keep the axes fixed but redraw events. These are just two different ways of illustrating the same transformation.

takes time to internalize[18]. But it will pay off in terms of the power and mathematical beauty of the construction.

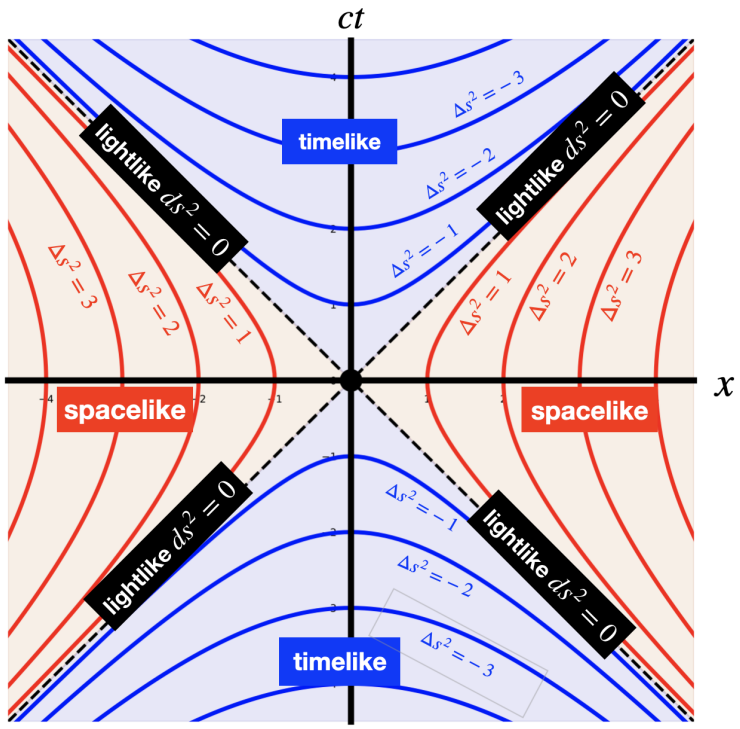


Figure 1.10: Spacetime diagram showing the regions where the spacetime interval (measured here from the origin) is either timelike ($\Delta s^2 < 0$), spacelike ($\Delta s^2 > 0$) or lightlike ($\Delta s^2 = 0$) . All points along a hyperbola have the same spacetime distance from the origin. All points along the diagonal have a null distance $\Delta s^2 = 0$ from the origin.

1.9 *Causal Structure of Spacetime*

The Minkowski metric captures essentially all of the content of special relativity. But why did Nature choose this seemingly unintuitive metric? The common explanation given is: "because this metric assures that the speed of light is constant in all inertial reference frames". This reason is not very satisfying, as it merely pushes back the question – why then is the speed of light constant?

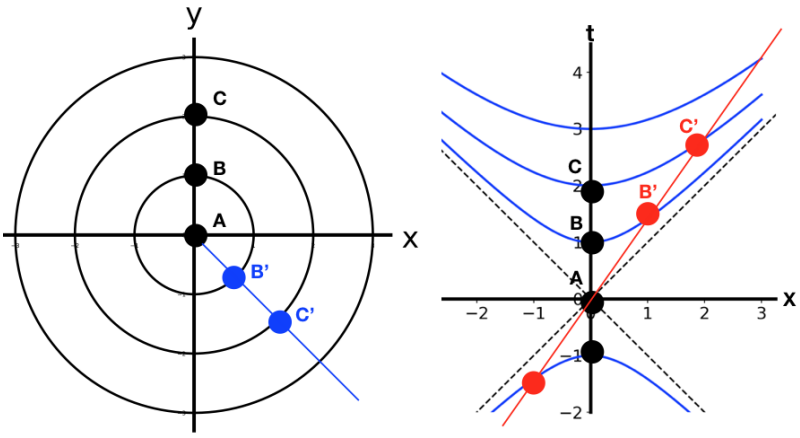
Perhaps a better motivation for the spacetime metric is that it seems readymade to define a *casual structure* to spacetime. Causality is an important concept in physics – we believe that some event A can cause another event B. For causality to make sense, the cause must always *precede* the effect. Moreover, this time ordering must hold in every reference frame. If we could find some reference frame in which the effect B came before the cause A we would be forced to either abandon the notion of causality, or abandon the principle of relativity which holds that all inertial reference frames are valid[19]

Thus for causality to make sense we need the time ordering of causal events to be *absolute*. The familiar Euclidean metric does not

[18] Recall the Mercator map of the earth shown in Figure 1.1 where Antarctica looks huge to our eye. To properly interpret the map, we need to keep in mind how the metric behaves. Similarly, when looking at spacetime diagrams we need to keep in mind the metrical structure to convert what is displayed to what are actual physical distances.

[19] Abandoning the principle of relativity would be equivalent to abandoing a certain symmetry of spacetime, as the laws of physics would no longer remain unchanged under a boost transformation. Since we like symmetry, this would be unappealing.

permit such an absolute ordering. As Figure 1.11 shows, the ordering of a sequence of points in space changes under ordinary circular rotations. For spatial dimensions there is no absolute notion of "point A is in front of B", since we can always rotate to some coordinate system where B is in front of A.



Euclidean metric does not preserve the ordering of points in a spatial coordinates (e.g., the y-coordinate) under rotations.

Lorentzian metric preserves the ordering of *timelike* events in time under any Lorentz transformation (i.e., hyperbolic rotation)

Figure 1.11: In space described by a Euclidean metric $dl^2 = dx^2 + dy^2$ there is no absolute sense of ordering. While point B may be in front of point A in one frame, we can always rotate or coordinate system such that A is in front of B. In spacetime described by the Minskowski metric, $ds^2 = -c^2dt^2 + dx^2$ there is a sence of absolute time ordering. For points that are timelike separated, transformation

If we were to add the time dimension to our metric in the same way that we would add a space dimension (i.e., with a plus instead of a minus sign in front of dt) we could have no absolute time ordering and hence no causality. But Figure 1.11 shows that for a Lorentz transformation (i.e., a hyperbolic rotation) events with timelike separation do indeed have an absolute time ordering. While the time interval between event A and B will be relative to a particular coordinate system frame, event A will precede event B in *every* frame.

Absolute time ordering only holds for events with a timelike separation (i.e., with $\Delta s^2 < 0$). For two events with a spacelike separation, the time ordering is indeed relative to the reference frame[20]. However, to get from an event A to an event B with a spacelike separation, one would have to travel faster than the speed of light. If we add a postulate that *nothing can travel faster than the speed of light* then causality makes sense, since there is no mechanism by which spacelike events could ever cause each other[21].

Thus, the Minkowski metric provides the simplest and most symmetric way to incorporate space and time together into a 4D spacetime that permits a sense of causality. One consequence of such a metric is that spacetime intervals ds^2 can be either negative or positive with a special critical case $ds^2 = 0$ in between. Since ds^2 is an

[20] Look at the spacelike hyperbolas of Figure !1.10 and convince yourself that an event one of these hyberbolas could happen before or after the event at the origin.

[21] For example, imagine your friend in L.A. and you in Berkeley both woke up at 7 AM on the dot (in some reference frame). There is no way to claim that your waking up *caused* your friend to wake up, since there is no way you could have communicated to L.A. in time to wake her (even a phone call only travels at most at speed c). In other frames, one of these events can occur before the other, but this relativity of time ordering for spacelike events is of no concern as there is never a causal connection between them.

invariant the critical case $ds^2 = 0$ implies that $dx/dt = c$ in every inertial frame. Instead of thinking of the constancy of the speed of light as the *reason* that spacetime is the way it is, we can think of it as the *mathematically consequence* of choosing a metric that allows for causality[22].

In fact, instead of the "speed of light" it would be better to call c the "critical speed". It just so happens that light travels at this speed, but other massless particles would also move at c. The fact that c takes on some seemingly arbitrary value ($c \approx 2.99 \times 10^8$ m s^{-1}) is because we typically use an arbitrary unit of length (i.e., somebody just decided to call some rod of a certain length "one meter"). It would be more natural to use units where light moves one unit of space per one unit of time, e.g., to measure lengths in light-seconds. Then the critical speed of spacetime is $c = 1$.

[22] An alternative, of course, would to consider time as a completely separate dimension than space. This is the approach of Newtonian physics, for which time (and hence time ordering) is absolute. However there is something more symmetric and beautiful about combining space and time together into a 4D spacetime.

2
Black Holes

2.1 The Equivalence Principle

Einstein said his "happiest thought" was the realization thats someone freely falling in gravity appears to be an *inertial reference frame*. An inertial frame is one in which Newton's first law holds, i.e., "an undisturbed body stays at rest or moves at a constant velocity". If you are in a elevator freely falling towards earth and you let go of your phone, it will float stationary in front of you, exactly as it would if you were sitting at rest far out in empty space.

The converse is also true – an elevator accelerating through empty space (i.e., in a region free from gravity) is not an inertial frame. You feel an effective "g-force" in an accelerating frame, and if you let go of your phone it will fall as if something is pulling it down. In Newtonian physics, such forces are often called "fictitious", since they are a consequence of the acceleration of the frame, and not any actual physical force being applied by some external object. How can you tell if a force was fictitious or not? An important clue is that a fictitious force should accelerate all objects *at the same rate*, regardless of their mass, shape, charge, etc...

One force of nature – gravity – appears to behave like a fictitious force. Using Newton's law of gravity, and the second law, $F = m_I a$, the acceleration of a mass by another mass M is

$$m_I a = \frac{GMm_G}{r^2} \implies a = \frac{GM}{r^2}\frac{m_G}{m_I} \tag{2.1}$$

Here we have made a conceptual distinction between the *inertial mass*, m_I (which describes how resistant an object is to acceleration) and the *gravitational charge*, m_G (which describes how strongly an object feels a gravitational force)[1]. . According to all measurements done to date, these quantities are equal

$$m_I = m_G \tag{2.2}$$

and so both m_I and m_G are usually just called the "mass". This equal-

[1] The name "gravitational charge" draws an analogy with the electric force, where the charge q describes how strongly an object couples to the electric field. The charge q is not necessarily proportional to the inertial mass m_I so all objects do *not* accelerate the same way due to an electric force. Gravity appears special in this regard.

ity is one form of the *equivalence principle*. As a result, gravity accelerates all objects at the same rate, $a = GM/r^2$, resembling a fictitious force. Einstein felt that this was too remarkable to be a coincidence.

These insights led to the General theory of Relativity. Einstein's philosophical instinct was that all frames of references should be equally valid in describing physics. Because his theory of Special Relativity only applied to inertial frames, he set out to generalize those ideas to accelerating frames and in doing so realized this connection to gravity. The theory of General Relativity thus developed it a completely new theory of gravity. Einstein proposed that a massive object like the Earth warps the spacetime metric around it. An object that follows a "natural" path – or *geodesic* – in the curved spacetime[2] would find itself to be locally in an inertial frame. In contrast, an object that deviated from any geodesic would be in a non-inertial frame and so notice apparent forces. An inertial frame was no longer one that "moved at constant velocity" but rather one that "moved along a geodesic in spacetime". An elevator plummeting in freefall to the earth appears to be accelerating from the ground, but in fact that elevator is just following its natural geodesic and is an inertial frame. A person standing still on the earth, on the other hand, is *not* on a geodesic, and so is effectively "accelerating" relative the natural paths of the curved space time. As a result of being in this non-inertial frame, the person feels an effective force that we call "gravity".

[2] We will define what is meant by a "natural path", but it is the generalization of a straight line to a curved spacetime.

2.2 *Curved Spacetime*

Though we will not give a complete description of curved spacetime (which requires the full machinery of tensors and differential geometry) we can provide a bit of quantitative discussion. It is perhaps easiest to visualize curvature by "embedding" a space in a higher dimension. For example, imagine a 1D space (e.g., ants walking along a string). We can draw this 1D linear space on the 2D plane as shown in Figure 2.1. The ants, of course, are not aware of a second dimension, they can only move backwards and forwards along the string. We call placing the 1D space in the 2D space an "embedding".

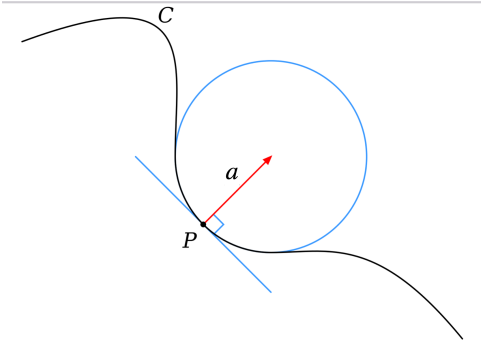


Figure 2.1:

The 1D string shown in Figure 2.1 appears curved and the curvature varies from point to point. We can quantify the degree of curvature in the following way. At some point P, draw a circle that follows the bending (i.e., second derivative) of the string at that point. We define the linear curvature, κ_L, at point P to be $\kappa_L = 1/a$, where a is the radius of that circle. A circle of smaller radius a implies a more sharply curving line, which is why we define curvature as the *inverse* of the radius.

Consider now the 2D space shown in Figure 2.2. The 2D surface

has been embedded in 3D space in which we can see it curving. Drawing a single circle will not fully characterize the curvature; instead we must fit two circles at any point P: one along the direction of maximum curvature and one along the direction of minimum curvature. Doing so gives two numbers, $\kappa_1 = 1/a_1$ and $\kappa_2 = 1/a_2$, which are the minimum and maximum curvatures corresponding to circles of radius a_1 and a_2. From these we define the Gaussian curvature, K_G, as

$$K_G = \kappa_1 \kappa_2 = \frac{1}{a_1}\frac{1}{a_2} \tag{2.3}$$

The sign of κ_1 and κ_2 are set by the direction in which the circles bend, as seen in Figure 2.3. If the two circles bend in the same direction (as on the surface of a sphere) the curvature is positive. If the two circles bend in opposite directions (as on the surface of a saddle) the curvature is negative. If one of the curvatures is zero (as on a cylinder) the Gaussian curvature is zero.

These embedding diagrams illustrate curvature by showing the space "bending" in the higher dimension. But could ants confined to the space itself determine that the space has curvature? For ants living on a 2D surface, the answer is "Yes". In a curved space the familiar rules of Euclidean geometry do not hold, e.g., the circumference of a circle of radius R is $2\pi R$ and the angles of a triangle add up to 180^o. Local measurements can thus determine the curvature[3].

To see how this can work, imagine the 2D surface of a sphere of radius a, for which the Gaussian curvate is $K_G = 1/a^2$ at all points. We embed this surface in 3D flat (i.e., uncurved) space where the regular Euclidean metric is, using spherical-polar coordinates

$$dl^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta^2 d\phi^2 \tag{2.4}$$

For ants leaving on the 2D surface of the sphere, the radius is fixed at $r = a$ and $dr = 0$ since we do not allow displacements in the radial direction. The 3D metric above then reduces to the 2D metric that the ants would use for calculating distances on the sphere

$$dl^2 = a^2 d\theta^2 + a^2 \sin^2\theta^2 d\phi^2 \tag{2.5}$$

Imagine on ant that seeks to measure the circumference of a circle in its 2D world. The ant starts at the North Pole ($\theta = 0$) and lays out a measuring tape due south to a location θ_0. Along this path $d\phi = 0$ (since the ϕ coordinate does not change) and so the metric Eq. 2.5 becomes $dl^2 = a^2 d\theta^2$. To determine the length, R, of the measuring tape path we integrate $dl = a d\theta$ over the path

$$R = \int dl = \int_0^{\theta_0} a d\theta = a\theta_0 \tag{2.6}$$

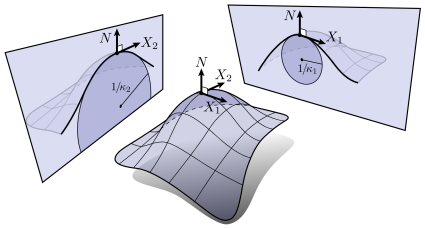


Figure 2.2: Illustration of the curvature of a 2D surface. Taken from this website .

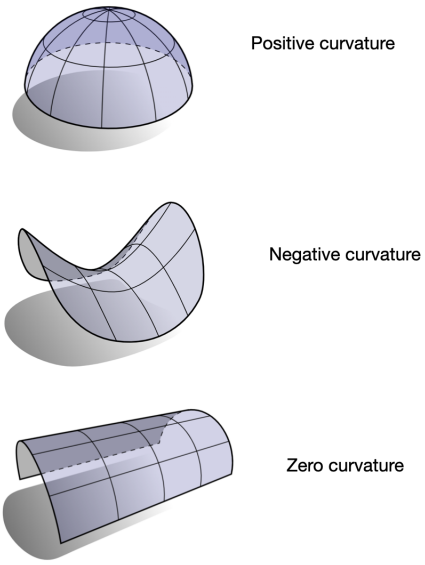


Figure 2.3:

[3] For the case of ants living in a 1D space, it turns out there is no way they could determine that the space is curved. There is no measurement an ant confined to the string could make to determine that his string was curving and not straight. We say that such a curvature is not "instrinsic" to the 1D space, but "extrinsic" to it (i.e., relates to how it is embedded in the higher dimensional 2D space.

Next the ant keeps θ_0 fixed and lays a measuring tape all the way around the ϕ direction (a line of latitude) to make a complete circle. Since $d\theta = 0$ on this path, the metric Eq. 2.5 becomes $dl^2 = a^2 \sin^2\theta_0 d\phi^2$. The length of this path is

$$C = \int dl = \int_0^{2\pi} a \sin\theta_0 d\phi = 2\pi a \sin\theta_0 \tag{2.7}$$

Since Eq. 2.6 gives $\theta_0 = R/a$ we have

$$C = 2\pi a \sin\left(\frac{R}{a}\right) \tag{2.8}$$

We see that $C \neq 2\pi R$ as is the case in Euclidean geometry (i.e., flat space). If the size of the circle is small relative to the radius of the sphere, $R \ll a$, we can do a Taylor expansion of $\sin(R/a)$

$$C = 2\pi a \left[\frac{R}{a} - \frac{1}{3!}\frac{R^3}{a^3} + ...\right] \tag{2.9}$$

where the ... indicate terms of higher order in R/a in the Taylor expansion. We rewrite this as

$$C = 2\pi R \left[1 - \frac{1}{3}\frac{R^2}{a^3} + ...\right] \tag{2.10}$$

Thus, in this curved space, the circumference of a circle is *smaller* than $2\pi R$ by the factor in brackets. In the limit that the sphere is very big $a \to \infty$, the term in brackets goes to 1 and we approach the flat space geometry.

Having measured R and C, our surveying ant can now go ahead and quantify the curvature of the space it lives in. Rearranging Eq. 2.10 to solve for $1/a^2$, which is the Gaussian curvature of a sphere, gives

$$K_G = \lim_{R\to 0}\left[\frac{3}{\pi r^3}(2\pi R - C)\right] \tag{2.11}$$

where we write limit as $R \to 0$ since we have used a Taylor expansion approximation keeping only the leading term, which is only valid for very small circles. While we derived this relation for a sphere, it turns out that it holds for any 2D space. For a space of negative curvature ($K_G < 0$) C is *larger* than $2\pi R$. For a general 2D space, like that of Figure 2.4, we can imagine our ant walking around the space, drawing tiny circles at each point and quantifying the Gaussian curvature, which varies from place to place.

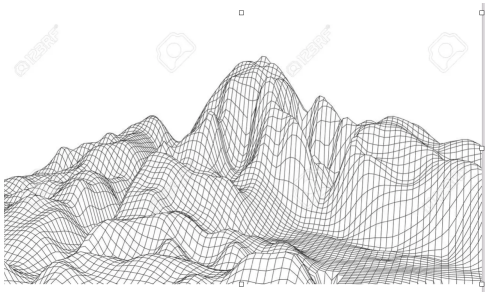


Figure 2.4:

While we often visualize the curved 2D surface by bending it within a 3D space, the metric Eq. 2.5 uses only 2 coordinates (θ, ϕ) and makes no reference to a 3rd dimension. Such a curvature is *intrinsic* to the space; i.e., it is encoded within the metric rule for

measuring distances. In contrast, *extrinsic* curvature refers to the bending of a space in an extra dimension which does *not* modify the geometry within the space itself. If we roll a flat piece of paper into a cylinder, we do not affect the intrinsic geometry of triangles or circles written on the paper (which remain that of Euclidean space); the inhabitants of this 2D surface would not see any difference (locally) between a flat plane and a cylinder[4]. However, if we try to bend the paper into a sphere, there is be no way to do so without stretching and distorting the shapes drawn on it. This would introduce intrinsic curvature, which would be reflected in the metric.

For 2D surfaces, the Gaussian curvature κ_G at all points fully describes the geometry of spacetime. But in higher dimensions, more than one number is needed to quantify the curvature at each point. For example, in 3D space you could at any point draw circles around each of the x, y, and z axis, and may find different values of $C/2\pi R$ for each one. In higher dimensions the curvature is described by the *Riemann Tensor* which in our 4D spacetime, has 20 independent components. However, if we restrict ourselves to simple spacetimes, like a perfectly homogenous and isotropic universe, then the imposed symmetry reduces the Riemann tensor to only one free parameter, which is essentially the Gaussian curvature.

[4] There are global measurements the ant could make, since it could walk around the cylinder and notice that it wound up back where it started. This feature, however, refers to the *topology* of the surface (i.e., how it is connected) rather than it's curvature.

2.3 *The Schwarzchild Metric*

The Einstein equations of General Relativity allow one to calculate the metric of spacetime given some distribution of mass/energy. We will not attempt to solve these equations, and instead will merely quote the resulting metric for certain cases, then explore the astrophysical consequences.

One of the most important solutions in GR is the *Schwarzchild metric* which describes the spacetime around a spherical mass M. Written using "Schwarzchild coordinates" (t, r, θ, ϕ) the metric is

$$ds^2 = -\left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 + \frac{dr^2}{(1 - 2GM/c^2 r)} + r^2 d\theta^2 + r^2 \sin\theta^2 d\phi^2$$

We notice a combination of parameters, $2GM/c^2$ which has dimensions of length. This is called the *Schwarzchild radius*

$$r_s = \frac{2GM}{c^2} \approx 3 \text{ km} \left(\frac{M}{M_\odot}\right) \tag{2.12}$$

The Schwarzchild radius turns out to be the *event horizon* of a (non-spinning) black hole[5] . Writing the metric in terms of r_s

$$ds^2 = -\left(1 - \frac{r_s}{r}\right) c^2 dt^2 + \frac{dr^2}{(1 - r_s/r)} + r^2 d\Omega^2 \tag{2.13}$$

[5] The radius $r_s = 2GM/r^2$ actually appeared as long ago as 1783 in Newtonian physics. A particle in Newton's physics will escape the gravitational pull of a spherical mass M if its total energy (kinetic plus gravitational potential) is positive

$$E = \frac{1}{2}mv^2 - \frac{GMm}{r} > 0$$

Considering a photon moving at speed $v = c$ and solving for r, the photon can escape only if

$$r > \frac{2GM}{c^2} = r_s$$

So in Newtonian gravity if you compress a star of M to a radius $r < r_s$, light will not be able to escape from its surface, rendering it a "dark star". This derivation is of course completely incorrect as it ignores relativity, which is essential at speeds near c. The fact that it gives the same answer as GR is coincidental (or rather the factor of 2 is coincidental; the GM/c^2 part has to appear by dimensional analysis as it is the only way to combine those quantities to get units of length).

where we used the shorthand $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$. The Schwarzchild metric only applies to the region *outside* the mass M. The spacetime outside the Sun, for example, is described by Eq. 2.13 but inside the sun ($r < R_\odot$) the metric differs. The Schwarzchild radius for the sun $r_s < R_\odot$, so the Sun has no event horizon. Black holes appear when masses are compressed to radii smaller than r_s.

Far away from the black hole ($r \gg r_s$), the Schwarzchild metric Eq. 2.13 becomes

$$ds^2 = -c^2dt^2 + dr^2 + r^2d\Omega^2 \qquad (r \gg r_s) \tag{2.14}$$

which is just the Minkowski metric of flat spacetime. So spacetime becomes flat far away from the black hole, as we would expect. It is the factors of $(1 - r_s/r)$ in Eq. 2.13 that indicate curvature of the metric close to the massive object.

2.3.1 *Measuring Lengths in the Schwarzchild Metric*

The coordinates used in the Schwarzchild metric Eq. 2.13 are called *Schwarzchild coordinates*, or sometimes "far-away" or "bookkeeper" coordinates, because they are a natural coordinate system for an observer situated far away from the black hole. The coordinates θ and ϕ are equivalent to the angles we use in ordinary spherical coordinates. The coordinate r labels concentric spheres in the radial direction. It is tempting to think of the coordinate r as telling us the distance from the black hole center, but this is not quite correct as we shall see.

If we consider a slice of fixed time (i.e., $dt = 0$) then the spacetime interval is a measure of *proper distance*, and we can use the notation $dl = \sqrt{ds^2}$, where l is a proper spatial distance. If we lay out a measuring tape along the radial direction and measuring proper radial distance, then $dt = d\theta = d\phi = 0$ and the metric becomes

$$dl = \sqrt{ds} = \frac{dr}{(1 - r_s/r)^{1/2}} \tag{2.15}$$

The proper length dl is is a factor of $(1 - r_s/r)^{-1/2} > 1$ larger than the coordinate interval dr. For $r \gg r_s$ we can use a binomial expansion to write this as

$$dl = dr\left[1 + \frac{1}{2}\frac{r_s}{r} + ...\right] \tag{2.16}$$

where the ... represents higher order terms in r/r_s. Integrating to find the proper length between two coordinates r_1 and r_2 we find

$$l = \int_{r_1}^{r_2} dl = \Delta r + \frac{r_s}{2}\log\left(\frac{r_2}{r_1}\right) + ... \tag{2.17}$$

where $\Delta r = (r_2 - r_1)$. We see that the proper distance is greater than Δr by an amount that will vary depending on how close we are to

r_s. The coordinate labels r are not equally spaced, rather the proper distance between labels gets larger as we approach the Schwarzchild radius.

If instead we hold the radial coordinate fixed ($dr = 0$) and lay a measuring tape around the equator ($\theta = \pi/2$) we would find a proper length

$$\int dl = \int_0^{2\pi} r d\phi \implies l = 2\pi r \tag{2.18}$$

And so the circumference of a circle is $C = 2\pi r$. This is actually how r is *defined* in Schwarzchild coordinates– it is not defined as the distance from the center, but rather by measuring the circumference at fixed r and dividing by 2π. The coordinate $r = C/2\pi$ is thus sometimes called a *circumferential* coordinate.

2.3.2 *Gravitational Time Dilation of the Schwarzchild Metric*

If we imagine a clock held at a fixed position (i.e., $dr = d\theta = d\phi = 0$) the spacetime interval becomes a measure of *proper time*, and we can write $c^2 d\tau = \sqrt{-ds^2}$ where τ is the proper time. In this case the metric becomes

$$d\tau = \sqrt{-ds^2/c^2} = \left(1 - \frac{r_s}{r}\right)^{1/2} dt \tag{2.19}$$

If we want we can integrate this over some finite interval, $\Delta t = t_2 - t_1$

$$\Delta\tau = \left(1 - \frac{r_s}{r}\right)^{1/2} \Delta t \tag{2.20}$$

We see that an interval of proper time $\Delta\tau$ is smaller than an interval of coordinate (or "faraway") time Δt. Thus a clocke held fixed at coordinate r around a mass M will run slowly compared to a clock very far away. This effect is known as *gravitational time dilation*.

An effect closely related to gravitational time dilation is *gravitational redshift*. If light is emitted with a wavelength λ_s within a gravitational field, it will be observed far away from the gravitational field to have a longer (i.e., redder) wavelength λ_o. This is because the period of the light wave $\Delta T = \lambda/c$ is subject to gravitational time dilation, such that Eq. 2.20 implies

$$\lambda_s = \left(1 - \frac{r_s}{r}\right)^{1/2} \lambda_o \tag{2.21}$$

and so $\lambda_o > \lambda_s$. As an object approaches the event horizon at r_s, the gravitational time dilation and redshift approach infinity.

2.3.3 *The Event Horizon and the Singularity*

The Schwarzchild Metric Eq. 2.13 appears to a have a *singularity* (i.e., terms in the metric that go to infinity) at coordinate $r = r_s$. The turns

out to be a feature of the particular coordinates being used, and can be eliminated by using a different set of coordinates. The curvature at r_s remains finite and an observer moving through r_s would not notice anything particularly unusual about that point in space. Such a point is called a *coordinate singularity.*

Eq. 2.13 has another singularity at $r = 0$. Unlike the one at $r = r_s$, this is a true singularity that cannot be gotten rid of by transforming coordinates. The spacetime curvature becomes infinite at $r = 0$ and we can no longer use the metric to derive sensible (i.e., finite) results for what happens when an object hits $r = 0$. It is believed that before hitting $r = 0$, quantum mechanical effects (which are neglected in GR) come into play and these may tame the singularity. However, no successful theory of quantum gravity has yet been developed.

The boundary $r = r_s$, while not a physical singularity, is important as the *event horizon* of a black hole. Matter and light can go inside the event horizon, but can never come out from it. Notice that for $r < r_s$, the terms in front of the dt^2 and dr^2 in Eq. 2.13 flip signs. We can rewrite the Schwarzchild metric inside r_s as

$$ds^2 = -\frac{dr^2}{(r_s/r - 1)} + \left(\frac{r_s}{r} - 1\right) c^2 dt^2 + r^2 d\Omega^2 \qquad \text{(inside) } r < r_s \quad (2.22)$$

Since the minus sign in front of a metric term indicates a timelike coordinate, inside r_s the r coordinate becomes a timelike coordinate and the t coordinate becomes spacelike.

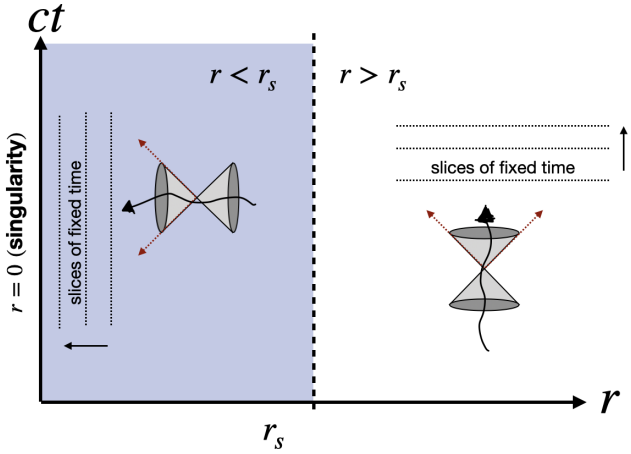


Figure 2.5: Material objects move on timelike path, which are those that within the future light cone bounded by the paths of light (red dashed lines). Outside of the event horizon ($r > r_s$), the timelike paths move forward in t. Inside the event horizon ($r < r_s$), the r coordinate becomes timelike and the light cone "flips on its side" and the timelike paths move inward in r.

Recall that material objects are required to move on *timelike* paths ($ds^2 < 0$). In the Minkowski metric, this means that an object stay within its future light cone bounded by the paths of light along the 45^0 lines. Inside the event horizon of a black hole, however, the r and t coordinates switch roles and so the light cone "flips onto its side" (see Figure 2.5. Since r has become a timelike coordinate, we can no-

longer avoid moving inwards in the r direction any more than we could avoid moving forward in t outside the event horizon.

The inevitable inward motion of the timelike paths for $r < r_s$ carries all objects to the singularity at $r = 0$. Because r is a timelike coordinate inside the event horizon, $r = 0$ represents not a point in space, but a *moment in time*. In flat space, we take it for granted that time continues on forever; but inside the event horizon of a black hole no matter how we move around in t, we inevitably encounter the "end of time" at $r = 0$.

While the "flipping of space and time" inside the event horizon of a black hole seems odd, it is really just an artifact of the particular Schwarzchild coordinates being used, and can be avoided by using other coordinates. As an analogy, consider azimuthal projection maps of the earth (Figure 2.6, which use separate maps for the Northern and Southern hemispheres. Imagine moving radially outward from the North Pole, at the center of the Northern map. When we reach the equator, the mapping breaks down, and we must "jump" to the Southern map. Of course, nothing strange *actually* happens at the equator; the breakdown of the maps there is just a feature of the particular coordinates being used, and is avoided in other maps.

Once we cross the equator, we would see that the sense our radial coordinate has "flipped". While previously moving South was radially outward, now moving south is radially *inward*. Obviously, the flip is a consequence of our particular coordinate mapping.

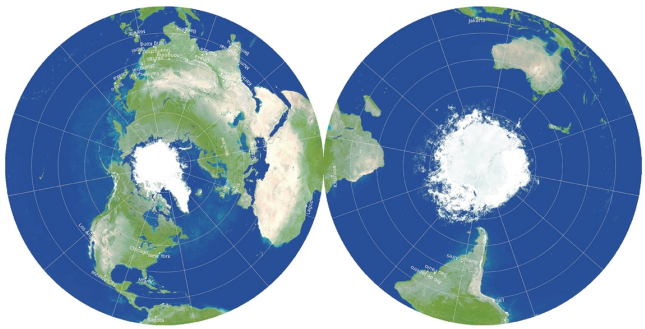


Figure 2.6: Azimuthal hemisphere maps, which split the globe into the Northern hemisphere map. (left) and Southern hemisphere map (right).

The mapping of Schwarzchild coordinates has analogous peculiarities, where outside of r_s we use r to label space but inside we use it to label time. There are other coordinate systems, though, that do not exhibit such a switch.

2.4 Geodesics and Motion in Curved Spacetime

General Relativity postulates that objects that move freely (i.e., without any forces put on them) follow *geodesics*. The geodesics are the generalization of a straight line to curved spacetime. A familiar example is the "great circle routes" on the surface of the earth. When drawn on a flat 2D map, such routes do not look like straight lines, but they in fact are (when taking the metric into account) the shortest distance between the points. We will explore a couple of ways of inferring the geodesics for a given metric.

2.4.1 Gravitational Lensing - Huygen's Principle

Light moves on null geodesics with $ds^2 = 0$. Consider a light ray moving radially towards or awary from a spherical mass M. The Schwarzhild metric with $ds^2 = 0$ gives

$$\left(1 - \frac{r_s}{r}\right) c^2 dt^2 = \frac{dr^2}{(1 - r_s/r)} \tag{2.23}$$

From which we find

$$\frac{dr}{dt} = \pm c \left(1 - \frac{r_s}{r}\right) \tag{2.24}$$

Thus the radial speed of light, dr/dt, is *not* always equal to c. This is not a violation of relativity, as dr/dt is a *coordinate velocity*; it merely says how many spatial coordinate labels r an object passes in a unit of coordinate time t. We already noted that the Schwarzchild r labels are not *not* equally spaced; rather, the proper distance between two r coordinates gets larger as we get closer to the central mass. Thus we should not expect light to move at constant dr/dt in such a coordinate system. The invariant (i.e., coordinate-free) property of light is not $dr/dt = c$, but rather that light moves on null paths ($ds^2 = 0$). [6]

Eq. 2.24 says that $v = dr/dt < c$ for light moving near a central mass. It is as if light has an effective "index of refraction" given by

$$n = \frac{c}{v} = \left(1 - \frac{r_s}{r}\right)^{-1} \tag{2.25}$$

The index of refraction in a medium like glass arises because matter is made up of charged particles that absorb and re-radiate electromagnetic radiation passing through, and so cause the net light beam to move at $v < c$. In contrast, $v < c$ in the Schwarzchild metric because of the curvature of spacetime, and the value of v will depend on the particular coordinates being used. Despite the different physical origin, we can use the principles of optics to determine the bending of light in the vicinity of a mass M.

Consider a electromagnetic wave moving in the z direction and passing by a mass M. Eq. 2.24 indicates that the part of the light

[6] The coordinate speed of light is still the "speed limit" for motion, it is just that the value of this limit will depend on the coordinates. If we write a more general metric in the form

$$ds^2 = -g_{tt}dt^2 + g_{xx}dx^2$$

(where we require $g_{tt} > 0$ so that the dt^2 term in the metric is negative as it should be for a time coordinate) then the general coordinate speed of light (using $ds^2 = 0$) is

$$|v_L| = \frac{dx}{dt} = \left(\frac{g_{tt}}{g_{xx}}\right)^{1/2}$$

So the coordinate speed of light will depend on the metric terms which are the "scale bars" that set physical lengths. For example, if you were to follow light moving across the earth, it would not appear to move at constant speed on a Mercator map (Figure 1.1) since some regions of that coordinate map are rendered bigger than they physically are (e.g., the polar regions). Rather, the coordinate speed with which light traversed the map would depend on the scale bar in any particular region.

For objects that aren't light ($ds \neq 0$) we find from the metric a coordinate speed

$$\frac{dx}{dt} = \pm |v_L| \left[1 + \frac{ds^2}{dt^2}\frac{1}{g_{tt}}\right]^{1/2}$$

Since material objects move on timelike paths, $ds^2 < 0$, the term in brackets is < 1. So material objects always move at a slower coordinate speed than does light.

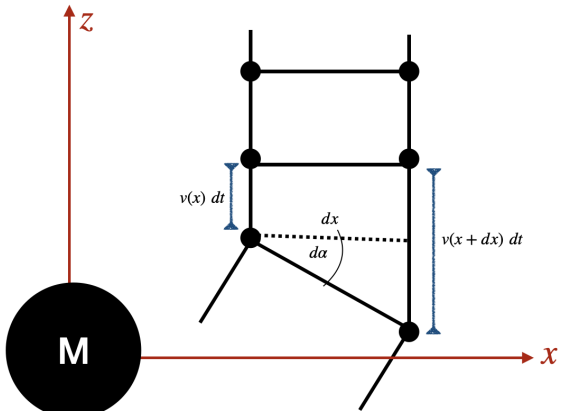


Figure 2.7: Schematic of Huygens principle for the bending of light.

closer to the mass moves more slowly than that farther away. *Huygens Principle* then implies that the light ray will naturally bend and turn to keep the wavefronts straight. Figure 2.7 shows the turning of wavefronts near a mass M due to the fact that end slightly closer to M does not make as much headway as the end a distance dx further away. If an infinitesmal time dt the geometry of Fig 2.7 shows that the beam turns an angle

$$\tan d\alpha \approx d\alpha = \frac{v_z(x+dx)\ dt - v_z(x)\ dt}{dx} = dt\left(\frac{v_z(x+dx) - v_z(x)}{dx}\right) \tag{2.26}$$

where we assumed the angle is small, so $\tan d\alpha \approx d\alpha$. In the limit $dx \to 0$ (i.e., a small wavefront) the term in parenthesis is just the definition of the derivative and we have[7]

$$d\alpha = dt\left(\frac{dv_z}{dx}\right) \tag{2.27}$$

To determine dv_z/dx we need an expression for the coordinate velocity of light moving in the z direction in Cartesian-like coordinates. We found above (Eq. 2.24) the coordinate velocity of light in the r direction. To get the equivalent expression in the z direction we we can use the so-called "isotropic" coordinates, which in the approximation of light passing far from the mass is (see Appendix)

$$v_z = \frac{dz}{dt} = c\left(1 - \frac{r_s}{r'}\right) \tag{2.28}$$

where $r'^2 = x^2 + y^2 + z^2$. From Fig. 2.7 we choose the light moving in the z direction in the $x - z$ plane (so $y = 0$) so

$$\frac{dv_z}{dx} = c\frac{r_s x}{(x^2+z^2)^{3/2}} \tag{2.29}$$

And the angular deviation $d\alpha$ for light moving a time step dt is

$$d\alpha = cdt\ r_s \frac{x}{(x^2+z^2)^{3/2}} \tag{2.30}$$

[7] The expression Eq. 2.27 for the deflection angle $d\alpha$ assumes the light is moving in the z direction, in which case $d\alpha$ is set by the gradient of velocity perpendicular to the direction of motion (i.e., the x direction). More generally, if the light propagates in an arbitrary direction, the deflection angle will be

$$d\alpha = dt\,(\nabla_\perp v)$$

where $\nabla_\perp v$ is the derivative of the velocity field v perpendicular to the direction of light propagation.

To calculate the total angle, α, by which light is bent when moving past a mass M, we need to integrate this over the path of the light (see Figure 2.8). To simplify the integral, we can make the approximation that the bending of the light is small ($\alpha \ll 1$) so that the path of light is very nearly a straight line along z. Then we can integrate a path along z at a fixed distance (impact parameter) $x = b$ away from the mass. In a time step dt the light moves a distance $dz = cdt$ so the integral becomes

$$\alpha = \int_{-\infty}^{\infty} dz\, r_s \frac{b}{(b^2 + z^2)^{3/2}} \tag{2.31}$$

Looking up the integral, we find

$$\alpha = \frac{2r_s}{b} = \frac{4GM}{bc^2} \tag{2.32}$$

This is a fundamental result in the theory of gravitational lensing. It only holds for small angle lensing ($\alpha \ll 1$) but a more accurate approach can be carried out by doing the integration without approximation (typically requiring numerical integration).

Our approach to this problem was to assign an variable "index of refraction" to spacetime and follow the "refraction" of a light ray as it moved through this medium. Had we used a different coordinate system to describe Schwarzchild spacetime, the expressions for the coordinate velocity of light (and hence index of refraction) may have looked different. Our final result, however, would be the same, as we have calculated the deviation α at points infinitely far away from the mass, where the spacetime is flat.

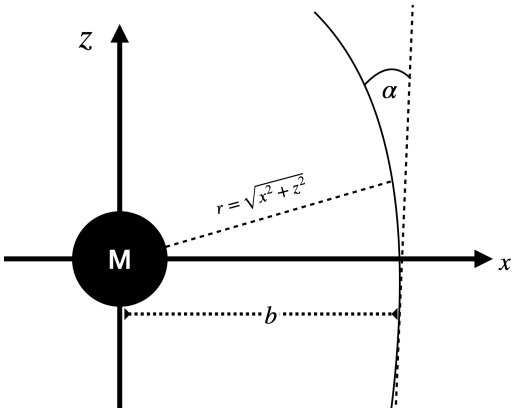


Figure 2.8: Light that passes by a mass M with impact parameter b is deflected by an angle α relative to the path the light would have taken if the mass was not there and spacetime is flat. To determine α, we need to integrate up the tiny deflection $d\alpha$ that occurs at each step along the path. In the limit of small deflection, ($\alpha \ll 1$ which occurs when light passes far away from the mass, $b \gg r_s$) we can do the integral by approximating the true path of light (solid line) by the straight dashed line.

2.4.2 *Geodesics and Constants of Motion*

In the previous section we visualized the path of a light ray as being "bent" as it tried to make its way through a curved spacetime. While this may be a useful heuristic for doing the calculation, in reality the light ray's path is only "bending" relative to what our expectations are for flat spacetime. The path of the light ray simply follows the natural contours of the curved spacetime – a geodesic – which is the generalization of a straight line. The deeper way to approach the problem[8] is to identify these geodesics for a given spacetime.

In flat Euclidean space, you can draw an infinite number of paths connecting two points. Of these, one path – a straight line – is singled out as having the *minimum* distance along it. In Newtonian physics, a mass undergoing "natural" motion (i.e., not being acted on by any force) moves in a straight line. We generalize this idea to curved spacetime by saying that the "natural" motion of a mass is the *geodesic* where the spacetime distance is an extremum.

[8] For example, imagine you are floating in a boat at the equator. Your friend is also on the equator 1 mile to the east of you. You both start navigating your boats due North (along lines of lattitude). Over time, you would notice that the distance between the two of you was decreasing, and as you both approached the North pole your two paths would intersect. If you were to draw these paths on a 2D mercator map, they would look like curved lines that bend towards each other. But in reality neither of you has "bent" or deviated your path – you both headed straight North without turning. It is simply because the 2D surface you are on is curved that the natural paths (geodesics) do not behave as you would expect for a flat 2D plane.

To calculate the spacetime time distance of a path, we break it up into numerous straight line segments of length ds^2 (see Figure 2.9). Massive objects are restricted to *time-like* paths (where each step has $ds^2 < 0$) so it is convenient to write ds^2 as a proper-time, $d\tau^2 = -ds^2/c^2$, which is a positive quantity. We can add up all the segments to get the total proper-time of the path[9]

$$\tau = \int d\tau = \int \sqrt{-ds^2/c^2} \tag{2.33}$$

We want to find the geodesics, which are paths through spacetime that *maximize*[10] the proper time τ. In flat Minkowski space, we know that the geodesics are straight lines (i.e., moving at constant velocity). What makes a line "straight" in this context is that its slope is constant, i.e., for each segment $d\tau$ of the geodesic path, the changes dx and dt are the same (see Fig 2.9). So in Minkowski space the geodesics are curves where the slopes are *constants of motion*

$$\frac{dx}{d\tau} = \text{ constant} \qquad \frac{dt}{d\tau} = \text{ constant} \quad \text{(in Minkowski space)} \tag{2.34}$$

To put it another way, in flat space an object will not change its spacetime slope (i.e., its velocity) unless some force acts upon it. Absent any external forces, the natural paths are of constant slope.

In a curved spacetime, the geodesics will not have constant slope, but is there some generalization of this idea of something staying constant in the motion? Yes there is. We now show that in some cases geodesics are curves where the slope of the path *times a metrical term* is constant. To derive this, let's consider for simplicity just 1 spatial dimension and write the general metric as

$$ds^2 = -g_{tt}dt^2 + g_{xx}dx^2 \tag{2.35}$$

where g_{tt} and g_{xt} are components of some general metric[11] that encodes the geometry of the spacetime. For simplicity, let's analyze just two line segments of a path, as shown in Figure 2.10. The path begins at an initial spacetime point P_i with coordinates $(t, x) = (0, 0)$ and moves to a final point P_f at $(\Delta t, \Delta x)$, passing through a middle point P_m along the way. We fix the end-points P_i and P_f and ask: "How should we pick P_m to maximize the proper time'?'

Writing the above metric in terms of proper time

$$d\tau^2 = -\frac{ds^2}{c^2} = \frac{g_{tt}}{c^2}dt^2 - \frac{g_{xx}}{c^2}dx^2 \tag{2.36}$$

The proper time interval for the first segment of the path shown in Figure 2.10 is then

$$\Delta\tau_1^2 = g_{tt}c^{-2}\Delta t_1^2 - c^{-2}\Delta x_1^2 \tag{2.37}$$

[9] The τ of a path is an invariant (i.e., has the same value in every coordinate system). In particular, in the coordinate system of someone moving along the path, position does not change ($dx' = 0$) and so τ is just the time elapsed in that frame (hence the name "proper time").

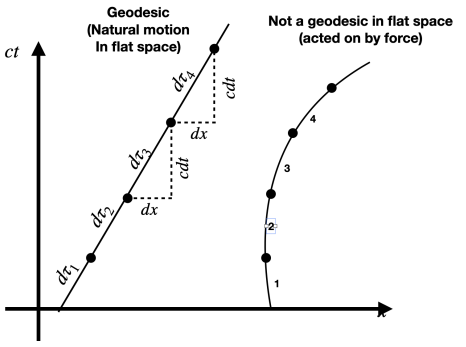


Figure 2.9: An arbitrary path through spacetime could be broken down into many small straight line segments. Each pair of segments follow the relations derived

[10] To be precise, time-like geodesics correspond to *extrema* of proper time. But in basically all cases we will consider, the extrema are maxima, not minima. Note that reducing the spatial distance traveled tends to *increase* the proper time, since the spatial terms have a minus sign in the expression for $d\tau$. Thus maximizing the proper time is a way of generalizing the idea of the "shortest path" between two spacetime events.

[11] In general, mixed terms like $dxdt$ may appear in the metric. We won't bother with those here, but the argument that follows can be easily extended to include them.

or

$$\Delta\tau_1 = c^{-1}\left[g_{tt}\Delta t_1^2 - g_{xx}\Delta x_1^2\right]^{1/2} \tag{2.38}$$

And similarly for the second segment

$$\Delta\tau_2 = c^{-1}\left[g_{tt}\Delta t_2^2 - g_{xx}\Delta x_2^2\right]^{1/2} \tag{2.39}$$

We can eliminate Δt_2 and Δx_2 using the fact that $\Delta t = \Delta t_1 + \Delta t_2$ and $\Delta x = \Delta x_1 + \Delta x_2$. This becomes

$$\Delta\tau_2 = c^{-1}\left[g_{tt}(\Delta t - \Delta t_1)^2 - g_{xx}(\Delta x - \Delta x_1)^2\right]^{1/2} \tag{2.40}$$

We want to find the path between the starting point $(0,0)$ and endpoint $(\Delta t, \Delta x)$ that maximizes the total proper time $\Delta\tau = \Delta\tau_1 + \Delta\tau_2$. We first vary Δx_1 to find the value that maximizes $\Delta\tau$ while holding Δt_1 (and the endpoints) fixed. The derivative to take to find the minimum is

$$\frac{d\Delta\tau}{d\Delta x_1} = \frac{d\Delta\tau_1}{d\Delta x_1} + \frac{d\Delta\tau_2}{d\Delta x_1} = 0 \tag{2.41}$$

Calculating the first term we have

$$\frac{d\Delta\tau_1}{d\Delta x_1} = \frac{1}{2}\frac{c^{-1}}{\left[g_{tt}\Delta t_1^2 - g_{xx}\Delta x_1^2\right]^{1/2}}\left[-2g_{xx}\Delta x_1 + \frac{dg_{tt}}{dx}\Delta t_1^2 - \frac{dg_{xx}}{dx}\Delta x_1^2\right]$$

Now if we make the assumption that *the metric is independent of the coordinate x*, i.e., that the metrical terms are constant with respect to x, then the derivatives $dg_{tt}/dx = 0$ and $dg_{xx}/dx = 0$, and the above simplifies to

$$\frac{d\Delta\tau_1}{d\Delta x_1} = -\frac{g_{xx}\Delta x_1 c^{-1}}{\left[g_{tt}\Delta t_1^2 - g_{xx}\Delta x_1^2\right]^{1/2}} = -\frac{g_{xx}}{c}\frac{\Delta x_1}{\Delta\tau_1}$$

Carrying out the same calculation for $\Delta\tau_2$ gives similarily

$$\frac{d\Delta\tau_2}{d\Delta x_1} = \frac{c^{-1}}{2}\frac{-2g_{xx}(\Delta x - \Delta x_1)(-1)}{\left[(\Delta t - \Delta t_1)^2 - (\Delta x - \Delta x_1)^2\right]^{1/2}} = \frac{g_{xx}}{c}\frac{\Delta x_2}{\Delta\tau_2}$$

And so our condition that the path maximizes proper time is

$$\frac{d\Delta\tau}{d\Delta x_1} = \frac{d\Delta\tau_1}{d\Delta x_1} + \frac{d\Delta\tau_2}{d\Delta x_1} = -\frac{g_{xx}}{c}\frac{\Delta x_1}{\Delta\tau_1} + \frac{g_{xx}}{c}\frac{\Delta x_2}{\Delta\tau_2} = 0 \tag{2.42}$$

The above implies that

$$g_{xx}\frac{\Delta x_1}{\Delta\tau_1} = g_{xx}\frac{\Delta x_2}{\Delta\tau_2} \tag{2.43}$$

An arbitrary path through spacetime is built of many small straight line segments (see Fig. 2.9) and so applying the same argument to line segments 2 and 3, and every subsequent pair thereafter implies

$$g_{xx}\frac{\Delta x_1}{\Delta\tau_1} = g_{xx}\frac{\Delta x_2}{\Delta\tau_2} = g_{xx}\frac{\Delta x_3}{\Delta\tau_3} = g_{xx}\frac{\Delta x_4}{\Delta\tau_4} = ... \tag{2.44}$$

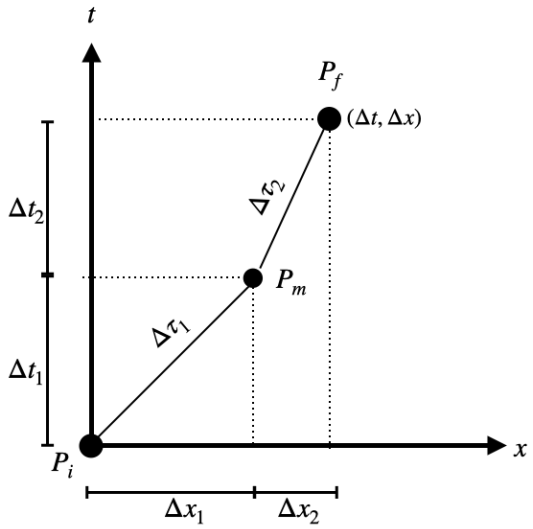


Figure 2.10: Two small segments of a path through spacetime. We hold the endpoints of the path P_i and P_f fixed, and ask "how should we choose P_m such that the proper time is maximized? This allows us to calculation properties of the geodesics.

where the metric term g_{xx} should be evaluated at the point in space at which each segment of the path is at. We can take the limit that the path segments are infinitesimally small by changing notation from Δ to d

$$g_{xx}\frac{dx}{d\tau} = \text{ constant} \quad \text{(if metric is independent of } x\text{)} \tag{2.45}$$

We have found a "constant of the motion". In flat space, the metric term $g_{xx} = 1$ and so the constant of motion is just the slope $dx/d\tau$ of the path. This reflects what we said out front, that the geodesics in flat space are straight lines. We now see that generalization of the notion of a "straight line" to curve space is straightfoward, it is the slope *times the metrical term* that is a constant of motion.

If we carry out the exact same analysis to maximize proper time with respect to Δt_1 (while holding Δx_1 fixed) we find a similar constant of motion $g_{tt}dt/d\tau$. Summarizing, for geodesics

$$\frac{p_x}{m} = g_{xx}\frac{dx}{d\tau} = \text{ constant} \quad \text{(if } g \text{ is independent of } x\text{)} \tag{2.46}$$

$$\frac{E}{m} = g_{tt}\,\frac{dt}{d\tau} = \text{ constant} \quad \text{(if } g \text{ is independent of } t\text{)} \tag{2.47}$$

We give these constants of motion suggestive names and call p_x/m the *x-momentum per unit mass* and call E/m the *energy per unit mass*. It doesn't really matter what we call them, the point is that they are constants along a geodesic path. But we will see that they are indeed related to our traditional notions of momentum and energy.

Our derivation of the constants of motion required that the metric components be independent of the associated coordinate. The property of being independent of a coordinate is an expression of symmetry. For example, if g does not depend on x, a translation of the coordinate ($x' = x + D$, where D is a constant) leaves the metric unchanged. Thus we have seen how space translational symmetry leads to a constant of motion that we call linear momentum. Similarly, time translational symmetry leads to a constant of motion that we call energy. This connection between conservation laws with space-time symmetries is an expression of the more extensive mathematical result called *Noether's Theorem*.

2.4.3 *Schwarzchild Geodesics: Falling into a Black Hole*

The constants of motions we just derived can be used to determine the geodesics around a mass M. Consider first an object falling radially inward into a black hole (so $d\Omega = 0$). The Schwarzchild metric written in terms of proper time $d\tau^2$ is

$$d\tau^2 = \left(1 - \frac{r_s}{r}\right)dt^2 - \frac{1}{c^2}\frac{dr^2}{(1 - r_s/r)} \tag{2.48}$$

The metric is independent of t so we have an energy constant of motion

$$\frac{E}{m} = g_t \frac{dt}{d\tau} = \left(1 - \frac{r_s}{r}\right) c^2 \frac{dt}{d\tau} \tag{2.49}$$

The metric does depend on r, so the radial momentum will not be a constant of motion[12]. To make use of our energy constant of motion, we divide the metric through by $d\tau^2$ and multiply through by $(1 - r_s/r)$ to get

$$\left(1 - \frac{r_s}{r}\right) = \left(1 - \frac{r_s}{r}\right)^2 \left(\frac{dt}{d\tau}\right)^2 - \frac{1}{c^2}\left(\frac{dr}{d\tau}\right)^2 \tag{2.50}$$

We recognize the first term on the right hand side as just E^2/m^2c^4 find

$$\frac{E^2}{m^2c^4} - 1 = \frac{1}{c^2}\left(\frac{dr}{d\tau}\right)^2 - \frac{r_s}{r} \tag{2.51}$$

By eliminating t We now have a differential equation for $dr/d\tau$ that we can solve to determine the path $r(\tau)$. Before doing, we note that the equation becomes even more suggestive if we multiply through by $m/2$ and plug in the expression $r_s = 2GM/c^2$

$$\frac{mc^2}{2}\left[\left(\frac{E}{mc^2}\right)^2 - 1\right] = \frac{1}{2}m\left(\frac{dr}{d\tau}\right)^2 - \frac{GMm}{r} \tag{2.52}$$

This looks a lot like the equation for Newtonian energy, where the first term on the right hand side is kinetic energy and the second term gravitational potential energy. The left hand side is not just the energy E but it is still a constant, and we can show [13] that it in the limit it reduces to

$$E_N = \frac{1}{2}m\left(\frac{dr}{dt}\right)^2 - \frac{GMm}{r} \tag{2.53}$$

where E_N is the Newtonian energy. Thus we have derived the Newtonian equation for radial motion using just the Schwarzchild metric and the principle of maximum proper time! In Netwonian physics, the first term on the right hand side is kinetic energy and the second is potential energy. In GR there is no real distinction between kinetic and gravitational potential energy. We instead have a constant of motion E/m that incorporates both.

Imagine now an astronaut that falls into a black hole starting from very far away, $r_0 \gg r_s$. If the person starts out at rest, the energy is just the rest mass energy, $E = mc^2$. Eq. 2.52 can then be written

$$\frac{1}{2}m\left(\frac{dr}{d\tau}\right)^2 = \frac{GMm}{r} \implies \left(\frac{dr}{d\tau}\right)^2 = \frac{c^2 r_s}{r} \tag{2.54}$$

[12] By fixing the mass M at $r = 0$ we have broken the space translational symmetry from the point of view of the infalling object, since now all r points are not the same. Of course, space translational symmetry still holds for the combined mass M plus object m system, and momentum is conserved of the entire system.

[13] The Newtonian energy does not include rest mass energy so $E_N = E - mc^2$, so putting this into the left hand we get

$$\frac{mc^2}{2}\left[(\frac{E_N}{mc^2} + 1)^2 - 1)\right]$$

In the Newtonian limit objects move with speed $v \ll c$ and so $E \ll mc^2$, so we can binomial expand the term with $E_n/mc^2 \ll 1$

$$\frac{mc^2}{2}\left[1 + 2\frac{E_N}{mc^2} + ... - 1)\right] = E_N$$

So the left hand side is just E_N to leading order. In the Newtonian limit where all speeds are $\ll c$ we also have $d\tau = dt$ (i.e., proper time is equal to coordinate time since time dilation is negligible).

This differential equation can be solved to find the elapsed proper time to fall from r_0 to a radius r

$$\Delta\tau = \frac{2}{3}\frac{r_s}{c}\left[\left(\frac{r_0}{r_s}\right)^{3/2} - \left(\frac{r}{r_s}\right)^{3/2}\right] \tag{2.55}$$

This result shows that an astronaut in free fall will pass through the event horizon of a black hole and reach the center in a finite proper time. In particular, the proper time to fall from the event horizon ($r_2 = r_s$) to the singularity $r_1 = 0$ is simply $(2/3)r_s/c$. For a stellar mass black hole, $r_s \approx 3$ km, this time is only about 6 microseconds. For a supermassive black hole of $M = 10^9 M_\odot$, the time is a bit less than 2 hours. The freefall inside a black hole is a geodesic which maximizes proper time, so if you try to accelerate inside the event horizon to avoid your fate, you will only hit the singularity sooner.

2.4.4 *Schwarzchild Geodesics – Circular Orbits*

The last section considered purely radial motion into a black hole. Let's consider now the orbits of objects around a spherical mass M, which could be a black hole or a star. We will consider orbits that lie in the equatorial plane, so $\theta = \pi/2$ and $d\theta = 0$. For comparisons sake, first recall that in Newtonian physics the energy equation for an orbiting object is

$$E = \frac{1}{2}m\left(\frac{dr}{dt}\right)^2 + \frac{1}{2}mr^2\left(\frac{d\phi}{dt}\right)^2 - \frac{GMm}{r} \tag{2.56}$$

The first terms on the right hand side are the radial and angular kinetic energy. We have two constants of motion in Newtonian orbits, the energy E and the angular momentum

$$L = mr\left(r\frac{d\phi}{dt}\right) \tag{2.57}$$

Using this definition of L the energy equation becomes

$$E = \frac{1}{2}m\left(\frac{dr}{dt}\right)^2 + \frac{L^2}{2mr^2} - \frac{GMm}{r} \tag{2.58}$$

which we can rewrite

$$E = \frac{1}{2}m\left(\frac{dr}{dt}\right)^2 + V_{\text{eff}}(r) \tag{2.59}$$

where V_{eff} is an *effective potential* that includes both the gravitational potential term $-GMm/r$ (which tends to pull objects to the center) and the centrifugal term $L/2mr^2$ (which tends to push objects out). We have thus isolated the radial motion of the orbit.

A circular orbit is one where the object stays at a fixed radius r_c. This occurs at a minimum of $V_{\text{eff}}(r)$, which we can find by setting $dV_{\text{eff}}/dr = 0$. One finds

$$r_{c,N} = \frac{L^2}{GMm^2} \tag{2.60}$$

Circular orbits of smaller radius have smaller L. Note that in Newtonian physics there are stable orbits all the way down to $r = 0$. This will change in general relativity.

To figure out orbits in GR, we use the Schwarzchild metric which for $\theta = \pi/2$, $d\theta = 0$ is

$$d\tau^2 = \left(1 - \frac{r_s}{r}\right) dt^2 - \frac{1}{c^2}\frac{dr^2}{(1 - r_s/r)} - \frac{r^2}{c^2} d\phi^2 \tag{2.61}$$

Since the metric is independent of t and ϕ, we can identify two constants of motion

$$\frac{E}{m} = c^2 \left(1 - \frac{R_s}{r}\right) \frac{dt}{d\tau} \qquad\qquad \frac{L}{m} = r^2 \frac{d\phi}{d\tau} \tag{2.62}$$

where E/m is the energy (per unit mass) discussed in the last section, and L/m is an angular momentum-like quantity.

Our approach to these problems is to use the constants of motion to eliminate the $d\phi$ and dt terms in the metric and thus get an expression that just relates dr and $d\tau$. This we can integrate to get $r(\tau)$. Dividing the metric through by $d\tau^2$ and using the constants of motion to replace $dt/d\tau$ and $d\phi/d\tau$ we find after some algebra

$$\frac{mc^2}{2}\left[\left(\frac{E}{mc^2}\right)^2 - 1\right] = \frac{1}{2} m \left(\frac{dr}{d\tau}\right)^2 + V_{\text{eff}} \tag{2.63}$$

where the effective potential is

$$V_{\text{eff}}(r) = -\frac{GMm}{r} + \frac{L^2}{2mr^2} - \frac{L^2 GM}{r^3 mc^2} \tag{2.64}$$

The Schwarzchild orbit equation looks very similar to the Newtonian one, with the notable exception that there is an extra terms in the effective potential that goes like $-1/r^3$. This is sometimes called the "pit in the potential", because it causes the effective potential to turn over and become negative at small r.

Taking $dV_{\text{eff}}/dr = 0$ we can find the equilibrium points corresponding to circular orbits. Doing so results in a quadratic equation which has a solution

$$r_c = \frac{1}{2}\frac{L^2}{GMm^2} \pm \frac{1}{2}\sqrt{\frac{L^4}{G^2M^2m^4} - 4\frac{3L^2}{m^2c^2}} \tag{2.65}$$

This equation can be made more instructive if we write things in terms of the characteristic length scale of the problem, r_s, and a characteristic angular momentum scale defined as

$$L_c = r_s mc \tag{2.66}$$

Then the solution for the circular orbits can be written

$$r_c = r_s \left(\frac{L}{L_c}\right)^2 \left[1 \pm \sqrt{1 - \frac{3L_c^2}{L^2}}\right] \tag{2.67}$$

We see that for the Schwarzchild metric there are *two* positions where circular orbits are possible. Only the outermost circular orbit is stable. If an object was on the innermost circular orbit, and small perturbation would cause it to fly either into the black hole, or off into space. In the limit $L \gg L_c$ we just get the Newtonian solution of $r_c = r_N$ (and the other circular orbit is at $r = 0$).

From Eq. 2.67 we see that there is a minimum angular momentum allowed for a circular orbit, because for $L < \sqrt{3}L_c$ the solution is not real. Thus there is a minimum radius of a stable circular orbit in a Schwarzchild metric which (plugging in $L = \sqrt{3}L_c$) is

$$r_{\rm isco} = 3r_s \qquad \text{(innermost stable circular orbit)} \tag{2.68}$$

Any object that tries to orbit at $r < r_{\rm isco}$ is destined to fall into the black hole. For example, a disk of gas accreting onto a black hole will only extend down to $r_{\rm isco}$; below that, the gas plunges into the black hole. Similarly, a binary system of two black holes will only remain stable when their orbital separation is greater than $r_{\rm isco}$; below that, the fall into each other and merge.

2.4.5 Schwarzchild Geodesic – Black Hole Accretion

We can use the above results to estimate how much energy can be released by gas spiraling into a black hole. Gas in an accretion disk will follow nearly circular orbits, with the gas nearer the center swirling around faster than gas farther out. If there were little friction in the disk, the gas could orbit around indefinitely, like the rings of Saturn. However, as the annuli of gas at different radii are sliding and shearing against each other, the viscosity of the gas releases heat and slows down the inner layers, which fall closer to the black hole. Once gas drops below $r_{\rm isco}$ it can no longer orbit stably and will fall quickly below the event horizon.

Consider a bit of mass m in a disk around a black hole. We previously found the energy equation of Schwarzchild orbits

$$\frac{mc^2}{2}\left[\left(\frac{E}{mc^2}\right)^2 - 1\right] = \frac{1}{2}m\left(\frac{dr}{d\tau}\right)^2 - \frac{GMm}{r} + \frac{L^2}{2mr^2} - \frac{L^2 GM}{r^3 mc^2} \tag{2.69}$$

For a circular orbit the radius is constant, so we can set $dr/d\tau = 0$. Rewriting everything in terms of r_s and L_c the equation becomes

$$\left(\frac{E}{mc^2}\right)^2 = 1 - \frac{r_s}{r} + 3\frac{L^2}{L_c^2}\left(\frac{r_s}{r}\right)^2\left[1 - \frac{r_s}{r}\right] \tag{2.70}$$

From this equation we find that the energy of a bit of mass orbiting at the innermost stable orbit (where $L = L_c$ and $r = 3r_s$) is $E = \sqrt{8/9}(mc^2)$. If this bit of mass started off very far from the black hole, $r \gg r_s$,its initial energy is just the rest mass energy $E = mc^2$. The energy difference is

$$\Delta E = \left(1 - \sqrt{\frac{8}{9}}\right) mc^2 \approx 0.052\ mc^2 \tag{2.71}$$

Thus for a bit of mass to move from far out in the disk to a circular orbit at $r_{\rm isco}$ it must loose about 5% of its energy. It can do this by radiating away the energy as light. The viscosity in the disk turns some of the kinetic energy of the orbiting gas into heat, and the hot gas can then emit light. The remaining 95% or so of the mass/energy winds up being eaten by the black hole[14]

For rotating black holes, the innermost stable orbit turns out to be even closer in. For a maximally spinning black hole, one finds $r_{\rm isco} = r_s$ (assuming the disk rotates in the same direction as the black hole). In this case, even greater energy release is possible, reaching around 12% of mc^2. Not all accretion onto black holes leads to such a large release of energy. If gas lacks angular momentum it may plunge directly into the black hole rather then remaining on circular orbits. In this case it may only radiate a small amount of energy away, with almost all of the energy going into the black hole (or being blown out in winds or jets). This is the case with the supermassive black hole in the center of our own Galaxy, which radiates quite inefficiently.

[14] We can compare this energy release to other processes. Typical chemical processes (e.g., burning coal) involves changes in the atomic and molecular structure, where the energy levels are of order electron volts (eV). The masses of typical atoms are of order 1 GeV (the mass of a proton), so the chemical processes release around eV/GeV $\sim 10^{-9}$ of the rest mass energy. Nuclear reactions such as those that power the energy of the Sun release about an MeV per nucleus, so the efficiency is around 10^{-3}. These are all well below the possible energy release efficiency of black hole accretion.

2.5 *Appendix: Kruskal–Szekeres coordinates (optional)*

While we typical dealt with the Schwarzchild metric using Schwarzchild coordiantes, writing the metric in other coordinate systems can be helpful. Choosing different coordinates is like choosing different ways of rendering the 2D curved surface of the earth on a 2D flat map – different maps distort different aspects of the space.

Beginning with the Schwarzchild metric in common Schwarzchild coordinates (t, r, θ, ϕ)

$$ds^2 = \left(1 - \frac{r_s}{r}\right)c^2dt^2 + \frac{dr^2}{\left(1 - \frac{r_s}{r}\right)} + r^2d\Omega^2$$

We define new KS coordinates (T, X, θ, ϕ) where (θ, ϕ) are the same as in Schwarzchild coordinates, while X and T are defined by

$$T = \frac{r_s}{c}\left|\frac{r}{r_s} - 1\right|^{1/2} e^{r/2r_s} \sinh\left(\frac{ct}{2r_s}\right)$$

$$X = r_s\left|\frac{r}{r_s} - 1\right|^{1/2} e^{r/2r_s} \cosh\left(\frac{ct}{2r_s}\right)$$

After some algebra you can show that the metric becomes in the KS coordinates

$$ds^2 = \frac{4r_s}{r} e^{-r/r_s}\left(-c^2 dT^2 + dX^2\right) + r^2 d\Omega^2 \tag{2.72}$$

where the Schwarzchild radial coordinate r is related to X and T by

$$-c^2T^2 + X^2 = (r - r_s)\, e^{r/r_s}$$

In principle we could solve Eq. 2.5 to eliminate r in the KS metric Eq. 2.72 and get an explicit metric soley in KS coordinates. There is no pretty way to do this, though, so it is better to leave r as implicit function of X and T.

We point out two interesting properties of the Schwarzchild metric written in KS coordinates. First, there is no longer a singularity at $r = r_s$. This confirms by construction that the event horizon singularity we previously saw in the Schwarzchild metric is only a *coordinate singularity* (i.e., a breakdown of the particular coordinate labeling being used). On the other hand, the KS Schwarzchild metric still has a singularity at $r = 0$. The is a physical singularity (i.e., a point of infinite curvature) that cannot be removed by a change of coordinates.

Second, for radial lightlike paths, $ds^2 = 0$, Eq. 2.72 implies

$$0 = -c^2 dT^2 + dX^2 \implies \frac{dX}{dT} = \pm c \tag{2.73}$$

So the coordinate speed of light in KS coordinates is always c, and lightlike paths always along the diagonal. The future light cone does not change shape or "flip on its side" as we had in Schwarzchild coordinates.

Figure 2.11 illustrates the Szchwarzdhild metric in KZ coordinates. The path of an observer moving in this spacetime (shown as the blue solid line) always move upward within its future light cone, which does not close or flip. We can imagine the observer to be riding in a toy car that always moves forwards and can at most turn 45 degrees right or left. The event horizon is seen to be along the diagonal (a lightlike surface) and so if the observer path passes the event horizon, they will never be able to get out (as their "toy car" can never turn further than the 45 degrees of their future light cone). Instead, the observer is destined to hit singularity (labeled $r = 0$) which is seen to be a timelike surface.

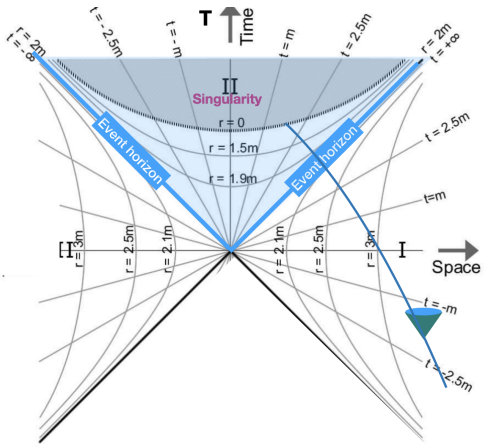


Figure 2.11: F

2.6 *Appendix - Isotropic Coordinates (optional)*

To write the Schwarzchild metric in "isotropic" coordinates, we define a new radial coordinate, r', related to the standard one r by

$$r = r'\left(1 + \frac{r}{r'}\right)^2 \tag{2.74}$$

Plugging this into Eq.2.13 we find, after some algebra

$$ds^2 = -\left(\frac{1 - r_s/4r'}{1 + r_s/4r'}\right)^2 c^2 dt^2 + \left(1 + \frac{r_s}{4r'}\right)^4 \left(dr'^2 + r'^2 d\Omega^2\right) \tag{2.75}$$

The last term in parenthesis is just that of flat space, so we can write this as

$$ds^2 = -\left(\frac{1 - r_s/4r'}{1 + r_s/4r'}\right)^2 c^2 dt^2 + \left(1 + r_s/4r'\right)^4 \left(dx^2 + dy^2 + dz^2\right) \tag{2.76}$$

where $r'^2 = x^2 + y^2 + z^2$. Note that in isotropic coordinates we no longer have a singularity at $r = r_s$, which confirms this is only a coordinate singularity. In the section on gravitational lensing, we considered light passing by at great distance to the mass, such that $r' \gg r_s$. Then, performing binomial expansions and only keeping terms up to order r_s/r

$$ds^2 = -\left(1 - \frac{r_s}{r'}\right) c^2 dt^2 + \left(1 + \frac{r_s}{r'}\right)\left(dx^2 + dy^2 + dz^2\right) \tag{2.77}$$

For light ($ds^2 = 0$) moving along the z direction ($dx = dy = 0$), we then find

$$v_z = \frac{dz}{dt} = c\left(\frac{1 - r_s/r'}{1 + r_s/r'}\right)^{1/2} \approx c\left(1 - \frac{r_s}{r'}\right) \tag{2.78}$$

where we used the binomial approximation again.

3
Cosmic Dynamics

3.1 The Robertson-Walker Metric

General relativity provides a foundation of modern cosmology by providing a means to calculate the spacetime metric of the entire universe. This may seem like a daunting task, given the vastness and complexity of the universe. However, we can simplify the problem enormously by imposing certain symmetry properties, namely that on large enough scales, the universe is *homogeneous and isotropic*[1]. This assumption (often called the *cosmological principle*) limits the possible metrics to ones that are the same at all points in space and in all directions in space. Current observations of large scale structure validate the cosmological principle, but it still remains a theoretical assumption, especially when we apply it to distance scales that are too large for us to directly observe.

[1] Homogeneous means the universe is the same at all points in space. Isotropic means that the universe is the same in all directions. In the language of symmetries, homogeneity refers to the invariance of the contents of the universe under translations in space (i.e., shifting of coordinates). Isotropy refers to the invariance of the contents of the universe under rotations,

Clearly the universe is not smooth and homogeneous on small scales (such as that of planets, stars or galaxies) but if we average over large enough regions the clumpy structure is washed out. Similarly, the granular nature of water can be ignored when you examine scales much larger than the individual water molecules that compose it.

The spacetime of a homogeneous, isotropic universe may still be curved, but the curvature must be the same at all points. It will then be described by one number, which is essentially the Gaussian curvature. The Gaussian curvature will be positive for a sphere-like curvature, negative for a saddle-like curvature, and zero for no curvature (i.e., a flat universe). It is conventional in cosmology texts to use a notation which separates out the sign of the curvature from the degree of curvature,

$$\text{Gaussian Curvature} = \frac{\kappa}{R_0^2} \tag{3.1}$$

where R_0 is the "radius of curvature" of the universe (at the current time), and κ gives the sign of the curvature, and is equal[2] to either +1, -1 or 0. Keep in mind that R_0 does not actual refer to any distance between points in the universe, but rather is a convenient way to express the degree of curvature of spacetime.

[2] In lecture, I have often written κ as k, because it is easier for me to draw. The Ryden textbook uses κ so I will continue with notation here.

To get some intuition into a homogeneous, isotropic space with curvature, consider first the 2D surface of a sphere. The surface is

curved, and each point on the surface is equivalent to any other. To derive the metric, we can embedded this 2D surface within a 3-dimensional flat space, where we know the metric is (in Cartesian coordinates, and adding in the time component)

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 \tag{3.2}$$

The surface of the 2D sphere in subset of the points in the 3D space, and in particular those that obey the constraint

$$x^2 + y^2 + z^2 = R_0^2 \tag{3.3}$$

where R_0 is the radius of the sphere[3]. We can use this to eliminate one of the variables; for example this equation implies the variable z is

$$z = \sqrt{R_0^2 - (x^2 + y^2)} \tag{3.4}$$

and so we can use this to eliminate dz from the metric Eq. 3.2, and write the metric as a function of two coordinates.

We can generalize this approach to derive the metric of true universe of 3 spatial dimensions. We imagine embedding our 3D space in the metric of flat 4D space (plus time)

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 + dw^2 \tag{3.5}$$

where w is a new fictitious variable which we can eliminate by applying the constraint of a *hypersphere* (the higher dimensional analog of a sphere)

$$x^2 + y^2 + z^2 + w^2 = R_0^2 \quad \text{(spherical curvature)} \tag{3.6}$$

We can eliminate the fictitious variable w by solving for it

$$w = \sqrt{(R_0^2 - (x^2 + y^2 + z^2)} = \sqrt{R_0^2 - \sigma^2} \tag{3.7}$$

where we defined[4]. the coordinate $\sigma = \sqrt{x^2 + y^2 + z^2}$. From this we can calculate the differential dw using the chain rule

$$dw = \frac{-\sigma d\sigma}{\sqrt{R_0^2 - \sigma^2}} \tag{3.8}$$

and plugging this into the metric Eq: 3.5 we get

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 + \frac{\sigma^2 d\sigma^2}{R_0^2 - \sigma^2} \tag{3.9}$$

Recalling that $dx^2 + dy^2 + dz^2$ can be written in terms of the 3D spherical-polar (σ, θ, ϕ) coordinates as

$$dx^2 + dy^2 + dz^2 = d\sigma^2 + \sigma^2 d\theta^2 + \sigma^2 \sin^2\theta \, d\phi^2 \tag{3.10}$$

[3] Clearly the radius of a sphere (i.e., the line connecting the center of the sphere to the surface) does not exist within the 2D surface of the sphere itself. Thus, for ants crawling on the surface of the sphere, R_0 is not a distance that can be measured by laying out a measuring tape. Instead, it just quantifies how curved the surface is (which could be determined by drawing small circles on the surface and measuring how their circumference differs from $2\pi r$.)

[4] Usually when we use spherical-polar coordinates, the coordinate $\sigma = \sqrt{x^2 + y^2 + z^2}$ is called r. However, in Ryden's notation, the symbol r is used for a different (though related) coordinate, which we will discuss shortly. What we have called σ Ryden calls x. For the moment I use the variable σ to avoid confusion with the ordinary dimension x in Cartesian coordinates.

Making this substitution and collecting terms we find

$$ds^2 = -c^2dt^2 + \frac{d\sigma^2}{1 - \sigma^2/R_0^2} + \sigma^2 d\Omega^2 \tag{3.11}$$

where we used the shorthand $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$. This is the spacetime metric for the case where 3D space has constant positive (i.e., spherical curvature). The other possibility is *hyperbolically* (or negatively) curved space, in which the constraint on the coordinates is that of a three dimensional hyperboloid

$$x^2 + y^2 + z^2 - w^2 = -R_0^2 \quad \text{(hyperbolic curvature)} \tag{3.12}$$

The metric for hyperbolic space is

$$ds^2 = -c^2dt^2 + \frac{d\sigma^2}{1 + \sigma^2/R_0^2} + \sigma^2 d\Omega^2 \tag{3.13}$$

We can combine the spherical and hyperbolic and flat possibilities in a single equation

$$ds^2 = -c^2dt^2 + \frac{d\sigma^2}{1 - \kappa\sigma^2/R_0^2} + \sigma^2 d\Omega^2 \tag{3.14}$$

where again $\kappa = -1$ for the hyperbolic (or negative curvature) case, $\kappa = +1$ for the spherical (or positive curvature) case, and $\kappa = 0$ for flat space.

One of the remarkable predictions of general relativity is that the metric of spacetime is in general not static, but may either expand or contract over time. To maintain our assumption of homogeneity and isotropy, this expansion has to be the same everywhere. This can be accomplished by having the radius of curvature, R_0, vary over time, $R_0 \rightarrow R(t)$. Then conventional notation splits off the time-dependence as

$$\text{Radius of Curvature} : R(t) = a(t)\ R_0 \tag{3.15}$$

where R_0 is the radius of curvature at the present time (which we label t_0) and $a(t)$ is called the *scale factor*. By our assumption of homogeneity, $a(t)$ is the same everywhere in space and only varies in time. By definition we set the scale factor at the present time, t_0, equal to unity, $a(t_0) = 1$, such that the scale factor describes how much the scale of universe has changed *relative to today* (e.g., if at some earlier time, t, the universe was a factor of 2 smaller, then $a(t) = \frac{1}{2}$).

It is conventional to write the metric of the dynamical universe using *comoving coordinates*. These coordinates move along with the expansion, such that a distant galaxy remains at fixed comoving coordinate even as the universe expands (see Figure 3.1). Replacing

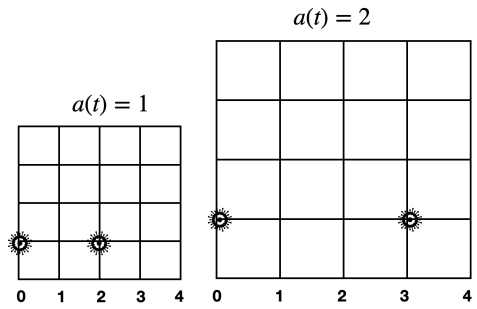


Figure 3.1: Example of comoving coordinates. We can imagine labeling points using graph paper. In the figure on the right, two galaxy are located at (horizontal) comoving coordinates $\sigma_0 = 0$ and $\sigma_0 = 2$. As the universe expands, the boxes of the graph paper increase in size, but the labeling of the nodes remains the same. The galaxies thus remain at fixed comoving coordinates, however the physical distance between them changes with the scale factor, $d_p = \Delta\sigma_0\ a(t)$.

Of course, things in the universe need not remain fixed at a node on the graph paper. Photons, for example, travel through spacetime and so their comoving coordinate r changes over time. Galaxies too may have motions on top of the overall cosmic expansion – the gravitational attraction of the Milky Way and the Andromeda Galaxy, for example, is causing them to falling into each other. This kind non-cosmological motion is called its "peculiar velocity" of a galaxy. When talking about distant galaxies, we can often neglect the peculiar velocities, since these motions are small compared to the cosmic expansion.

$\sigma \rightarrow \sigma_0\, a(t)$, where σ_0 is the comoving coordinate, and replacing $R_0 \rightarrow R_0\, a(t)$ we find

$$ds^2 = -c^2dt^2 + a(t)^2\left[\frac{d\sigma_0^2}{1-\kappa\sigma_0^2/R_0^2} + \sigma_0^2 d\Omega^2\right] \tag{3.16}$$

This is the *Robertson-Walker metric*[5] . The form of the metric is determined by symmetry conditions alone (i.e., a homogenous, isotropic spacetime), but additional information will be needed determine the curvature variables (κ and R_0) and what the dynamical scale factor, $a(t)$, is. We will soon explore the addition input (the Friedmann equations) in detail.

The form of the FRW metric given in Eq. 3.16 is not the one we will use most frequently in the class. We next derive a more convenient choice of coordinates for the FRW metric.

[5] Ryden uses x instead of σ_0 for the comoving coordinate in this metric, writing it as

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

Other textbooks use different notation, so it is important to always check what the conventions are in any text you are reading.

3.2 *Proper Distances in the RW metric*

The proper distance – or "measuring-tape" distance – is the distance measured at a fixed time, i.e., $dt = 0$. If we are considering distances measured in the radial direction, then $d\Omega = 0$ and the RW metric (Eq. 3.16) gives

$$ds^2 = a(t)^2 \frac{d\sigma_0^2}{1-\kappa\sigma_0^2/R_0^2} \tag{3.17}$$

We typically will choose our coordinates such that we sit at $\sigma_0 = 0$. If another galaxy is at coordinate σ_0 we can calculate the radial proper distance to it by integrating ds above

$$d_p = \int \sqrt{ds^2} = a(t)\int_0^{\sigma_0} \frac{d\sigma_0'}{\sqrt{1-\kappa\sigma_0'^2/R_0^2}} \tag{3.18}$$

The integral in this expression is not trivial (though it is easy enough to look up), which indicates that the coordinate σ_0 is not a direct measure of radial distance. We may then find it convenient to switch to a new comoving coordinate, r, that is a more direct measure of radial distance. We define the coordinate r such that

$$dr = \frac{d\sigma_0}{\sqrt{1-\kappa\sigma_0' 2/R_0^2}} \implies r = \int_0^{\sigma_0} \frac{d\sigma_0'}{\sqrt{1-\kappa\sigma_0'^2/R_0^2}} \tag{3.19}$$

Looking up this integral we find that the relationship between the coordinates σ_0 and r is

$$\sigma_0 = S_k(r) \quad \text{where} \quad S_k(r) = \begin{cases} R_0 \sin(r/R_0) & \text{if } \kappa = +1 \\ r & \text{if } \kappa = 0 \\ R_0 \sinh(r/R_0) & \text{if } \kappa = -1 \end{cases} \tag{3.20}$$

Now we can write the FRW metric (Eq. 3.16) in terms of the comoving coordinate r instead of σ_0

$$ds^2 = -c^2dt^2 + a(t)^2\left[dr^2 + S_k(r)^2 d\Omega^2\right] \tag{3.21}$$

This is the form of the FRW metric we will most often use, as the proper radial distance is easy to calculate

$$d_p = \int_0^r a(t)dr = a(t)r \tag{3.22}$$

In other words, for a galaxy at comoving coordinate r, the proper distance now is simply $d_p(t_0) = r$ and at some other time the proper distance is just scaled by the scale factor, $d_p(t) = ra(t)$. Figure 3.2 provides a 2D analog to help understand the distinction between the comoving coordinates r and σ_0.

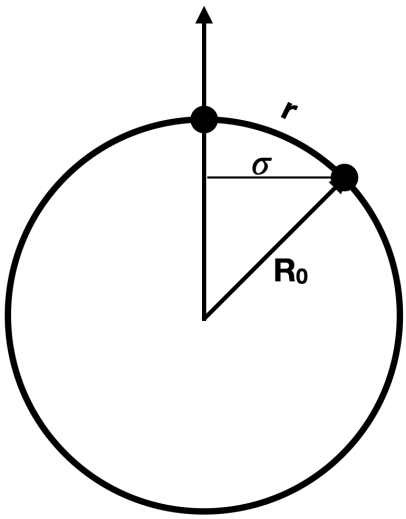


Figure 3.2: 2-dimensional analogue to help explain the coordinates r and σ_0 of the RW metric. By convention we label the north pole as $r = \sigma_0 = 0$. We can label some other point on the surface using either coordinate r or σ. The r coordinate is convenient since it gives the distance from the origin measured along the surface. The utility of the σ_0 coordinate is that it gives the circumference of a circle drawn around the ϕ direction, i.e., $C = 2\pi\sigma_0$. From the geometry, the two coordinates are related by $\sigma_0 = R_0 \sin(r/R_0)$. If we were to consider the surface of a hyperbola instead of a sphere, then the relationship would be $\sigma_0 = R_0 \sinh(r/R_0)$

3.3 *Redshift and Hubble's Law*

The first detection of the expansion of the universe came from observing the redshift of light from distance sources. Consider a photon emitted by a source moving away from you at speed v. If in the source frame the photon has wavelength λ_e, you will observe a photon with a wavelength, λ_0, that is modified by the *Doppler shift* formula[6]

$$\lambda_0 = \left(1 + \frac{v}{c}\right)\lambda_e \tag{3.23}$$

For such a source moving away from you, $\lambda_0 > \lambda_e$ and so the observed light is shifted to longer, or redder, wavelengths. We define the *redshift*, z, of the photon as

$$z = \frac{\lambda_0. - \lambda_e}{\lambda_e} = \frac{v}{c} \tag{3.24}$$

The redshift of light from a cosmic source like a galaxy is often readily observable by measuring the observed wavelength of known atomic line features (e.g., the Lyman alpha atomic line, which occurs at $\lambda_e \approx 912$ angstroms). Edwin Hubble observed the photons from (not too distant) galaxies and found that they were systematically redshifted, with the redshift proportional to the distance to the galaxy

$$z = \frac{H_0}{c}d_p \quad \text{or} \quad v = zc = H_0 d_p \tag{3.25}$$

where d_p is the proper distance to the galaxy, and H_0 is a constant called the *Hubble constant*. In other words, galaxies appear to be moving away from us with a speed v that is proportional to the proper distance to them. This was the first evidence of the expansion of the

[6] The relativistic expression for Doppler shift is, in general

$$\lambda_0 = \gamma(1 - \frac{v}{c}\cos\theta)$$

where $\gamma = (1 - v^2/c^2)^{1/2}$ is the Lorentz factor, and θ is the angle between the source motion and the emission direction. Often we consider cases where $v \ll c$ and so $\gamma \approx 1$, and the source is moving away from us, so $\theta = 180^o$ and so $\cos\theta = -1$. Then the Doppler shift formula reduces to Eq. 3.23.

universe. We will see later that this is just a linear approximation to the full relationship between redshift and distance, $d_p(z)$.

If we consider a galaxy at a fixed comoving coordinate r away from us is $d_p = a(t)r$, and so if we use the variable v to describe the rate at which this proper distance changes, then

$$v = \frac{d}{dt}d_p = \frac{d}{dt}a(t)r = \frac{da}{dt}r \tag{3.26}$$

And so $v = \dot{a}\, r$, where a dot over a variable indicates a time derivative. Hubble's law, $v = H_0 d_p$ can then be written

$$\dot{a}(t_0)r = H_0 a(t_0)r \implies H_0 = \frac{\dot{a}(t_0)}{a(t_0)} \tag{3.27}$$

So the Hubble constant tells us the rate at which the scale factor is changing today. More generally, we can define the *Hubble parameter* at a time t as

$$H(t) = \frac{\dot{a}(t)}{a(t)} \tag{3.28}$$

The Hubble parameter describes the rate at which the scale factor changes, *scaled by the value of the scale factor at that time*.

These days, we do not usually think of the galaxies as moving away from us, but rather think of the spacetime of the universe expanding, as described by the scale factor $a(t)$. The wavelength of a photon can be imagined as "stretching" along with the scale factor. A proper analysis using the FRW metric (see Ryden, section 3.4) shows that for a photon emitted with wavelength λ_e at time t_e and observed at time t_0 with wavelength λ_0

$$\frac{\lambda_e}{a(t_e)} = \frac{\lambda_0}{a(t_0)} \tag{3.29}$$

Given the definition of the redshift, z, (Eq. 3.24) this equation implies

$$1 + z = \frac{a(t_0)}{a(t_e)} \tag{3.30}$$

Very often the observed time is now and by definition $a(t_0) = 1$, so we can write the above relation as

$$1 + z = \frac{1}{a} \qquad \text{(observed today)} \tag{3.31}$$

where it is understood that by a we mean $a(t_e)$. That is, observations of the redshift z tells us directly how much the universe has expanded since the time the photon was emitted.

3.4 *Matter-Energy in the Expanding Universe*

The dynamical evolution of the universe (i.e., the time dependence of the scale factor $a(t)$) will turn out to be determined by the energy-density content of the universe. We will describe the contents of

the universe as a perfect fluid with energy density ϵ and pressure P which, given the assumption of homogeneity, have the same value everywhere in space. A component of energy will evolve in time in a way given by the *fluid equation*[7]

$$\frac{d\epsilon}{dt} = -3\frac{\dot{a}}{a}(\epsilon + P) \tag{3.32}$$

From the cosmic dynamics perspective, different types of energy are distinguished by the relationship between their energy density and pressure, which is called the *equation of state*

$$P = w\epsilon \tag{3.33}$$

We will usually consider cases where the equation of state parameter, w, is a dimensionless constant[8].

From the fluid equation and the equation of state, we can determine how the energy density of some component varies as the universe expands. Replacing $p = w\epsilon$ and writing out $\dot{a} = da/dt$ the fluid equation can be writtine

$$\frac{d\epsilon}{da} = -\frac{\epsilon}{a}3(1+w) \tag{3.34}$$

When w is constant, this can be solved by seperating variables and integrating

$$\int \frac{d\epsilon}{\epsilon} = -3(1+w)\int \frac{da}{a} \tag{3.35}$$

Which has a solution

$$\epsilon = \epsilon_0 a^{-3(1+w)} \qquad \text{constant } w \tag{3.36}$$

where we use the standard notation that the subscript 0 means that quantities are evaluated at the present time, t_0, i.e., $\epsilon_0 = \epsilon(t_0)$. The equation of state parameter thus determines how a certain component of energy changes as the universe expands. We will see below that w also affects how the scale factor $a(t)$ changes with time.

We will consider three different types of energy content of the universe:

Matter: In the context of cosmic expansion, "matter" refers to a pressureless fluid, with $w = 0$. The energy density of the fluid is dominated by the rest mass energy of the particles composing it, or $\epsilon = \rho c^2$, where ρ is the mass density. The matter energy density drops as the volume of the universe expands, and scales (by Eq. 3.36) as $\epsilon_m \propto a^{-3}$. We will often split up the matter component into *dark matter*, and *baryons*[9].

To further understand the equation of state of matter, imagine that the matter is ideal fluid at temperature T and number density n. The

[7] The fluid equation can be derived from the first law of thermodynamics

$$\frac{dE}{dt} = -P\frac{dV}{dt}$$

where the total energy, $E = \epsilon V$ where V is the volume. This equation is an expression of conservation of energy when the only change in energy comes from the work done ($P\,dV/dt$) by the system (or on the system). Plugging in $E = \epsilon V$ we get

$$V\frac{d\epsilon}{dt} + \epsilon\frac{dV}{dt} = -P\frac{dV}{dt}$$

and rearranging gives

$$\frac{d\epsilon}{dt} = -\frac{1}{V}\frac{dV}{dt}(\epsilon + P)$$

Writing the volume of some cube of space $V = r^3a(t)^3$ and taking the time derivative we get $dV/dt = 3\dot{a}a(t)^2r^3$. Plugging this in gives

$$\frac{d\epsilon}{dt} = -3\frac{\dot{a}}{a}(\epsilon + P)$$

which is the fluid equation. Although we derived this starting from the first law of thermodynamics, it actually arises from Einstein's equations of GR.

[8] More generally, w, need not be constant and may be a function of ϵ. This does not really change the approach we use for calculating for the dynamics of the Universe, but the integrations become more difficult and generally have to be done numerically.

[9] By baryons we mean protons and neutrons (since other types of baryons are unstable and will decay away) and often we lump electrons into this component since they combine with the protons and neutrons to make atoms (although the electrons are leptons, not baryons, and their mass is much less than that of protons, so they don't contribute much to the rest mass energy density.

pressure is given by the ideal gas law $P = nk_BT$ (where k_B is the Boltzmann constant) and the kinetic energy density is $\epsilon = \frac{3}{2}nk_BT$ (i.e., the average energy of a gas particle is $3/2k_BT$). The equation of state parameter is then

$$w = \frac{P}{\epsilon} = \frac{nk_BT}{\frac{3}{2}nk_BT + nmc^2} \tag{3.37}$$

where we included the rest mass energy density of the gas, nmc^2 (where m is the mass of the particles composing the fluid) since rest mass energy contributes to cosmic dynamics. Dividing top and bottom by nk_BT gives

$$w = \frac{1}{\frac{3}{2} + mc^2/k_BT} \tag{3.38}$$

We see that when $k_BT \ll mc^2$, the equation of state parameter $w \approx 0$. This is basically the assumption we are making when we talk about "cold matter" in the universe[10]. Since k_BT is roughly the average kinetic energy of particles, this condition $k_BT \ll c^2$ implies $\frac{1}{2}mv^2 \ll mc^2$ which means $v \ll c$, that is the particle of are non-relativistic. In the opposite limit, $k_BT \gg mc^2$, the particles become relativistic ($v \approx c$) and the non-relativistic ideal gas law no-longer applies. We next discuss this limit of *radiation*.

[10] Of course, the temperature of the universe changes as it expands and cools. Thus, a fluid that is currently cold and pressureless (i.e., has $k_BT \ll mc^2$) may have at some point earlier in the universe been hot and have non-zero pressure (i.e., at some point in the past $k_BT \gg mc^2$). Thus in certain cases where we are examining a fluid over cosmic time, we may have to be a bit careful about our assumption $w = 0$. We generally won't worry about this when we solve the Friedmann equations of cosmic dynamics, but is something to keep in mind.

Radiation: In this context, radiation refers to a fluid made up of relativistic particles, i.e., particles with $k_BT \gg mc^2$. The significant components of radiation in the universe are photons (which are massless) and neutrinos (which have a small, though non-zero rest mass). For photons in a thermal (i.e., blackbody) distribution at temperature T, the energy and pressure are given by[11]

$$\epsilon = a_RT^4 \qquad P = \frac{1}{3}a_RT^4 \tag{3.39}$$

where a_R is the *radiation constant* (not to be confused with the scale factor $a(t)$; here a_R is just a physical constant). The equation of state parameter is thus $w = 1/3$ for radiation. The energy density of radiation scales as $\epsilon_r \propto a^{-4}$ since in addition to the volume of the universe expanding, the wavelength of photons gets stretched by a factor of $a(t)$, and so the energy of each photon drops by a factor of $a(t)$, since $E = h\nu = hc/\lambda \propto a^{-1}$.

[11] The relation $\epsilon = a_RT^4$ holds for bosons (i.e., particles with integer values of spin) such as photons. For fermions (particles with half-integer values of spin) like neutrinos, the expression is $\epsilon = \frac{7}{8}a_RT^4$. The pressure of a fermion fluid is $P = \epsilon/3$, just like for bosons.

Dark Energy: For fluids composed of ordinary particles, the above results indicate that the limiting values of the equation of state parameter are $w = 0$ for "cold" (i.e., non-relativistic) matter and $w = 1/3$ for "hot" (i.e., relativistic) radiation. For "warm" matter in-between these limits (i.e., $k_BT \sim mc^2$) the equation of state parameter will be between $w = 0$ and $w = 1/3$.

However, certain more unusual forms of energy – in particular, quantum scalar fields – can have have w values outside the range 0 to 1/3. In fact, such scalar fields can have *negative* pressure, such that $w < 0$. We will see that a fluid with equation of state parameter $w < -1/3$ has effectively a repulsive gravitational effect in cosmology, and so serves to *accelerate* the expansion of the universe. Any substance with $w < -1/3$ is typically called "dark energy". Of particular interest is the cosmological constant, Λ, which has $w = -1$, and has an energy density that stays constant as the universe expands, $\epsilon_\Lambda \propto a^0$.

The "dominant energy condition" of general relativity, specifies that the maximum value of the pressure is $|P| \leq \epsilon$ which implies $w \leq 1$. One motivation for this is that the (isothermal) sound speed of a perfect fluid can be shown to be

$$v_s = c\sqrt{\frac{|P|}{\epsilon}} = c\sqrt{|w|} \tag{3.40}$$

so if $|w| > 1$ sound waves would move faster than the speed of light, which would pose causality problems. Thus we typically only consider substances with $-1 \leq w \leq 1$, although there are some interesting theoretical explorations into what would happen if the universe had a component which violated the dominant energy condition[12].

[12] You explored one such example in the homework, the case of "phantom energy" with $w < -1$ which led to a "big rip".

component	w	ϵ	scaling
matter (non-relativistic particles)	0	ρc^2	$\epsilon \propto a^{-3}$
radiation (relativistic boson particles)	1/3	$a_R T^4$	$\epsilon \propto a^{-4}$
radiation (relativistic fermion particles)	1/3	$\frac{7}{8} a_R T^4$	$\epsilon \propto a^{-4}$
cosmological constant	-1	$\frac{c^2}{8\pi G}\Lambda$	$\epsilon \propto a^0$

3.5 *Equations of Cosmic Dynamics*

The dynamical evolution of a homogenous, isotropic Universe is determined by solving the Einstein equations of General Relativity, which leads to the *Friedman equation* describing the evolution of the scale factor $a(t)$

$$\frac{\dot{a}(t)^2}{a(t)^2} = \frac{8\pi G}{3c^2}\epsilon(t) - \frac{\kappa c^2}{R_0^2 a(t)^2} \tag{3.41}$$

where the dot above a variable indicates a time derivative. This equation expresses the relationship among the dynamics of the universe (described by the Hubble parameter $H = \dot{a}/a$), the energy density of the universe (described by $\epsilon(t)$) and the curvature of the universe (described by κ and R_0).

It is often useful when approaching an equation to look for characteristic quantities and rescale the variables. Since the Hubble parameter is defined by $H(t) = \dot{a}/a$, if we divide both sides by $H(t)^2$ we get

$$1 = \frac{8\pi G}{3c^2H^2}\epsilon(t) - \frac{\kappa c^2}{R_0^2 a^2 H^2} \tag{3.42}$$

Now that we have made the left hand side of this equation dimensionless, each term on the right hand side of the equation must also be dimensionless. It follows that the jumble of constants in front of the $\epsilon(t)$ must have units of one over energy density, so we define a "critical energy density"

$$\epsilon_c(t) = \frac{3c^2H(t)^2}{8\pi G} \tag{3.43}$$

We will see why this is called the "critical" energy density shortly. Having found a characteristic energy density of the problem, it makes sense to rescale the actual energy density ϵ by this value, so we define the "dimensionless" (or "scaled") energy density

$$\Omega(t) = \frac{\epsilon(t)}{\epsilon_c(t)} \tag{3.44}$$

We can then write the Friedmann equation as

$$H^2 = H^2\Omega(t) - \frac{\kappa c^2}{R_0^2 a^2} \tag{3.45}$$

Rearranging this makes apparaneter how the curvature of the universe is related by the Friedmann equation

$$\frac{kc^2}{R_0^2 a^2} = H^2(\Omega - 1) \tag{3.46}$$

Evaluating this equation at the present time, when $a(t_0) = 1$ and $H(t_0) = H_0$ (where H_0 is the Hubble *constant*) and $\Omega(t_0) = \Omega_0$ (where Ω_0 is the scaled energy density today) we find

$$H_0^2 = H_0^2\Omega_0 - \frac{\kappa c^2}{R_0^2} \tag{3.47}$$

and rearranging we find

$$\frac{\kappa c^2}{R_0^2} = H_0^2(\Omega_0 - 1) \tag{3.48}$$

from which we see that the curvature of the universe is determined by the scaled energy density. If $\Omega_0 > 0$ the curvature is positive ($\kappa = +1$). If $\Omega_0 < 0$ the curvature is negative ($\kappa = -1$) and if $\Omega_0 = 1$ the curvature is zero ($\kappa = 0$).

Using Eq. 3.48 to replace the curvature term in the Friedmann equation Eq. 3.45, we can rewrite the Friedmann equation as

$$H^2 = H^2\Omega(t) + H_0^2(1-\Omega_0)a^{-2} \tag{3.49}$$

We should be a little careful with how $\Omega(t) = \epsilon(t)/\epsilon_c(t)$ evolves with time, since both $\epsilon(t)$ and $\epsilon_c(t)$ change in distinct ways over time. The value of the critical density now is

$$\epsilon_c(t_0) = \frac{3c^2H_0^2}{8\pi G} \tag{3.50}$$

So the value of the critical density at some other time can be written

$$\epsilon_c(t) = \frac{3c^2H(t)^2}{8\pi G} = \epsilon_{c,0}\frac{H(t)^2}{H_0^2} \tag{3.51}$$

For matter, we found previously the scaling $\epsilon_m = \epsilon_{m,0}a^{-3}$. It follows that

$$\Omega_m(t) = \frac{\epsilon_m(t)}{\epsilon_c(t)} = \frac{\epsilon_{m,0}}{a(t)^3}\frac{H_0^2}{\epsilon_{c,0}H(t)^2} \tag{3.52}$$

and since the scaled energy density of matter today is $\Omega_{m.0} = \epsilon_{m.0}/\epsilon_{c,0}$ this can be written as

$$\Omega_m(t) = \frac{\Omega_{m,0}}{a(t)^3}\frac{H_0^2}{H(t)^2} \tag{3.53}$$

Plugging this into the scaled form of the Friedmann equation (Eq. 3.49) then gives

$$H^2 = \frac{\dot{a}^2}{a^2} = H_0^2\left[\frac{\Omega_{m,0}}{a^3} + \frac{1-\Omega_{m,0}}{a^2}\right] \quad \text{(matter only)} \tag{3.54}$$

If we want to include other components in our universe, we can similarly write how the scaled energy densities of radiation and lambda change

$$\Omega_r(t) = \frac{\Omega_{r,0}}{a(t)^4}\frac{H_0^2}{H(t)^2} \qquad \Omega_\Lambda(t) = \Omega_{\Lambda,0}\frac{H_0^2}{H(t)^2} \tag{3.55}$$

And so the Friedmann equation can be written

$$H^2 = \frac{\dot{a}^2}{a^2} = H_0^2\left[\frac{\Omega_{m,0}}{a^3} + \frac{\Omega_{r,0}}{a^4} + \Omega_{\Lambda,0} + \frac{1-\Omega_{m,0}}{a^2}\right] \tag{3.56}$$

where $\Omega_0 = \Omega_{m,0} + \Omega_{r,0} + \Omega_{\Lambda,0}$.

For a substance with arbitrary equation of state parameter, w, (assumed to be a constant) the scaled energy density is

$$\Omega_w(t) = \frac{\Omega_{w,0}}{a(t)^{3(1+w)}}\frac{H_0^2}{H(t)^2} \tag{3.57}$$

And so a general form of multiple components which different values of w is

$$H^2 = \frac{\dot{a}^2}{a^2} = H_0^2 \left[\sum_i \frac{\Omega_{i,0}}{a^{3(1+w_i)}} + \frac{1-\Omega_0}{a^2} \right] \tag{3.58}$$

$$\text{where } \Omega_0 = \sum_i \Omega_{0,i} \tag{3.59}$$

This equation can be solved analytically for certain simple scenarios (e.g., single component flat universes). More generally it can be solved numerically.

3.6 *Effective Potential Approach*

We can get some intuition into the solutions to the Friedmann equation by using an "effective potential" approach. Consider first the Friedmann equation for matter only written in the form

$$\frac{\dot{a}^2}{a^2} = H_0^2 \left[\frac{\Omega_{m,0}}{a^3} + \frac{1-\Omega_{m,0}}{a^2} \right] \tag{3.60}$$

If we multiply this equation through by $a^2/2$ and move the first term on the right hand side to the left, we find

$$\frac{1}{2}\dot{a}^2 - \frac{H_0^2 \Omega_{m,0}}{2a} = \frac{1}{2} H_0^2 (1-\Omega_0) \tag{3.61}$$

The mathematical form of this equation is exactly the same as the Newtonian equation for the energy (per unit mass) of a mass rolling in a potential well[13]. To be even more explicit, we can write the equation as

$$\frac{1}{2}\dot{a}^2 + V_{\text{eff}}(a) = E_{\text{eff}} \tag{3.62}$$

where the first term on the left hand side is like the "kinetic energy", the second term

$$V_{\text{eff}} = -\frac{H_0^2 \Omega_{m,0}}{a} \tag{3.63}$$

is like an "effective potential" term, and $E_{\text{eff}} = H_0^2(1-\Omega_0)/2$ is a constant which plays the role of the "total energy".

Because this equation is mathematically identical in form to the energy equation, we can intuit the solutions by thinking about how a ball rolls around in a potential well.

$$V_{\text{eff}} = -\frac{H_0^2}{2}\left[\frac{\Omega_{m,0}}{a}\right] \quad \text{(matter only universe)} \tag{3.64}$$

Figure 3.3 plots the effective potential for a matter only universe. The initial conditions of a "big bang" corresponds to the ball being "shot out" like a pinball from $a = 0$ at $t = 0$. The ball will then decelerate as it climbs up the potential well. There are two possible outcomes: if

[13] Of course, we are not using Newtonian mechanics here; the Friedmann equation comes out of GR. But the point is that the equation we have is mathematically identical in form to a Newtonian energy equation that we can intuit easily, and this helps us "see" the solutions to the Friedmann equation in an intuitive way.

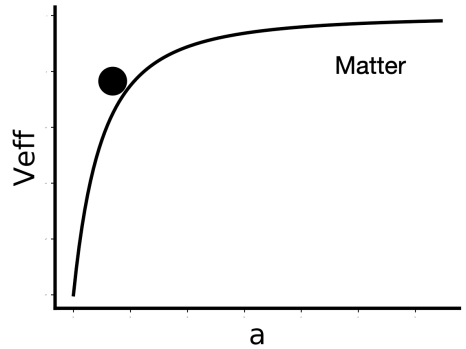


Figure 3.3: Effective potential for a matter dominated universe. The ball here is a marker for how the scale factor $a(t)$ changes with time.

$E_{\text{eff}} < 0$ (i.e., $1 - \Omega_0 < 0$) the expansion is "bound" and the ball will eventually halt, turn around, and collapse back to the center in a "big crunch". If $1 - \Omega_0 > 0$ the expansion is "unbound" and the universe will continue growing and cooling forever in a "big chill". The case $\Omega_0 = 1$ is the crossover point, where the universe will *just* barely expand to infinity.

For a universe with only a dark energy (specifically a cosmological constant Λ, with $w = -1$) the effective potential term is

$$V_{\text{eff}} = -\frac{H_0^2}{2}\left[\Omega_{\Lambda,0} a^2\right] \qquad \text{(matter only universe)} \tag{3.65}$$

which is plotted in Figure 3.4. We see that if the universe is initially expand, it will continue to expand forever and at an *accelerated* rate.

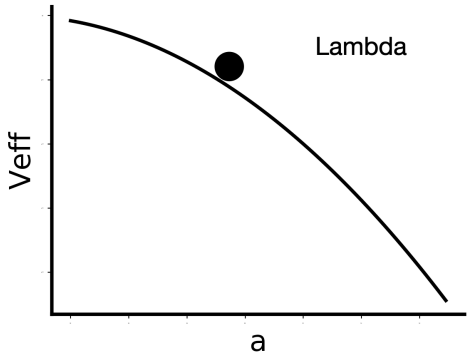


Figure 3.4: Effective potential for a positive lambda (cosmological constant) dominated universe.

For a universe with *both* matter and a cosmological constant, the effective potential is

$$V_{\text{eff}} = -\frac{H_0^2}{2}\left[\frac{\Omega_{m,0}}{a} + \Omega_{\Lambda,0} a^2\right] \tag{3.66}$$

which is plotted in Figure 3.5. This potential allows for multiple possibilities; we notice an equilibrium point on the top of the potential "hill" where in principle $a(t)$ could remain fixed. This was Einstein's motivation for introducing the cosmological constant so that he could produce a static universe. However, we can easily see that this equilibrium point is *unstable* and slight perturbations will lead either to expansion or contraction. Therefore Einstein's static universe had a serious flaw. If you tuned the parameters just right, you could get a universe that perhaps stalled for some period of time at the top of the hill, but eventually it would either recollapse or expand out (such a scenario is sometimes called a *loitering universe*).

There are several possible qualitative behaviors for the universe with both matter and dark energy shown in Figure 3.5.

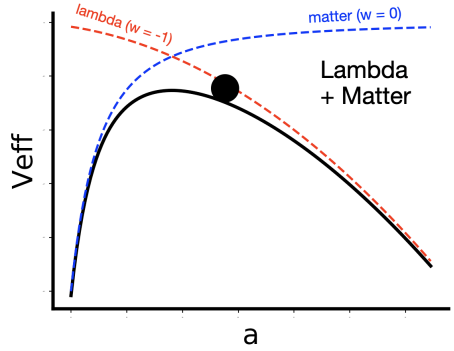


Figure 3.5: Effective potential for a lambda (cosmological constant) plus matter universe.

1. Big Crunch – The universe starts at $a = 0$, but is not expanding fast enough to make it over the potential "hill", and so eventually halts, turns around and collapses back to $a(t) = 0$.

2. Big Chill – The universe starts at $a = 0$, and makes it over the top of the hill, then proceeds to accelerate at a faster and faster rate, under the influence of dark energy.

3. Static or Loitering universe – the scale factor of the universe is either initially at the equilibrium point at the top of the potential hill, or it begins at $a = 0$ with an expansion rate just tuned such that the universe *almost* comes to a halt at the peak, before either contracting inward or expanding outward.

4. Big Bounce – the scale factor starts at $a(0) \gg 1$ and is contracting inward. The ball now approaches the potential hill from the right; as it starts to climb the potential hill it is not fast enough to get over it, therefore it stalls, "bounces" and starts expanding outward. While the initial conditions here might seem strange (the universe starting out infinitely large and shot inward in a contraction) at least they avoid the singularity of the standard big bang model which starts out at $a(0) = 0$.

Cosmological observations (e.g., of Type Ia supernovae) indicate that we are in a universe with both matter and dark energy, and have just recently passed the "peak" of the potential hill, such that dark energy now dominates and the cosmic expansion is accelerating, with our fate being a "big chill".

3.7 *Proper Distance to redshift z*

Say we observe the light from a distance galaxy at redshift z. How far away is this galaxy now, and how long did it take to get to us? To answer this, we can use the fact that light follows a null geodesic where $ds^2 = 0$; the RW metric (written in terms of the r coordinate) then becomes

$$ds^2 = -c^2dt^2 + a(t)^2\left[dr^2 + R_0^2 \,\mathrm{sink}^2\left(\frac{r}{R_0}\right) d\Omega^2\right] = 0 \tag{3.67}$$

if the light travels radially, $d\Omega = 0$, this equation reduces to

$$dr = \pm c\frac{dt}{a(t)} \tag{3.68}$$

where the plus sign applies if light moves in the positive r direction, while the minus sign applies if light moves in the negative r direction. Imagine that a light beam was emitted by a galaxy at comoving coordinate r at time t_e and was received by us at a coordinate $r = 0$ at time t_0. The light is moving in the negative r direction, so integrating both sides of the above we find

$$\int_r^0 dr = -c\int_{t_e}^{t_0} \frac{dt}{a(t)} \implies r = c\int_{t_e}^{t_0} \frac{dt}{a(t)} \tag{3.69}$$

To do this integral over dt we would need to know the time dependence of the scale factor, $a(t)$. Alternatively, we can change variables from dt to da using the chain rule identity

$$dt = dt\left(\frac{da}{da}\right) = da\frac{dt}{da} = \frac{da}{\dot{a}} \tag{3.70}$$

Using this in the integral above we have

$$r = c\int_{a_e}^{a_0} \frac{da}{a}\frac{1}{\dot{a}} = c\int_{a_e}^{a_0} \frac{da}{a^2}\frac{a}{\dot{a}} = c\int_{a_e}^{a_0} \frac{da}{a^2}\frac{1}{H} \tag{3.71}$$

where when changing the integral to one over a we also changed the limits to $a_e = a(t_e)$ and $a_0 = a(t_0)$. It is often convenient to write this as an integral over z instead of a. From the relationship $1 + z = 1/a$ we have

$$dz = -\frac{da}{a^2} \tag{3.72}$$

and so the integral becomes

$$r(z) = c \int_0^z \frac{dz}{H(z)} \tag{3.73}$$

To expression for $H(z)$ is given by the Friedmann equation, which can write by replacing $1/a = (1 + z)$

$$H(z)^2 = H_0^2 \left[\Omega_{m,0}(1+z)^3 + \Omega_{r,0}(1+z)^4 + \Omega_{\Lambda,0} + (1 - \Omega_0)(1+z)^2 \right] \tag{3.74}$$

The proper distance is related to the comoving coordinate as $d_p = a(t)r$. At the present time when $a(t_0) = 1$, the proper distance from us to a source at redshift z is then

$$d_{p.0}(z) = r(z) = c \int_0^z \frac{dz}{H(z)} \tag{3.75}$$

Thus the proper distance to a source at redshift z depends on the expansion history of the universe (reflected by the integral over $H(z)$.

3.8 Distance-Redshift Relation

If we knew the distance to some cosmic light source at redshift z, we could presumably constrain the expansion history of the universe, since the distance in Eq. 3.75 is given by an integral over $H(z)$. We usually can easily measure the redshift of an astronomical source (by measuring the the wavelength shift of atomic spectral lines). But how are we going to determine the distance to the object?

A common approach to measuring distances is the *standard-candle method*. If the intrinsic luminosity, L_0 of a certain kind of source is known, then the flux we measure on earth is $F = L_0/A$, where L_0 is the observed luminosity and A is the area over which that luminosity has spread out. For a flat, static universe $A = 4\pi d^2$ and so

$$F = \frac{L_0}{4\pi d^2} \implies d = \sqrt{\frac{L_0}{4\pi F}} \quad \text{(flat, static space)} \tag{3.76}$$

Thus, if we knew L_0 and measured F we could calculate d. Two cosmological effects, however, modify this simple relation. First, in a curved spacetime the area of a sphere is not $4\pi d^2$ but is instead (by integrating the angular part of the FRW metric)

$$A = 4\pi R_0^2 \, \text{sink}^2 \left(\frac{r}{R_0} \right) \tag{3.77}$$

where

$$\text{sink}(x) = \begin{cases} \sin(x) & \text{if } \kappa = +1 \\ x & \text{if } \kappa = 0 \\ \sinh(x) & \text{if } \kappa = -1 \end{cases} \tag{3.78}$$

Second, as the photons from the source travel through space, their energy is reduced because expansion of the universe stretches the wavelength by a factor of $(1+z)$. The expansion also dilates the rate at which photons arrive at us by another factor of $(1+z)$. Thus the luminosity observed now (L_0) is related to the intrinsic source luminosity measured at the source (L) by

$$L_0 = \frac{L}{(1+z)^2} \tag{3.79}$$

Putting these effects together, the observed flux $F = L_0/A$ is written

$$F = \frac{L}{4\pi R_0^2 \,\text{sink}^2\,(r/R_0)} \frac{1}{(1+z)^2} \tag{3.80}$$

Now we *define* the *luminosity distance* to be a quantity that relates flux and luminosity in the ordinary way

$$F = \frac{L}{4\pi d_L^2} \implies d_L = \sqrt{\frac{L}{4\pi F}} \tag{3.81}$$

Comparing with Eq. 3.80 we see that

$$d_L = R_0 \,\text{sink}\left(\frac{r}{R_0}\right)(1+z) \tag{3.82}$$

So d_L accounts for the affects of curvature and photon stretching.

We would like to find an expression for the luminosity distance to a source at redshift z. To do so, we first use Eq. 3.48 to rewrite the radius of the universe, R_0

$$-\frac{\kappa c^2}{R_0^2} = H_0^2(1-\Omega_0) \implies R_0 = \frac{c}{H_0|1-\Omega_0|^{1/2}} \tag{3.83}$$

Second we use Eq. 3.73 to replace the comoving coordinate $r(z)$. Putting it all together we have

$$d_L(z) = \frac{c(1+z)}{H_0|1-\Omega_0|^{1/2}} \text{sink}\left[H_0|1-\Omega_0|^{1/2}\int_0^z \frac{dz}{H(z)}\right] \tag{3.84}$$

This expression appears to give division by zero for a flat universe when $1-\Omega_0 = 0$. However, recall that for a flat universe $\text{sink}(x) = x$, so in the flat case the luminosity distance is

$$d_L(z) = c(1+z)\int_0^z \frac{dz}{H(z)} \qquad \text{(flat universe)} \tag{3.85}$$

Comparing to our previous equation for proper distance (Eq. 3.75) we see that for a flat universe, $d_L = (1+z)d_p$. The factor of $1+z$ difference is a result of time dilation and energy loss of photons in the expanding universe. For non-flat universe, the curvature effects lead to the more complicated expression for $d_L(z)$ of Eq. 3.84.

If we measure the redshift z, and flux F of a source of known intrinsic luminosity L, we can determine the luminosity distance of the source d_L. Doing this for multiple sources allows us to observationally map out $d_L(z)$ which will then constrain the cosmological parameters $(\Omega_{m,0}, \Omega_{r,0}, \Omega_{\Lambda,0})$, that appear in the expression (Eq. 3.84) through $H(z)$. An object of known intrinsic luminosity is called a *standard candle*. Type Ia supernovae were used as standard candles to map out the expansion history of the universe, finding that the expansion was currently accelerating.

4
Thermal Evolution of the Universe

4.1 The Cosmic Soup (optional)

So far we have described components of the universe (e.g., matter, radiation, lambda) by their energy density, ϵ (assumed to have the same value everywhere in space) and equation of state parameter, w. These bulk properties are what are relevant in the dynamical equations of the universe (the Friedmann equations) but we'd like to develop a more detailed picture of the makeup of these components and how they came to be.

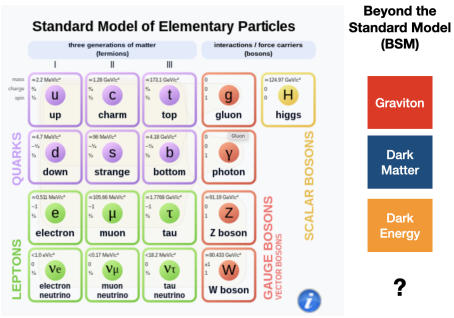


Figure 4.1: The particles of the standard model. We suspect there are additional particles beyond the standard model. For example, the dark matter may be some new particle (or particles) not included in the standard model.

According to the Standard Model of physics, our universe is built out of the fundamental particles shown in Figure 4.1. There are six types of *leptons* (from the Greek word for "light" or "thin", since they are relatively light) and six types of quarks. There are also particles that act as mediators of interactions (called "gauge" bosons"); photons are mediators of the electromagnetic force, the W and Z bosons are mediators of the weak force, and gluons are mediators of the strong force. The Standard Model also adds the Higgs boson, which plays a role in setting particle's mass.

Given the bewildering number[1] in the standard model, a microscopic description of the universe seems unbearably complicated. Fortunately, things wind up simpler in practice. At high temperatures $\gtrsim 10^{12}$ K, one can have a soup of all particles in the standard model. But at lower temperatures quarks and gluons do not exist in isolation, but always clump together into composite particles (called hadrons). A hadron composed of an even number of quarks is called a *meson* (e.g., a u quark and a $\bar{d}$ quark form a π^+-meson). A hadron composed of an odd number of quarks is called a *baryon* (e.g., 2 u quarks and 1 d quark form a proton, while 2 d quarks and 1 u quark form a neutron).

Clearly, a huge number of composite particles can be enumerated (even 4-quark "tetraquark" mesons and 5-quark "pentaquark" baryons are possible). However, almost all of these hadrons are unstable and

[1] The standard model includes 17 types of fundamental particles (6 leptons + 6 gauge quarks + 4 gauge bosons + Higgs boson). A more careful accounting gives a larger number: each lepton has an antiparticle, giving $6 \times 2 = 12$ distinct leptons. Each quark can come in 3 different "colors" (which relates to their strong force interactions) and each has an antiparticle, giving $6 \times 3 \times 2 = 36$ quarks. The gluons come in 8 different "colors", while the W^+ boson has an antiparticle (the W^- boson). The Z-boson and photon their own antiparticle, so we have $8 + 2 + 1 + 1 = 12$ gauge bosons. Adding in the Higgs (which is its own antiparticle) the total comes to 12 + 36 + 12 + 1 = 61 distinct particles.

can decay to lighter particles. There is exactly one stable free hadron: the proton. Even a free neutron will decay via $n \rightarrow p + e^- + \bar{\nu}_e$ with a half life of about 30 minutes. Neutrons can be stable, though, when they are bound together with protons in an atomic nucleus[2].

Similarly, the only stable leptons are electrons and neutrinos (and their antiparticles). The heavy leptons τ and μ leptons can decay, for example, via $\tau^- \rightarrow e^- + \nu_\tau + \bar{\nu}_e$ and $\mu^- \rightarrow e^- + \nu_\mu + \bar{\nu}_e$. Thus, at low enough temperatures, the only main particles we are concerned with are electrons and neutrons, and protons and neutrons (plus any dark matter particles). Moreover, if we require the universe to be charge neutral, the number density of electrons must equal the number density of protons[3]

[2] Some atomic nuclei are unstable (i.e., radioactive) and the neutron will eventually beta decay into a more bound state. But for stable atomic nuclei, there is no accessible nucleon state to decay into, so the neutron is stuck there.

[3] This discussion may raise a deep question – why is the standard model so complicated, with so many different types of fundamental particles and interactions? And what sets the properties (e.g., masses, spins) of all these particles? The honest answer is, "We don't know...". Perhaps someday we will discover a deeper theory of physics that allows us to derive everything in Figure 4.1. Alternatively, some theories (e.g., string theory) suggest that there are many parallel universes (composing the "multiverse"), and each one randomly samples different sets of particles and interactions. A very simple universe – say, one with a single kind of particle – would presumably lack the complex structures (e.g., nuclei, atoms, molecules, etc...) that seem required for intelligent life. The fact there we are here to ask these questions implies we must be in one of the universes with a richer fundamental structure. There is then no predicting the properties of the fundamental particles, other than requiring that they fulfill this "anthropic" constraint. Such anthropic arguments are seen by some as compelling and natural, and by others as a pathetic resignation of the search for a deeper theory.

4.2 *The Temperature of the universe*

When we look out into space, we see a nearly uniform blackbody field of photons in all directions. This is called the *cosmic background radiation* or the *cosmic microwave background* (CMB). This light was produced in the earlier hot universe, and has been cooling with the expansion of the universe ever since.

We measure the blackbody temperature of the CMB today to be $T_0 \approx 2.7$ K, which implies that the energy density of the CMB today is $\epsilon_0 = a_R T_0^4$. We showed in the previous summary notes that as the universe expands, the energy density of radiation changes varies

$$\epsilon = \epsilon_0 a^{-4} = \epsilon_0 (1+z)^4 \tag{4.1}$$

So the temperature of the CMB at some other other epoch would be

$$a_R T^4 = a_R T_0^4 (1+z)^4 \implies T(z) = T_0(1+z) \tag{4.2}$$

or equivalently $T = T_0/a$. If we want to know how the temperature changes with time, we would have to solve the Friedmann equation to determine $a(t)$.

In the early universe, photons in the CMB could exchange energy with matter through absorption, scattering and emission processes. We should then expect that the matter and radiation will come into equilibrium at the same temperature, and so $T(z)$ describes the temperature of matter as well as radiation[4]. At the present time, the radiation is not well coupled to matter (see below) and so the temperature of matter can deviate from that of the CMB (as it clearly does in the gas in galaxies and stars). When we refer to the temperature of the universe, we usually mean the temperature of the CMB.

[4] A more careful treatment takes into account that photons are not the only type of radiation (i.e., relativistic particles) and we should write the energy density of radiation as

$$\epsilon_r = \frac{g_\star}{2} a_R T^4$$

where $g_\star$ is the effective statistical weight (or number of degrees of freedom). If we just have photons, which have two polarization states, then $g_\star = 2$. But if there are other relativistic particles (e.g., neutrinos), $g_\star$ is greater. As long as $g_\star$ is constant, it doesn't affect the temperature scaling $T = T_0/a$. But in general $g_\star$ may change over time as particles decay away, or cool and become non-relativistic. We won't worry about this effect in this class.

4.3 *Thermodynamic Equilibrium*

Given the number of different particles in the standard model and the diverse array of interactions between them, it would appear extremely complex to calculate the evolution of the components of the universe. Incredibly, when the rate of interactions is rapid enough to efficiently shuffle energy around between different particle types, a well-defined equilibrium distribution is established, and out of the complexity emerges simplicity. This is the foundational idea of statistical mechanics.

As an analogy, imagine that we distribute 100 units of "energy" (say 100 sugar cubes) to each of the students in the classroom. We then play a game where people walk around and "interact", such that every time you bump into someone you randomly exchange some energy (e.g., you can play a game of rock-paper-scissors and the loser gives the winner a sugar cube). After some time playing the game, we do not expect that everyone will remain with the exact same amount of energy. Rather, we expect a *distribution*, where some people have more and some people have less energy.

Initially, the distribution of energy among people will evolve as we play this energy-exchange game. But after some time we will reach a *steady-state*. Remarkably, we can write down a simple equation for the fraction of people with energy E in the steady state.

$$f(E) \propto e^{-E/k_B T} \quad \text{(Boltzmann distribution)} \tag{4.3}$$

When we reach this state, we say we have reached *thermodynamic equilibrium*. We will not prove the Boltzmann distribution here (refer to a statistical mechanics text) but it arises by calculating the distribution that maximizes the entropy.

The Boltzmann distribution is quite general. One of the challenges in statistical mechanics is to learn how to apply it in different contexts. In this course, we will consider only two cases – ideal classic gases that composed of either relativistic or non-relativistic particles (and with chemical potential $\mu = 0$). If there are reactions that effectively convert particles into each other such that energy is shuffled around and we reach thermodynamic equilibrium, then the number density of a certain type of particle is

$$n = g\, b_R T^3 \quad \text{(relativistic particles)} \tag{4.4}$$

$$n = \frac{g}{\lambda_T^3} e^{-mc^2/k_B T} \quad \text{(non-relativistic particles)} \tag{4.5}$$

where where b_R is a physical constant[5] and g is the statistical weight which counts the possible internal spin states of the particle (for spin 1/2 particles, $g = 2$). The factor λ_T is the *thermal DeBroglie wavelength*

[5] The constant b_R is different for bosons and fermions. For bosons, the constant is

$$b_R = \frac{\zeta(3)}{\pi^2} \frac{k_B^3}{\hbar^3 c^3} g \approx 20.2$$

For fermions, the constant is

$$b_R = \frac{3\zeta(3)}{4\pi^2} \frac{k_B^3}{\hbar^3 c^3} g \approx 15.2$$

where $\zeta(3) \approx 1.2$ is the Reimann zeta function and we used $g = 2$ for the statistical weight.

given by

$$\lambda_T = \frac{h}{\sqrt{2\pi k_B T}} \tag{4.6}$$

The above expressions for number density assume that chemical potential, μ, is zero (which we will always assume in this class). This assumption applies when particles can be created and destroyed without restriction, which will typically be the case in the cosmological situations we are interested in here[6].

[6] More generally, if there is some conservation law that restricts particles from being created or destroyed, then the chemical potential $\mu \neq 0$. For example, there are no processes that create or destroy molecules of air in a room, and therefore the number of air molecules will not simply be a function of of temperature as in Eq. 4.5. The equation for the number density of a non-relativistic ideal classical gas with non-zero chemical potential is

$$n = \frac{g}{\lambda_T^3} e^{-mc^2/k_B T} e^{\mu/k_B T}$$

We can solve this equation for μ

$$\mu = mc^2 + k_B T \ln\left(n\Lambda_T^3/g\right)$$

So in cases where the number density n is fixed (as in the air molecules in a room) we can calculate the chemical potential μ. On the other hand, in the case where particles are created and destroyed without restriction, we have $\mu = 0$ and we can solve for n based only on T.

4.4 *Condition for Thermodynamic Equilibrium and Decoupling*

In order for the expressions above for the number density n to apply, we need to be assured that the system is in thermodynamic equilibrium, i.e., that interactions are frequent enough to shuffle energy among different particles. How long does it take for thermodynamic equilibrium to be established? Well, say Δt_c is the average time between collisions between particles in which energy is shared. To shuffle energy around we need many collisions, so the timescale is several times Δt_c.

To calculate Δc, imagine a particle as a solid ball of radius R moving at speed v through a swarm of other particles of number density n (see Figure 4.2). After a time Δt the ball will have moved a distance $v\Delta t$ and so swept out a cylindrical volume $V = v\Delta t\sigma$ where $\sigma = \pi R^2$ is the cross-sectional area of the ball. The number of collisions experienced by the ball in this time is then just the number of target particles in this volume

$$N = nV = n\sigma v\Delta t \tag{4.7}$$

and the rate of collisions (i.e., number of collisions per unit time) is

$$\Gamma_c = \frac{N}{\Delta t} = n\sigma v \tag{4.8}$$

The average time between collisions is determined by asking how long Δt does it take to get on average one collision, $N = 1$. From Eq. 4.7 we have $N = 1$ when the time Δt is

$$\Delta t_c = \frac{1}{\Gamma_c} = \frac{1}{n\sigma v} \tag{4.9}$$

The cross-section of a microscopic particle is of course not given by the geometrical cross-section πR^2, but rather depends on the strength of the coupling force that mediates the interaction (e.g., the electromagnetic force, or the weak force). Such cross-sections can typically be calculated from the fundamental particle theory describing the interaction.

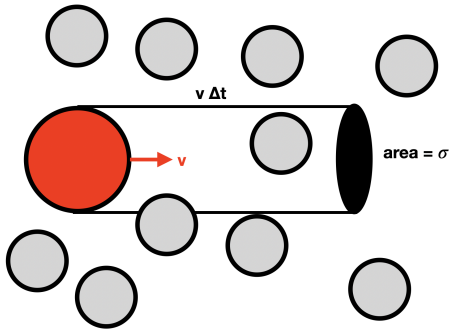


Figure 4.2: A particle moving at speed v and with cross-sectional area σ sweeps out a volume $V = v\Delta t\sigma$ in time Δt. The expected number of collisions experienced in this time is V times the number density n of the target particles

For particles to interact readily, the average time between collisions must be shorter than the timescale for the universe to expand (e.g., for the scale factor to double in size). Otherwise the space between particles will be growing faster than the particles can find each other and collide. As $\dot{a}$ is the rate of change of a, in a time Δt the scale factor changes by a factor

$$\Delta a \approx \dot{a}\, \Delta t \tag{4.10}$$

For the scale factor to double in size we want $\Delta a \approx a$, so

$$a \approx \dot{a}\, \Delta t \implies \Delta t = \frac{a}{\dot{a}} = \frac{1}{H} \tag{4.11}$$

So the time scale for the universe to expand (double in size) is just $\Delta t_{\rm ex} \approx 1/H$, where the Hubble parameter H is the relevant rate of expansion. The condition for particles to interact faster than the universe expands and hence be in thermodynamic equilibrium is

$$\Delta t_c \lesssim \Delta t_{\rm ex} \implies \frac{1}{n\sigma v} \lesssim \frac{1}{H} \tag{4.12}$$

We typically write this condition in terms of rates

$$\Gamma_c \gtrsim H \implies n\sigma v \gtrsim H \tag{4.13}$$

As the universe expands, the number density of particles n will decrease, and hence so will Γ_c. We thus expect to find a time when the condition $\Gamma_c \approx H$ is reached. After this, the rate of collisions is no longer sufficient to establish thermodynamic equilibrium, and we say that the particles have "decoupled" or "frozen out".

4.5 *Decoupling of the CMB*

Let us calculate when the CMB photons decoupled from matter. Prior to decoupling, the CMB photons are largely trapped in matter due to electron scattering. After decoupling, the photons can largely move freely. The decoupling thus represents the "release" of the CMB. The standard condition for decoupling is

$$n\sigma v \approx H \tag{4.14}$$

For the case of CMB photons scattering off of electrons, the speed of the photons is $v = c$, the cross-section is the Thomson cross-section $\sigma = \sigma_T$, and n is the number density of free electrons, given by

$$n_{\rm e,free} = x_{\rm ion} n_b = x_{\rm ion} n_{b,0} a^{-3} \tag{4.15}$$

where n_b is the number density of baryons[7], and X is the fraction

[7] We are assuming here that all baryons are protons, and since the universe is neutral, the number density of electrons is equal to that of protons. More generally, some of the baryons will be neutrons and so the results will be slightly modified.

of baryons (i.e., protons) that are ionized and so have released a free-electron. We have used the fact that the baryon number density scales like a^{-3} to write n_b in terms of the present number density of baryons $n_{b,0}$.

To determine the Hubble parameter, we use the Friedmann equation. Let's assume a flat universe (which is reasonable for our universe) in which case the Friedmann equation is

$$H^2 = H_0^2 \left[\Omega_{m,0} a^{-3} + \Omega_{r,0} a^{-4} + \Omega_{\Lambda,0}\right] \tag{4.16}$$

Let's guess that the decoupling of CMB photons happens when a is just small enough that the matter term dominates in this equation (but not so small that the radiation term dominates). We can check this assumption once we have solved for the a of decoupling. Ignoring the terms with $\Omega_{r,0}$ and $\Omega_{\Lambda,0}$ above, we find for the Hubble parameter is

$$H = H_0 \sqrt{\Omega_{m,0}} a^{-3/2} \tag{4.17}$$

Settng the expansion rate H equal to the collision rate Γ_c gives us our condition for decoupling

$$x_{\text{ion}} n_{b,0} a^{-3} \sigma_T c \approx H_0 \sqrt{\Omega_{m,0}} a^{-3/2} \tag{4.18}$$

and solving for a we find

$$a_{\text{dec}} \approx \left[\frac{x_{\text{ion}} n_{b,0} \sigma_T c}{H_0 \sqrt{\Omega_{m,0}}}\right]^{2/3} \tag{4.19}$$

To determine the baryon number density $n_{b,0}$ we can use the fact that the scaled baryon energy density today has been measured to be $\Omega_{b,0} \approx 0.04$. The energy density of baryons comes from their rest mass energy, so $\epsilon_{b,0} = n_{b,0} m_p c^2$ and so the scaled energy density is

$$\Omega_{b,0} = \frac{n_{b,0} m_p c^2}{\epsilon_{c,0}} = n_{b,0} m_p c^2 \frac{8\pi G}{3c^2 H_0^2} \tag{4.20}$$

where we used the formula for the critical energy density at the present time $\epsilon_{0,c}$. Solving for $n_{b,0}$

$$n_{b,0} = \Omega_{b,0} \frac{3H_0^2}{8\pi G m_p} \approx 0.2 \text{ m}^{-3} \tag{4.21}$$

and plugging this in to our expression of a_{dec} we find

$$a_{\text{dec}} \approx \left[\frac{3H_0 \sigma_T c}{8\pi G m_p} \frac{X \Omega_{b,0}}{\sqrt{\Omega_{m,0}}}\right]^{2/3} \tag{4.22}$$

Plugging in the constants, where $\sigma_T \approx 6.6 \times 10^{-29}$ m and using $\Omega_{m,0} = 0.3, \Omega_{b,0} = 0.04$ and a Hubble constant of $H_0^{-1} \approx 14 \times 10^9$ years, we find

$$a_{\text{dec}} \approx 0.02 x_{\text{ion}}^{2/3} \tag{4.23}$$

or in terms of the redshift $1 + z_{\rm dec} = 1/a_{\rm dec}$

$$z_{\rm dec} \approx 43 x_{\rm ion}^{-2/3} \tag{4.24}$$

(where we used $1 + z \approx z$ for $z \gg 1$). If the universe remained fully ionized ($x_{\rm ion} = 1$) the decoupling of the CMB would occur at a redshift of $z \approx 43$. We will see next that the recombination of electrons and protons to neutral hydrogen atoms reduces $x_{\rm ion}$ by a large factor, so that decoupling of the CMB actually occurs at much higher redshift than this.

4.6 Recombination

The Thomson cross-section σ_T used in the last section only applies to photons scattering off of *free* electrons. When electrons bind to a proton to form neutral hydrogen atom, the cross-section for photon interactions drops by orders of magnitude. The process of forming hydrogen is called *recombination*[8].

[8] The term is something of a misnomer, as prior to "recombination" the electrons and protons had never been bound together before. So perhaps we should call it "combination" as it is the first time atoms formed. The term is borrowed from laboratory experiments where you can ionize (split apart) an hydrogen atom and then wait for the electron and proton to "recombine".

The process of recombination (and its inverse process, ionization) can proceed for example by the reaction

$$p^+ + e^- \leftrightarrow H + \gamma \tag{4.25}$$

i.e., where a proton and electron recombine to hydrogen and emit the excess binding energy in a photon. It can be shown that the rates of these reactions are relatively fast compared to the expansion rate of the universe. Thus, the number densities of hydrogen atoms (n_H) and *free* electrons (n_e) and protons (n_p) will reach a state of *thermodynamic equilibrium* given by our expression for non-relativistic particles

$$n_p = \frac{g_p}{\lambda_{T,p}^3} \exp\left[-\frac{m_p c^2}{k_B T}\right] \tag{4.26}$$

$$n_e = \frac{g_e}{\lambda_{T,e}^3} \exp\left[-\frac{m_e c^2}{k_B T}\right] \tag{4.27}$$

$$n_H = \frac{g_H}{\lambda_{T,H}^3} \exp\left[-\frac{m_H c^2}{k_B T}\right] \tag{4.28}$$

$$\tag{4.29}$$

The rest mass of a hydrogen atom, m_H, is a bit smaller than the sum of the proton and electron rest masses due to the binding energy

$$m_p c^2 + m_e c^2 - m_H c^2 = Q \approx 13.6\ eV \tag{4.30}$$

where Q is the absolute value of the binding energy (also called the ionization potential or ionization energy).

The statistical weight of the electron and proton are $g_e = g_p = 2$ (since either can be either spin up or spin down). The statistical weight of a hydrogen atom (assuming it is in the ground state) is $g_H = 4$ since the proton and electron each can be either spin up or down. Dividing the expressions for n above, we find

$$\frac{n_p n_e}{n_H} = \frac{1}{\lambda_{T,e}^3} \frac{\lambda_{T,H}^3}{\lambda_{T,p}^3} e^{-Q/k_B T} \tag{4.31}$$

Since the mass of the hydrogen atom is roughly equal to the mass of the proton, we can set $\lambda_{T,H} = \lambda_{T,p}$ and our equation becomes

$$\frac{n_p}{n_H} = \frac{1}{n_e \lambda_{T,e}^3} e^{-Q/k_B T} \tag{4.32}$$

This is called the *Saha equation*, and allows us to calculate (n_p/n_H), i.e., the ratio of ionized hydrogen atoms to combined neutral hydrogen atoms, given a system in thermodynamic equilibrium at temperature T.

Given the definition of the ionization fraction, $x_{\rm ion}$ as the fraction of baryons that are ionized, we have $n_p = x_{\rm ion} n_b$ and $n_H = (1 - x_{\rm ion}) n_b$. By charge neutrality $n_e = n_p$ so we can write the Saha equation as

$$\frac{x_{\rm ion}^2}{(1 - x_{\rm ion})} = \frac{1}{n_b \lambda_{T,e}^3} e^{-Q/k_B T} \tag{4.33}$$

This quadratic equation can be solved analytically for $x_{\rm ion}$. For now, we are interested in estimating the temperature at which recombination takes place. We thus ask when $x_{\rm ion} \approx 1/2$ (which marks the point where half of the protons have recombined). The Saha equation gives

$$\frac{1}{2} = \frac{1}{n_b \lambda_{T,e}^3} e^{-Q/k_B T} \implies e^{Q/k_B T} = \frac{2}{n_e \lambda_{T,e}^3} \tag{4.34}$$

Taking the natural log of both sides and rearranging we find

$$T_{\rm rec} = \frac{Q}{k_B} \frac{1}{\xi} \quad \text{where } \xi = \ln\left(2 n_e^{-1} \lambda_{T,e}^{-3}\right) \tag{4.35}$$

This is not actually an explicit solution for $T_{\rm rec}$ since the thermal DeBroglie wavelength $\lambda_{T,e}$ (and hence ξ) itself depends on temperature.However, because n_b and $\lambda_{T,e}$ appear in the logarithm, the expression depends only weakly upon them. For the values of n_b and T that arise near recombination, $\xi \approx 40$. Thus

$$T_{\rm rec} \approx \frac{13.6 \text{ eV}}{k_B} \frac{1}{40} \approx 3750 \ K \tag{4.36}$$

Given our expression for the temperature evolution of the universe $T(z) = T_0(1+z)$ where $T_0 \approx 2.7$ the redshift of recombination is

$$1 + z_{\rm rec} = \frac{1}{a_{\rm rec}} \approx 1380 \tag{4.37}$$

If we want to find the time at which recombination occurs, we would need to solve the Friedmann equation to get $a(t)$. Doing so with benchmark values of the cosmological parameters, results in $t_{\rm rec} \approx 240,000$ years after the big bang.

We have defined the "time of recombination" here to be when half of the hydrogen has recombined. Decoupling happens shortly thereafter, when the vast majority of hydrogen has recombined and so photons very rarely scatter off of free electrons. We found above that decoupling occurs at $z_{\rm dec} \approx 43 x_{\rm ion}^{-2/3}$. Using the Saha equation, we would find that around $z \approx 1100$ the ionization fraction has dropped to $x_{\rm ion} \approx 0.008$ and that decoupling occurs (this corresponds to a time $t_{\rm dec} \approx 350,000$ years).

4.7 *CMB Anisotropies*

The CMB we observe is remarkably uniform, appearing as a $T_0 \approx 2.7$ K blackbody in any direction that you look. But if you look closely, you will see small variations of about 10^{-5} K across the sky (see Figure 4.3). These small deviations represent density inhomogeneity that will act as seeds for the growth of the structures that will form galaxies and galaxy clusters. CMB photons emitted from a over-dense region will suffer greater gravitational redshift than the average, and so will appear to have a slightly cooler temperature than T_0, while CMB photons from an under-dense region will appear slightly hotter than average.

Maps of the CMB anisotropies look a lot like noise (see Figure 4.3), but by doing a Fourier-type analysis we can quantify the significance of fluctuations on a particular angular scale. We write the fluctuations in temperature across the sky as a sum

$$\frac{T(\theta,\phi) - T_0}{T_0} = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} a_{lm} Y_{lm}(\theta,\phi) \tag{4.38}$$

where the spherical harmonics, $Y_{lm}(\theta,\phi)$ behave as 2D analogs of sine and cosines. The coefficient $a_{\rm lm}$ gives the contribution of the spherical harmonic Y_{lm} to the temperature fluctuation map. The power on a particular angular scale is often defined as

$$C_l = \frac{1}{2l+1} \sum_{m=-l}^{m=l} |a_{lm}|^2 \tag{4.39}$$

where the scale of anisotropies are described by the multipole moment l. The thing to know is that angular size, θ of anisotropies are related to multipole moment via

$$\theta \approx \pi / l \tag{4.40}$$

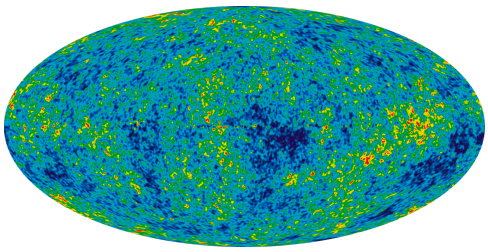


Figure 4.3: Map of the CMB temperature across the sky, from the WMAP experiment. The average CMB temperature $T_0 \approx 2.7$ K has been subtracted off so we can see the small fluctuations in temperature on various scales.

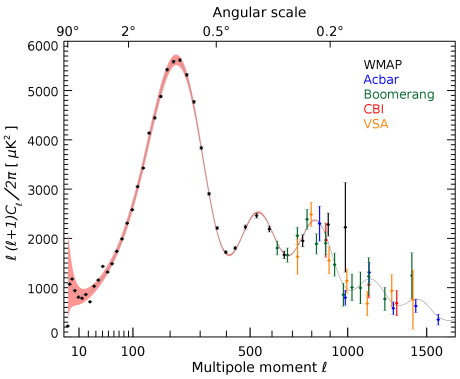


Figure 4.4: CMB power spectrum, derived by doing a Fourier-type analysis of the CMB temperature such as that in Figure 4.3. The plot shows the significance of fluctuations on a particular angular scale.

Figure 4.4 shows a measured CMB power spectrum. For large angular sizes ($\theta \gtrsim 2^o$, or $l \lesssim 100$) the observed power spectrum is relatively flat. These are believed to be *primordial fluctuations* that were seeded soon after the big bang, perhaps through random quantum fluctuations that formed regions of slightly higher and lower dark matter density. These primordial fluctuations do not demonstrate a preferred angular scale (i.e., are the same for all l) and so correspond to a "white noise".

At small angular scales ($\theta \lesssim 2^o$, or $l \gtrsim 100$) the power spectrum shows a series of peaks that appear to be at integer multiples in l. These peaks are believed to be the result of waves, or *acoustic oscillations*, in the early universe. Prior to decoupling, the CMB photons were trapped by scattering off of electrons, so that the electrons, baryons, and photons formed a kind of coupled fluid, called a *baryon-photon fluid*. This fluid will tend to fall into the potential wells of high density regions in the primordial fluctuations; but because the fluid has pressure (provided primarily by the fast moving photons) it will also resist compression and rebound in acoustic oscillations.

Unlike the primordial fluctuations, the acoustic oscillations do pick out preferred angular scales. To understand the origin of these power spectrum peaks, imagine the analogy of throwing a handful of pebbles into a lake. The disturbances of each pebble will create a circular wave that spreads out at the sound speed, c_s. If we were to somehow magically freeze the lake a time Δt later, each pebble would have produced a ripple in the lake of radius $c_s \Delta t$. The combination of many waves from many pebbles would make a complicated looking pattern, but if we did a Fourier-type analysis we would find a power spectrum peak at the length scale $c_s \Delta t$.

A similar effect occurs for the CMB. The primordial fluctuations act as disturbances (the "pebbles") which induce waves in the baryon-photon fluid. When the CMB decouples from the baryons/electrons, the photons free stream away and the pressure in this fluid disappears, and the over- and under-densities of these waves are effectively "frozen" in. This will produce a peak in the CMB power spectrum at angular scales corresponding to the distance that sound waves have traveled before decoupling. Figure 4.5 shows that if $r_{\rm sh}$ is the maximum distance sound waves can travel before decoupling (the "sound horizon") and $r_{\rm ls}$ is the distance from us to the *last scattering surface* at which the CMB was released, then there should be a peak in the power spectrum at angles $\theta \approx r_{\rm sh}/r_{\rm ls}$. In addition, we expect a series of harmonic peaks at integer multiples due the higher oscillation modes of the baryon-fluid in the potential wells.

We can quantify the angular scale, θ, of the acoustic oscillations by calculating $r_{\rm sh}$ and $r_{\rm ls}$. From hydrodynamics theory, the isothermal

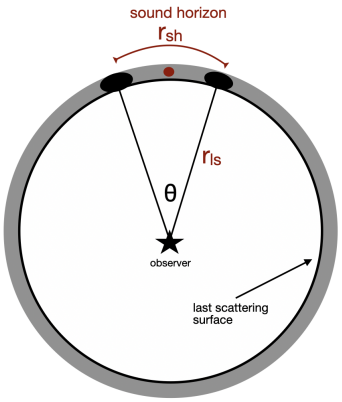


Figure 4.5: Illustration of the geometry of CMB acoustic oscillations. Initial fluctuations lead to the propagation of waves in the baryon-photon fluid, which travel a distance $r_{\rm sh}$ (the sound horizon) before the CMB decouples and the wave freezes out leaving an overdensity. To the observer, the angular size of this fluctuation is $\theta = r_{\rm sh}/r_{\rm ls}$ where $r_{\rm ls}$ is the distance to the location that the CMB was released (e.g., the last scattering surface where photons decoupled).

sound speed of a fluid is

$$c_s = c\sqrt{\frac{P}{\epsilon}} \tag{4.41}$$

where for our photon-baryon fluid the pressure P is provided by the radiation, P_r (the baryons and electrons are taken to be pressureless) and the energy density ϵ is the sum of both baryon energy density ϵ_b and radiation ϵ_r, giving

$$c_s = c\sqrt{\frac{P_r}{\epsilon_r + \epsilon_b}} = c\sqrt{\frac{P_r/\epsilon_r}{1 + \epsilon_b/\epsilon_r}} \tag{4.42}$$

For radiation at temperature T we have $P_r = a_R T^4/3$ and $\epsilon_r = a_R T^4$, and so $P_r/\epsilon_r = 1/3$ so this becomes

$$c_s = \frac{c}{\sqrt{3}}\left(1 + \frac{\epsilon_b}{\epsilon_r}\right)^{-1/2} \tag{4.43}$$

At these epochs the energy density of radiation exceeds that of baryons ($\epsilon_r \gg \epsilon_b$) so $c_s \approx c/\sqrt{3}$ with the factor in parenthesis being a small correction that depends on the density of baryons.

In a tiny time interval dt a sound wave travels a distance $dl \approx c_s dt$ The relation between proper distance and the comoving coordinate r is $dl = a(t)dr$, so

$$a(t)dr = c_s dt \implies dr = c_s \frac{dt}{a(t)} \tag{4.44}$$

We define the *sound horizon* to be the maximum distance that a sound wave could have traveled from the start of the universe ($t = 0$) to the time of decoupling, $t_{\rm dec}$. We calculate the sound horizon in comoving coordinates, $r_{\rm sh}$ by integrating the above equation

$$r_{\rm sh} = \int dr = \int_0^{t_{\rm dec}} c_s \frac{dt}{a(t)} \tag{4.45}$$

To determine the scale factor $a(t)$ we need the Friedmann equation. As a simple approximation, will assume a flat matter only universe[9] ($\Omega_{m,0} = 1$) for which the familiar solution of the Friedmann eq. is

$$a(t) = \left(\frac{3}{2}\frac{t}{t_H}\right)^{2/3} \tag{4.46}$$

where $t_H = H_0^{-1}$ is the Hubble time. Plugging this into Eq. 4.45 and integrating gives

$$r_{\rm sh} = 3\left(\frac{2}{3}\right)^{2/3} c_s t_H \left(\frac{t_{\rm dec}}{t_H}\right)^{1/3} \tag{4.47}$$

Using $t_{\rm dec} \approx 350,000$ years and $t_H \approx 14 \times 10^9$ years, and taking $c_s = c/\sqrt{3}$, the sound horizon is

$$r_{\rm sh} \approx 150 \text{ Mpc} \tag{4.48}$$

[9] A slightly better approximation would be to only assume that the matter term dominates over the radiation, dark energy, and curvature terms in the Friedmann eq., such that

$$\left(\frac{\dot{a}}{a}\right)^2 \approx H_0^2 \Omega_{m,0} a^{-3}$$

with the solution

$$a(t) = \left(\frac{3}{2} H_0 \sqrt{\Omega_{m,0}}\, t\right)^{2/3}$$

For $\Omega_{m,0} = 0.3$ this differs from our assumption by the factor $\Omega_{m,0}^{1/3} \approx 0.67$.

Recall that the proper distance is related to the comoving distance by $d_p = ra(t)$. Since $a(t_0) = 1$ now, the current proper distance of the sound horizon is just $r_{\rm sh} \approx 150$ Mpc. At the time of decoupling, the universe was smaller by the factor $a_{\rm dec} \approx 1/1100$ and the sound horizon then had a size more like 135 kpc.

To determine the comoving distance to the last-scattering surface, $r_{\rm ls}$ we can similarly calculate the distance that CMB photons moving at speed c have traveled from the time they were released, $t_{\rm dec}$, to the present time when they are observed, $t_0 = (2/3)t_H$

$$r_{\rm ls} = \int dr = \int_{t_{\rm dec}}^{t_0} \frac{c}{a(t)} = 2ct_H \approx 8.8 \text{ Gpc} \tag{4.49}$$

where we used[10] Eq. 4.46 for $a(t)$ and took $t_{\rm dec} \ll t_H$. Thus the angular scale of the sound horizon is

$$\theta_{\rm sh} = \frac{r_{\rm sh}}{r_{\rm ls}} \approx 0.017 \text{ rad} \approx 1^o \tag{4.50}$$

or in terms of multipole moment $l_{\rm sh} = \pi/\theta_{\rm sh} \approx 200$. This is indeed roughly where the first peak is seen in the observations. Because the angular scale of this (and other peaks) are dependent on the expansion history of the universe from early times to today, we can use the CMB power spectrum to constrain a variety of cosmological parameters. In particular, the first peak allows us to draw a large triangle on the sky (see Figure 4.5) and we can use the observed geometry of this triangle to constrain the curvature of the universe.

[10] Using the $a(t)$ of a flat matter only universe is not a great approximation; really we should include the effects of dark energy which is currently dominated the expansion.

4.8 Thermal Dark Matter Relics

Cosmological observations (of supernova and the CMB) indicate that the matter content of the universe is $\Omega_{m,0} \approx 0.27$. Other cosmological observations (of the big bang nucleosynthesis and CMB anisotropies) indicate the baryonic content of the universe is $\Omega_{b,0} \approx 0.04$. This implies that that the majority of matter in the universe is *non-baryonic*, i.e., not composed of protons and neutrons (and hence not made up of the ordinary kind of stuff we see around us).

One possible explanation of dark matter is a new type of particle[11] (call it the X-particle) not in the standard model. Although we have no direct experimental evidence about this particle, we can make some remarkable arguments from cosmology that tell one possible story about how the dark matter in the universe came to be. This theory will also allow us to estimate the mass, m_x, of the X-particle.

[11] In principle there could be a number of particles forming an entire "dark sector". For simplicity we will just consider a single particle here. The basic ideas could be generalized for a richer dark sector phenomenology.

The basic idea is to presume that interactions exist that convert standard model particles into dark matter particles, and vice versa

$$S + \bar{S} \leftrightarrow X + \bar{X} \tag{4.51}$$

where S is some standard model particle (e.g., a quark, photon, or a lepton like an electron or neutrino) and $\bar{S}$ is its antiparticle, and X is a dark matter particle and $\bar{X}$ its antiparticle. When these interactions happen rapidly the X-particles should come into *thermodynamic equilibrium* and we have formulae that give the number density of X particles given their temperature and mass.

As the universe expands, the density drops and there will come a point when the rate of creation/annihilation interactions of the X-particles becomes slower than expansion of the universe (i.e., the condition $\Gamma_c > H$ ceases to hold) . The number of the X-particles will then stop changing, or *freeze-out*. These left-over X-particles should still be around today – a so-called "thermal relic".

Let's see if a thermal relic can explain the dark matter inferred in our universe, which has a scaled density today of $\Omega_{dm,0} \approx 0.23$. We will have to separately consider the cases where the X-particle is "hot" (i.e., relativistic at the time of freeze-out, with $m_x c^2 \ll k_B T_F$) and when the X-particle is "cold" (i.e., non-relativistic at freeze-out, with $m_x c^2 \gg k_B T_F$).

4.8.1 A Hot Dark Matter Thermal Relic

Consider first a hot dark matter relic. We would like to calculate when freeze-out occurs by comparing the collision rate $\Gamma_c = nv\sigma_T$ to the Hubble expansion. For relativistic particles, the number density in thermodynamic equilibrium is

$$n = b_R T^3 \tag{4.52}$$

where b_R is a physical constant[12] and we assumed zero chemical potential. We will assume the X-particle interact via the weak interaction, in which case the cross-section for relativistic particles is

$$\sigma_X \approx G_F^2 (k_B T)^2 \tag{4.53}$$

where G_F is the Fermi constant, which sets the scale of weak interactions[13]

Putting these together to determine the interaction rate $\Gamma_c = n_x \sigma_x v$ and taking $v = c$ because the particles are assumed to be relativistic

$$\Gamma_c = (b_R T^3) \times (G_F^2 k_B^2 T^2) \times c = b_R c G_F^2 k_B^2 T^5 \tag{4.54}$$

To get H we use the Friedmann equation

$$H^2 = \frac{\dot{a}^2}{a^2} = H_0^2 \left[\frac{\Omega_{0,m}}{a^3} + \frac{\Omega_{0,r}}{a^4} + \Omega_{0,\Lambda} + \frac{1 - \Omega_0}{a^2} \right] \tag{4.55}$$

Adopting a flat universe ($\Omega_0 = 1$) and assuming we are in the radiation dominated phase of the early universe (so keeping only the

[12] Recall that for radiation like photons, the energy density $\epsilon_R = a_R T^4$ and since each particle has energy of order $k_B T$ the number density of particles goes like $n \sim \epsilon / k_B T \propto T^3$. A careful calculation which integrates the blackbody distribution finds $n = b_R T^3$ where

$$b_R = \frac{\zeta(3)}{\pi^2} \frac{k_B^3}{\hbar^3 c^3} g \approx 20.2 \quad \text{(for bosons)}$$

and

$$b_R = \frac{3\zeta(3)}{4\pi^2} \frac{k_B^3}{\hbar^3 c^3} g \approx 15.2 \quad \text{(for fermions)}$$

where $\zeta(3) \approx 1.2$ is the Reimann zeta function and we have assumed a statistical weight $g = 2$.

[13] It is possible that the interactions that convert X-particles to standard model particles are not a result of the weak force. Then there may be a different constant than G_F^2 in the cross-section, but the general form $\sigma_X \propto (k_B T)^2$ may still hold for relativistic particles for kinematic reasons.

radiation term) gives

$$H^2 = H_0^2 \frac{\Omega_{0,r}}{a^4} \implies H = \frac{H_0\sqrt{\Omega_{r,0}}}{a^2} \tag{4.56}$$

The temperature is related to scale factor by $T = T_0/a$ where $T_0 \approx 2.7$ K is the current temperature of the CMB. Then since $a = T_0/T$ we have

$$H = \frac{H_0\sqrt{\Omega_{r,0}}}{T_0^2} T^2 \tag{4.57}$$

We can now write down the condition for freeze-out by setting $\Gamma_c \approx H$ which gives

$$bG_F^2 k_B^2 c T^5 = \frac{H_0\sqrt{\Omega_{r,0}}}{T_0^2} T^2 \tag{4.58}$$

Solving for T gives the temperature at freeze-out

$$T_F = \left[\frac{H_0\sqrt{\Omega_{r,0}}}{bG_F^2 k_B^2 c T_0^2}\right]^{1/3} \tag{4.59}$$

If we plug in values for all of the terms in this expression, we find $T_F \approx 10^{10}$ K, which corresponds to $k_B T_F \approx 1$ MeV. This freeze-out would have occurred at a redshift of $(1+z) = T_F/T_0 \approx 3.7 \times 10^9$, which corresponds to about 1 second after the big-bang. The assumption that the X-particle is "hot" at freeze-out then means that its mass must be $m_x c^2 \ll 1$ MeV.

Given that the total number of X particles, remains constant after freeze-out, we can determine the number density of X-particles left over today. Just prior to freeze-out, the particles are in thermodynamic equilibrium (For the last time) and so their number density $n_{x,F} = b_R T_F^3$. The total number of particles in a volume V is $N_x = n_x V \propto n_x a^3$ where we used the fact that volume scales as $V \propto a^3$ as the universe expands. The fact that the number of X particles is frozen out gives

$$n_{x,0} a_0^3 = n_{x,F} a_F^3 \implies n_{x,0} = n_{x,F} a_F^3 \tag{4.60}$$

where we used $a_0 = 1$. Since the universe was smaller in the past, the scale factor at freeze-out is $a_F < 1$, so we confirm that the number density today, $n_{x,0}$ is less than it was at freeze-out $n_{x,F}$, as expected due to the expanding universe.

We can relate the scale factor at freeze-out to the temperature at freeze-out using the relation $T = T_0/a$ where $T_0 = 2.7$ is the temperature of radiation today. Thus $a_F = T_0/T_F$ and

$$n_{x,0} = n_{x,F} \frac{T_0^3}{T_F^3}$$

Using the fact that in thermodynamic equilibrium $n_{x,F} = b_R T_F^3$ we have

$$n_{x,0} = b_R T_F^3 \frac{T_0^3}{T_F^3}. = b_R T_0^3$$

While we have assumed the X-particles were relativistic at the time of freeze-out, today they should be non-relativistic if they are to behave as matter (and not radiation). The energy of each X-particles should then just be the rest mass, $m_x c^2$. The energy density of X-particles today will be $\epsilon_{x,0} = m_x c^2 n_{x,0}$. Dividing by the critical density gives the scaled Ω

$$\Omega_{x,0} = \frac{\epsilon_{x,0}}{\epsilon_{c,0}} = \frac{m_x c^2 n_{x,0}}{\epsilon_{c,0}} = \frac{m_x c^2 b T_0^3}{\epsilon_{c,0}} \tag{4.61}$$

Plugging in the standard expression for the critical density today

$$\Omega_{x,0} = \left(\frac{8\pi G b T_0^3}{3H_0^2} \right) m_x \tag{4.62}$$

Now if we plug in numbers for the constants (and assuming the X-particle is a spin 1/2 fermion so $g = 2$ and $b_R \approx 15.2$) we find

$$\Omega_{x.0} \approx 0.06 \left(\frac{m_x c^2}{1 \text{ eV}} \right) \tag{4.63}$$

Which indicates that the mass of the X-particles needs to be around $m_x c^2 \approx 4$ eV to have $\Omega_{x,0} \approx 0.23$ and explain all of the dark matter. This is consistent with our assumption that the dark matter particle was "hot" at freeze-out (since the freeze-out temperature we found was $k_B T_F \approx 1$ MeV) but "cold" today (since today $k_B T_0 \approx 0.00023$ eV). More detailed analyses[14]. find that the mass $m_x c^2 \approx 9$ eV. If there was more than one relativistic dark-matter particle, the sum of their rest masses must add to this value.

The summed rest mass of all types neutrinos has been constrained to be $m_\nu \lesssim 0.3$ eV. So the thermal relics of neutrinos (which we call the cosmic neutrino background, or CNB) apparently cannot explain the dark matter in the universe[15]. So apparently we need to invoke some new particle to be the dark matter.

There are arguments from cosmic structure formation that disfavor a hot dark matter particle. Primordial over-densities in the dark matter distribution are the "seeds" for structure formation, but if these seeds are composed of hot particles, these "seeds" will initially smear out as the particles free-stream in all directions at the speed of light. By the time gravity can pull things together, the perturbations have dispersed to large sizes, while smaller scale structures (e..g., galaxy-scale structures) will have been washed out. This poses problems for explaining the evolution of structures we see in the universe.

[14] The more detailed analyses account for the fact that the temperature of the early universe is not quite equal to $T = T_0/a$. Instead we need to account for the fact that their are other relativistic particles besides the putative dark matter particles, such as photons and neutrinos.

[15] The expression Eq. 4.63 allows us estimate the neutrino contribution. Using the upper limit $m_\nu \lesssim 0.3$ eV. constrains $\Omega_{\nu.0} \lesssim 0.02$

4.8.2 Cold Dark Matter Thermal Relic

We can now consider now the case of a cold dark matter relic. The key differences from the previous hot thermal relic analysis, is that the number density for non-relativistic particles in thermodynamic equilibrium (and zero chemical potential) is

$$n = \frac{g}{\lambda_T^3} e^{-m_x c^2/k_B T} \tag{4.64}$$

We can now proceed to calculate the condition for freeze-out. The collisional rate is

$$\Gamma_c = n_x v_x \sigma_x = \frac{g_x}{\lambda_{T,x}^3} e^{-m_x c^2/k_B T} v_x \sigma_x \tag{4.65}$$

Setting this equal to the previous expression for H (Eq. 4.57) gives

$$\frac{g v_x \sigma_X}{\lambda_T^3} e^{-m_x c^2/k_B T} = \frac{H_0\sqrt{\Omega_{r,0}}}{T_0^2} T^2 \tag{4.66}$$

Rearranging this gives

$$e^{m_x c^2/k_B T} = \frac{g v_x \sigma_X}{\lambda_T^3} \frac{T_0^2}{T^2} \frac{1}{H_0\sqrt{\Omega_{r,0}}} \tag{4.67}$$

And taking the natural log of and solving for temperature we find

$$k_B T_F = \frac{m_x c^2}{\ln \xi} \tag{4.68}$$

where

$$\xi = \left[\frac{g v_x \sigma_X}{\lambda_T^3} \frac{T_0^2}{T^2} \frac{1}{H_0\sqrt{\Omega_{r,0}}} \right] \tag{4.69}$$

This is not an explicit solution for T_F, as temperature also appears in the factor ξ. However, ξ only varies weakly with temperature as the dependence is logarithmic,

Having determined the freeze-out temperature, we can determine the number density at freeze-out by assuming thermodynamic equilibrium (which hold just up to freeze-out)

$$n_{x,F} = \frac{g}{\lambda_T^3} e^{-m_x c^2/k_B T_F} = \frac{g}{\lambda_T^3} e^{-ln\xi} = \frac{g}{\lambda_T^3} \frac{1}{e^{\ln \xi}} = \frac{g}{\lambda_T^3} \frac{1}{\xi} \tag{4.70}$$

Plugging in the expression[16] for ξ

$$n_{x,F} = \frac{1}{v_x \sigma_x} \frac{T_F^2}{T_0^2} H_0 \sqrt{\Omega_{r,0}} \tag{4.71}$$

After freeze-out, the number of X-particles stays constant, while the density scales as a^3 the number density today is then

$$n_{x,0} = n_{x,F} a_F^3 = n_{x,F} \frac{T_0^3}{T_F^3} \tag{4.72}$$

[16] Note that we could have gotten the result Eq. 4.71 directly from the freeze-out condition, $\Gamma_c = H$ since that can be written out as

$$n_{x,F} v_x \sigma_X = \frac{H_0\sqrt{\Omega_{r,0}}}{T_0^2} T^2$$

which implies

$$n_{x,F} = \frac{1}{v_x \sigma_x} \frac{T_F^2}{T_0^2} H_0 \sqrt{\Omega_{r,0}}$$

and using Eq. 4.71 for $n_{x,F}$ we have

$$n_{x,0} = \frac{H_0\sqrt{\Omega_{r,0}}}{v_x\sigma_x}\frac{T_0}{T_F} \tag{4.73}$$

The energy of each particle is just the rest mass energy (since these are cold particles) so the energy density today is $\epsilon_{x,0} = m_x c^2 n_{x,0}$, or

$$\epsilon_{x,0} = m_x c^2 \frac{H_0\sqrt{\Omega_{r,0}}}{v_x\sigma_x}\frac{T_0}{T_F} \tag{4.74}$$

Using our result for the freeze-out temperature $k_B T_F = m_x c^2/\ln\xi$, this expression implies a dimensionless density today, $\Omega_{x,0} = \epsilon_{x,0}/\epsilon_{c,0}$ of

$$\Omega_{x.0} \approx \left[\frac{8\pi G k_B T_0\sqrt{\Omega_{r,0}}}{3H_0 c^3}\right]\frac{\ln\xi}{v_x\sigma_x} \tag{4.75}$$

A numerical treatment is needed to determine ξ, but using typical values one finds $\ln\xi \approx 20$ (since it is a logarithimic function, the result is not to sensitive to the exact values used). This value implies[17] $v_x \approx c/3$. Using this $\xi = 20$ and plugging in numbers for the constants gives

$$\Omega_{x.0} \approx \frac{2\times 10^{-40}\ \mathrm{m}^2}{\sigma_x} \tag{4.76}$$

Interestingly, this expression for $\Omega_{x,0}$ depends only on the cross-section for X-particle interactions. To explain the dark matter in the universe requires $\sigma_X \approx 10^{-39}$ m^2. Remarkably, this is the ballpark of the cross-section for the weak interactions, which suggests that a weakly interacting massive particle (WIMP) could naturally explain the dark matter as a thermal relic. The fact that the numbers work out this way is sometimes called the "WIMP miracle"[18].

In the theory of weak interactions, the cross-section for interactions with a non-relativistic massive particles can be written to order of magnitude approximation as

$$\sigma_x \approx G_F^2 (m_x c^2)^2 \tag{4.77}$$

where G_F is the Fermi constant, with $G_F^2 \approx 3\times 10^{-42}$ m^2/Gev^2. If we assume this cross-section applies. we can write the predicted $\Omega_{x,0}$ in terms of the X-particle mass

$$\Omega_{x.0} \approx 66\left(\frac{m_x c^2}{1\ \mathrm{GeV}}\right)^{-2} \tag{4.78}$$

And so to get $\Omega_{x,0} \approx 0.23$ implies a dark matter particle mass of $m_x c^2 \approx 16$ GeV.

The arguments you have just carried out have inspired millions of dollars of experiments to try to directly detect dark matter particles that have cross-sections and masses roughly in this range (e.g.,

[17] To find the value of v_x, use the fact that the particle is non-relativistic and the kinetic energy is of order $k_B T$ so

$$\frac{1}{2}m_x v_x^2 \approx k_B T_F = \frac{m_x c^2}{\ln\xi}$$

which solving for v_x gives

$$v_x \approx c\sqrt{\frac{2}{\ln\xi}}$$

and so for $\ln\xi \approx 20$ we have $v_x \approx c\sqrt{1/10} \approx c/3$

[18] Actually, the result for $\Omega_{x,0}$ doesn't depend only on σ_x, it also depends on the mass m_x through the factor $\ln\xi$. While the dependence on m_x is weaker in that it appears in the log, if one varies m_x by orders of magnitude it can change the implied σ_x significantly and make the WIMP miracle seem less miraculous.

$m_x c^2 \sim 1 - 100$ GeV). So far nothing has been found, but the sensitivities are continually improving.

Eq. 4.78 appears to rule out cold dark matter particles with masses $\lesssim 1$ GeV as these would give too high a dark matter density today. However theories of light dark matter particles (e.g., $m_x c^2 \ll 1$ GeV) may work if one modifies their interactions or the physics of their production. In some theories (e.g., in particular those that suggest dark matter is made up of *axion* particles) the dark matter particle is never in thermal equilibrium at temperature T but is instead produced as a condensate during a phase transition of the universe. Such non-thermal processes allows particles to be produced cold in the early universe, even if they have a very small mass (e.g., fractions of an eV). Thus the assumptions underlying our calculations here of *thermal relics* no longer apply.