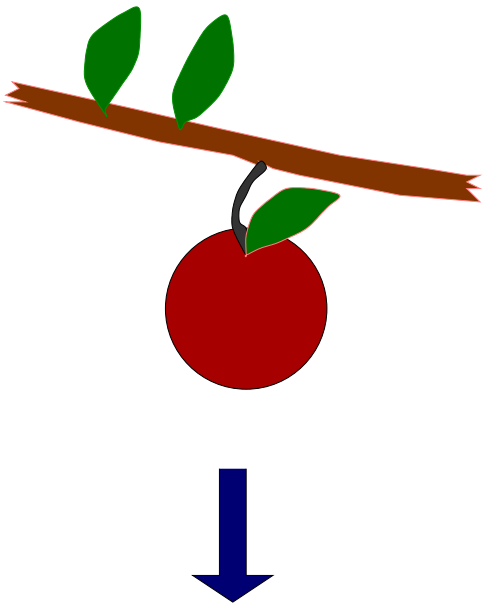
1. Newtonian and non-Newtonian gravity

This course is about General Relativity (GR) and its application to the Universe as a whole. GR is a theory of gravity, and the first lecture attempts to convince you that GR is necessary — a theory like GR is needed to resolve the inconsistencies in Newtonian gravitation, which are horrible and egregious.

This lecture is intended to spell out some of the problems with Newtonian gravitation, and to introduce the basic ideas that lead to a different and internally consistent view of gravitation.

1.1. Newtonian gravitation: force

Suppose we look at an apple suspended by its stem from the branch of a tree. The stem breaks and the apple drops to the ground. Then the question that Newton asked is "Why does the apple fall?". Let's ask the same question and try to find some answers.



Newton's answer can be paraphrased as "The apple falls because the Earth's attractive gravitational force pulls it down." This force is proportional to the mass of the Earth and the apple, and can be expressed as

$$\mathbf{F} = -\frac{G\,M_{\rm E}\,m_{\rm apple}}{r^3}\,\mathbf{r} \quad ,$$

where $\mathbf{r}$ is the position vector to the apple from the center of the Earth (of length $r = R_{\rm E}$), $m_{\rm apple}$ is the mass of the apple, the other quantities are

$$\begin{array}{lll} G & = \text{universal gravitational constant} & = (6.67259 \pm 0.00085) \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}} \\ M_{\rm E} & = \text{mass of Earth} & = (5.9729 \pm 0.0008) \times 10^{24}\ \mathrm{kg} \\ R_{\rm E} & = \text{radius of Earth} & = (6.3782 \pm 0.0000) \times 10^{6}\ \mathrm{m} \end{array}$$

Of course, this assumes a spherically-symmetrical Earth, but this will do for now, since we don't want to get bogged down with J_2. It also ignores modifications to the apparent force because of the rotation of the Earth, but likewise I'll ignore that.

The force then causes acceleration, according to

$$\mathbf{F} = m_{\rm apple}\ddot{\mathbf{r}}$$

so that the acceleration of the apple is

$$\ddot{\mathbf{r}} = -\frac{G\,M_{\rm E}}{r^3}\,\mathbf{r} \quad .$$

Two problems with this answer should occur to physicists. First, since the apple reacts to $G\,M_{\rm E}$ at relative position $-\mathbf{r}$, it needs to know the instantaneous components of the vector $\mathbf{r}$. But how can it know the location of the centre of the Earth "instantly", that is, faster than the speed of light? Furthermore, since the Earth isn't a perfect sphere, the apple needs to know "instantly" not just the location of the centre of the Earth, but the locations of all of its matter.

Second, who measures $\mathbf{r}$? The apple and the Earth do *not* move with the same speed, and so they see their separation with different Lorentz contractions. They therefore will not agree on the same value of r as an observer at rest in the centre of mass frame. Who is right?

1.2. Newtonian gravitation: potential

Let's try to avoid these difficulties by being more sophisticated. We introduce a new function, the potential, $\Phi(\mathbf{r})$ and use it to calculate the force, as

$$\mathbf{F} = -m_{\rm apple}\nabla\Phi(\mathbf{r}).$$

Then we no longer have instantaneity problems — we say that the apple reacts to the local properties of Φ, by moving in the direction of steepest slope and accelerating at a rate determined by the gradient of Φ.

We now have to add another ingredient into the mix — an equation for Φ. The usual form is the Poisson equation which relates Φ to the density distribution, $\rho(\mathbf{r})$

$$\nabla^2\Phi = 4\pi G\rho \quad ,$$

for which we can use the Greens function solution for Φ

$$\Phi = -\int d^3\mathbf{r}' \, \frac{G}{|\mathbf{r}-\mathbf{r}'|} \, \rho(\mathbf{r}') \quad .$$

This, of course, depends on the mass distribution being static. If the density depends on time, we must also take dynamics into account, and we might use an analogy with the retarded (Liénard-Wiechert) potentials of electromagnetism

$$\Phi(\mathbf{r},t) = -\int d^3\mathbf{r}' \, \frac{G}{|\mathbf{r}-\mathbf{r}'|} \, \rho\big(\mathbf{r}', t - |\mathbf{r}-\mathbf{r}'|/c\big),$$

where we have used the retarded time, taking into account the time that information (about where the masses are) needs to get to point $\mathbf{r}$ from $\mathbf{r}'$.

For a point mass we can describe the density using a delta function (note it is not really a function but a distribution, but as long as we use it in an integral, we are safe),

$$\rho = m\,\delta(\mathbf{r} - \underline{\xi}(t'))$$

where $\underline{\xi}(t')$ is the path of the particle. Then we get an expression for the potential

$$\Phi(\mathbf{r},t) = -\frac{Gm}{\big|\mathbf{r}-\mathbf{r}'\big| + \frac{1}{c}\dot{\underline{\xi}}(t').(\mathbf{r}-\mathbf{r}')}$$

where $t' = t - \frac{|\mathbf{r}-\mathbf{r}'|}{c}$ is the retarded time.

This is a good guess, but unfortunately wrong in detail (though correct to low order). One place where we can see it must fail is that in electromagnetism from which we drew the analogy, for the static electric field there is a dynamic magnetic field: static charges give an electric field, moving charges give a magnetic field. What is the equivalent of magnetic field for gravity? Where is the gravitomagnetic force? It must be there, but there's nothing describing it in Newtonian theory.

However, even if this analogy with electromagnetism and this use of potentials fails, it does illustrate a general approach to avoiding the simultaneity problem — we've broken the calculation into two parts: a force law

$$\mathbf{F} = -m_{\rm apple}\nabla\Phi$$

and the dynamics of the gravitational field, which would come from something like

$$\nabla^2\Phi = 4\pi G\rho$$

(or its d'Alembertian equivalent), which could be obtained from an action principle if we wanted.

1.3. Inertial and gravitational mass

The use of potential is moderately satisfactory: we've recovered a local theory. However, it leaves an important hole. Let's look at the motion of a point mass again.

In the force equation, the masses that we use are the gravitational masses, $m_{\rm g}$. The gravitational mass is that property of a body which responds to the gravitational potential to give the force:

$$\mathbf{F} = -m_{\rm g}\nabla\Phi \qquad .$$

The acceleration, derived from Newton's second law, is proportional to the inertial mass, $m_{\rm i}$, which describes a different property of a body, how it reacts to a force to acquire its acceleration:

$$\ddot{\mathbf{r}} = \frac{\mathbf{F}}{m_{\rm i}} \qquad .$$

Combining these two equations, we can write the acceleration of a body as

$$\ddot{\mathbf{r}} = -\frac{m_{\rm g}}{m_{\rm i}}\,\nabla\Phi \qquad .$$

Now, it is found experimentally that **all** bodies have the same ratio $m_{\rm g}/m_{\rm i}$. This comes out of (for example) the Eötvös experiment, especially in its more modern incarnations (and despite a flurry of activity a few years ago about a pattern in the residuals in that experiment that were supposed to show evidence for a fifth force — the Fischback conjecture, Fischback, E. *et al.*, 1986. Phys. Rev. Lett., **56**, 3-6). The measurements of Roll, Krotkov and Dicke in the early 1960s (see Misner, Thorne & Wheeler, pages 14-17) showed that the variation

$$\delta\left(\frac{m_{\rm g}}{m_{\rm i}}\right) < 10^{-12}$$

over all bodies. That is, the gravitational acceleration is independent of the mass of the body being accelerated.

I want to emphasize how remarkable this is. For no other force is the amplitude of the acceleration caused independent of the "charge" (e.g., electric charge, or colour). A higher electric charge on a particle would cause it to accelerate faster in an electric field. A larger colour would cause stronger interactions in the strong force. But for gravitation, doubling the mass of a body has no effect on its acceleration.

Since acceleration in a gravitational field is independent of mass, and independent of what a body is made of, we can absorb any ratio $m_{\rm g}/m_{\rm i}$ into the definition of G, write $m_{\rm g} = m_{\rm i} = m$, and

$$\ddot{\mathbf{r}} = -\nabla\Phi$$

and say that gravitation causes an acceleration of a body which is a function of all other bodies' locations and histories.

The most important consequence is that *we can transform to a frame of reference (a freely falling frame) such that all bodies which are sufficiently close together in that frame feel NO external gravitational force.* In particular, we can use no local dynamical experiment to measure Φ (or any other indicator of gravity) in a freely falling frame. Said another way, an observer would find it impossible to tell whether he/she is in a freely falling frame in a gravitational field, or in a force-free part of space.

i.e., "a maggot in a falling apple is unaware of the external gravitational field" – and the same will apply to weightlessness in an orbiting spacecraft (which is "freely falling" about the Earth).

This is, effectively, the **weak principle of equivalence** — the dynamics of moving bodies are independent of whether the frame of reference is accelerating or the bodies are being affected by an external gravitational field (at a single point). Newtonian gravitation gives no clue as to why inertial and gravitational mass are so accurately proportional — which means that we can take the overall constant of proportionality into the gravitational constant, G, and call them equal.

1.4 Natural states of motion

This idea can be extended to the **strong principle of equivalence**, which says that *at every point it is possible to choose a local reference frame such that all laws of nature have the same form as in an unaccelerated frame in the absence of gravitation.*

That is, in the correct frame of reference (a freely-falling frame), gravitation vanishes. This leads to the central concept of GR:

the natural state of motion is free fall

no force needs to be applied to create this state of motion

forces must be applied to stop free fall

This gives us our post-Newtonian interpretation of the falling apple.

original question: what causes the apple to fall?

original answer: the gravitational force of the Earth

improved answer: the local gravitational field at the apple

better question: what stopped the apple from falling freely all the time?

answer: the obvious, and non-gravitational, forces in the stem of the apple

Where Newtonian physics points to the gravitational force as the cause of the acceleration which causes the apple to move downwards, our new and more sophisticated

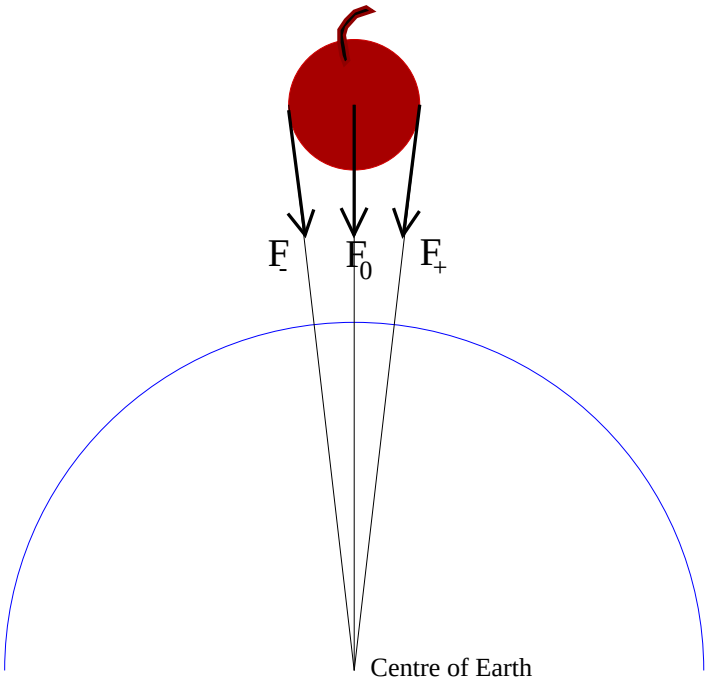
viewpoint says that the apple would naturally be moving downwards, and points to the existence of the stem as the reason (the outside force, based on the electromagetic force) why the apple wasn't originally in its natural state of free-fall.

This is the viewpoint that we will adopt — the General Relativistic viewpoint. Every body continues in a state of free-fall unless there is a non-gravitational force acting on it to distort its path away from free-fall. Making this concept mathematical, and deciding how the shape of the free-fall path will be perceived by observers who may not themselves be in the equivalent state of free-fall, is what makes some of the general relativistic calculations difficult.

1.5. General relativity and equivalence

The strong principle of equivalence tells us that we can get rid of gravitational effects at a single point by transforming coordinates to a freely-falling frame. But this cannot be done **except** at a single point.

For example, a falling apple sees different sizes and directions of gravitational acceleration at different points around its skin



and we generally call the difference between these forces the "tidal stress" on the object.

General relativity provides

(1) a method of transferring coordinates from an observer's frame to a freely-falling frame

(2) relationships between this transformation (a local coordinate change) and the global matter distribution

and relates the coordinate change needed to the tidal stresses that indicate the shapes of the paths of freely-falling bodies.

In what follows, remember that we are trying to relate observer positions, times, and physical quantities to those same parameters in a frame where gravitation **plays no role** — i.e., where the physics is simple.

The corollary is that the paths followed by freely falling bodies are straight lines.

This is true even for (e.g.) spaceships orbiting the Earth — they move in straight lines. The reason that we see their motion as curved is that we are not freely-falling observers, but are standing on the surface of the Earth (being pushed out of our natural state of free-fall by electromagnetic forces). If we were free-falling and in a region local to the spacecraft, it would move in a straight line. Relating the coordinates of a body freely-falling here (in Bristol) looking at a spacecraft freely-falling hundreds of miles away would require a non-local transformation.

The non-local effects are related to the tidal stresses caused by lumpy matter. And to work with these effects in a proper way, we will have to learn how to deal with geometry in a new way, a way that expands on the four-vector notation that you learned when dealing with special relativity.

In later lectures we must

- develop a consistent vector notation;
- explain the ideas of curvature and how they relate to tidal effects; and
- do some geometry with related algebra.

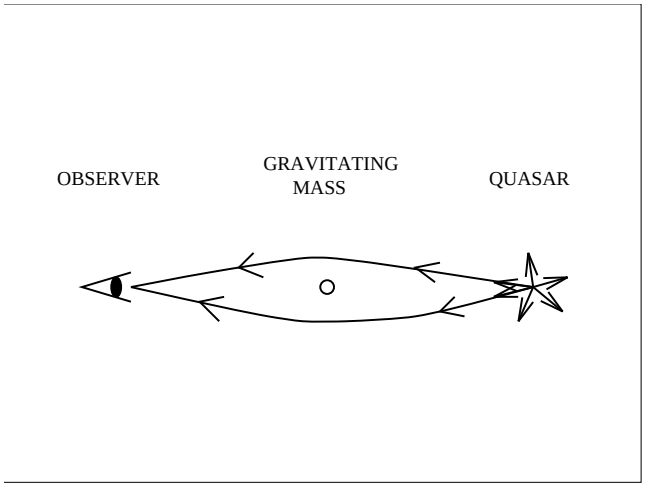
We will do that after using the next lecture to examine the sizes of GR effects.

2. Equivalence and the gravitational redshift

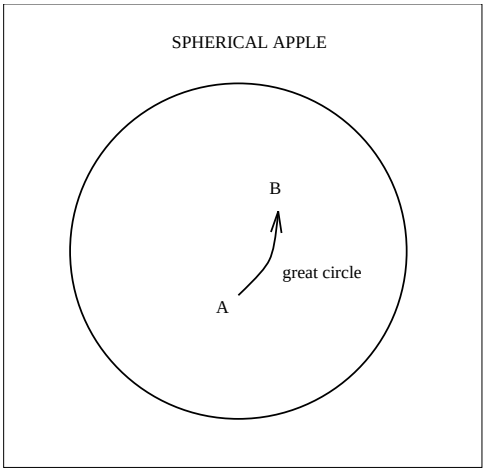
2.1. Geodesics

I have said that freely-falling particles follow straight-line motion: the take the path of shortest distance between two points (the start and end points of their path). These shortest paths are called **geodesics**, and have a central role in GR because everything, including light, travels on a geodesic unless deflected by a non-gravitational force. Thus in looking at a distant object in the Universe, we see light that came to us along a geodesic.

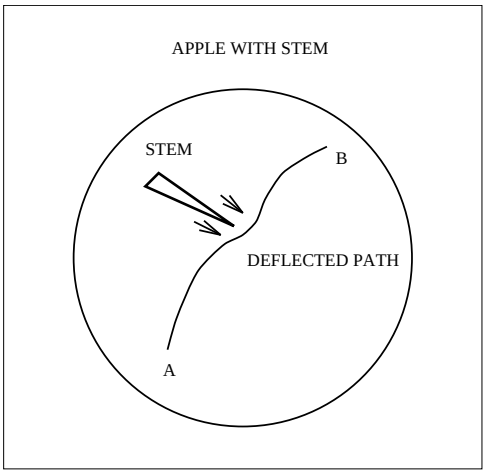
Of course, from our point of view the geodesics need not be straight. So, for example, we may see the same object in two different directions — as in gravitational lensing which causes multiple images. Both paths are geodesics, both are straight, but they are not the same. And both paths are shortest, as defined by **small** deformations of the path.



An analogy (from Misner, Thorne & Wheeler) that's useful here is to think about ants moving on the surface of an apple, trying to get from one place to another. For efficiency, each ant will take the shortest possible route — and hence those routes will be great circles if the apples are spherical. These are the geodesics, the "straight lines" on the surface of a sphere.



If the apple is non-spherical, perhaps having a dimple near the stem, it's clear that the shortest routes are no longer great circles, but will distort (i.e., bend) near the stem. The geodesics are still "straight", but appear distorted when seen from outside the apple.



Indeed, two ants that start by walking geodesic paths on either side of the stem, will find their paths converged by the effect of the stem, and hence that their geodesics intersect sooner than they would expect by virtue of the spherical shape of the apple alone. They might attribute this to an attractive force from the stem of the apple — but that's not what's doing it, it's just that the curvature of the surface deforms their geodesic paths towards one another.

On a grander scale, even if the ants start by moving on parallel paths, these paths will eventually cross because of the large-scale curvature of the apple.

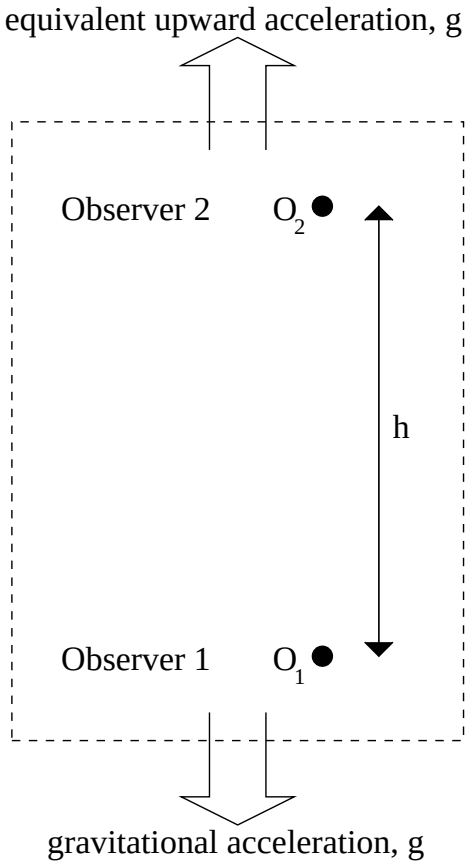
- small-scale deformation: analogy with Sun's gravitational field
- large-scale deformation: analogy with cosmological curvature

In GR,

- ◇ mass causes curvature
- ◇ particles follow the shortest paths between points (if unaffected by other forces)
- ◇ and we have to change the definition of "distance" so that "shortest paths" makes sense.

2.2. Gravitational redshift

Let's look at a simple use of the equivalence principle to get an estimate of the sizes of GR effects — by making an *exact* calculation of the gravitational redshift.



The two observers, O_1 and O_2, separated by h, can't distinguish between the effects of gravitation acting on them and a corresponding upward acceleration. If the gravitational acceleration is g, then we can think of a box encompassing O_1 and O_2 accelerating upwards at g being the equivalent of a downward gravitational force producing an acceleration g.

So, now suppose O_1 emits a photon of wavelength λ at time t_0, and that O_2

receives that photon at time $t_0 + \frac{h}{c}$.

In time interval $\frac{h}{c}$, O_2 has increased its velocity (taking the view that the box is accelerating upwards) by

$$\Delta v = g\frac{h}{c}$$

and so the Doppler shift of the wavelength of the emitted photon as seen by O_2, relative to the emitted wavelength, is

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta v}{c} = \frac{gh}{c^2} \quad .$$

But this must be the same as if O_1 and O_2 are at rest in a gravitational field with acceleration g, and hence in such a case and for an experiment on the surface of the Earth

$$\frac{\Delta\lambda}{\lambda} \equiv z = \frac{gh}{c^2} = \frac{GM_{\rm E}}{c^2 R_{\rm E}^2} h \quad .$$

This prediction has been tested (in the Pound-Rebka-Snider experiment, which involved the emission of X-rays and their detection using the Mossbauer effect), with total success. So there really is a gravitational redshift.

Of course, in the relatively feeble gravity of the Earth, the effect is small. But in some astrophysical situations it can be substantial. Consider, for example, the gravitational redshift of emission lines from the surface of a neutron star, which might have radius $R_{\rm ns} \approx 10$ km and mass $M_{\rm ns} \approx 1.4\ M_\odot$. Then

$$z = \int_0^\infty dh \frac{GM_{\rm ns}}{c^2(R_{\rm ns}+h)^2}$$

adding up all the elemental gravitational redshifts, which integrates to

$$z = \frac{GM_{\rm ns}}{c^2 R_{\rm ns}} = -\frac{\Phi_{ns}}{c^2} \approx 0.2$$

where $\Phi_{\rm ns}$ is the Newtonian gravitational potential at the surface of the star. This means that a line which would appear at 500 nm in the laboratory is shifted to 600 nm in the spectrum of a neutron star.

2.3. Sizes of Effects in GR

Let us use dimensional analysis to check this result and to estimate the conditions under which GR effects will appear.

2.3.1. Point mass

For a point mass, the physical quantities which are important will be

M: the mass of the object, with dimension $[M]$; and

R: the distance of the object, with dimension $[L]$.

In addition, we are dealing with gravitational effects, so the problem must involve

G: the "inflexibility of spacetime", with dimension $[M^{-1}L^3T^{-2}]$, and

c: the "scaling between distance and time", with dimension $[LT^{-1}]$.

From these variables we can create only one dimensionless quantity,

$$\delta_1 = \frac{GM}{c^2 R}$$

which will be the fractional size of any GR effects relative to the ordinary Newtonian effects. Putting in the numbers, we get

$$\begin{aligned}\delta_1 &= 7.4 \times 10^{-28}\,(M/\mathrm{kg})\,(R/\mathrm{m})^{-1} \\ &= 2.1 \times 10^{-6}\,(M/M_\odot)\,(R/R_\odot)^{-1}\end{aligned}$$

so that to get a large effect, with $\delta_1 \approx 1$, we would need a very massive and dense object: for even the Sun, the GR effects are pretty small.

2.3.2. Diffuse mass

For a diffuse mass, the physical quantities which are important will be

ρ: the density of the object, with dimension $[ML^{-3}]$; and

R: the distance or scale of the object, with dimension $[L]$.

Once again the problem must involve

G: the "inflexibility of spacetime", with dimension $[M^{-1}L^3T^{-2}]$, and

c: the "scaling between distance and time", with dimension $[LT^{-1}]$.

and a single dimensionless quantity

$$\delta_2 = \frac{G\,\rho R^2}{c^2}$$

can be formed. Thus the general relativistic effects are of the order of

$$\begin{aligned}\delta_2 &= 7.4 \times 10^{-28}\,(\rho/\mathrm{kg\,m^{-3}})\,(R/\mathrm{m})^2 \\ &= 0.003\,(\rho/\rho_{\rm crit})\,h_{50}^{-2}\,(R/\mathrm{Gpc})^2\end{aligned}$$

where

$$\rho_{\rm crit} = \frac{3H_0^2}{8\pi G} = 4.7 \times 10^{-27}\,h_{50}^2 \quad \mathrm{kg\,m^{-3}}$$

is the "critical density of the Universe", a quantity which will appear later, and is a characteristic density for matter in the Universe. $H_0 = 50h_{50}$ km s^{-1} Mpc^{-1} is the Hubble constant (and $h_{50} = 0.5 - 1.0$ is a dimensionless scaled Hubble constant), which tells us how rapidly the Universe is currently expanding.

What we see is that GR effects in the Universe as a whole become large ($\delta_2 \approx 1$) only on the largest scales, $R \gtrsim 10$ Gpc (where 1 pc $= 3.086 \times 10^{16}$ m). What we also see is that we can form *no* dimensionless number if we leave R out of the reckoning — this tells us that a simple Universe cannot be both *static* and *uniform* (i.e., have no length scale R). In describing the Universe, there must be a scale size or time, or new physics (which we will ignore as a possibility, here!).

2.4. Goodbye, c

In GR, as in special relativity, we are always encountering factors of c, c^2, c^3, and so on. So it's conventional to change from units where

$$c = 2.99792458 \times 10^8 \quad \mathrm{m\,s^{-1}} \quad \text{(exactly)}$$

to units where $c = 1$. This corresponds to measuring distances in space and distances in time both in the same units, metres.

In all that follows, I'll take $c = 1$, and scale back from the corresponding special-relativitic units to everyday SI units as appropriate. This I will write the SI velocity $v_{\rm SI}$ and acceleration, $a_{\rm SI}$ as

$$\begin{aligned}(v_{\rm SI}/\mathrm{m\,s^{-1}}) &= v \times (2.99792458 \times 10^8) \\ (a_{\rm SI}/\mathrm{m\,s^{-2}}) &= (a/\mathrm{m^{-1}}) \times (2.99792458 \times 10^8)^2\end{aligned}$$

and use the quantities v (dimensionless) and a (units of m^{-1}) in all calculations.

Later on, we will also take $G = 1$, and sometimes in extreme cases, you will see units with $\hbar = 1$, too. But we won't need to take this last step in the present discussion, since we won't be doing too much with quantum cosmology.

3. Special relativity as geometry

3.1. Frames, events, coordinates

Special relativity (SR) is a *geometrical theory*, like general relativity, but with an almost trivial geometry which is why it is usually described algebraically in early courses on relativity. We now need to upgrade your understanding, and make the parallel between

Special relativity	$\longrightarrow$	**General relativity**
principle of relativity	$\longrightarrow$	principle of equivalence
velocity unmeasurable	$\longrightarrow$	acceleration same as gravity

In special relativity we focus on "inertial frames" and "inertial observers", who are in special states of motion — with no relative acceleration, and hence no general relativistic worries. This means that the observers and frames have constant relative velocity.

In any single inertial frame, the distance between points P_1 at (x_1, y_1, z_1) and P_2 at (x_2, y_2, z_2) is

$$s_{12} = \left((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \right)^{1/2}$$

and is constant with time. For infinitesimal separations, $\Delta x_{12} = x_1 - x_2$, for example, the separation can be written as

$$\Delta s_{12}^2 = \Delta x_{12}^2 + \Delta y_{12}^2 + \Delta z_{12}^2$$

and is also constant. All points in the single inertial frame have synchronized clocks. And so we can think of coordinates in an inertial frame as being defined by a set of rigid rods, with clocks at each rod-rod intersection, with all the clocks showing the same time.

In such frames we talk about **events**. For example, event $\mathcal{A}$ which occurs at time t and location (x, y, z), can be described as

$$x(\mathcal{A}) = (t, x, y, z)$$

in coordinate terms — that is, we describe the event's location and time by a quadruplet of numbers, which constitute a four-component vector, which we call a **four-vector**. Remember that each component of $x(\mathcal{A})$ is measured in the same units (metres, for example), since we have rescaled $c \to 1$.

Different coordinates could be used for the same event — for example by deciding to measure (x, y, z) in miles rather than metres, but this doesn't change the event. That is

event	= physically real quantity
coordinates	= convenient description of where event is

For convenience, the four components (t, x, y, z) are often referred to by the four-component notation

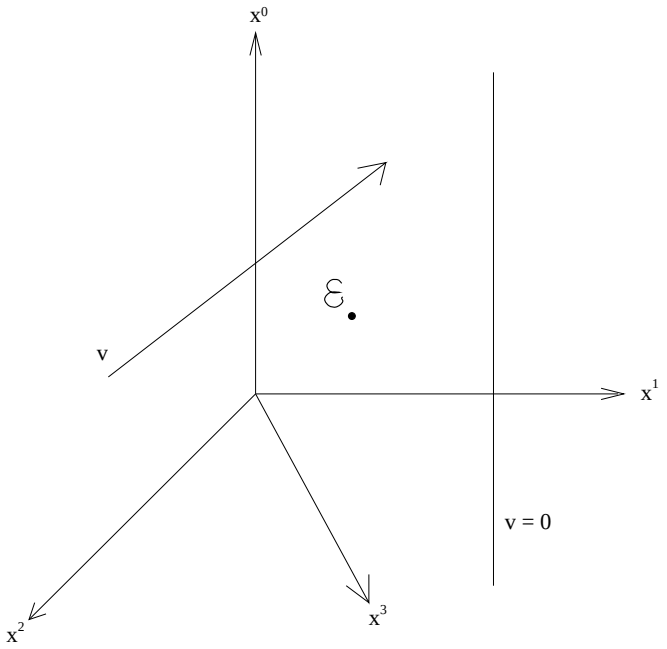
$$(x^0, x^1, x^2, x^3)$$

where $x^0 \equiv t$, $x^1 \equiv x$, $x^2 \equiv y$, $x^3 \equiv z$: the components x^α, with $\alpha = 0, 1, 2, 3$ are merely alternative names for the t, x, y, z coordinates.

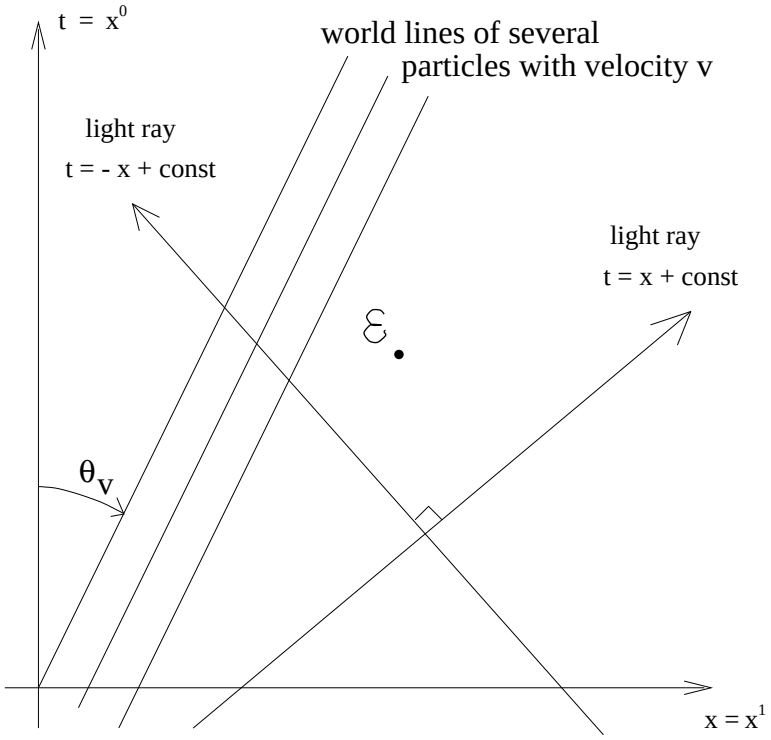
It is convenient shorthand to refer to all the (x^0, x^1, x^2, x^3) as the x^α. And I will always use Greek indices to label components with run from 0 to 3 (i.e., over all spatial components and the time component). If I want to describe time only, I will use x^0, and if I want to describe one of the space components, I will use a Latin index, as $x^{\rm k}$, for example.

3.2. Spacetime Diagrams

We can plot the position of any event $\mathcal{E}$ at coordinates $x(\mathcal{E}) \equiv (t^\mathcal{E}, x^\mathcal{E}, y^\mathcal{E}, z^\mathcal{E})$ on a four-dimensional diagram, as shown.



Here I also show the *world line* of a particle that is at rest (and hence at constant (x^1, x^2, x^3) for all time) and of a particle that is moving at some speed v. However, it's difficult to sketch four dimensions on a two-dimensional page, so we usually suppress the (y, z) coordinates, and plot the location of $\mathcal{E}$ at $(t^\mathcal{E}, x^\mathcal{E}, 0, 0)$ only. Then the diagram takes the simpler form below.



Also shown on this diagram are the *world lines* of light rays (or photons) which move along the positive and negative x axis, and so which have equations

$$t = \pm x + \text{constant}$$

with a positive sign if the particle is moving up the x axis. Clearly we can generalize this: a particle moving at velocity v in this diagram will have equation

$$t = \frac{x}{v} + \text{constant}$$

and hence appear on a line of slope $\frac{1}{v}$. This velocity v is that measured by the observer who "owns" this frame (and so is at rest in it).

Now, let's suppose that there is a second observer, who is at rest in a frame moving at velocity v in the x-direction. How are the coordinates measured by this observer related to the coordinates measured by the observer who is at rest? Let the moving observer be $\bar{\mathcal{O}}$, and the stationary observer be $\mathcal{O}$. Let an event at $x = (t, x, y, z)$ for $\mathcal{O}$ appear at $\bar{x} = (\bar{t}, \bar{x}, \bar{y}, \bar{z})$ for $\bar{\mathcal{O}}$.

By definition, the t axis is the locus of events at constant $(x, y, z) = (0, 0, 0)$ and the $\bar{t}$ axis is the locus of events at constant $(\bar{x}, \bar{y}, \bar{z}) = (0, 0, 0)$. Therefore the $\bar{t}$ axis is just the world line of the moving particle, as drawn above, and the angle between this axis and the t axis, $\theta_{\rm v}$ is given by

$$\tan\theta_{\rm v} = v$$

(since the slope of the world line is $\frac{1}{v}$). Now the harder question is where to draw the $\bar{x}$ axis. Before I simply drew the x axis perpendicular to the t axis — but it would have been better to come up with a real reason for this. In fact, it is possible to construct the location of the $\bar{x}$ axis by geometrical argument, but I won't do that here, but simply invoke the Lorentz Transform (LT), which tells us that the coordinates are related by

$$\begin{aligned}\bar{t} &= \gamma(t - vx)\\ \bar{x} &= \gamma(x - vt)\\ \bar{y} &= y\\ \bar{z} &= z\end{aligned}$$

if the axes are initially aligned (at $t = \bar{t} = 0$ at $x = \bar{x} = 0$) and observer $\bar{\mathcal{O}}$ is moving at velocity v along the x axis of observer $\mathcal{O}$. The quantity

$$\gamma = \left(1 - v^2\right)^{-\frac{1}{2}}$$

is usually called the Lorentz factor.

Notice how prettily symmetrical the equations are in this representation (with $c = 1$) — the time and x relationships look the same. We can make it look even better by making use of our 4-vector notation, and writing the LT as

$$x^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_{\alpha}\, x^{\alpha}$$

with the usual (Einstein summation convention) implied sum over the α index. The LT appears here simply as a 4×4 matrix of transformation coefficients, Λ. We will return to this notation later, when we introduce vectors and tensors.

Using the LT we can see that the equation of the $\bar{t}$ axis is $\bar{x} = 0$, or

$$x - vt = 0$$

which is a line at angle $\theta_{\rm v}$ given by

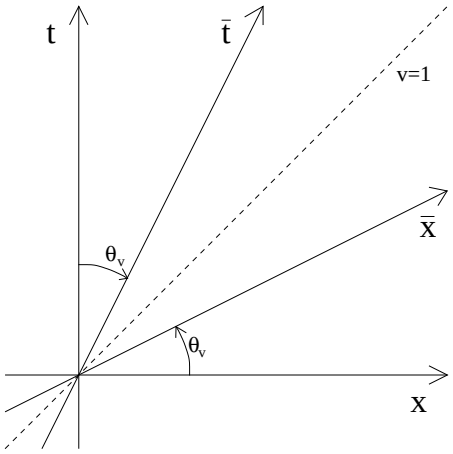
$$\tan\theta_{\rm v} = v$$

to the t axis, as before. Similarly, the $\bar{x}$ axis has $\bar{t} = 0$, or

$$t - vx = 0$$

which is (clearly, by symmetry) a line at angle $\theta_{\rm v}$ to the x axis.

Thus the $\bar{x}$ and $\bar{t}$ axes lie symmetrically with respect to the x and t axes, and to the $x = t$ line (which corresponds to the world line of a photon emitted from the origin at $t = 0$), as shown below.

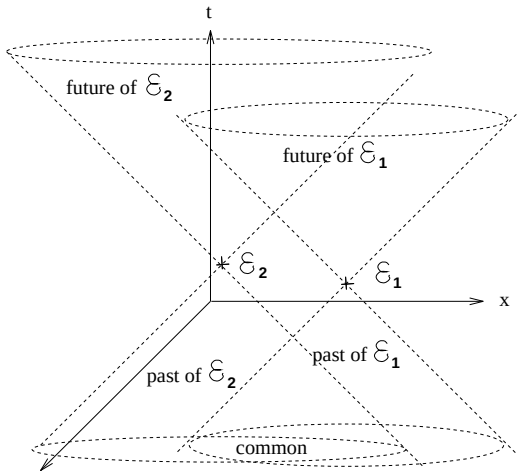


3.3. Spacetime interval

One of the consequences of the LT (which can be proved independently of the LT) is that any two events $\mathcal{A}$ and $\mathcal{B}$ with coordinates $x(\mathcal{A})$, $x(\mathcal{B})$ as seen by observer $\mathcal{O}$ or coordinates $\bar{x}(\mathcal{A})$, $\bar{x}(\mathcal{B})$ as seen by observer $\bar{\mathcal{O}}$ have a spacetime interval

$$\begin{aligned}\Delta s^2 &= \Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2 \\ &= \Delta \bar{x}^2 + \Delta \bar{y}^2 + \Delta \bar{z}^2 - \Delta \bar{t}^2\end{aligned}$$

which is a sort of distance, *the same* expressed by either observer. This is easily proved from the LT, and suggests that the LT is a transformation which preserves *interval* in just the same way that a rotation of the coordinate system in 3-space preserves ordinary distance. A consequence of the **invariance** of spacetime interval (i.e., its identical value as measured by observers at any velocity) is that Δs provides a classification of the relationship between pairs of events.



The **SR-invariant** Δs classifies the various chunks of spacetime as

$\Delta s^2 > 0$ spacelike separated
$\Delta s^2 = 0$ null separated (can be linked by light ray)
$\Delta s^2 < 0$ timelike separated (can be linked by particle moving slower than the speed of light)

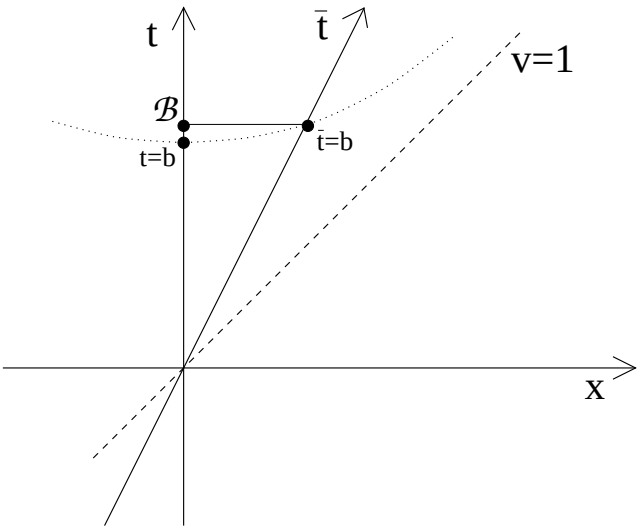
The invariance of Δs then means that all observers will agree about the past and future of any event $\mathcal{E}$ — and also about those events which are elsewhere and causally unconnected with it.

3.4. Time dilation and length contraction

Let's use the spacetime interval and the symmetry of the $\bar{t}$ and $\bar{x}$ axes to see what is implied for the observation of moving clocks and moving rods. We won't need to use the LT itself.

3.4.1. Time dilation

The spacetime diagram that we need to consider here is



Events at times $\bar{t} = 0$ and b fixed at the origin of the moving frame $\bar{\mathcal{O}}$ are observed by observer $\mathcal{O}$ at $t = 0$ and t at $x = 0$ and x. Spacetime interval is an invariant, so we can write

$$\bar{t}^2 - \bar{x}^2 \equiv b^2 = t^2 - x^2$$

which is the equation of a hyperbola, as shown, in the (t, x) spacetime diagram, which

crosses the $\bar{t}$ axis at $\bar{t} = b$, and the t axis at $t = b$. Note that this implies that intervals in a spacetime diagram don't behave the same way as distances in 3-D space: the appearance is that the length of the $\bar{t}$ axis from the origin to the $\bar{t} = b$ point is longer than the length of the t axis from the origin to the $t = b$ point, but this isn't true. We know that the $\bar{t}$ axis is the line

$$t = \frac{x}{v}$$

so the intersection of this line with the hyperbola lies at

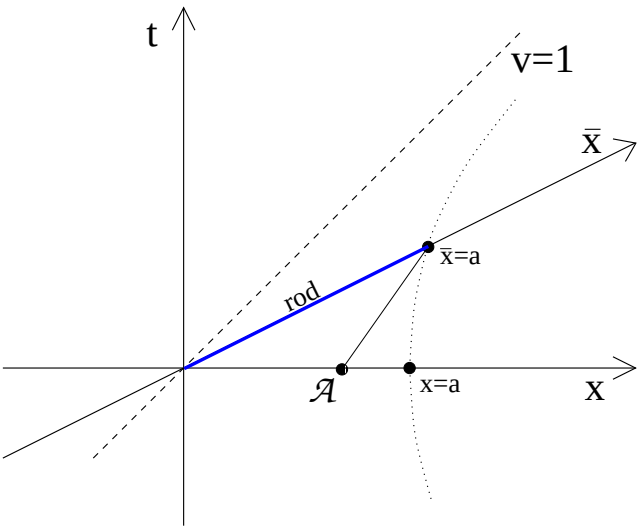
$$t = \gamma b$$
$$x = bv\gamma$$

where $\gamma = (1 - v^2)^{-1/2}$. That is, a time interval b on a clock at rest in the $\bar{\mathcal{O}}$ frame is seen as a time interval γb in the $\mathcal{O}$ frame which sees the clock moving. That is, there is a time dilation in the sense that

$$\Delta t_{\mathcal{O}} = \gamma \Delta \bar{t}_{\bar{\mathcal{O}}} \quad .$$

3.4.2. Lorentz contraction

The spacetime diagram that we need to consider here is



Now the rod is defined by its ends, which lie at $\bar{x} = 0$ and $\bar{x} = a$ at all times $\bar{t}$ in its rest frame $\bar{\mathcal{O}}$ which moves at velocity v along the x axis in the frame of observer $\mathcal{O}$. We can proceed in the same way as before. Spacetime interval is an invariant, so we can write

$$\bar{t}^2 - \bar{x}^2 \equiv -a^2 = t^2 - x^2$$

which is the equation of a hyperbola, as shown, in the (t, x) spacetime diagram (and the same comment about scale of the diagram applies). But we know that the $\bar{x}$ axis is the line

$$t = vx$$

so the intersection of this line with the hyperbola lies at

$$t = av\gamma$$
$$x = a\gamma$$

which apparently corresponds to a length dilation. But the end of the rod at $\bar{x} = 0$ is observed at $t = 0$ while the front of the rod is being observed at $t = av\gamma$. This is **not** what we usually mean by length — we want to measure the length of the rod at a particular time in the $\mathcal{O}$ frame.

Therefore we need to know, instead, where the front of the rod was at $t = 0$, so that we can get a $\mathcal{O}$-instantaneous measurement of the rod's length. Since the rod moves at speed v, we know that the front of the rod lies on a line of slope $\frac{1}{v}$, and the world line of the end of the rod crosses the x axis at point $\mathcal{A}$. The x-coordinate of event $\mathcal{A}$ is therefore

$$a\gamma - v(av\gamma) = a(1 - v^2)\gamma = \frac{a}{\gamma}$$

and so there is a length contraction in the sense that

$$\Delta x_{\mathcal{O}} = \frac{\Delta \bar{x}_{\bar{\mathcal{O}}}}{\gamma} \quad .$$

I want to emphasize that it is the asymmetry in the definition of time and length is what causes the difference between time *dilation* and length *contraction*, since without the back-extrapolation to the x axis we would get a length dilation.

4. Cartesian tensors in special relativity

4.1. Vectors

The special-relativistic transformation law for vectors (the Lorentz Transform, LT) can be written in the compact form I wrote earlier

$$x^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_{\alpha}\, x^{\alpha}$$

where the Greek indices, $\bar{\alpha}$ and α can be any of the $0, 1, 2, 3$ corresponding to t, x, y, z, so I can write for the vector $\vec{x}$

$$\begin{aligned}\vec{x} &\equiv (x^0, x^1, x^2, x^3) \\ &\equiv (t, x, y, z) \quad \text{in } \mathcal{F} \\ &\equiv (x^{\bar{0}}, x^{\bar{1}}, x^{\bar{2}}, x^{\bar{3}}) \\ &\equiv (\bar{t}, \bar{x}, \bar{y}, \bar{z}) \quad \text{in } \bar{\mathcal{F}}\end{aligned}$$

and $\Lambda(v)$ is the LT matrix

$$\Lambda^{\bar{\alpha}}{}_{\alpha} = \begin{cases} \gamma & \alpha = \bar{\alpha} = 0 \text{ or } 1 \\ -\gamma v & \alpha = 1,\ \bar{\alpha} = 0 \text{ or } \alpha = 0,\ \bar{\alpha} = 1 \\ 1 & \alpha = \bar{\alpha} = 2 \text{ or } 3 \\ 0 & \text{otherwise} \end{cases}$$

for "normal alignment" of the axes. More generally, for frame $\bar{\mathcal{F}}$ moving at velocity v in direction $\mathbf{n}$ in frame $\mathcal{F}$, we can write the general LT

$$\Lambda^{\bar{\alpha}}{}_{\alpha}(v, \mathbf{n}) = \begin{cases} \gamma & \alpha = \bar{\alpha} = 0 \\ -\gamma v n^{\mathrm{i}} & \alpha = 0, \bar{\alpha} = i \text{ or } \alpha = i, \bar{\alpha} = 0 \\ (\gamma - 1) n^{\mathrm{i}}\, n^{\mathrm{j}} + \delta^{\mathrm{ij}} & \alpha = i, \bar{\alpha} = j \end{cases}$$

where $\mathbf{n}$ is a unit 3-vector with

$$\left(n^1\right)^2 + \left(n^2\right)^2 + \left(n^3\right)^2 = 0$$

There are some important points to make about the notation.

(1) $\vec{x}$ is a (pseudo)-vector, a physical/mathematical quantity. It can be represented by some set of coordinates $\{x^{\alpha}\}$ or another set $\{x^{\bar{\alpha}}\}$, but has a reality over and above the coordinate representation. Thus we say

> $\vec{x}$ can be written in the coordinates of $\mathcal{F}$, using (x^0, x^1, x^2, x^3), or in the coordinates of $\bar{\mathcal{F}}$, using $(x^{\bar{0}}, x^{\bar{1}}, x^{\bar{2}}, x^{\bar{3}})$, but it is the **same** $\vec{x}$ in either case ... and hence the bar is shown over the index, and not over the vector $\vec{x}$.

(2) The index α runs over $0 \rightarrow 3$, as explained before, with 0 indicating the time component (some books use $1 \rightarrow 4$, but I believe the 0 spells out time's uniqueness better).

(3) We use the Einstein summation convention, that repeated indices be summed over. Thus

$$\begin{aligned} x^{\bar{\alpha}} &= \Lambda^{\bar{\alpha}}{}_{\alpha}\, x^{\alpha} \\ &= \Lambda^{\bar{\alpha}}{}_{0}\, x^{0} + \Lambda^{\bar{\alpha}}{}_{1}\, x^{1} + \Lambda^{\bar{\alpha}}{}_{2}\, x^{2} + \Lambda^{\bar{\alpha}}{}_{3}\, x^{3} \\ &= \Lambda^{\bar{\alpha}}{}_{0}\, x^{0} + \Lambda^{\bar{\alpha}}{}_{\mathrm{i}}\, x^{\mathrm{i}} \end{aligned}$$

which represents a set of **four** results, for $\bar{\alpha} = 0, 1, 2, 3$, for example

$$x^{\bar{2}} = \Lambda^{\bar{2}}{}_{0}\, x^{0} + \Lambda^{\bar{2}}{}_{1}\, x^{1} + \Lambda^{\bar{2}}{}_{2}\, x^{2} + \Lambda^{\bar{2}}{}_{3}\, x^{3} \quad .$$

The summed ("dummy") index, α, can equally be called β, γ, or whatever. That is, $\Lambda^{\bar{\alpha}}{}_{\alpha}\, x^{\alpha}$ is identical to $\Lambda^{\bar{\alpha}}{}_{\gamma}\, x^{\gamma}$ – the $\bar{\alpha}^{\mathrm{th}}$ component of the transformed $\vec{x}$.

> Note that summations are **always** over an "up" and a "down" index (what in the old language were called the contravariant and covariant indices, respectively ... but I'll not use that older, and worse, nomenclature).

x^{α} = component of a vector
x_{α} = component of a *one-form*
$\Lambda^{\bar{\alpha}}{}_{\beta}$ = component of a matrix transforming a vector to another vector

> Note carefully the distinction between $k^{\alpha}x_{\alpha}$, which is a single quantity (a scalar), and $k^{\alpha}x_{\beta}$, which is one of 16 numbers (a component of a tensor).

(4) A vector is **defined** as a quantity that transforms like $\vec{x}$. That is, the set of four numbers (a^0, a^1, a^2, a^3) represent a vector $\vec{a}$ if the components transform as

$$a^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_{\alpha}\, a^{\alpha}$$

when converting from values in the $\mathcal{F}$ frame to values in the $\bar{\mathcal{F}}$ frame (under Lorentz transform and also, perhaps, rotation of the coordinate system).

4.2. Basis Vectors

There are four special vectors in $\mathcal{F}$, which describe the four axes. These are the **basis vectors**, $\vec{e}_{\alpha}$. Written in frame $\mathcal{F}$ coordinates,

$$\begin{aligned} \vec{e}_0 &= (1, 0, 0, 0) \\ \vec{e}_1 &= (0, 1, 0, 0) \\ \vec{e}_2 &= (0, 0, 1, 0) \\ \vec{e}_3 &= (0, 0, 0, 1) \end{aligned}$$

and similarly the four basis vectors in $\bar{\mathcal{F}}$ are

$$\begin{aligned} \vec{e}_{\bar{0}} &= (1, 0, 0, 0) \\ \vec{e}_{\bar{1}} &= (0, 1, 0, 0) \\ \vec{e}_{\bar{2}} &= (0, 0, 1, 0) \\ \vec{e}_{\bar{3}} &= (0, 0, 0, 1) \end{aligned}$$

which I can write in shorthand as

$$(\vec{e}_\alpha)^\beta = \delta_\alpha{}^\beta$$
$$(\vec{e}_{\bar{\alpha}})^{\bar{\beta}} = \delta_{\bar{\alpha}}{}^{\bar{\beta}}$$

where $\delta_\alpha{}^\beta$ is the Kronecker delta, which is

$$\delta_\alpha{}^\beta = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{if } \alpha \neq \beta \end{cases}$$

We can express any vector $\vec{a}$ in two ways

$$\vec{a} = a^\alpha \, \vec{e}_\alpha = a^{\bar{\alpha}} \, \vec{e}_{\bar{\alpha}}$$

which tells us that $\vec{a}$ is a linear superposition of the basis vectors, with the components saying how much of each basis vector is contained in $\vec{a}$. But we know that if $\vec{a}$ is a vector,

$$a^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_\alpha \, a^\alpha$$

and hence

$$\vec{a} = a^\alpha \, \vec{e}_\alpha = a^{\bar{\alpha}} \, \vec{e}_{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_\beta \, a^\beta \, \vec{e}_{\bar{\alpha}}$$

Change the order in the finite sum and relabel the dummy index β:

$$\begin{aligned} a^\alpha \, \vec{e}_\alpha &= a^\beta \, \Lambda^{\bar{\alpha}}{}_\beta \, \vec{e}_{\bar{\alpha}} \\ &= a^\alpha \, \Lambda^{\bar{\alpha}}{}_\alpha \, \vec{e}_{\bar{\alpha}} \end{aligned}$$

or

$$a^\alpha \left(\Lambda^{\bar{\alpha}}{}_\alpha \, \vec{e}_{\bar{\alpha}} - \vec{e}_\alpha \right) = 0$$

But $\vec{a}$ was an arbitrary vector, so the a^α can be assumed to be non-zero, and hence I can write the basis vectors in $\mathcal{F}$ as a weighted sum of the basis vectors in $\bar{\mathcal{F}}$ as

$$\vec{e}_\alpha = \Lambda^{\bar{\alpha}}{}_\alpha \, \vec{e}_{\bar{\alpha}}$$

This is the transformation law for basis vectors, it states how the basis vectors in $\mathcal{F}$ and $\bar{\mathcal{F}}$ are related. Contrast the transformation of components,

$$a^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_\alpha \, a^\alpha$$

which expresses a vector component in $\bar{\mathcal{F}}$ as a weighted sum of vector components in $\mathcal{F}$. The components and the basis vectors transform "oppositely".

The meaning of this can be seen by considering the components of a vector $\vec{a}$ expressed in a frame $\bar{\bar{\mathcal{F}}}$ which moves at velocity $-\mathbf{v}$ relative to frame $\bar{\mathcal{F}}$ which moves at velocity $\mathbf{v}$ in frame $\mathcal{F}$. These components are

$$a^{\bar{\bar{\alpha}}} = \Lambda^{\bar{\bar{\alpha}}}{}_{\bar{\alpha}}(-\mathbf{v}) \, \Lambda^{\bar{\alpha}}{}_\alpha(\mathbf{v}) \, a^\alpha$$

but $\bar{\bar{\mathcal{F}}}$ is identical with $\mathcal{F}$, so

$$a^{\alpha} = \Lambda^{\alpha}{}_{\bar{\alpha}}(-\mathbf{v})\,\Lambda^{\bar{\alpha}}{}_{\beta}(\mathbf{v})\,a^{\beta}$$

which must be true for any $\vec{a}$, so that

$$\Lambda^{\alpha}{}_{\bar{\alpha}}(-\mathbf{v})\,\Lambda^{\bar{\alpha}}{}_{\beta}(\mathbf{v}) = \delta^{\alpha}{}_{\beta}$$

and so $\Lambda(-\mathbf{v})$ is the inverse matrix to $\Lambda(\mathbf{v})$: the inverse LT is simply the LT with reversed sign of the velocity.

4.3. Scalar products

So far we've talked only about the transformation of quantities like $\vec{x}$ under the LT, and the addition of scaled quantities like $\vec{x}$, for example in the summation

$$\vec{a} = a^{\alpha}\,\vec{e}_{\alpha}$$

which says that the sum of a scaled set of vectors is itself a vector. Nothing very surprising there.

The next step simply reproduces something that we did earlier, and extends the idea. Earlier, I talked about the invariance of the interval,

$$\Delta s^2 = -\left(\Delta x^0\right)^2 + \left(\Delta x^1\right)^2 + \left(\Delta x^2\right)^2 + \left(\Delta x^3\right)^2$$

under Lorentz transforms, where $\vec{\Delta x}$ is a vector distance. This is just the simplest type of **scalar product**, $\vec{\Delta x}.\vec{\Delta x}$, and the LT was defined so that Δs^2 is a scalar and itself invariant under LT. But *all* vectors transform like $\vec{\Delta x}$: this is the *definition* of a vector. And therefore for all vectors $\vec{a}$,

$$-\left(a^0\right)^2 + \left(a^1\right)^2 + \left(a^2\right)^2 + \left(a^3\right)^2 = \text{invariant}$$

and the vector $\vec{a}$ is described as spacelike, null, or timelike according to whether this invariant is positive, zero, or negative.

We can extend this idea, and define the scalar product of two vectors $\vec{a}$ and $\vec{b}$ by

$$\vec{a}.\vec{b} = -a^0 b^0 + a^1 b^1 + a^2 b^2 + a^3 b^3$$

and it is possible to prove that $\vec{a}.\vec{b}$ is invariant by considering

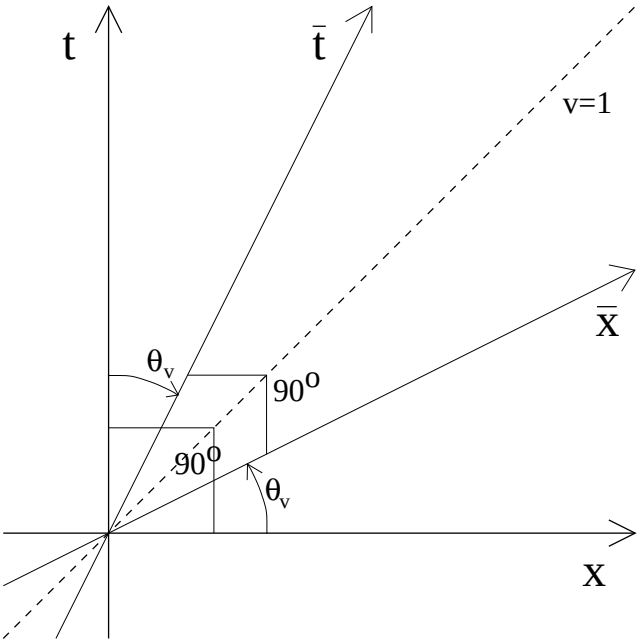
$$\begin{aligned}\left(\vec{a}+\vec{b}\right).\left(\vec{a}+\vec{b}\right) &= \text{invariant} \\ &= (\vec{a}.\vec{a}) + \left(\vec{b}.\vec{b}\right) + 2\left(\vec{a}.\vec{b}\right) \\ &= \text{invariant} + \text{invariant} + 2\left(\vec{a}.\vec{b}\right)\end{aligned}$$

and hence $\vec{a}.\vec{b}$ is invariant.

We can also do this for the basis vectors of $\mathcal{F}$, the $\vec{e}_\alpha$, when

$$\begin{aligned} \vec{e}_0.\vec{e}_0 &= -1 \\ \vec{e}_1.\vec{e}_1 = \vec{e}_2.\vec{e}_2 &= \vec{e}_3.\vec{e}_3 = 1 \\ \vec{e}_\alpha.\vec{e}_\beta &= 0 \qquad \text{if } \alpha \neq \beta \end{aligned}$$

That is, the $\vec{e}_\alpha$ define a set of mutually orthogonal axes of unit magnitude. This is also true for the $\vec{e}_{\bar{\alpha}}$, so on a spacetime diagram, despite what the diagram looks like, the angles as marked below are correct!



The distorted geometry in this figure is entirely because the definition of distance (and hence the amplitude of a vector, $\vec{a}$), is

$$-\left(a^0\right)^2 + \left(a^1\right)^2 + \left(a^2\right)^2 + \left(a^3\right)^2$$

rather than

$$\left(a^0\right)^2 + \left(a^1\right)^2 + \left(a^2\right)^2 + \left(a^3\right)^2$$

It is convenient to summarize the results for $\vec{e}_\alpha.\vec{e}_\beta$ as

$$\vec{e}_\alpha.\vec{e}_\beta \equiv \eta_{\alpha\beta} = \begin{cases} -1 & \alpha = \beta = 0 \\ +1 & \alpha = \beta \neq 0 \\ 0 & \alpha \neq \beta \end{cases}$$

where the quantity g whose components are $\eta_{\alpha\beta}$ gives us information about the *geometrical structure of special relativity.*

We can put g to use immediately, in the calculation of the scalar product. In frame $\mathcal{F}$, I can write any vector in the form

$$\vec{a} = a^\alpha \vec{e}_\alpha$$

and hence

$$\vec{a}.\vec{b} = (a^\alpha \vec{e}_\alpha) \,.\, \left(b^\beta \vec{e}_\beta\right) = a^\alpha b^\beta \left(\vec{e}_\alpha.\vec{e}_\beta\right) = a^\alpha b^\beta \eta_{\alpha\beta}$$

which tells us that the set of quantities $\eta_{\alpha\beta}$, defined by $\vec{e}_\alpha.\vec{e}_\beta$, is what is needed to combine the $\mathcal{F}$-frame components of vectors into the **frame-invariant** scalar product.

$\eta_{\alpha\beta}$ = components of **metric tensor** g in SR
metric tensor = quantity that combines two vectors to produce a scalar

A **scalar** is a frame-invariant number.

4.4. Tensors

g, with components $\eta_{\alpha\beta}$, is your first non-trivial **tensor**, a quantity that takes two vectors into a scalar. This is an example of a **general rule**.

a tensor of type $\begin{pmatrix} 0 \\ N \end{pmatrix}$ *is a* **linear** *function of* N *vectors into real numbers (i.e., scalars)*

This is a component-independent statement: it means that we get the same real number by doing this operation whatever frame the component of the N vectors or the tensor are specified in.

The simplest tensor, of type $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is clearly a scalar, since when fed no vectors it returns a real number.

In order for a tensor to return a pure number, when fed particular components of a vector in frame $\mathcal{F}$, that tensor must also have components in $\mathcal{F}$. Then

the components of a $\begin{pmatrix} 0 \\ N \end{pmatrix}$ *tensor in* $\mathcal{F}$ *are the numbers obtained when that tensor is fed* N *basis vectors.*

For example, for the tensor g,

$$\begin{aligned} g(\vec{e}_\alpha, \vec{e}_\beta) &\equiv \vec{e}_\alpha.\vec{e}_\beta \\ &\equiv \eta_{\alpha\beta} \end{aligned}$$

using the definition of $g(\ ,\ )$, and expressing the components of g as $g_{\alpha\beta} = \eta_{\alpha\beta}$, as is appropriate for special relativity (but not for GR ...). Clearly based on this argument,

g is a $\begin{pmatrix}0\\2\end{pmatrix}$ tensor, since it takes two vectors into a scalar (the scalar product of basis vectors).

4.4.1 One-forms

If $\begin{pmatrix}0\\0\end{pmatrix}$ tensors are scalars, and $\begin{pmatrix}0\\2\end{pmatrix}$ tensors are things like g, what are $\begin{pmatrix}0\\1\end{pmatrix}$ tensors?

These **one-forms** are defined as quantities which can be fed vectors to return scalars: if $\tilde{p}$ is a one-form, then

$$\tilde{p}(\vec{a}) \quad \text{is a real number.}$$

In particular, if $\tilde{p}$ is fed a basis vector, then the real number produced will be a component of $\tilde{p}$

$$p_\alpha \equiv \tilde{p}(\vec{e}_\alpha)$$

The notation here is important: compare

$$\begin{aligned} p_\alpha &= \alpha^{\rm th} \text{ component of a one-form } \tilde{p} \text{ in frame } \mathcal{F} \\ a^\alpha &= \alpha^{\rm th} \text{ component of a vector } \vec{a} \text{ in frame } \mathcal{F} \end{aligned}$$

Now, $\tilde{p}$ is a linear function of vectors, with

$$\tilde{p}(k\,\vec{a}) = k\,\tilde{p}(\vec{a})$$

for scalar k, so

$$\begin{aligned} \tilde{p}(\vec{a}) &= \tilde{p}(a^\alpha \vec{e}_\alpha) \\ &= a^\alpha\,\tilde{p}(\vec{e}_\alpha) \\ &= a^\alpha\,p_\alpha \\ &\equiv a^0 p_0 + a^1 p_1 + a^2 p_2 + a^3 p_3 \end{aligned}$$

This **contraction** of $\vec{a}$ and $\tilde{p}$ differs from a scalar product in having all positive signs — there's no statement about the metric tensor g in the formulae above (and note that we can't make a scalar product of $\tilde{p}$ and $\vec{a}$ because they aren't both vectors!).

How does a one-form transform under the LT? We can write

$$\begin{aligned} p_{\bar\alpha} &= \tilde{p}(\vec{e}_{\bar\alpha}) \\ &= \tilde{p}(\Lambda^\alpha{}_{\bar\alpha}\,\vec{e}_\alpha) \\ &= \Lambda^\alpha{}_{\bar\alpha}\,\tilde{p}(\vec{e}_\alpha) \\ &= \Lambda^\alpha{}_{\bar\alpha}\,p_\alpha \end{aligned}$$

that is, the components of a one-form transform like the basis vectors, using the **inverse** LT to vector components:

one-form component transform	$p_{\bar{\alpha}} = \Lambda^{\alpha}{}_{\bar{\alpha}} p_{\alpha}$
vector component transform	$a^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_{\alpha} a^{\alpha}$
basis vector transform	$\vec{e_{\bar{\alpha}}} = \Lambda^{\alpha}{}_{\bar{\alpha}} \vec{e_{\alpha}}$

Just as the vectors $\vec{a}$ have basis vectors $\vec{e}_\alpha$, so the one-forms $\tilde{p}$ have basis one-forms $\tilde{\omega}^\alpha$ so that

$$\tilde{p} = p_\alpha \, \tilde{\omega}^\alpha$$

These basis one-forms must be consistent with $\tilde{p}(\vec{a}) = p_\alpha a^\alpha$, so

$$\begin{aligned} \tilde{p}(\vec{a}) &= p_\alpha \, a^\alpha \\ &= (p_\alpha \, \tilde{\omega}^\alpha) \left(a^\beta \, \vec{e}_\beta \right) \\ &= p_\alpha \, a^\beta \, \left(\tilde{\omega}^\alpha \, \vec{e}_\beta \right) \end{aligned}$$

which, since $\vec{a}$ and $\tilde{p}$ are arbitrary, requires that

$$\tilde{\omega}^\alpha \vec{e}_\beta = \delta^\alpha{}_\beta$$

which defines the one-forms $\{\tilde{\omega}^\alpha\}$ in terms of the $\{\vec{e}_\alpha\}$, and from which it can be proven that the basis one-forms transform like vector components under LT,

$$\tilde{\omega}^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_{\alpha} \, \tilde{\omega}^\alpha$$

What, physically, is a one-form? It is what a vector crosses, and tells you how much a quantity changes under changes of position. The derivative of a scalar function is a one-form.

That is, it helps to think of a one-form as a local approximation to

a set of contour lines	in two dimensions (2D)
a set of planes of equal value	in 3D
a set of volumes of equal value	in 4D

Forming the quantity $\tilde{p}(\vec{a})$ then tells you how much the quantity $\tilde{p}$ changes in direction $\vec{a}$. For example, in 2D, a close-packed set of contour lines corresponds to a large $\tilde{p}$, since a small vector displacement $\vec{a}$ gives a large change in value of whatever is being contoured. Well-separated contours would correspond to a small $\tilde{p}$, since it takes a large vector displacement $\vec{a}$ to get much change.

Clearly this is closely related to the gradient of a field ϕ, $\nabla\phi$, that you encountered when doing vector calculus. We'll see exactly how close later (in lecture 5)

4.4.2 $\begin{pmatrix} 0 \\ N \end{pmatrix}$ tensors

A one-form is a linear mapping of a vector to a real number

$$\tilde{p}(\vec{a}) = \text{scalar} \quad .$$

A $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ tensor is a linear mapping of two vectors to a real number

$$P(\vec{a}, \vec{b}) = \text{scalar} \quad .$$

And clearly a $\begin{pmatrix} 0 \\ \mathrm{N} \end{pmatrix}$ tensor takes N vectors to a scalar

$$P(\vec{a}, \vec{b}, \vec{c}, ...) = \text{scalar} \quad .$$

Just as for a one-form, the components of a $\begin{pmatrix} 0 \\ \mathrm{N} \end{pmatrix}$ tensor can be obtained by feeding the tensor the basis vectors

$$P_{\alpha\beta\gamma...} = P(\vec{e}_\alpha, \vec{e}_\beta, \vec{e}_\gamma, ...)$$

so that

$$\begin{aligned} P(\vec{a}, \vec{b}, \vec{c}, ...) = P(a^\alpha \vec{e}_\alpha, b^\beta \vec{e}_\beta, c^\gamma \vec{e}_\gamma, ...) &= a^\alpha b^\beta c^\gamma P(\vec{e}_\alpha, \vec{e}_\beta, \vec{e}_\gamma, ...) \\ &= a^\alpha b^\beta c^\gamma P_{\alpha\beta\gamma...} \end{aligned}$$

If $\mathrm{P}(\vec{\mathrm{a}}, \vec{\mathrm{b}}, \vec{\mathrm{c}}, ...)$ is to be invariant under LT, then it also follows that

$$P_{\bar{\alpha}\bar{\beta}\bar{\gamma}...} = \Lambda^\alpha{}_{\bar{\alpha}} \, \Lambda^\beta{}_{\bar{\beta}} \, \Lambda^\gamma{}_{\bar{\gamma}} \, ... P_{\alpha\beta\gamma...}$$

and

$$P = P_{\alpha\beta\gamma...} \, \tilde{\omega}^\alpha \otimes \tilde{\omega}^\beta \otimes \tilde{\omega}^\gamma ...$$

where $\otimes$ means $\tilde{p} \otimes \tilde{q}(\vec{a}, \vec{b}) = \tilde{p}(\vec{a})\tilde{q}(\vec{b})$. Note, in general, $\tilde{p} \otimes \tilde{q} \neq \tilde{q} \otimes \tilde{p}$: the tensor product, $\otimes$ is **not** commutative.

4.4.3 $\begin{pmatrix} M \\ N \end{pmatrix}$ tensors

Let's stop doing special cases and launch to the general case.

> *a tensor of type* $\begin{pmatrix} M \\ N \end{pmatrix}$ *is a* **linear** *function of M one-forms and N vectors into real numbers (i.e., scalars)*

Thus a $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ tensor is a vector, a $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ tensor is a one-form, and so on.

For example, consider a tensor function of one one-form and three vectors, $R(\ ,\ ,\ ,\ )$. For this tensor

$$\begin{aligned} R(\tilde{p},\vec{b},\vec{c},\vec{d}) &= \text{scalar} \\ R^{\alpha}{}_{\beta\gamma\delta} &= R(\tilde{\omega}^{\alpha},\vec{e}_{\beta},\vec{e}_{\gamma},\vec{e}_{\delta}) \\ R^{\bar{\alpha}}{}_{\bar{\beta}\bar{\gamma}\bar{\delta}} &= \Lambda^{\bar{\alpha}}{}_{\alpha}\,\Lambda^{\beta}{}_{\bar{\beta}}\,\Lambda^{\gamma}{}_{\bar{\gamma}}\,\Lambda^{\delta}{}_{\bar{\delta}}\,R^{\alpha}{}_{\beta\gamma\delta} \end{aligned}$$

and so on. R may seem like a complicated quantity, but like all tensors it really isn't — it's merely a linear function of four other quantities to a real number. And the R quantity is defined so that it has a physical reality independent of the observer who happens to observe it.

5. The flat-space metric tensor, velocity, momentum

5.1. The flat-space metric tensor

Earlier we turned up a quantity g, with components $\eta_{\alpha\beta}$ in special relativity

$$\eta_{\alpha\beta} \equiv \vec{e}_\alpha.\vec{e}_\beta = \begin{cases} -1 & \alpha = \beta = 0 \\ +1 & \alpha = \beta \neq 0 \\ 0 & \alpha \neq \beta \end{cases}$$

and I said that g tells us about the *geometrical structure of special relativity.* Let's investigate further.

Clearly I can always choose g to be symmetric, since

$$g(\vec{a},\vec{b}) = g(\vec{b},\vec{a}) = \vec{a}.\vec{b}$$

by definition. So what is $g(\vec{a},\ )$? It must be a quantity which, when fed a vector like $\vec{b}$, returns a scalar. Therefore it's a one-form. Since it's clearly a one-form closely related to $\vec{a}$, let's call it $\tilde{a}$. That is,

$$\tilde{a}(\vec{b}) \equiv g(\vec{a},\vec{b}) = \vec{a}.\vec{b} \quad .$$

Since g is symmetric, $g(\ ,\vec{a})$ must also be $\tilde{a}$, since when fed $\vec{b}$ in the first slot the same scalar $\vec{a}.\vec{b}$ is returned.

The components of $\tilde{a}$ are $\tilde{a}(\vec{e}_\alpha)$ by definition, and these

$$\begin{aligned} a_\alpha \equiv \tilde{a}(\vec{e}_\alpha) = g(\vec{a},\vec{e}_\alpha) &= \vec{a}.\vec{e}_\alpha \\ &= (a^\beta \vec{e}_\beta).\vec{e}_\alpha \\ &= a^\beta(\vec{e}_\beta.\vec{e}_\alpha) \\ &= \eta_{\alpha\beta}a^\beta \end{aligned}$$

so that g, with components $\eta_{\alpha\beta}$ is exactly what is needed to change from vector to one-form components

$$a_\alpha = \eta_{\alpha\beta}a^\beta \quad .$$

We say that **the metric tensor** g is the quantity which converts a vector to its equivalent one-form by "lowering the index".

In the same way, the inverse of the metric tensor g^{-1}, with components $\eta^{\alpha\beta}$, can be used to convert a one-form to its equivalent vector by "raising the index": since

$$\begin{aligned} a_\alpha &= \eta_{\alpha\beta}a^\beta \\ \eta^{\gamma\alpha}a_\alpha &= \eta^{\gamma\alpha}\eta_{\alpha\beta}a^\beta \\ &= \delta^\gamma{}_\beta a^\beta \\ &= a^\gamma \end{aligned}$$

Strictly this is only possible if $\det g \neq 0$, as it is in special relativity. The components of the metric g and its inverse are

$$\eta_{\alpha\beta} = \begin{cases} -1 & \alpha = \beta = 0 \\ +1 & \alpha = \beta \neq 0 \\ 0 & \alpha \neq \beta \end{cases}$$
$$\eta^{\alpha\beta} = \begin{cases} -1 & \alpha = \beta = 0 \\ +1 & \alpha = \beta \neq 0 \\ 0 & \alpha \neq \beta \end{cases}$$

and it's easy enough to confirm that these are inverses.

So now given any one-form, $\tilde{p}$, or vector $\vec{a}$, we can easily make the associated vector or one-form.

It follows that the magnitude of a one-form, defined as equal to the equivalent vector magnitude, is given by

$$\begin{aligned} \tilde{p}.\tilde{p} &\equiv \vec{p}.\vec{p} \\ &= \eta_{\alpha\beta} p^\alpha p^\beta \\ &= \eta_{\alpha\beta} (\eta^{\alpha\gamma} p_\gamma) (\eta^{\beta\delta} p_\delta) \\ &= \eta^{\alpha\gamma} \delta^\delta{}_\alpha p_\gamma p_\delta \\ &= \eta^{\gamma\delta} p_\gamma p_\delta \end{aligned}$$

and with the help of g we can use either the one-form or the vector components to calculate magnitudes.

Finally, just as for vectors the scalar product is

$$\vec{a}.\vec{b} = -a^0 b^0 + a^1 b^1 + a^2 b^2 + a^3 b^3$$

so for one-forms it is

$$\tilde{p}.\tilde{q} = -p_0 q_0 + p_1 q_1 + p_2 q_2 + p_3 q_3$$

and

$$\vec{a}.\vec{b} = \eta_{\alpha\beta} a^\alpha b^\beta = a^\alpha b_\alpha = a_\alpha b^\alpha$$

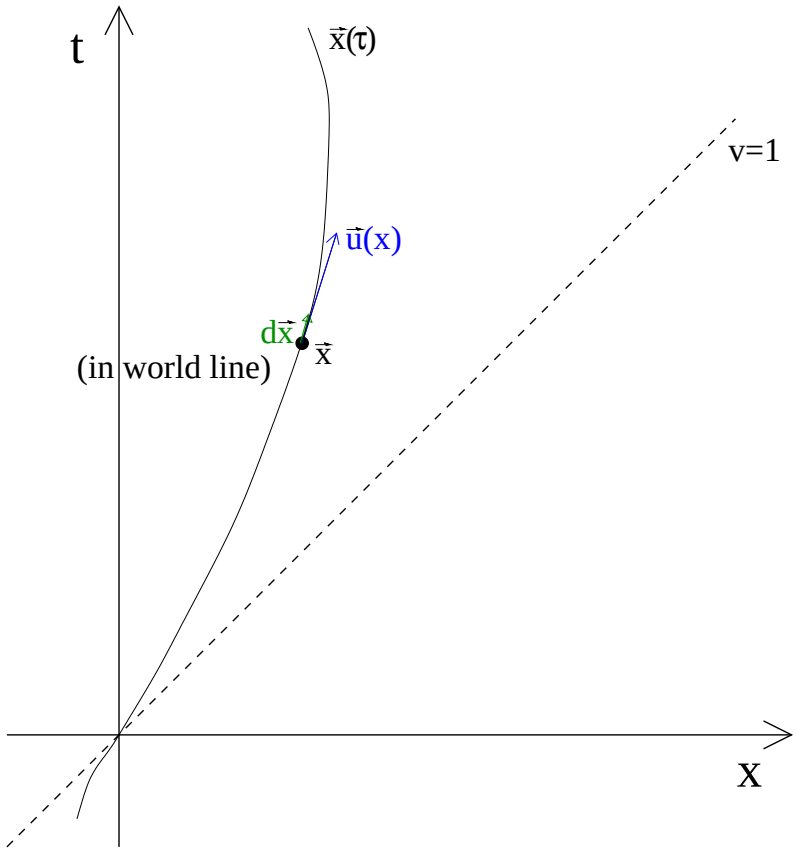
or the identical quantity can be calculated using the one-forms

$$\tilde{a}.\tilde{b} = \eta^{\alpha\beta} a_\alpha b_\beta = a_\alpha b^\alpha = a^\alpha b_\alpha \quad .$$

5.2. Velocity

Let's use some of these ideas to talk about velocity in more general terms than before. So far, we've used the expression "velocity" to refer to the velocity of one frame relative

to another in the Lorentz Transform. We can regard this as a simple version of a more complicated issue — how to define the velocity of any one object (for example a dust particle) as observed in some frame $\mathcal{F}$.



We're used to velocity as the rate of change of position. We must extend this idea for SR, since position is now a four-vector rather than the old 3-vector. On a space-time diagram, velocity is the local gradient of the line — in more than two dimensions we have to generalize slightly,

the velocity of a particle is the tangent vector to that particle's world line.

That's OK for the direction of the vector, which is now clearly geometrical and coordinate-independent. But what of the length? This is defined by the unit of time that we choose — and we choose to measure time in the local comoving reference frame of the particle, usually referred to as the **MCRF, momentarily comoving reference frame** since the particle may be accelerating. This means that the velocity of a particle is the **time basis vector in the MCRF of the particle** — i.e., it's $\vec{e}_{\bar{0}}$ in the $\bar{\mathcal{F}}$, where $\bar{\mathcal{F}}$ is the MCRF.

This is, essentially, obvious. In the MCRF, the "velocity" of the particle is one second per second on the time axis, since the particle is (by definition) at rest. This is a geometrical definition, and hence is true in any other reference frame.

Thus we write for the particle velocity

$$\vec{u} = \vec{e}_{\bar{0}} \quad .$$

But we know that $\vec{e}_{\bar{0}} = \Lambda^{\alpha}{}_{\bar{0}}\, \vec{e}_{\alpha}$, and so

$$\vec{u} = \Lambda^{\alpha}{}_{\bar{0}}\, \vec{e}_{\alpha} \quad .$$

Using the general LT, given earlier

$$\Lambda^{\bar{\alpha}}{}_{\alpha}(v, \mathbf{n}) = \begin{cases} \gamma & \alpha = \bar{\alpha} = 0 \\ -\gamma v n^{\mathrm{i}} & \alpha = 0, \bar{\alpha} = i \text{ or } \alpha = i, \bar{\alpha} = 0 \\ (\gamma - 1) n^{\mathrm{i}}\, n^{\mathrm{j}} + \delta^{\mathrm{ij}} & \alpha = i, \bar{\alpha} = j \end{cases}$$

for a boost at velocity v in direction $\mathbf{n}$,

$$\begin{aligned} \vec{u} &= \gamma \vec{e}_0 + \gamma v n^1 \vec{e}_1 + \gamma v n^2 \vec{e}_2 + \gamma v n^3 \vec{e}_3 \\ &= \gamma \vec{e}_0 + \gamma \vec{v} \end{aligned}$$

as usual, so that

$$u^0 = \gamma \quad ; \quad u^1 = \gamma v^1 \quad ; \quad u^2 = \gamma v^2 \quad ; \quad u^3 = \gamma v^3 \quad .$$

We can also get at the velocity in a different way from the world line, using more sophisticated language than the LT, now that we have the metric tensor g to give us the scalar product. Since velocity is the **unit tangent vector** of a particle's world line, if we consider a small displacement $d\vec{x}$ of the particle, near $\vec{x}$, such that the particle remains on the world line,

$$ds^2 = d\vec{x}.d\vec{x} = \text{invariant}$$

and real particle displacements are timelike, so that in the MCRF of the particle the spacelike displacements $dx^{\bar{i}} = 0$. Thus I can write

$$ds^2 = -d\tau^2$$

where τ is the time increment in the particle's MCRF. Thus

$$\left(\frac{d\vec{x}}{d\tau} \right)^2 = \left(\frac{d\vec{x}.d\vec{x}}{d\tau^2} \right) = -1$$

so that $\frac{d\vec{x}}{d\tau}$ is a unit vector (a vector of unit magnitude, even if negative in the square), is constructed to be tangent to the particle's world line, and in the MCRF coincides with $\vec{e}_{\bar{0}}$. Therefore it *is* the particle velocity

$$\text{particle velocity } , \vec{u} = \frac{d\vec{x}}{d\tau} = \frac{d\vec{x}}{(-ds^2)^{1/2}}$$

which tells us exactly how to calculate the particle's velocity given its world-line.

Why is this important? Because it's our first example of a fundamental relativistic principle — that the equations of physics **must** be written (or be capable of being written) in a form that's independent of the frame of reference. The result

$$\vec{u} = \frac{d\vec{x}}{d\tau}$$

obeys this dictum: $\vec{u}$ is a physical quantity, a coordinate-independent vector, $\vec{x}$ is a physical quantity, another coordinate-independent vector, and τ is a third physical quantity, a coordinate-independent scalar that can be constructed from $\vec{x}$ using the metric g.

5.3. One-forms as derivatives

We can use this idea of velocity further, to illuminate the interpretation of one-forms a bit more. It turns out that a one-form is a **derivative**. Define a scalar field $\phi(\vec{x})$ over all events $\vec{x}$. Let this field be observed by a particle on some world-line, and parameterize that world line by a scalar, τ, which we might as well make the proper time. Of course, we can't use τ for a photon, but there is another quantity we can use instead. The velocity of a particle on this world line (see figure) is

$$\vec{u} = \frac{d\,\vec{x}}{d\,\tau} = \left(\frac{dt}{d\tau}, \frac{dx}{d\tau}, \frac{dy}{d\tau}, \frac{dz}{d\tau}\right)$$

note, here we have *exact differentials*, not *partial*.

How much does ϕ change on a world line?

$$\frac{d\phi}{d\tau} = \frac{d}{d\tau}\phi(\vec{x}) = \frac{\partial\phi}{\partial t}\frac{d\,t}{d\,\tau} + \frac{\partial\phi}{\partial x}\frac{d\,x}{d\,\tau} + \frac{\partial\phi}{\partial y}\frac{d\,y}{d\,\tau} + \frac{\partial\phi}{\partial z}\frac{d\,z}{d\,\tau}$$

Therefore

$$\frac{d\phi}{d\tau} = \frac{\partial\phi}{\partial t}\,u^0 + \frac{\partial\phi}{\partial x}\,u^1 + \frac{\partial\phi}{\partial y}\,u^2 + \frac{\partial\phi}{\partial z}\,u^3$$

where $\frac{d\phi}{d\tau}$ is a real number, uniquely associated with any event on the curve, and therefore a scalar, and the u^α are the components of a vector. Now, something which combines with a vector to give a scalar is a one-form, therefore $\tilde{d}\phi$ is a one-form with components

$$\tilde{d}\phi = \left(\frac{\partial\phi}{\partial t}, \frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right)$$

We call $\tilde{d}\phi$ is the gradient of ϕ. Let's check on the Lorenz transformation of $\tilde{d}\phi$

$$\tilde{d}\phi_{\bar{\alpha}} = \tilde{d}\phi_\alpha \frac{\partial x^\alpha}{\partial x^{\bar{\alpha}}} = \tilde{d}\phi_\alpha \frac{\partial}{\partial x^{\bar{\alpha}}}\left(\Lambda^\alpha{}_{\bar{\beta}} x^{\bar{\beta}}\right) = \Lambda^\alpha{}_{\bar{\alpha}}\,\tilde{d}\phi_\alpha$$

so $\tilde{d}\phi$ transforms as we expect. It is conventional to write $\phi_{,\alpha}$ for the partial derivative

$$\phi_{,\alpha} = \frac{\partial \phi}{\partial x^\alpha}$$

and you should note, that α is a *subscript* on the left hand side (since $\phi_{,\alpha}$ is the $\alpha^{\rm th}$ component of the one-form $\tilde{d}\phi$) and a *superscript* on the right hand side (since x^α is the $\alpha^{\rm th}$ component of vector $\vec{x}$).

We can use this to look at the basis one-forms in a different way, too. Suppose that we take ϕ as being one of the x^α. Then

$$x^\alpha{}_{,\beta} = \delta^\alpha{}_\beta$$

but for basis one-forms $\tilde{\omega}^\alpha$, we have

$$\tilde{\omega}^\alpha(\vec{e}_\beta) = \delta^\alpha{}_\beta \quad ,$$

and therefore we can *identify* the basis one-forms as the gradients of the coordinate functions

$$\tilde{\omega}^\alpha \equiv \tilde{d}\, x^\alpha$$

so that the $\beta^{\rm th}$ component of $\tilde{\omega}^\alpha$ is $\delta^\alpha{}_\beta$. Thus, for example, we can write

$$\tilde{d}\phi \equiv \phi_{,\alpha}\, \tilde{\omega}^\alpha \equiv \phi_{,\alpha}\, \tilde{d}x^\alpha \equiv \frac{\partial \phi}{\partial x^\alpha}\, \tilde{d}x^\alpha \quad .$$

5.4. Momentum

We define momentum exactly as in ordinary dynamics — it's the velocity times a mass,

$$\vec{p} = m\vec{u}$$

where m is sometimes referred to as the "rest mass". Perhaps it's better to think of it purely as a scalar quantity — something that is invariant under Lorentz transform — intrinsic to the particle being discussed.

We manipulate momentum just as we would manipulate it in ordinary dynamics: the total momentum of a system of particles is obtained by a vector sum: if (i) labels the particles (not any index!)

$$\vec{p}_{\rm total} = \sum_{(i)} \vec{p}_{(i)}$$

and the non-relativistic result that the total momentum is conserved suggests that total 4-vector momentum should also be conserved.

And so it is, as demonstrated by experiment. **But that need not have been the case**, since there are a number of ways that we might have defined momentum so that the correct non-relativistic limit was achieved ... for example, by adding higher-order terms in the velocity, such as

$$\vec{p} = m\vec{u} + ku^3\vec{u}$$

where k might be another quality of the particle other than mass.

Note that p^0 has the significance of being called "energy": that is, energy in SR in a particular reference frame is **defined** as the $0^{\rm th}$ component of the momentum 4-vector in that reference frame. And it's easy to prove that in the low-v limit

$$p^0 \rightarrow m\left(1 + \frac{1}{2}v^2\right)$$

corresponding to a "rest-mass energy" of m (or mc^2, putting back the c factors) and a "kinetic energy" of $\frac{1}{2}mv^2$.

6. The stress-energy tensor, conservation laws

6.1. Particle flux

We've seen that the velocity vector $\vec{u} = \vec{e}_{\tilde{0}}$, the time basis vector of the particle's world line in its MCRF, and that the particle's momentum vector is $\vec{p} = m\vec{u}$ where m is the particle's (rest) mass. What about a set of particles moving with velocity **v**? What is their collective number per second passing some point (or the number per metre, in geometrical units)?

What sort of quantity are we talking about? For a given **surface**, we want the number of particles per unit time per unit area crossing the surface. That is, for a given 3-D surface (which will have 3D of space, or 2D of space and 1D of time, or some combination), what is the number of particles?

To make progress, we need to say what a **surface** is. We define a surface as some boundary which obeys $\phi(\vec{x}) =$ constant for some choice of the function ϕ. That is, $\phi_{,\alpha}$, the gradient of $\phi(\vec{x})$ is some non-zero value, and so we can define a unit one-form normal to the surface of constant ϕ by

$$\tilde{n} \equiv \frac{\tilde{d}\phi}{\left|\tilde{d}\phi\right|}$$

where

$$\left|\tilde{d}\phi\right| = \sqrt{\left|\tilde{d}\phi.\tilde{d}\phi\right|} \quad .$$

Now, for the particle flux we want something which combines with $\tilde{n}$ to give a scalar (a count of number of particles, clearly a real number). This **must be a vector**, call it $\vec{N}$. Then $\tilde{n}(\vec{N})$ will be the number of particles per unit "volume" across the surface defined by $\tilde{n}$.

In the particle MCRF, there is no motion, so $\vec{N}$ has no spatial but only a time-like part. And outside the MCRF we expect

$$N \propto v$$
$$N \propto n_{\rm p}$$

where $n_{\rm p}$ is the particle density, since this is true at small velocities. So **try**

$$\vec{N} = n_{\rm p}\vec{u}$$

where $\vec{u}$ is the velocity vector, and $n_{\rm p}$ is the number of particles per unit volume in the MCRF (a unique, scalar, number which all observers can agree about).

Since this is the product of a scalar and a vector, it is a vector as required. And at small v,

$$N^0 \rightarrow n_{\rm p} \quad ; \quad N^1 \rightarrow n_{\rm p}v^{\rm x} \quad ; \quad N^2 \rightarrow n_{\rm p}v^{\rm y} \quad ; \quad N^3 \rightarrow n_{\rm p}v^{\rm z}$$

also as required. At high speeds,

$$N^0 \rightarrow n_{\rm p}\left(1-v^2\right)^{-\frac{1}{2}}$$

corresponding to the particle density increasing because of Lorentz contraction, just as we would expect. Therefore $\vec{N}$ has the right sort of properties, and is a sensible flux vector. Whether it's really the **correct** flux vector or not depends on logical and experimental tests — but in fact it all works out well, and again we have been able to build a logical 4-vector based on the non-relativistic limit.

Now, if we take $\phi = z$ (say, in $\mathcal{F}$ coordinates), then

$$\tilde{n} \equiv \tilde{dz} = (0,0,0,1) \quad \text{in } \mathcal{F} \text{ coordinates}$$

and so

$$\tilde{n}(\vec{N}) = N^\alpha n_\alpha = N^{\rm z} = \gamma n_{\rm p} v^{\rm z}$$

which is the number of particles per unit time per unit (x,y) area flowing across the $\bar{z}$ surface.

If we choose $\phi = t$, then following the same logic

$$\tilde{n} \equiv \tilde{dt} = (1,0,0,0) \quad \text{in } \mathcal{F} \text{ coordinates}$$

and

$$\tilde{n}(\vec{N}) = N^\alpha n_\alpha = N^{\rm t} = \gamma n_{\rm p}$$

which is the number of particles per unit (x,y,z) volume flowing across the time surface, which is simply the number of particles per unit volume.

Thus the number of particles per unit volume is simply a time-like flux. In the spatial directions, "unit volume" converts to "unit area × unit time", and we have the ordinary spatial flux.

6.2. The stress-energy tensor

This tensor is of particular importance, since it will turn out to be the source of gravitation in GR. And its form in SR will carry over directly to GR.

6.2.1. The stress-energy tensor for dust

"Dust" in GR is a gas without pressure — a set of particles which move, but have no internal energy density (no random velocities or inter-particle forces) in their MCRF.

Then we can construct the stress-energy tensor from its definition:

the stress-energy tensor is the flux of momentum in some direction ($\alpha = 0, 1, 2, 3$) across a surface in some direction ($\beta = 0, 1, 2, 3$).

That is, the stress-energy tensor T is a second-rank tensor, with components such that when fed a surface one-form it returns the (vector) flux of momentum across the surface.

For dust,

$$T = \vec{p} \otimes \vec{N} \quad .$$

Now, it's obvious that T defined like this is a second-rank tensor, since it's constructed from two vectors

$$\begin{aligned} \vec{p} &= m\vec{u} \quad \text{momentum/particle} \\ \vec{N} &= n_{\rm p}\vec{u} \quad \text{particle flux} \end{aligned}$$

and in the MCRF of the particles, where only the time component of $\vec{u}$ is non-zero,

$$\begin{aligned} p^{\bar{0}} &= m \\ N^{\bar{0}} &= n_{\rm p} \\ T^{\bar{0}\bar{0}} &= mn_{\rm p} = \rho \quad \text{(all other terms zero)} \end{aligned}$$

where ρ is the density of particles in the MCRF. That is, in this frame there is no momentum flux (since all the particles move together). We can use ρ to rewrite

$$T = \rho\vec{u} \otimes \vec{u}$$

so that the components of T are

$$\begin{aligned} T^{\alpha\beta} &= T(\tilde{\omega}^\alpha, \tilde{\omega}^\beta) \\ &= \rho\vec{u}(\tilde{\omega}^\alpha)\vec{u}(\tilde{\omega}^\beta) \\ &= \rho u^\alpha u^\beta \end{aligned}$$

and therefore in a frame in which the particles move at velocity v in direction $\mathbf{n}$,

$$T^{\mu\nu} = \begin{cases} \rho\gamma^2 & \mu = \nu = 0 \\ \rho v n^{\rm i}\gamma^2 & \mu = 0,\ \nu = i \text{ or } \mu = i,\ \nu = 0 \\ \rho v^2 n^{\rm i} n^{\rm j}\gamma^2 & \mu = i,\ \nu = j \end{cases}$$

which shows the symmetry of T explicitly (clearly $T^{\mu\nu} = T^{\nu\mu}$), and also that the apparent density of particles in a frame where the particles are moving with speed v is

$$\rho_{\rm apparent} = \rho\gamma^2$$

which is, if you like, the LT for density. One γ factor arises from the Lorentz contraction of volume in the direction of the flow, the other from the "relativistic transformation of mass" (a horrible concept).

6.2.2. Relativistic thermodynamics

We do thermodynamics in relativity in the MCRF (the "simplest possible frame"). Then the laws of thermodynamics are:

0th law There is an empirical temperature relating to heat content: i.e., there is a meaningful quantity called temperature in the MCRF.

1st law The law of conservation of energy:

$$\Delta Q = \Delta E + P\Delta V$$

where E is the total energy content of some fluid element, ΔQ is the heat flow into that heat element, and $P\Delta V$ is the work done by the fluid element by changing its volume by ΔV (i.e., a loss of internal energy if the fluid expands).

Now, if fluid of volume V contains N particles of mass m, and the particles aren't created or destroyed, then the particle density is $n_{\rm p} = \frac{N}{V}$, and so

$$V = \frac{N}{n_{\rm p}}$$
$$\Delta V = -\frac{N}{n_{\rm p}^2}\Delta n_{\rm p}$$

and so if we write the energy of the fluid element as

$$E = \rho V$$

including the rest-mass energy in the total density,

$$\Delta E = \rho\Delta V + V\Delta\rho$$

and back in the first law,

$$\begin{aligned}\Delta Q &= (\rho\Delta V + V\Delta\rho) + P\Delta V \\ &= \frac{N}{n_{\rm p}}\Delta\rho - \frac{N}{n_{\rm p}^2}\Delta n_{\rm p}(\rho + P)\end{aligned}$$

so that the heat absorbed per particle is

$$\begin{aligned}\delta q = \frac{\Delta Q}{N} &= \frac{1}{n_{\rm p}}\left(\Delta\rho - \frac{\Delta n_{\rm p}}{n_{\rm p}}(\rho + P)\right) \\ &= T_{\rm f}\Delta S \qquad \text{by definition}\end{aligned}$$

where S is the entropy per particle and $T_{\rm f}$ is the temperature of the fluid.

2nd law Statement about the properties of S — since this is true in the MCRF, it is true everywhere (all observers can identify the MCRF).

3rd law Statement about the properties of S — since this is also true in the MCRF, it is true everywhere (all observers can identify the MCRF).

6.2.3. The stress-energy tensor for a perfect fluid

Now how does this dollop of thermodynamics help? In the MCRF we know that

$T^{\bar{0}\bar{0}}$ = energy density, ρ

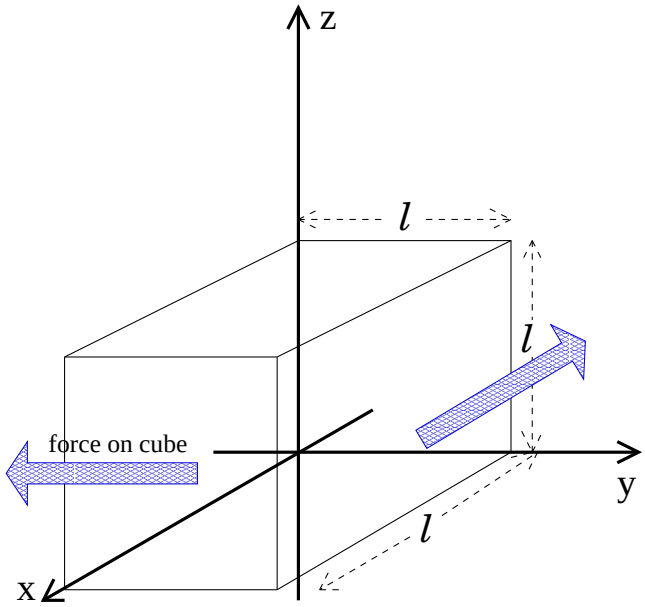
$T^{\bar{0}\bar{i}}$ = flux of $\bar{0}$-momentum (energy) across $\bar{i}$ surface; related to heat conduction
= 0 in a perfect fluid

$T^{\bar{i}\bar{0}}$ = flux of $\bar{i}$-momentum across $\bar{0}$ surface
= 0, nothing flows in the MCRF

$T^{\bar{i}\bar{j}}$ flux of $\bar{i}$-momentum across $\bar{j}$ surface, **stress**

We can show that $T^{\bar{i}\bar{j}}$ must be symmetric, since otherwise a fluid would go into very rapid rotation: consider the cube of fluid below. Then $T^{\rm ix}$ is the rate of transfer of momentum by fluid in the cube on the $+x$-face in the i-direction — that is, it's the force in the i-direction that fluid inside the cube exerts on its neighbours on this face. So the torque about the z axis exerted by the neighbouring cube because of this face is the opposite, and is

$$\Gamma_{\rm z} = -T_{\rm yx} l^2 \left(\frac{1}{2} l\right)$$



and the same torque arises from the $-x$ face. And there are similar torques due to the $+y$ and $-y$ faces of the cube. Summing them all up

$$\begin{aligned}\Gamma_{\rm z} &= -2T_{\rm yx}l^2\left(\frac{1}{2}l\right) + 2T_{\rm xy}l^2\left(\frac{1}{2}l\right)\\ &= l^3\left(T^{\rm xy} - T^{\rm yx}\right)\\ &= I\ddot{\theta}_{\rm z}\end{aligned}$$

where I is the moment of inertia of the cube about the z axis and $\ddot{\theta}_{\rm z}$ is the angular acceleration. But we know that $I \propto Ml^2 \propto \rho l^5$, where M is the mass of the cube and ρ is its density, so the angular acceleration is

$$\ddot{\theta}_{\rm z} \propto \frac{T^{\rm xy} - T^{\rm yx}}{l^2}$$

which tends to infinity as $l \to 0$ unless $T^{\rm xy} = T^{\rm yx}$. Thus $T^{\bar{i}\bar{j}}$ is symmetric, as asserted.

Furthermore, in a perfect fluid there is no viscosity, so all forces must be perpendicular to the faces of the mass elements and

$$T^{\bar{i}\bar{j}} = 0 \qquad \text{unless } \bar{\rm i} = \bar{\rm j}.$$

Also, all directions $\bar{i} = 1,\ 2,\ 3$ must be the same, so

$$T^{\bar{i}\bar{j}} = \text{constant} \times \delta^{\bar{i}\bar{j}} \quad .$$

We call the constant here the pressure, P: it is the normal momentum flux. Both P and ρ are defined in the MCRF (so that they are relativistic scalars). Thus, in the MCRF,

$$T^{\bar{\alpha}\bar{\beta}} = \begin{cases} \rho & \bar{\alpha} = \bar{\beta} = 0 \\ P & \bar{\alpha} = \bar{\beta} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

We can LT this to another frame (or find a frame-invariant expression which reduces to this in the MCRF and minimally couples) to obtain

$$T^{\alpha\beta} = (\rho + P)u^\alpha u^\beta + P\eta^{\alpha\beta}$$

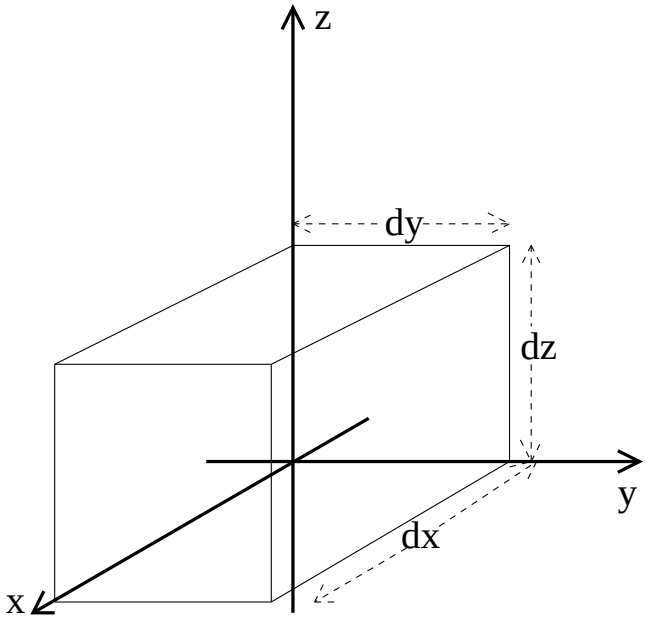
which can be written in a way which is manifestly frame-invariant and hence valid in all frames

$$T = (\rho + P)\vec{u} \otimes \vec{u} + Pg^{-1}$$

since g as we defined it acts on vectors, while T acts on one-forms, so we have to use the "raised" version of g, which is its inverse. Note that this reduces properly to the dust result if $P = 0$.

6.3. Conservation laws

Consider a chunk of fluid, as shown below.



Then the rate of flow of energy into the volume is

$$\left(T^{0\mathrm{x}}(x=0) - T^{0\mathrm{x}}(x=dx)\right) dy\, dz + \left(T^{0\mathrm{y}}(y=0) - T^{0\mathrm{y}}(y=dy)\right) dz\, dx + \left(T^{0\mathrm{z}}(z=0) - T^{0\mathrm{z}}(z=dz)\right) dx\, dy$$

since $T^{0\mathrm{i}}$ is the flux of 0-momentum (energy) in the $+i$ direction on the face perpendicular to the i axis. If the sides of the volume are small this becomes

$$-\left(\frac{\partial T^{0\mathrm{x}}}{\partial x} + \frac{\partial T^{0\mathrm{y}}}{\partial y} + \frac{\partial T^{0\mathrm{z}}}{\partial z}\right) dx\, dy\, dz$$

which is the rate of increase of the energy density contained within the cube, or

$$\frac{\partial}{\partial t}\left(T^{00}\,dx\,dy\,dz\right)$$

Setting these equal,

$$\frac{\partial T^{00}}{dt}+\frac{\partial T^{0\mathrm{x}}}{dx}+\frac{\partial T^{0\mathrm{y}}}{dy}+\frac{\partial T^{0\mathrm{z}}}{dz}=0$$

or, in snappier language,

$$T^{0\alpha}{}_{,\alpha}=0 \quad .$$

This is the law of conservation of energy — and clearly if such a law applies to the 0 component of the stress-energy tensor it must also apply to the i-components (i.e., there must be conservation of momentum too). Then

$$T^{\alpha\beta}{}_{,\beta}=0$$

or in even snappier language

$$\nabla.T=0$$

Note that, strictly, $\nabla.T=T^{\alpha\beta}{}_{,\alpha}$: the divergence applies to the first slot of T. But T is symmetric, so $T^{\alpha\beta}=T^{\beta\alpha}$, so all is well.

We can do exactly the same thing for particle flux: the rate of flow of particles into the cubic volume is

$$\begin{aligned}\left(N^{\mathrm{x}}(x=0)-N^{\mathrm{x}}(x=dx)\right)dy\,dz&+\left(N^{\mathrm{y}}(y=0)-N^{\mathrm{y}}(y=dy)\right)dz\,dx\\&+\left(N^{\mathrm{z}}(z=0)-N^{\mathrm{z}}(z=dz)\right)dx\,dy\\&=-\left(\frac{\partial N^{\mathrm{x}}}{\partial x}+\frac{\partial N^{\mathrm{y}}}{\partial y}+\frac{\partial N^{\mathrm{z}}}{\partial z}\right)dx\,dy\,dz\\&=\frac{\partial}{\partial t}\left(dx\,dy\,dz\,N^0\right)\end{aligned}$$

since $N^0\,dx\,dy\,dz$ is the number of particles in the volume. Rearranging, and using the tidier notation

$$N^{\alpha}{}_{,\alpha}=0 \qquad \text{or}$$
$$\nabla.\vec{N}=0$$

is the law of conservation of particle number.

We derived these two conservation laws with no assumptions about the fluids being "perfect" — they will be true for other types of fluid too, unless work is being done from another source (e.g., viscosity), when $\nabla.T\neq 0$, or if particles are created or destroyed, with $\nabla.\vec{N}\neq 0$. Even then, we could extend the definition of T, or add source and sink terms in the derivations, to include proper treatment of where the energy or particles go, although this leads to more complicated expressions for the conservation laws.

It's possible to use these two conservation laws to derive the equations of fluid motion and mass conservation for relativistic gases, but although that's exciting and a good exercise in manipulation, it's too long for a lecture.

6.5. Integral forms of conservation laws, Gauss' theorem

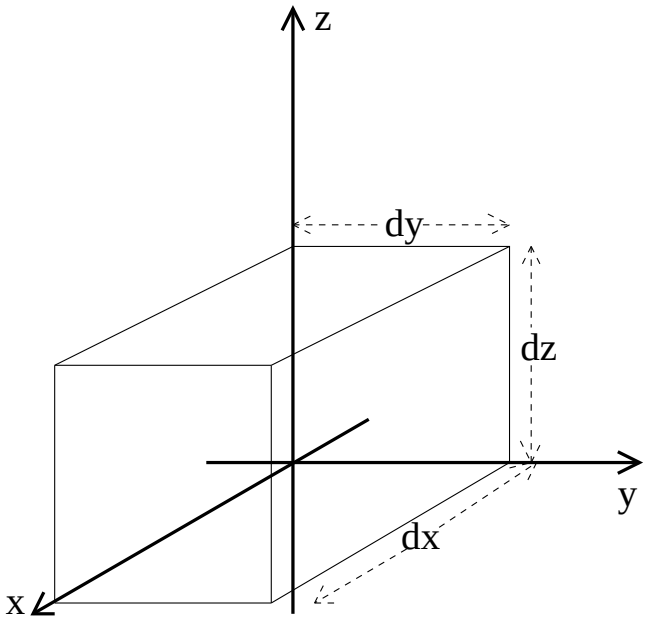
We're dealing with a 4-D space here, so the analogue of Gauss' theorem that we would expect is something like

$$\int_\Omega \nabla.W \, d^4\Omega = \oint_{\partial\Omega} W^\gamma \, d^3\Sigma_\gamma$$

where $d^4\Omega$ is an element of volume Ω, but now four-dimensional. $d^3\Sigma$ is an element of the surface bounding Ω, $\partial\Omega$. This surface element is **outward-directed**. And $\nabla.W \equiv W^\gamma{}_{,\gamma}$ for vector W (or simply referring to one of the indices of a more complicated W).

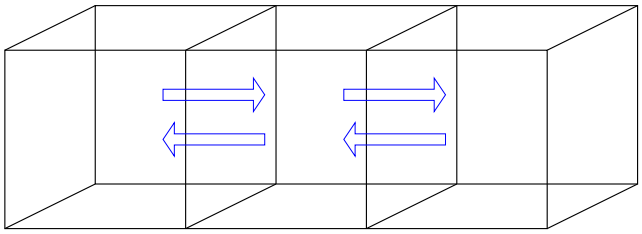
This four-dimensional version of Gauss' theorem is derived exactly like the three-dimensional version is derived: over the elemental (3D) volume depicted below, the change in W^γ inside Ω because of the flux through the surface $\partial\Omega$ is

$$\begin{aligned}\oint_{\partial\Omega} W^{\rm i} \, d^2\Sigma_{\rm i} &= (W^{\rm x}(x=dx) - W^{\rm x}(x=0)) \; dy\, dz \\ &+ (W^{\rm y}(y=dy) - W^{\rm y}(y=0)) \; dx\, dz \\ &+ (W^{\rm z}(z=dz) - W^{\rm z}(z=0)) \; dx\, dy \\ &= W^{\rm i}{}_{,{\rm i}} \, dx\, dy\, dz \\ &= \int_\Omega W^{\rm i}{}_{,{\rm i}} \, d^3\Omega\end{aligned}$$



Now for an arbitrary volume Ω, we fill the volume with lots of little cuboids. All the surface integrals cancel over the interior surfaces, and we're left only with the outer surfaces. Therefore the change in W^γ within the volume

$$\oint_{\partial\Omega} W^{\rm i} d^2\Sigma_{\rm i} = \int_\Omega W^{\rm i}{}_{,{\rm i}} d^3\Omega$$

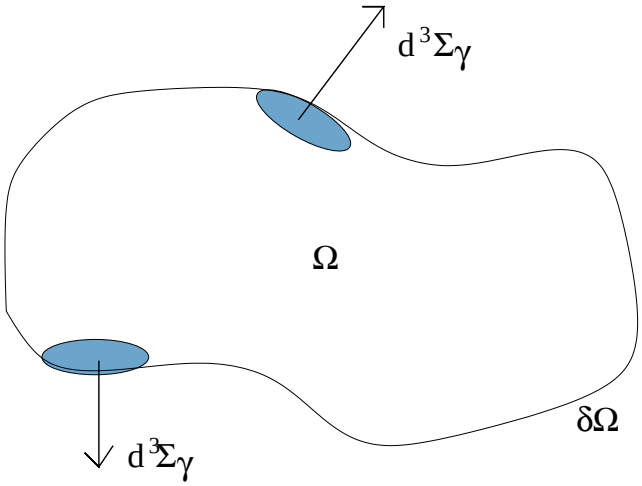


equal and opposite
(dxdy points out of volume)

and making an identical argument over the four dimensional space-time (and considering a more complicated cuboid) will lead to

$$\oint_{\partial\Omega} W^\gamma \, d^3\Sigma_\gamma = \int_\Omega W^\gamma{}_{,\gamma} \, d^4\Omega$$

We are left asking what $d^4\Omega$ is — it's the 4-space volume element, $dx^0\, dx^1\, dx^2\, dx^3$. And the surface element $d^3\Sigma$ is a one-form surface element directed outwards from the four-space volume, as sketched below.



7. Non-Cartesian tensors

7.1. Polar coordinates and tensors

We need to develop a way of dealing with non-orthogonal axes to be able to handle GR, where the coordinates are arbitrary and can become pretty distorted. So I'll develop the type of notation that we need by working first with polar coordinates in 2D as an illustration before **asserting** that this works in more complicated coordinates and applying it to the case that we're interested in.

Consider a system of coordinates (ξ, η) related to (x, y) by

$$\xi = \xi(x, y)$$
$$\eta = \eta(x, y)$$

then small changes $(\Delta x, \Delta y)$ cause changes in (ξ, η) as

$$\Delta\xi = \frac{\partial\xi}{\partial x}\Delta x + \frac{\partial\xi}{\partial y}\Delta y$$
$$\Delta\eta = \frac{\partial\eta}{\partial x}\Delta x + \frac{\partial\eta}{\partial y}\Delta y$$

Specifically for polar coordinates (r, θ),

$$r = \left(x^2 + y^2\right)^{\frac{1}{2}}$$
$$\theta = \arctan\frac{y}{x}$$
$$\Delta r = \frac{x}{r}\Delta x + \frac{y}{r}\Delta y = \cos\theta\,\Delta x + \sin\theta\,\Delta y$$
$$\Delta\theta = -\frac{y}{r^2}\Delta x + \frac{x}{r^2}\Delta y = -\frac{\sin\theta}{r}\Delta x + \frac{\cos\theta}{r}\Delta y$$

If the (ξ, η) or (r, θ) coordinates are to make sense, then the mapping from (x, y) must be unique (1:1), and we must require that the mapping has an inverse. In that case, the Jacobian of the transformation

$$\left|\frac{\partial(\xi,\eta)}{\partial(x,y)}\right| \equiv \det\begin{pmatrix} \frac{\partial\xi}{\partial x} & \frac{\partial\xi}{\partial y} \\ \frac{\partial\eta}{\partial x} & \frac{\partial\eta}{\partial y} \end{pmatrix} \neq 0$$

For (r, θ), this corresponds to the requirement that

$$\left|\frac{\partial(r,\theta)}{\partial(x,y)}\right| = \frac{1}{r} \neq 0$$

which is fine everywhere except the singular point $r = \infty$, where we might anticipate some problems.

Now, in an arbitrary set of coordinates (ξ, η), a differentiable scalar field $\Phi(x, y) \equiv \Phi(\xi, \eta)$ has derivatives $\frac{\partial \Phi}{\partial \xi}$ and $\frac{\partial \Phi}{\partial \eta}$. **Define** the one-form $\tilde{d\Phi}$ to have (ξ, η) components

$$\left(\frac{\partial \Phi}{\partial \xi}, \frac{\partial \Phi}{\partial \eta} \right)$$

then since Φ is (mostly) arbitrary, we can get a **huge** set of one-forms by simply choosing the Φ that we want. The components of $\tilde{d\Phi}$ transform as

$$\frac{\partial \Phi}{\partial \xi} = \frac{\partial \Phi}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial \Phi}{\partial y}\frac{\partial y}{\partial \xi}$$
$$\frac{\partial \Phi}{\partial \eta} = \frac{\partial \Phi}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial \Phi}{\partial y}\frac{\partial y}{\partial \eta}$$

as we switch from (x, y) to (ξ, η). In shorthand we can write this transformation law for the coordinate change as

$$d\Phi_{\beta'} = \Lambda^{\alpha}{}_{\beta'}\, d\Phi_{\alpha}$$

where α labels one of (x, y), β' labels one of (ξ, η), and the extension to more dimensions is obvious. The transformation matrix is

$$\Lambda^{\alpha}{}_{\beta'} = \frac{\partial x^{\alpha}}{\partial \xi^{\beta'}}$$

and Λ expresses a transformation into a potentially non-Cartesian coordinate system (note the analogy with SR). Λ has an inverse if its Jacobian is not zero. The inverse transformation then uses the inverse of Λ

$$d\Phi_{\alpha} = \Lambda^{\beta'}{}_{\alpha}\, d\Phi_{\beta'}$$

and

$$\Lambda^{\beta'}{}_{\alpha} = \frac{\partial \xi^{\beta'}}{\partial x^{\alpha}}$$

We can now define vectors as linear functions of one-forms into scalars. Their transformation laws are then the opposite of the transformation laws of the one-forms (by the same logic as we used to show this for the LT),

$$a^{\beta'} = \Lambda^{\beta'}{}_{\alpha}\, a^{\alpha}$$
$$a^{\alpha} = \Lambda^{\alpha}{}_{\beta'}\, a^{\beta'}$$

If these things are written out as matrices, then the vectors behave as column matrices with the one-forms behaving as row matrices, and it becomes clear that the forwards and backwards transforms are **scaled transposes** of one another.

Note in this case, unlike the LT, the transformation matrices need not be symmetric. In fact for the transformation $(x, y) \leftrightarrow (r, \theta)$,

$$\Lambda^r{}_x = \cos\theta \qquad \Lambda^r{}_y = \sin\theta$$
$$\Lambda^\theta{}_x = -\frac{\sin\theta}{r} \qquad \Lambda^\theta{}_y = \frac{\cos\theta}{r}$$

and the inverse matrix components are

$$\Lambda^x{}_r = \cos\theta \qquad \Lambda^x{}_\theta = -r\,\sin\theta$$
$$\Lambda^y{}_r = \sin\theta \qquad \Lambda^y{}_\theta = r\,\cos\theta$$

from which it's easy to prove

$$\Lambda^{\alpha'}{}_\beta\,\Lambda^\beta{}_{\gamma'} = \delta^{\alpha'}{}_{\gamma'}$$
$$\Lambda^\alpha{}_{\beta'}\,\Lambda^{\beta'}{}_\gamma = \delta^\alpha{}_\gamma$$

so that the inverse relationships work as required.

The basis vectors in (ξ, η) can then be calculated using the usual relationship

$$\vec{e}_{\alpha'} = \Lambda^\beta{}_{\alpha'}\,\vec{e}_\beta$$

and the (trivial) Cartesian basis vectors in (x, y), $\vec{\mathrm{e}}_\beta$. Therefore,

$$\vec{e}_\xi = \Lambda^x{}_\xi\,\vec{e}_x + \Lambda^y{}_\xi\,\vec{e}_y$$

and for example, for (r, θ) coordinates

$$\vec{e}_r = \Lambda^x{}_r\,\vec{e}_x + \Lambda^y{}_r\,\vec{e}_y = \cos\theta\,\vec{e}_x + \sin\theta\,\vec{e}_y$$
$$\vec{e}_\theta = \Lambda^x{}_\theta\,\vec{e}_x + \Lambda^y{}_\theta\,\vec{e}_y = -r\sin\theta\,\vec{e}_x + r\cos\theta\,\vec{e}_y$$

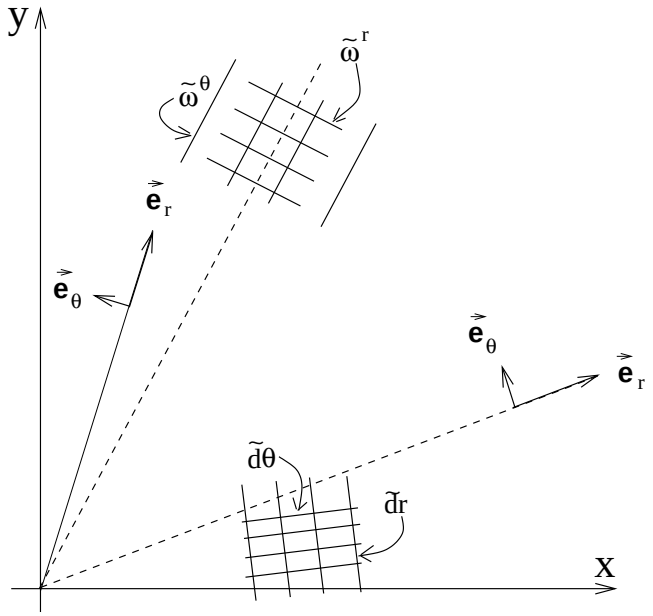
In the same way, the basis one-forms can be obtained from

$$\tilde{\omega}^{\alpha'} = \Lambda^{\alpha'}{}_\beta\,\tilde{\omega}^\beta$$

and are

$$\tilde{dr} \equiv \tilde{\omega}^r = \Lambda^r{}_x\,\tilde{\omega}^x + \Lambda^r{}_y\,\tilde{\omega}^y = \cos\theta\,\tilde{\omega}^x + \sin\theta\,\tilde{\omega}_y$$
$$\tilde{d\theta} \equiv \tilde{\omega}_\theta = \Lambda^\theta{}_x\,\tilde{\omega}^x + \Lambda^\theta{}_y\,\tilde{\omega}^y = -\frac{\sin\theta}{r}\,\tilde{\omega}^x + \frac{\cos\theta}{r}\,\tilde{\omega}^y$$

Notice that the basis one-forms and vectors change in orientation and magnitude from point to point. $\vec{e}_r$ has the same magnitude everywhere but varying orientation. $\tilde{d\theta} \equiv \tilde{\omega}^\theta$ has varying magnitude *and* orientation.



In fact, the magnitudes of the vectors and one-forms are

$$|\vec{e}_r| = 1 \qquad |\vec{e}_\theta| = r$$
$$\left|\tilde{dr}\right| = 1 \qquad \left|\tilde{d\theta}\right| = \frac{1}{r}$$

$|\vec{e}_\theta|$ increases with distance from the origin ($\propto r$), because a unit change in θ (e.g., 1 radian) moves a larger (x, y) distance at larger r.

$\left|\tilde{d\theta}\right|$ decreases with distance from the origin ($\propto 1/r$), because a unit (x, y) distance is covered by a smaller θ change than at small r.

These are *inverse* statements about the relationship between a coordinate change $\Delta\theta$ and the corresponding distance change Δs. The fact that $|\vec{e}_\theta|$ increases with distance **implies** that $\left|\tilde{d\theta}\right|$ decreases with distance, since the one-form and vector are in "dual" vector spaces.

To calculate the magnitude of one of the polar coordinate basis vectors, I use

$$\begin{aligned}
|\vec{e}_r| &= |\vec{e}_r \cdot \vec{e}_r|^{\frac{1}{2}} \\
&= \left((\cos\theta\vec{e}_x + \sin\theta\vec{e}_y) \cdot (\cos\theta\vec{e}_x + \sin\theta\vec{e}_y)\right)^{\frac{1}{2}} \\
&= \left(\cos^2\theta\,(\vec{e}_x \cdot \vec{e}_x) + 2\sin\theta\cos\theta\,(\vec{e}_x \cdot \vec{e}_y) + \sin^2\theta\,(\vec{e}_y \cdot \vec{e}_y)\right)^{\frac{1}{2}} \\
&= \left(\cos^2\theta g_{xx} + 2\sin\theta\cos\theta g_{xy} + \sin^2\theta g_{yy}\right)^{\frac{1}{2}} \\
&= 1
\end{aligned}$$

knowing that $g_{xx} = g_{yy} = 1$ and $g_{xy} = 0$ for the metric tensor in Cartesian coordinates (since the $\vec{e}_x$ and $\vec{e}_y$ basis vectors are perpendicular and unit).

To do this calculation I needed to use the metric tensor in Cartesian coordinates. Of course, I know that in *polar* coordinates

$$\begin{aligned} g_{r\,r} &= \vec{e}_r \cdot \vec{e}_r = 1 \qquad \text{(by above)} \\ g_{r\,\theta} &= \vec{e}_r \cdot \vec{e}_\theta = 0 \qquad \text{(similarly)} \\ g_{\theta\,\theta} &= \vec{e}_\theta \cdot \vec{e}_\theta = r^2 \qquad \text{(similarly)} \end{aligned}$$

It is often convenient to show the components of g in some basis by writing the **line element** (the interval in SR) as, for example

$$ds^2 = \left| dr\vec{e}_r + d\theta\vec{e}_\theta \right|^2 = dr^2 + r^2\, d\theta^2$$

which gives the separation squared of two points separated by dr and $d\theta$ at (r, θ). The inverse metric tensor g^{-1} has components

$$\begin{aligned} g^{rr} &= 1 \\ g^{r\theta} &= 0 \\ g^{\theta\theta} &= r^{-2} \end{aligned}$$

and so the gradient of a scalar field, Φ, becomes

$$\tilde{d}\Phi = \Phi_{,\alpha}\, \tilde{d}x^\alpha$$

which has components (in (r, θ) coordinates)

$$(\Phi_{,r}, \Phi_{,\theta})$$

with a corresponding vector

$$\begin{aligned} \vec{d}\Phi &= \left(g^{rr}\Phi_{,r} + g^{r\theta}\Phi_{,\theta} \right) \vec{e}_r + \left(g^{\theta r}\Phi_{,r} + g^{\theta\theta}\Phi_{,\theta} \right) \vec{e}_\theta \\ &= \left(\frac{\partial \Phi}{\partial r} \right) \vec{e}_r + \frac{1}{r^2} \left(\frac{\partial \Phi}{\partial \theta} \right) \vec{e}_\theta \end{aligned}$$

which has components (in (r, θ) coordinates)

$$\left(\Phi_{,r}, \frac{1}{r^2}\Phi_{,\theta} \right)$$

That is, the one-form components of $\tilde{d}\Phi$ and the vector components of $\vec{d}\Phi$ differ **even in a flat/Euclidean space**. This is because of the choice of coordinates — Cartesian coordinates (and only Cartesian coordinates) in Euclidean space have one-form and vector components which are the same. And this is why physics students have been able to duck the issue about what is really meant by a gradient for so long.

7.2. Tensor calculus in a non-Cartesian basis

Notice that $\vec{\mathrm{e}}_\mathrm{r}$ and $\vec{\mathrm{e}}_\theta$ change in direction and/or magnitude, $\vec{\mathrm{e}}_\mathrm{x}$ and $\vec{\mathrm{e}}_\mathrm{y}$ do not. This means that differentiating a vector may cause some trouble: consider differentiating

$$\vec{e}_x = \Lambda^r{}_x\,\vec{e}_r + \Lambda^\theta{}_x\,\vec{e}_\theta = \cos\theta\,\vec{e}_r - \frac{1}{r}\sin\theta\,\vec{e}_\theta$$

$\vec{e}_x$ is a constant vector and so has zero derivative — so when the RHS of this equation is differentiated we must expect that the derivatives of the coefficients (which are non-zero) are *cancelled* by the derivatives of the polar coordinate basis vectors, and we must learn how to differentiate these basis vectors.

We know that

$$\begin{aligned}\vec{e}_r &= \cos\theta\,\vec{e}_x + \sin\theta\,\vec{e}_y\\ \vec{e}_\theta &= -r\sin\theta\,\vec{e}_x + r\cos\theta\,\vec{e}_y\end{aligned}$$

and hence

$$\begin{aligned}\frac{\partial}{\partial r}\vec{e}_r &= \frac{\partial}{\partial r}\left(\cos\theta\,\vec{e}_x + \sin\theta\,\vec{e}_y\right) = 0\\ \frac{\partial}{\partial\theta}\vec{e}_r &= \frac{\partial}{\partial\theta}\left(\cos\theta\,\vec{e}_x + \sin\theta\,\vec{e}_y\right) = -\sin\theta\,\vec{e}_x + \cos\theta\,\vec{e}_y = \frac{1}{r}\,\vec{e}_\theta\\ \frac{\partial}{\partial r}\vec{e}_\theta &= \frac{\partial}{\partial r}\left(-r\sin\theta\,\vec{e}_x + r\cos\theta\,\vec{e}_y\right) = -\sin\theta\,\vec{e}_x + \cos\theta\,\vec{e}_y = \frac{1}{r}\,\vec{e}_\theta\\ \frac{\partial}{\partial\theta}\vec{e}_\theta &= \frac{\partial}{\partial\theta}\left(-r\sin\theta\,\vec{e}_x + r\cos\theta\,\vec{e}_y\right) = -r\cos\theta\,\vec{e}_x - r\sin\theta\,\vec{e}_y = -r\,\vec{e}_r\end{aligned}$$

These can now be used to show that

$$\frac{\partial}{\partial r}\vec{e}_x = \frac{\partial}{\partial r}\left(\cos\theta\,\vec{e}_r - \frac{1}{r}\sin\theta\,\vec{e}_\theta\right) = 0$$

as required, and similarly for $\frac{\partial\vec{e}_x}{\partial\theta}$ $\frac{\partial\vec{e}_y}{\partial r}$, and $\frac{\partial\vec{e}_y}{\partial\theta}$.

In general terms, a vector

$$\vec{a} = a^\alpha\vec{e}_\alpha$$

will differentiate as

$$\begin{aligned}\frac{\partial\vec{a}}{\partial x^\beta} &= \frac{\partial a^\alpha}{\partial x^\beta}\,\vec{e}_\alpha + a^\alpha\,\frac{\partial\vec{e}_\alpha}{\partial x^\beta}\\ &= \frac{\partial a^\alpha}{\partial x^\beta}\,\vec{e}_\alpha + a^\alpha\,\Gamma^\mu{}_{\alpha\beta}\,\vec{e}_\mu\end{aligned}$$

where the coefficients

$$\Gamma^\mu{}_{\alpha\beta} = \left(\frac{\partial\vec{e}_\alpha}{\partial x^\beta}\right)^\mu$$

are called **Christoffel symbols**. This particular one is the μth component of the β-gradient of the α basis vector. These give the geometry-dependent terms that must

be added to the simple derivatives of the components of a vector to obtain the correct (frame-independent) gradients.

This means that the Christoffel symbols **specify how physical quantities like vectors change with position because the coordinates in use are not the "best" coordinates** (e.g., Cartesian coordinates on a plane). In GR their physical meaning will be **the apparent accelerations that are felt because the coordinates in use are not the coordinates of a freely-falling frame**. That is, the Christoffel symbols contain *gravitational accelerations*.

For our 2D polar coordinates we have eight Christoffel symbols. These are given in the handout, but repeated here:

$$\Gamma^r{}_{rr} = \left[\frac{\partial \vec{e}_r}{\partial r}\right]^r = 0$$
$$\Gamma^r{}_{r\theta} = \left[\frac{\partial \vec{e}_r}{\partial \theta}\right]^r = 0$$
$$\Gamma^r{}_{\theta r} = \left[\frac{\partial \vec{e}_\theta}{\partial r}\right]^r = 0$$
$$\Gamma^r{}_{\theta\theta} = \left[\frac{\partial \vec{e}_\theta}{\partial \theta}\right]^r = -r$$
$$\Gamma^\theta{}_{rr} = \left[\frac{\partial \vec{e}_r}{\partial r}\right]^\theta = 0$$
$$\Gamma^\theta{}_{\theta r} = \left[\frac{\partial \vec{e}_\theta}{\partial r}\right]^\theta = \frac{1}{r}$$
$$\Gamma^\theta{}_{r\theta} = \left[\frac{\partial \vec{e}_r}{\partial \theta}\right]^\theta = \frac{1}{r}$$
$$\Gamma^\theta{}_{\theta\theta} = \left[\frac{\partial \vec{e}_\theta}{\partial \theta}\right]^\theta = 0 \quad .$$

Once we have got the Christoffel symbols, we can calculate derivatives of vectors in any chosen coordinate system without switching to or from Cartesians. This is a great savings in time, and lets us stick with a single chosen coordinate system (note that the expressions for the Christoffel symbols above make no mention of (x, y) coordinates).

Let's rearrange our expression for the derivative of a vector:

$$\begin{aligned}\frac{\partial \vec{a}}{\partial x^\beta} &= \frac{\partial a^\alpha}{\partial x^\beta}\vec{e}_\alpha + a^\alpha\,\Gamma^\mu{}_{\alpha\beta}\vec{e}_\mu \\ &= \frac{\partial a^\alpha}{\partial x^\beta}\vec{e}_\alpha + a^\mu\,\Gamma^\alpha{}_{\mu\beta}\vec{e}_\alpha \\ &= \left(\frac{\partial a^\alpha}{\partial x^\beta} + a^\mu\,\Gamma^\alpha{}_{\mu\beta}\right)\vec{e}_\alpha\end{aligned}$$

so that the 4 (in polar coordinates, or 16 in SR) $\alpha^{\rm th}$ and $\beta^{\rm th}$ components of $\frac{\partial \vec{a}}{\partial x^\beta}$ are

$$\frac{\partial a^\alpha}{\partial x^\beta} + \Gamma^\alpha{}_{\mu\beta}\, a^\mu \quad .$$

$\frac{\partial \vec{a}}{\partial x^\beta}$ can be regarded as the βth component of a $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ tensor field, $\nabla\vec{a}$. When fed a one-form (because of $\vec{a}$) and a vector (because of ∇), this returns a scalar. $\nabla\vec{a}$ is called the **covariant derivative** of $\vec{a}$, and has components

$$(\nabla\vec{a})^\alpha{}_\beta \equiv \nabla_\beta a^\alpha \equiv a^\alpha{}_{;\beta} \equiv a^\alpha{}_{,\beta} + a^\mu \Gamma^\alpha{}_{\mu\beta}$$

where I've introduced the notation

$$a^\alpha{}_{;\beta} \equiv \left(\frac{\partial \vec{a}}{\partial x^\beta} \right)^\alpha$$

for the components of the covariant derivative. Compare the notation that we use for the partial derivative

$$a^\alpha{}_{,\beta} \equiv \frac{\partial a^\alpha}{\partial x^\beta} \quad .$$

The difference between the covariant and partial derivative is the inclusion of terms involving the Christoffel symbols, which take account of the properties of the coordinates we are using, to create a the coordinate-invariant covariant derivative.

By doing the ";" rather than the "," operation, we get to calculate the components of $\nabla\vec{a}$ in whatever coordinate system we want. $\nabla\vec{a}$ is a physical quantity, which has real existence independent of its coordinate representation. To calculate the components of $\nabla\vec{a}$ in **any** coordinate system, we can either

1. do it in Cartesian coordinates (where the Christoffel symbols vanish), then transfer into the the coordinates we want; or
2. look up the Christoffel symbols and do it directly.

The second method is **much** more convenient, and this is why these definitions are so important.

And incidentally, for a scalar field Φ,

$$\nabla\Phi \equiv \tilde{d}\Phi$$

since the scalar field Φ doesn't depend on the basis vectors, so for a scalar $\Phi_{;\alpha} \equiv \nabla\Phi_\alpha = \left(\tilde{d}\Phi\right)_\alpha = \frac{\partial\Phi}{\partial x^\alpha} = \Phi_{,\alpha}$. That is, for a scalar the covariant and partial derivatives are the **same**.

Other operations are also possible. For example we calculate a divergence as the contraction of $\nabla\vec{a}$,

$$\nabla_\alpha a^\alpha \equiv \nabla\cdot\vec{a} = \text{scalar}$$

which therefore has the same value in any coordinate system — $\nabla.\vec{a}$ is frame independent. For our plane polar coordinate example,

$$\begin{aligned}\nabla\cdot\vec{a} &= a^{\alpha}{}_{;\alpha}\\ &= a^{\alpha}{}_{,\alpha} + \Gamma^{\alpha}{}_{\mu\alpha}a^{\mu}\\ &= a^{r}{}_{,r} + \left(\Gamma^{r}{}_{rr}a^{r} + \Gamma^{r}{}_{\theta r}a^{\theta}\right) + a^{\theta}{}_{,\theta} + \left(\Gamma^{\theta}{}_{r\theta}a^{r} + \Gamma^{\theta}{}_{\theta\theta}a^{\theta}\right)\\ &= a^{r}{}_{,r} + a^{\theta}{}_{,\theta} + \frac{1}{r}a^{r}\\ &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\,a^{r}\right) + \frac{\partial}{\partial\theta}a^{\theta}\end{aligned}$$

a result which may be familiar to you! Extending this idea, if $\vec{a} = \nabla\Phi$, where Φ is a scalar field, we get the Laplacian

$$\begin{aligned}\nabla\cdot\nabla\Phi &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\left(\nabla\Phi\right)^{r}\right) + \frac{\partial}{\partial\theta}\left(\left(\nabla\Phi\right)^{\theta}\right)\\ &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\left(g^{rr}\frac{\partial\Phi}{\partial r} + g^{r\theta}\frac{\partial\Phi}{\partial\theta}\right)\right) + \frac{\partial}{\partial\theta}\left(g^{\theta r}\frac{\partial\Phi}{\partial r} + g^{\theta\theta}\frac{\partial\Phi}{\partial\theta}\right)\\ &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\Phi}{\partial r}\right) + \frac{1}{r^{2}}\frac{\partial^{2}\Phi}{\partial\theta^{2}}\end{aligned}$$

This is where the Laplacian in plane polar coordinates (or any other coordinates) comes from — and gives an *easy* way to calculate any derivatives we want.

7.3.1. Covariant derivative of other tensors

For vectors we have an expression for covariant derivatives of components

$$a^{\alpha}{}_{;\beta} = a^{\alpha}{}_{,\beta} + \Gamma^{\alpha}{}_{\mu\beta}\,a^{\mu}$$

and we know that for a scalar field

$$\Phi_{;\beta} = \Phi_{,\beta}$$

What about one-forms? Use the same sort of arguments as when we were looking at the LT properties. Consider the quantity $\tilde{p}(\vec{a}) = \Phi$, where Φ is a scalar. Then

$$\begin{aligned}\left(\nabla\Phi\right)_{\beta} &= \Phi_{,\beta}\\ &= \frac{\partial p_{\alpha}}{\partial x^{\beta}}a^{\alpha} + p_{\alpha}\frac{\partial a^{\alpha}}{\partial x^{\beta}}\\ &= \frac{\partial p_{\alpha}}{\partial x^{\beta}}a^{\alpha} + p_{\alpha}\left(a^{\alpha}{}_{;\beta} - \Gamma^{\alpha}{}_{\mu\beta}a^{\mu}\right)\end{aligned}$$

and therefore the βth component of the tensor gradient of Φ is

$$\Phi_{,\beta} = \left(\frac{\partial p_\alpha}{\partial x^\beta} - \Gamma^\mu{}_{\alpha\beta} p_\mu \right) a^\alpha + p_\alpha a^\alpha{}_{;\beta}$$

But the second term on the RHS is the βth component of a tensor, and the LHS is certainly the βth component of a tensor. Therefore the first component on the RHS must also be the βth component of a tensor. Thus the quantity

$$\frac{\partial p_\alpha}{\partial x^\beta} - \Gamma^\mu{}_{\alpha\beta} p_\mu$$

must be a $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ tensor in component form. And therefore the tensor $\nabla \tilde{p}$ has components

$$\nabla_\beta p_\alpha = p_{\alpha;\beta} = p_{\alpha,\beta} - \Gamma^\mu{}_{\alpha\beta} p_\mu$$

Notice the *negative* sign for one-forms. In general, the covariant derivative of an object is equal to the partial derivative of the object plus terms containing the Christoffel symbols with positive sign for 'up' components (vectors) and negative sign for 'down' components (one-forms). For example the covariant derivative of a $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ tensor is

$$\nabla_\beta T_{\mu\nu} \equiv T_{\mu\nu;\beta} = T_{\mu\nu,\beta} - \Gamma^\alpha{}_{\mu\beta} T_{\alpha\nu} - \Gamma^\alpha{}_{\nu\beta} T_{\mu\alpha}$$

of a $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$ tensor is

$$\nabla_\beta S^{\mu\nu} \equiv S^{\mu\nu}{}_{;\beta} = S^{\mu\nu}{}_{,\beta} + \Gamma^\mu{}_{\alpha\beta} S^{\alpha\nu} + \Gamma^\nu{}_{\alpha\beta} S^{\mu\alpha}$$

and of a $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ tensor is

$$\nabla_\beta U^\mu{}_\nu \equiv U^\mu{}_{\nu;\beta} = U^\mu{}_{\nu,\beta} + \Gamma^\mu{}_{\alpha\beta} U^\alpha{}_\nu - \Gamma^\alpha{}_{\nu\beta} U^\mu{}_\alpha$$

The pattern of terms here is worth remembering — putting the "+" or "-" in the wrong place is the curse of beginning relativists. The keys are:

each index of the tensor in turn takes on a dummy value (α in the above), and this index matches an index on the Christoffel symbols (up or down as appropriate)

the final lower index of the Christoffel symbols is the index of the derivative (β in the above) in all cases

the other index on the Christoffel symbols is fixed by matching to the LHS of the equations

the Christoffel symbols have positive signs for derivatives of up indices, negative for derivatives of down indices

7.4. Calculating Christoffel Symbols

The Christoffel symbols tell us what extra terms we need in differentiation when we use non-Cartesian coordinates. The metric tensor g tells us what the distance is when we use coordinates of some type (i.e., it tells us something about the geometry). Therefore we expect that the Christoffel symbols are related to the metric tensor. The derivation of that relationship is given in the handout, and need not be memorized (the formula for the Christoffel symbols will always be given to you), but is repeated here for completeness.

For a vector $\vec{a}$, we get a one-form using the metric tensor, as $g(\vec{a},\ ) = \tilde{a}$. In Cartesian coordinates,

$$\nabla_\beta \tilde{a} = g\left(\nabla_\beta \vec{a},\ \right)$$

since the components of vectors and one-forms are the same in Cartesian coordinates, and covariant differentiation of components is what we mean by ∇_β. But

$$\nabla_\beta \tilde{a} = g\left(\nabla_\beta \vec{a},\ \right)$$

is a *valid* tensor equation derived in one coordinate system, and **must** be valid in **all** coordinate systems. Therefore

$$a_{\alpha;\beta} = g_{\alpha\mu} a^{\mu}{}_{;\beta}$$

(which will only work for "$;\beta$", not "$,\beta$", since "$;\beta$" is coordinate-independent, while "$,\beta$" isn't). Now we also know that $a_\alpha = g_{\alpha\beta}a^\beta$, and so

$$a_{\alpha;\beta} = g_{\alpha\mu;\beta}a^\mu + g_{\alpha\beta}a^{\mu}{}_{;\beta}$$

Comparing these two relations we see that

$$g_{\alpha\mu;\beta} = 0$$

in all coordinate systems. This is obvious anyway, since we know that $\nabla g = 0$ in Cartesian coordinates, and this equation is a frame-independent equation, and therefore is valid in all coordinate systems. But pressing on, we can write $g_{\alpha\mu;\beta}$ in full using the Christoffel symbols

$$g_{\alpha\mu;\beta} = g_{\alpha\mu,\beta} - \Gamma^{\nu}{}_{\alpha\beta}g_{\nu\mu} - \Gamma^{\nu}{}_{\mu\beta}g_{\alpha\nu} = 0$$

(which you can easily test using our plane polar coordinate example). We can use this expression to get an equation for the Christoffel symbols in terms of derivatives of the metric tensor.

Rewrite this as an equation for $g_{\alpha\mu,\beta}$, and then write two further equations with interchanged indices (β and ν), and (α and β). This gives the set of three equations

$$\begin{aligned}
g_{\alpha\mu,\beta} &= \Gamma^{\nu}{}_{\alpha\beta}\, g_{\nu\mu} + \Gamma^{\nu}{}_{\mu\beta}\, g_{\alpha\nu} \\
g_{\alpha\beta,\mu} &= \Gamma^{\nu}{}_{\alpha\mu}\, g_{\nu\beta} + \Gamma^{\nu}{}_{\beta\mu}\, g_{\alpha\nu} \\
g_{\beta\mu,\alpha} &= \Gamma^{\nu}{}_{\beta\alpha}\, g_{\nu\mu} + \Gamma^{\nu}{}_{\mu\alpha}\, g_{\beta\nu}
\end{aligned}$$

Add the first two and subtract the third, then we get

$$g_{\alpha\beta,\mu} + g_{\alpha\mu,\beta} - g_{\beta\mu,\alpha} = g_{\nu\beta}\left(\Gamma^{\nu}{}_{\alpha\mu} - \Gamma^{\nu}{}_{\mu\alpha}\right) + g_{\alpha\nu}\left(\Gamma^{\nu}{}_{\beta\mu} + \Gamma^{\nu}{}_{\mu\beta}\right) + g_{\nu\mu}\left(\Gamma^{\nu}{}_{\alpha\beta} - \Gamma^{\nu}{}_{\beta\alpha}\right)$$

But

$$\Gamma^{\nu}{}_{\alpha\beta} = \Gamma^{\nu}{}_{\beta\alpha}$$

(to be proven later), therefore the RHS becomes

$$2\, g_{\alpha\nu}\Gamma^{\nu}{}_{\beta\mu}$$

and so

$$g^{\alpha\gamma}g_{\alpha\nu}\Gamma^{\nu}{}_{\beta\mu} = \frac{1}{2}\left(g_{\alpha\beta,\mu} + g_{\alpha\mu,\beta} - g_{\beta\mu,\alpha}\right) g^{\alpha\gamma}$$

and therefore we get the final result that

$$\Gamma^{\gamma}{}_{\beta\mu} = \frac{1}{2}\, g^{\gamma\alpha}\left(g_{\alpha\beta,\mu} + g_{\alpha\mu,\beta} - g_{\beta\mu,\alpha}\right)$$

Note: this derivation works **only** if we're using a coordinate basis, so the derivatives commute. In a more general basis system the derivatives do not commute, and everything becomes much nastier — we need to include the so-called commutation coefficients, $c^{\alpha}{}_{\beta\gamma}$. I will always use a coordinate basis to avoid this problem!

To show that the Christoffel indices are symmetric in their lower two indices, let us start from a scalar field, Φ. $\nabla\Phi$, is a one-form, therefore $\nabla\nabla\Phi$ is a $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ tensor, with components

$$\nabla_{\alpha}\nabla_{\beta}\Phi = \Phi_{,\beta;\alpha}$$

(note the use of "," for a scalar derivative, and ";" for the subsequent vector derivative). Now, in Cartesian coordinates the partial derivatives commute, with

$$\frac{\partial}{\partial x^{\alpha}}\frac{\partial}{\partial x^{\beta}} = \frac{\partial}{\partial x^{\beta}}\frac{\partial}{\partial x^{\alpha}}$$

and therefore in Cartesian coordinates $\nabla\nabla\Phi$ is symmetrical. But "symmetry" under axis switches is a coordinate independent statement, and so $\nabla\nabla\Phi$ is symmetrical in all coordinates. Therefore

$$\Phi_{,\beta;\alpha} = \Phi_{,\alpha;\beta}$$

and so

$$\Phi_{,\beta,\alpha} - \Gamma^{\mu}{}_{\beta\alpha}\,\Phi_{,\mu} = \Phi_{,\alpha,\beta} - \Gamma^{\mu}{}_{\alpha\beta}\,\Phi_{,\mu}$$

But partial differentiation **always** has $\Phi_{,\alpha,\beta} = \Phi_{,\beta,\alpha}$, and Φ is arbitrary so that $\Phi_{,\mu}$ is not always zero, so that

$$\Gamma^{\mu}{}_{\beta\alpha} = \Gamma^{\mu}{}_{\alpha\beta}$$

as asserted earlier.

To repeat, the essential result is that we calculate the Christoffel symbols using

$$\Gamma^{\mu}{}_{\alpha\beta} = \frac{1}{2}\, g^{\mu\sigma}\left(g_{\sigma\alpha,\beta} + g_{\sigma\beta,\alpha} - g_{\alpha\beta,\sigma}\right)$$

in a coordinate basis.

8. Curvature

8.1. Pseudo-Riemannian manifolds

So, we have come to the point where we know how to do covariant differentiation — that is, how to deal with calculating the rates of changes of physical quantities in general coordinates.

What are these general coordinates? We are trying to describe the rates of change of some physical quantity with a continuous parameter, like time, or distance in the direction that we are moving. What is important here is that the parameter is *continuous* — that is, the spacetime in which we are describing our physics can be parameterized by continuous parameters (for example, the t, x, y, z coordinates).

A set that can be parameterized by a continuous parameter is called a *manifold*, and a manifold that also has a metric (so that we can talk about intervals between points with different parameters) and is differentiable (so that we can define one-forms and vectors) is a *Riemannian manifold*. Technically, we describe the spacetime of relativity as a *pseudo-Riemannian manifold*, because the interval between events is not positive definite. And since the space of events $\vec{x}$ requires $N = 4$ parameters to describe it, our physics is based on a 4-dimensional pseudo-Riemannian manifold.

As an illustration, the surface of the circle $x^2 + y^2 = a^2$, with $z = 0$, in Minkowski space has three dimensions — which we can take to be the coordinates (t, x, y), or the easier set for this geometry, (t, r, θ).

Notice that we require a metric, with components $g_{\alpha\beta}$ in some coordinate system. The $g_{\alpha\beta}$ may be complicated functions of location in the manifold. The metric g

- is always symmetric
- can always be tranformed to be diagonal, with components ± 1 or 0 on the diagonal.

That is, we can choose to regard g as transformable to η, the metric of special relativity, which can be written

$$\eta = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

with signature +2 (sum of terms on diagonal). The transformation to take g to η is local to the particular point at which we do the transformation. And what this means is that **sufficiently close to any point, it is possible to choose coordinates that make the metric** η, i.e., to make the spacetime appear locally **flat**.

Saying that another way, any spacetime, with any set of metric components, has a *tangent* spacetime which is flat. Such a tangent spacetime is called a *local Lorentz frame*, and is *inertial*.

Or, paraphrasing again, near any one point we can take choose coordinates to that the metric is flat, with all the derivatives of $g \rightarrow 0$ and so all the Christoffel symbols, $\Gamma^{\mu}{}_{\alpha\beta} = 0$.

What we cannot do, however, is force the second derivatives of g to zero: these measure the **amount of curvature in the spacetime**. And in GR, it is this curvature which is produced by the stress-energy tensor, T, and which corresponds to gravitation.

8.2. Curvature and parallel transport

I have been talking about the "flat spacetime" represented by the metric η, without really saying what I mean by "flat". Now is the time to rectify this omission. And to discuss what we mean by *curvature*.

There are two types of curvature:

(1) **extrinsic** curvature — the curvature of a manifold in relation to a higher-dimensional manifold in which it is embedded; and

(2) **intrinsic** curvature — the curvature of a manifold that can be defined entirely within the manifold itself.

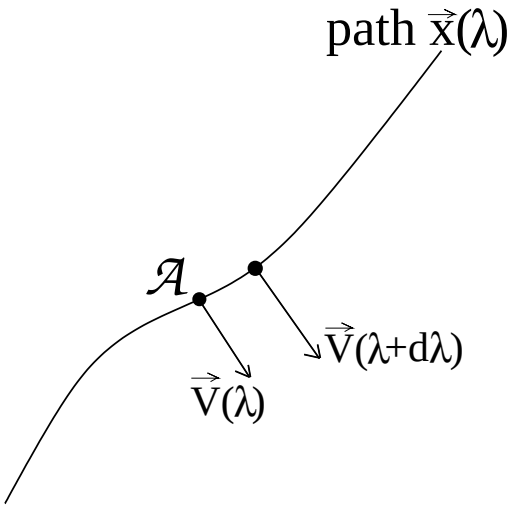
For example, consider a cylinder. This is a 2-D surface in our normal 3-D world, which has *extrinsic curvature* when viewed in a 3-D flat space, but *no intrinsic curvature*, since it can be made by continuously deforming a flat plane without tearing or crumpling. How would an ant that lives on the cylinder know that it has no intrinsic curvature? Because distances measured *in the surface of the cylinder* are the same as distances would be when measured on a normal 2-dimensional plane. We say that the cylinder *has flat geometry but non-trivial topology.*

In GR we need to discuss *intrinsic curvature* — this is what is related to the phenomenon that we usually call gravitation. Any extrinsic properties of spacetime would rely on a higher-dimension embedding space, but we don't need that. And we should be aware that there is **nothing** in GR which relates to the topology of the spacetime — only to the curvature.

$$\begin{aligned} \text{gravity} \rightarrow \text{local geometry} &\rightarrow \text{intrinsic curvature} \\ ? &\rightarrow \text{extrinsic curvature} \end{aligned}$$

8.2.1. Parallel transport

Draw a curve (which is defined as a path with some parameter, λ). At every point, draw a vector parallel to the vector at the previous point. If the vector field $\vec{V}$ has values at λ and $\lambda + d\lambda$ which are parallel and of equal length, we say that $\vec{V}$ is parallel-transported along the curve.



That is, in the neighbourhood of the point $\mathcal{A}$,

$$\frac{d\vec{V}}{d\lambda} = 0 \quad .$$

But

$$\frac{d\vec{V}}{d\lambda} = \frac{\partial V^\alpha}{\partial x^\beta}\frac{dx^\beta}{d\lambda} = u^\beta V^\alpha{}_{,\beta}$$

using the chain rule, and with the definition that

$$\vec{u} = \frac{d\vec{x}}{d\lambda}$$

is the tangent vector at $\mathcal{A}$. Now, in a locally-flat region near $\mathcal{A}$, where a local Lorentz frame can be defined,

$$u^\beta V^\alpha{}_{,\beta} = u^\beta V^\alpha{}_{;\beta}$$

since all the Christoffel symbols are zero in this inertial coordinate system, and therefore for parallel transport at $\mathcal{A}$,

$$u^\beta V^\alpha{}_{;\beta} = 0 \quad .$$

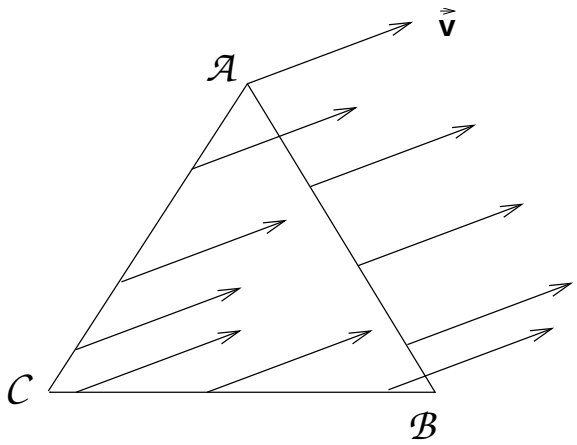
But this is a frame-invariant expression, and therefore will hold in any coordinate basis. Therefore for parallel transport, in any basis,

$$u^\beta V^\alpha{}_{;\beta} = 0$$
$$\nabla_{\vec{u}}\vec{V} = 0$$

where the second expression is written in coordinate-free language (and therefore looks prettier) but means the same thing as the first: the directional derivative of the vector $\vec{V}$ along the path with tangent vector $\vec{u}$ is zero.

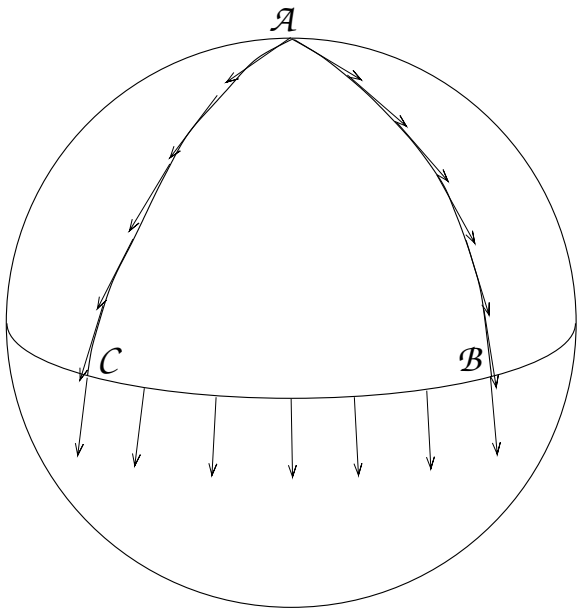
8.2.2. Parallel transport around a closed loop

Let's do this parallel-transport construction around two closed curves, and see what happens. In both cases I will do the transport around a triangle in a 2-D space. First, a triangle in a flat space.



As the vector $\vec{v}$ is taken along the path $\mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{C} \rightarrow \mathcal{A}$, under parallel transport so that at each point it is moved to stay parallel to itself, it returns to $\mathcal{A}$ in precise coincidence with its starting value.

Now do the same thing on the 2-D surface of a sphere.



As the vector $\vec{v}$ is taken along the path $\mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{C} \rightarrow \mathcal{A}$, under parallel transport so that at each point it is moved to stay parallel to itself, it returns to $\mathcal{A}$ at some angle

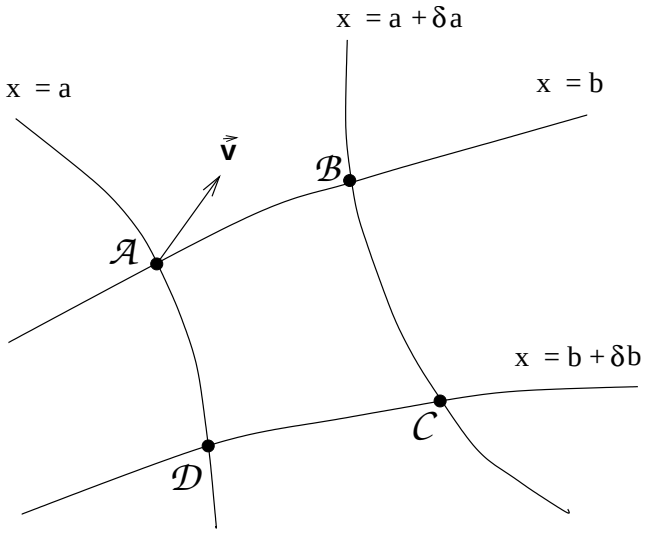
to its original direction. The change of angle arises because the space that we moved in is curved. And since this is an intrinsic construction (we didn't move outside the 2-D surface of the sphere), we have found a method of measuring the intrinsic curvature of the space.

Other methods can be found — for example, we could measure the change in area of a flat object the size and orientation of that object. For example, for a triangle we would measure the area as half product of the length of the base times the height only if the spacetime was flat. The triangle $\mathcal{ABC}$ on the surface of the sphere clearly doesn't have this for its area. But these other methods are related to the result we get from the parallel transport argument, so we'll stick with that here.

8.3. Measuring curvature: the Riemann tensor

We use this idea of parallel transport to obtain a measure of the curvature of a space from the difference in the value of a vector at the beginning and end of its parallel transport around a closed loop.

Consider a loop in the (x^1, x^2) surface (all other components will be similar).



Then we expect from the discussion of parallel transporting a vector that the mismatch between the value of the vector at the beginning and the end of its transit around the loop from $\mathcal{A}$ to $\mathcal{B}$ to $\mathcal{C}$ to $\mathcal{D}$ to $\mathcal{A}$ will depend linearly on

- the value of the vector itself,
- the size of the loop in the x^1-direction, and
- the size of the loop in the x^2-direction.

That is, we expect

$$\Delta v^\alpha \propto v^\beta \, dx^1 \, dx^2$$

but there is no reason to single out the x^1 and x^2 directions: we expect all directions to be taken equally into account. So the general result must be

$$\Delta v^\alpha \propto v^\beta \, dx^\mu \, dx^\nu \quad .$$

That is, a vector quantity is related linearly to the product of three other vector quantities. The most general way of making this relationship is if the constant of proportionality is a $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ tensor: we call it the *Riemann curvature tensor*, and write

$$\delta v^\alpha = -R^\alpha{}_{\beta\mu\nu} \, v^\beta \, \delta a^\mu \, \delta b^\nu$$

in component form, where the third and fourth indices on R refer to the area components, and the sign of R is a matter of convention: different books use different conventions. With the choice given here, it is possible to prove (after some longish manipulations, see the handout) that the components of R are given by

$$\boxed{R^\alpha{}_{\beta\mu\nu} = \Gamma^\alpha{}_{\beta\nu,\mu} - \Gamma^\alpha{}_{\beta\mu,\nu} + \Gamma^\alpha{}_{\sigma\mu}\Gamma^\sigma{}_{\beta\nu} - \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\beta\mu} \quad .}$$

You will never need to calculate R: but this is what it is. Notice that it has *exactly* the type of shape that you might expect. It depends on the Christoffel symbols, since these are the quantities that measure the tendancy of vectors to change direction and magnitude because of the choice of coordinates.

All the components of the Riemann curvature tensor are zero for a flat manifold. In a curved manifold the components are functions of position. Possibly complicated functions of position!

And since we can always choose coordinates so that the metric tensor looks like η (that is, if we choose to work in a local Lorentz frame) where the Christoffel symbols vanish **but the derivatives of the Christoffel symbols do not vanish**, it is clear that the Riemann tensor depends on the second derivatives of the components of the metric tensor — which is what you would expect for "curved coordinates".

8.3.1. Some properties of the Riemann tensor

In fact we can work through all the algebra in a locally inertial frame, we can to show that

$$R^\alpha{}_{\beta\mu\nu} = \frac{1}{2} g^{\alpha\sigma} \left(g_{\sigma\nu,\beta\nu} - g_{\sigma\mu,\beta\nu} + g_{\beta\mu,\sigma\nu} - g_{\beta\nu,\sigma\mu} \right)$$

which makes it explicit that the Riemann tensor depends on second derivatives of the metric tensor (and is non-linear: second order in the metric coefficients). This also shows up the symmetries of the Riemann tensor

$$\begin{aligned} R_{\alpha\beta\mu\nu} &= -R_{\beta\alpha\mu\nu} \\ R_{\alpha\beta\mu\nu} &= -R_{\alpha\beta\nu\mu} \\ R_{\alpha\beta\mu\nu} + R_{\alpha\nu\beta\mu} + R_{\alpha\mu\nu\beta} &= 0 \end{aligned}$$

These symmetry relations are frame-independent, valid in a local inertial frame, and therefore valid everywhere (unlike the expression for R in terms of partial derivatives of the metric tensor given earlier).

We can also use these results to prove the **Bianchi identities**

$$R_{\alpha\beta\mu\nu;\lambda} + R_{\alpha\beta\lambda\mu;\nu} + R_{\alpha\beta\nu\lambda;\mu} = 0$$

which it will turn out ensure that mass-energy is conserved in GR, $\nabla \cdot \vec{T} = 0$.

The Riemann tensor is a $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ tensor. We can therefore contract it, to obtain a $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ tensor, the *Ricci tensor*

$$R_{\mu\nu} = R^{\alpha}{}_{\mu\alpha\nu} \quad .$$

We can make one more contraction, after raising an index of the Ricci tensor, to get the *Ricci scalar*

$$R = R^{\mu}{}_{\mu} = g^{\mu\nu} R_{\mu\nu} \quad .$$

It turns out that these contractions are the *only* distinct contractions of the Riemann curvature tensor: the other contractions are zero, or related to the Ricci tensor or scalar by a sign change.

We can make a special combination of the Ricci tensor and Ricci scalar to create a new tensor

$$G^{\alpha\beta} = R^{\alpha\beta} - \frac{1}{2} g^{\alpha\beta} R$$

which has special significance. This is called the *Einstein tensor*, and appears in the field equations of General Relativity — it is the quantity "created" by the stress-energy tensor. Using earlier knowledge about the Riemann tensor, it's easy to prove that the Einstein tensor is symmetric

$$G^{\alpha\beta} = G^{\beta\alpha}$$

and only slightly harder to prove (with the assistance of the Bianchi identities) that it has zero divergence

$$G^{\alpha\beta}{}_{;\beta} = 0$$

which can also be written in coordinate-independent language as

$$\nabla \cdot G = 0 \quad .$$

This will turn out to be a requirement so that energy and momentum are conserved in GR.

9. Geodesics

9.1. Parallel-transport and geodesics

We can use the idea of parallel transport to construct *geodesics*, defined as curves that parallel-transport their own tangent vectors. That is, for a geodesic

$$\begin{aligned} & \nabla_{\vec{u}}\vec{u} = 0 \\ \text{i.e.} \quad & u^\beta u^\alpha{}_{;\beta} = 0 \\ \text{i.e.} \quad & u^\beta u^\alpha{}_{,\beta} + \Gamma^\alpha{}_{\beta\gamma} u^\beta u^\gamma = 0 \\ \text{or} \quad & \frac{d}{d\lambda}\left(\frac{dx^\alpha}{d\lambda}\right) + \Gamma^\alpha{}_{\beta\gamma}\frac{dx^\beta}{d\lambda}\frac{dx^\gamma}{d\lambda} = 0 \end{aligned}$$

where in the last of these expressions (often called the *geodesic equation*, though the first is also the geodesic equation), λ is the parameter of the curve.

We have some freedom to choose λ — if we choose it to be the proper time of a particle with the curve as its world line, then $\vec{u}$ is the velocity of the particle. However, λ is a more general quantity and can be used also for light rays with no proper time. Any linear transformation of λ, such as $\phi = a\lambda + b$ with a, b constants, has $\vec{x}(\phi)$ a valid solution of the geodesic equation (try the transformation $\lambda \to \phi$) — we refer to λ (or ϕ) as an **affine parameter**.

In a locally-flat region, where the Christoffel symbols vanish, clearly the geodesic equation reduces to

$$\frac{d^2x^\alpha}{d\lambda^2} = 0$$

which solves to the straight-line solution

$$x^\alpha = A^\alpha\lambda + B^\alpha \quad .$$

In fact we can say, in a very real sense, that all geodesics are **straight**. This definition about "parallel transport of the tangent vector" is the only sensible definition of a straight line — it means that the curve at each point keeps moving in the direction of its local tangent vector. No other frame-independent definition of "straight" makes sense.

A geodesic is also a line of *extremal length* between two points $\mathcal{A}$ and $\mathcal{B}$: small changes in the path cause no change in the distance $s(\mathcal{A},\mathcal{B})$ between the points, where

$$s(\mathcal{A},\mathcal{B}) = \int_{\lambda_\mathcal{A}}^{\lambda_\mathcal{B}} |\vec{u}.\vec{u}|^{1/2}\, d\lambda$$

and $\vec{u}$ is the tangent vector. That is, if you use the calculus of variations on this expression to solve for the extremal path $\vec{x}(\lambda)$, you will recover the geodesic equation.

Note that the geodesic equation is a set of four coupled, non-linear, second-order differential equations for the $\vec{x}(\lambda)$. To find a solution we must specify eight quantities, such as an initial position of the geodesic, $\vec{x}(\lambda_0)$, and an initial direction, $\vec{u}(\lambda_0) = \frac{d\vec{x}}{d\lambda}$.

All particles move on geodesics unless affected by non-gravitational forces. Therefore the geodesic equation is the "equation of motion" for any particle moving under the influence of gravitation alone.

9.2. Geodesics in the weak-field metric

It will turn out (when we look at the Einstein field equations) that in weak gravitational fields we can write the metric as

$$ds^2 = -(1+2\phi)dt^2 + (1-2\phi)(dx^2+dy^2+dz^2)$$

where ϕ is the usual Newtonian gravitational potential, of the form

$$\phi = -\frac{GM}{\left(x^2+y^2+z^2\right)^{1/2} c^2}$$

(where I throw the term c^2 in gratuitously, to make it clear that ϕ here is a dimensionless quantity).

What is the path of a particle that moves in such a metric? We can describe this path as a parameterized curve, $\vec{x}(\lambda)$, as described by the geodesic equation. Using the form

$$\nabla_{\vec{u}}\vec{u} = 0$$

where $\vec{u}$ is the tangent vector to the particle world line (that is, it's the particle velocity vector), we can now solve for the path. It's perhaps easier to work in terms of the particle momentum, $\vec{p} = m\vec{u}$, where m is the particle mass, when the geodesic equation takes the form

$$\nabla_{\vec{p}}\vec{p} = 0$$

which I can write in component form as

$$p^\alpha \left(p^\beta{}_{,\alpha} + \Gamma^\beta{}_{\mu\alpha}\, p^\mu\right) = 0 \quad .$$

Suppose that the particle motion is non-relativistic, with velocity $v \ll c$. Then the time component of the momentum 4-vector is much larger than any of the space components

$$p^0 \gg p^i \quad \forall\ i \quad .$$

We can also use

$$\begin{aligned} p^\alpha p^\beta{}_{,\alpha} &= m\, u^\alpha p^\beta{}_{,\alpha} \\ &= m\,\frac{d}{d\tau}p^\beta \end{aligned}$$

where τ is the particle proper time — the time as measured by a clock moving with the particle (which plays the role of the affine parameter of the particle path here).

Using these in the geodesic equation, the $\beta = 0$ component of that equation becomes

$$m\left(\frac{dp^0}{d\tau}\right) + \Gamma^0{}_{00}\, p^0\, p^0 = 0$$

to first order in v (since I've dropped the terms in the Christoffel symbols and the space-space and space-time components of the momentum 4-vector).

From the formula for the Christoffel symbols,

$$\begin{aligned}\Gamma^0{}_{00} &= \frac{1}{2} g^{0\alpha}\left(g_{\alpha 0,0} + g_{\alpha 0,0} - g_{00,\alpha}\right) \\ &= \phi_{,0} + O(\phi^2) \quad .\end{aligned}$$

Therefore, in the weak-field limit, where $\phi \ll 1$, we can take only the first term here, to obtain

$$m\frac{dp^0}{d\tau} = -m^2\,\phi_{,0}$$

or

$$\frac{dp^0}{d\tau} = -m\,\frac{\partial\phi}{\partial t} \quad .$$

The interpretation of this is that the particle energy (p^0) changes only if the gravitational potential is a function of time. If the gravitational potential is static, then the particle energy is a constant of the motion. This should be familiar to you from the normal theory of orbits — but we've gained a further insight about how what happens if the potential changes with time, and the quantity p^0 is a bit more subtle than the usual kinetic energy of a particle, since it includes something to do with the gravitational energy.

What about the space-like terms in the geodesic equation? Take the terms with $\beta = i$. Then to lowest order in velocity the geodesic equation becomes

$$m\frac{dp^i}{d\tau} = -\Gamma^i{}_{00} p^0 p^0$$

where I've again ignored the smaller space-time and space-space terms on the right-hand side. The Christoffel symbols are given by

$$\begin{aligned}\Gamma^i{}_{00} &= \frac{1}{2} g^{i\alpha}\left(g_{\alpha 0,0} + g_{\alpha 0,0} - g_{00,\alpha}\right) \\ &= \frac{1}{2}\left(1-2\phi\right)^{-1}\delta^{ij}\left(-g_{00,j}\right) \\ &= \phi_{,j}\,\delta^{ij} + O(\phi^2)\end{aligned}$$

so that in the weak-field limit the geodesic equation becomes

$$\frac{dp^i}{d\tau} = -m\phi^{,i} \quad .$$

This says that the rate of change of momentum arises from the gradient of the gravitational potential — which is the usual Newtonian result if we identify τ with t (which is OK for non-relativistic particle motions, as we assumed at the start).

So, by following through the logic of the geodesic equation and the assumed weak-field metric (which we should still demonstrate to be valid), we have recovered equations which are consistent with the Newtonian equations in the weak-field and low-speed limit. That is, we've demonstrated that GR is at least as good as Newtonian theory in explaining the motions of the planets and other gravitational motions of particles.

9.3. Radial null geodesics in the FRW metric, and redshift

To ram the activities of working with geodesics home a little further, let's consider the case of the metric of the Universe, the FRW metric

$$ds^2 = -dt^2 + a(t)^2 \left(\frac{dr^2}{1-kr^2} + r^2 \left(d\theta^2 + \sin^2\theta d\phi^2 \right) \right)$$

where $a(t)$ is the scale factor of the Universe, which is a function of time that we have to solve for from the field equations of General Relativity, and k is a constant.

It is a feature of the Universe that it is isotropic and homogeneous (at least as expressed by this metric), so we can take ourselves as being at the origin of coordinates. Any incoming light rays that we detect therefore have $d\theta = d\phi = 0$, by symmetry. That means that the equations of motion of the light rays can be obtained directly from the metric by putting $ds = 0$ (light rays are null geodesics),

$$-dt^2 + a(t)^2 \left(\frac{dr^2}{1-kr^2} \right) = 0 \quad .$$

For incoming light rays, which arrive at $r = 0$ (the location of the observer) at $t = t_0$, having been at r at t, we can write the equation of the path $r(t)$ as

$$\int_t^{t_0} \frac{dt}{a(t)} = \int_0^r \frac{dr}{(1-kr^2)^{1/2}}$$

where I've chosen the correct sign of the square root for incoming light rays.

Now suppose that the light originated at a galaxy with radial coordinate r_1 at time t_1. Then

$$\int_{t_1}^{t_0} \frac{dt}{a(t)} = \int_0^{r_1} \frac{dr}{(1-kr^2)^{1/2}} = f(r_1)$$

where $f(r_1)$ is some function which depends on the value of k (it may be a sin or a sinh function, or even simply the function r).

Suppose that the distant galaxy emits light with wavelength λ_1 and frequency $\nu_1 = c/\lambda_1$. Then adjacent peaks of the light wave are emitted at times t_1 and $t_1 + \frac{1}{\nu_1}$. Then we can calculate the arrival times of these peaks: they will be t_0 and $t_0 + \frac{1}{\nu_0}$ where ν_0 is the frequency at which the light is received. And using the above equation, we know that

$$\int_{t_1}^{t_0} \frac{dt}{a(t)} = f(r_1) = \int_{t_1+\frac{1}{\nu_1}}^{t_0+\frac{1}{\nu_0}} \frac{dt}{a(t)}$$

or, rearranging the limits on the integrals

$$\int_{t_1}^{t_1+\frac{1}{\nu_1}} \frac{dt}{a(t)} = \int_{t_0}^{t_0+\frac{1}{\nu_0}} \frac{dt}{a(t)}$$

So if $\nu_0 \gg \dot{a(t)}$,

$$\frac{1/\nu_1}{a(t_1)} = \frac{1/\nu_0}{a(t_0)}$$

or

$$\frac{\lambda_0}{\lambda_1} = \frac{a(t_0)}{a(t_1)}$$

That is, the observable wavelength that we see differs from the emitted wavelength by a factor which is the ratio of the scale factors of the Universe now and when the light was emitted.

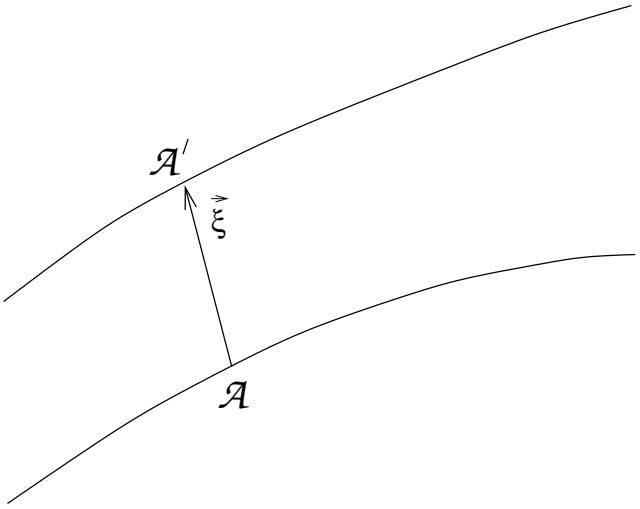
This is the origin of the redshift. Since the Universe is expanding, $a(t_0) > a(t_1)$, and so $\lambda_0 > \lambda_1$. We define the ratio of the emitted to the observed wavelength to be $1 + z$, where z is the *redshift* of light from the observed distant galaxy, so we have

$$1 + z \equiv \frac{\lambda_0}{\lambda_1} = \frac{a(t_0)}{a(t_1)}$$

or, in other words, the factor $1+z$ is a *direct measure of the relative sizes of the Universe now and when the light was emitted.*

9.4. Geodesic deviation and the Riemann tensor

In a flat space parallel lines stay parallel. In a curved space parallel lines do *not* stay parallel. The Riemann tensor describes the extent to which they want to converge or diverge.



Consider two geodesics, starting from two points that are close to each other, with a separation vector $\vec{\xi}$. Then if $\vec{v}$ is the tangent vector of the geodesics, it is possible to prove (with some effort) the equation of geodesic deviation

$$\nabla_{\vec{v}} \nabla_{\vec{v}} \vec{\xi} = R\left(\vec{v}, \vec{v}, \vec{\xi}\right) \quad .$$

That is, the acceleration of $\vec{\xi}$, the separation of the geodesics, with location along the geodesics, is proportional to the Riemann tensor. In a flat spacetime, where the Riemann tensor vanishes, $\vec{\xi}$ is a linear function of the affine parameter (the distance along the geodesic). In a curved spacetime, the deviation from this linear function is driven by the magnitude of the Riemann tensor: that is **the trajectories of adjacent particles are made to diverge or converge faster because of the curvature**.

In physical terms, where we think of gravitational effects, the Riemann tensor measures the "tidal force" — the tendency of adjacent particles to move closer to one another because of local clumping of matter.

10. Einstein Field Equations

10.1. Postulating the field equations

In Newtonian gravity, the effect of gravity is described by the potential, $\phi(\mathbf{r})$, and this potential is calculated by the *field equation*

$$\nabla^2\phi = 4\pi G\rho$$

where $\rho(\mathbf{r})$ is the density. For a point mass we can use this equation to show that

$$\phi = -\frac{G\,M}{r}$$

which is dimensionless if I insert the usual c^2 factor underneath to make it look relativistic. What we see here is that in Newtonian gravity

$$\text{mass density} \rightarrow \text{gravity}$$

In general relativity we must come up with a more general expression. We can't use ρ as the source of gravity, even if we make the obvious relativistic extension that ρ is the total energy density (including rest mass, kinetic, thermal, ... energies), since the apparent density is different for different observers and it doesn't make sense to use the density in the MCRF of the gravitating material (it should be something characteristic of the *observer's frame* that dictates the observer's motion). However, since we know that in the non-relativistic limit the theory should reduce to Newtonian gravity

- the source of gravity should be analogous to ρ
- it should reduce to ρ in the non-relativistic limit.

Earlier we constructed the stress-energy tensor $\mathbf{T}$, with T^{00} being the apparent total energy density in some specific frame, and equal to the density in the non-relativistic limit. However, $\mathbf{T}$ is a frame-independent quantity and more general than the density. So we **postulate that T is the source of gravity**.

In so doing, we *must* use the entire tensor and not only the T^{00} component, since only the entire tensor is frame-independent.

We are also trying to construct a theory where we understand the effects of gravity via the properties of the metric. So the field equations **must** take the form

$$\text{function of } \mathbf{g} = \text{scalar constant} \times \mathbf{T}$$

which would be a frame-independent equation (actually 16 equations). And for this to be valid, the function of the metric tensor that we use must be a $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$ tensor.

But in Newtonian gravity, we know that it's the second derivative of the potential that's related to one component of **T**. Hence in general relativity we require that the function should be a function of the second, first, and zero$^{\rm th}$ derivatives of **g**.

What functions of the metric tensor do we have that fit? The only tensor that we've constructed that's anything to do with second derivatives of the metric tensor is the Riemann curvature tensor, **R**. The Riemann tensor is a $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ tensor, and is *too big* to do what we want. However, from any large tensor we can always construct a smaller tensor by *contraction*, and the only unique contraction of the Riemann tensor is called the *Ricci tensor*

$$R_{\mu\nu} = R^{\alpha}{}_{\mu\alpha\nu} \quad .$$

Of course, we can always make $R^{\mu\nu}$ from the Ricci tensor defined like this by raising the indices

$$R^{\mu\nu} = g^{\mu\alpha}\, g^{\nu\beta}\, R_{\alpha\beta}$$

so we can choose the Ricci tensor as the quantity that is related to the stress-energy tensor.

If we're also allowed zeroth derivatives of the metric, however, we should be allowed to add some of the metric tensor to the Ricci tensor too, since the Ricci tensor and metric tensor are both second-rank tensors. And the amount of the metric tensor that we add will depend on some scalar. This might be a pure number, or it might be the scalar that can be made from the Riemann curvature tensor, the *Ricci scalar*, which is a contraction of the Ricci tensor

$$R = R^{\mu}{}_{\mu} = g^{\mu\nu}\, R_{\mu\nu} \quad .$$

No other independent contractions of the Riemann curvature tensor exist (the symmetries of **R** are such that the other possible contractions are zero, or related to the Ricci tensor or scalar by a sign change).

So, putting this all together, we can guess that the form of the field equations of general relativity should be

$$R^{\mu\nu} + \theta\, R\, g^{\mu\nu} + \Lambda\, g^{\mu\nu} = kT^{\mu\nu}$$

where k, θ, and Λ are constants (scalars) and the left-hand side involves the metric tensor, the Ricci tensor, and the Ricci scalar.

Now, let's suppose that we require **local conservation of energy and momentum**. This corresponds to

$$T^{\mu\nu}{}_{;\nu} = 0$$

(note the use of the covariant derivative, which is the general-relativistic generalization of the special-relativistic partial derivative, and we are assured that this is correct because the ";" version is correct in a local Lorentz frame).

Trying this out on our field equation, for energy conservation to be valid,

$$(R^{\mu\nu} + \theta\, R\, g^{\mu\nu} + \Lambda\, g^{\mu\nu})_{;\nu} = 0 \quad .$$

But ${g^{\mu\nu}}_{;\nu} = 0$ always (this is obvious, since $\nabla.\mathbf{g} = 0$ in a local Lorentz frame, and because this is a frame-invariant expression, it must be valid in any general frame). This means that the term in Λ vanishes. What about the covariant derivative of the Ricci scalar? A bit more algebra is needed here (a few pages), but the symmetries of the Riemann tensor cause

$${R^{\mu\nu}}_{;\nu} = \frac{1}{2}\,(g^{\mu\nu} R)_{;\nu}$$

from which we can see that energy conservation is enforced if

$$\theta = -\frac{1}{2}$$

We define the new tensor

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R$$

called the *Einstein tensor*. Then the field equation can be written (in frame-invariant form)

$$\mathbf{G} + \Lambda\, \mathbf{g} = k\, \mathbf{T} \quad .$$

The discussion that has led up to this point is not rigorous — there are *many* other choices possible that reduce correctly to Newtonian gravity and special relativity (for example, we might add any further powers of the Ricci scalar, since this is small for weak fields). However, the equation above is about the simplest equation which can reproduce the non-relativistic limits. The field equation as written here is **linear** in curvature, but non-linear in the metric (recall that the Riemann tensor itself is non-linear in the metric, so the contractions are non-linear, and the product of the metric tensor with the Ricci scalar is certainly non-linear).

What are the constants appearing in the field equation? They are

Λ the *cosmological constant*, and

$k = 8\pi G$ (or $8\pi G/c^4$ in physical units), which is shown by taking the weak-field limit of the field equation and enforcing agreement with Newtonian gravity.

10.1.1. Goodbye to G

At this point we become real relativists by losing the constant G. When we discussed special relativity, it became natural to scale physical variables so that time and distance are measured in the same units. We did this by

$$c \to 1$$

Now we can do the same thing for GR. Because we can see that G arises only as the scaling between curvature and mass density, it will be convenient to scale physical variables to that mass, time, and distance are all in the same units (metres). We do this by taking

$$G \to 1$$

Thus, in special relativity, a velocity v is

$$\left(v/\mathrm{m\,s^{-1}}\right) = v \times \left(2.99792458 \times 10^8\right)$$

where the numerical constant is $c/\mathrm{m\,s^{-1}}$. And similarly in general relativity

$$(M/\mathrm{kg}) = (M/\mathrm{m}) \times \left(2.99792458 \times 10^8\right)^2 \times \left(6.67259 \times 10^{-11}\right)^{-1}$$

where the second numerical constant is $G/\mathrm{m^3\,kg^{-1}\,s^{-2}}$. This works out to be

$$(M/\mathrm{kg}) = (M/\mathrm{m}) \times \left(1.347 \times 10^{27}\ \mathrm{kg\,m^{-1}}\right)$$

so, for example, the mass of the Sun, 1.989×10^{30} kg, is only 1.477 km.

In these units, the field equations of general relativity are

$$\mathbf{G} + \Lambda\,\mathbf{g} = 8\,\pi\,\mathbf{T} \quad .$$

These are 16 equations, which are reduced to 10 equations since $\mathbf{g}$, $\mathbf{G}$, and $\mathbf{T}$ are symmetrical. And four of these ten equations are redundant, since we know that energy is conserved, so there are four auxiliary equations

$$\nabla \cdot \mathbf{T} = 0 \quad .$$

10.1.2 The cosmological constant

The cosmological constant Λ is controversial. GR is simpler without it, but there's no particular reason to chuck it out: it comes in naturally enough in the derivation above. The meaning of Λ is interesting. Since

$$T^{\mu\nu} = (\rho + P)\,u^\mu u^\nu + P g^{\mu\nu}$$

the field equations effectively set the $G^{\mu\nu}$ component of the Einstein tensor equal to

$$8\,\pi\,T^{\mu\nu} - \Lambda\,g^{\mu\nu} = 8\,\pi\left((\rho + P)\,u^\mu u^\nu + \left(P - \frac{\Lambda}{8\pi}\right) g^{\mu\nu}\right)$$

so that Λ comes in as a sort of extra fluid in the Universe, with a negative pressure, so that the total pressure is

$$P' = P - \frac{\Lambda}{8\pi}$$

and a positive density, so that the total density is

$$\rho' = \rho + \frac{\Lambda}{8\pi} \quad .$$

We say that the Λ-term is like an extra *vacuum energy* field, with equation of state

$$P_\Lambda = -\rho_\Lambda \quad .$$

Compare this with the more normal radiation fields, which have

$$P_{\rm rad} = \frac{1}{3}\rho_{\rm rad}$$

or cold matter density field, for which

$$P_{\rm dust} = 0 \quad .$$

By contrast, $P = -\rho$ is an "exotic" equation of state. If we wish, we can lose the Λ term by absorbing it into the stress-energy tensor as a component of the mass density of the Universe, with this strange equation of state.

10.2. The weak-field metric

Earlier I talked about geodesics in an assumed weak-field metric of the form

$$ds^2 = -(1+2\phi)dt^2 + (1-2\phi)(dx^2+dy^2+dz^2) \quad ,$$

and showed that these were consistent with the Newtonian equations of motion. Now we have the field equations, let's prove that this is consistent with the Poisson equation for the potential, ϕ,

$$\nabla^2\phi = 4\pi\rho$$

(with $G = 1$) as is required to close the argument (and check the 8π factor in the field equations).

There is a sophisticated way to do this, but I'm going to use brute force and ignorance to illustrate that everything works through simply. And in all the equations I will work only to order ϕ, assume that the metric is static (all time derivatives of ϕ are small), and that the matter that's causing the curvature is moving non-relativistically. Then from the line element above,

$$g_{00} = -(1+2\phi) \qquad g^{00} = -(1-2\phi)$$
$$g_{ij} = (1-2\phi)\delta_{ij} \qquad g^{ij} = (1+2\phi)\delta^{ij}$$

with the other metric components zero. The Christoffel symbols are given by the usual expression

$$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2} g^{\alpha\sigma} \left(g_{\sigma\mu,\nu} + g_{\sigma\nu,\mu} - g_{\mu\nu,\sigma} \right)$$

which with a little manipulation leads to expressions (correct to first order in the gravitational potential)

$$\begin{aligned}
\Gamma^{0}{}_{00} &= \phi_{,0} \\
\Gamma^{0}{}_{0i} = \Gamma^{0}{}_{i0} &= \phi_{,i} \\
\Gamma^{0}{}_{ij} &= \delta_{ij}\phi^{,0} \\
\Gamma^{i}{}_{00} &= \phi^{,i} \\
\Gamma^{i}{}_{0j} = \Gamma^{i}{}_{j0} &= -\delta^{i}{}_{j}\phi_{,0} \\
\Gamma^{i}{}_{jk} &= \delta_{jk}\phi^{,i} - \delta^{i}{}_{j}\phi_{,k} - \delta^{i}{}_{k}\phi_{,j}
\end{aligned}$$

which are getting a little more complicated. For the Riemann curvature tensor,

$$R^{\alpha}{}_{\beta\mu\nu} = \Gamma^{\alpha}{}_{\beta\nu,\mu} - \Gamma^{\alpha}{}_{\beta\mu,\nu} + \Gamma^{\alpha}{}_{\sigma\mu}\Gamma^{\sigma}{}_{\beta\nu} - \Gamma^{\alpha}{}_{\sigma\nu}\Gamma^{\sigma}{}_{\beta\mu}$$

the assumption that the fields are weak allows us to lose the terms in products of the Christoffel symbols, and gives us non-zero components only for

$$\begin{aligned}
R^{0}{}_{i0j} &= \delta_{ij}\phi_{,00} - \phi_{,ij} \\
R^{i}{}_{0j0} &= \phi^{,i}{}_{,j} + \delta^{i}{}_{j}\phi_{,00} \\
R^{i}{}_{0jk} &= -\delta^{i}{}_{k}\phi_{,0j} + \delta^{i}{}_{j}\phi_{0k} \\
R^{i}{}_{kj0} &= \delta^{i}{}_{j}\phi_{,0k} - \delta_{jk}\phi^{,i}{}_{,0} \\
R^{i}{}_{kjl} &= -\delta^{i}{}_{l}\phi_{,jk} + \delta_{kl}\phi^{,i}{}_{,j} + \delta^{i}{}_{j}\phi_{,kl} - \delta_{jk}\phi^{,i}{}_{,l}
\end{aligned}$$

where the notation involves

$$\phi_{,ij} \equiv \frac{\partial^2 \phi}{dx^i\, dx^j} \quad .$$

These can be contracted to calculate the components of the Ricci tensor (again to order ϕ), giving

$$\begin{aligned}
R_{00} &= \nabla^2\phi + 3\phi_{,00} \\
R_{0i} = R_{i0} &= 2\phi_{,0i} \\
R_{ij} &= \delta_{ij}\left(\nabla^2\phi + \phi_{,00}\right)
\end{aligned}$$

where I'm writing

$$\nabla^2\phi \equiv \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2}$$

as usual. A final contraction gives the Ricci scalar

$$R = 2\,\nabla^2\phi \quad .$$

The Einstein tensor is then

$$\begin{aligned} G^{00} &= 2\,\nabla^2\phi + 3\phi_{,00} \\ G^{0i} = G^{i0} &= -2\phi^{,0i} \\ G^{ij} &= \delta^{ij}\,\phi_{,00} \end{aligned}$$

For matter which is moving non-relativistically, and which has zero pressure (i.e., has a pressure much less than ρc^2, which is always the case for Newtonian gravity), the stress-energy tensor

$$T^{\mu\nu} = (\rho + P)\,u^\mu u^\nu + P g^{\mu\nu}$$

reduces to only one non-zero term,

$$T^{00} = \rho$$

so that in the Einstein field equations, with zero cosmological constant (which is again an excellent approximation for Newtonian cases)

$$\mathbf{G} = 8\,\pi\,\mathbf{T}$$

leads to the 6 unique equations

$$\begin{aligned} 2\,\nabla^2\phi + 3\phi_{,00} &= 8\,\pi\,\rho \\ \phi_{,i0} &= 0 \\ \phi_{,00} &= 0 \end{aligned}$$

from which we can see that the potential ϕ is static (from the second equation: the spatial gradient of ϕ can't be zero if we're to have any sort of interesting gravitational effects). And substituting a zero second time derivative from the third equation ($\phi_{,00} = 0$) in the first, we get an equation for the potential function

$$\nabla^2\phi = 4\,\pi\,\rho$$

which is exactly the Poisson equation from Newtonian theory.

This confirms the consistency of general relativity and Newtonian gravity, and the use of the 8π factor in the Einstein field equations.

Technically, I've missed out an important step, which is to demonstrate that the coordinates I used (t, x, y, z) can be interpreted in a Newtonian sense as time and positions. However, the metric can be seen to be close to Minkowski in form, and we'd predict that the differences between the (t, x, y, z) coordinates and the observable times and positions are small (of order ϕ), so that to order ϕ the above argument is correct and no further development is needed. Refer to GR textbooks for a more complete treatment.

11. The equations of motion of the Universe

11.1. The Friedmann-Robertson-Walker metric

On the large scale the Universe is **isotropic and homogeneous**. That is,

- on sufficiently-large scales the Universe has roughly the same density everywhere — *i.e., the Universe is homogeneous on the large scale*
- in all directions, averaged over sufficiently-large scales, the Universe looks the same — *i.e., the Universe is isotropic*
- and we assume that we are not in any special position, so that *all observers will see roughly the same isotropic, homogeneous, Universe as we do.* This is called the **cosmological principle**.

What this means is that to a reasonable approximation we can describe the large-scale structure of the Universe by **isotropic, homogeneous, models**, and this will dictate the form of the metric. We may also choose to include small-scale structures to this, to describe local variations in density — but this is a frill which shouldn't affect the overall description of the Universe, which is what we're after at the moment.

What sort of metric is implied? In the metric we describe this physics using the mathematical statements:

(1) spacetime must contain homogeneous and isotropic subspaces at constant time (constant-time hypersurfaces are homogeneous and isotropic); and

(2) the rest frame of "average galaxies" defines constant-time hypersurfaces.

So our time, t, coordinate can be chosen to be the proper time for galaxies. The general form of the metric has a line element

$$ds^2 = g_{00}\,dt^2 + 2\,g_{0i}\,dt\,dx^i + g_{ij}\,dx^i\,dx^j \quad .$$

With the choice that t is the proper time, we know that $g_{00} = -1$, since a world line with $dx^i = 0$ (i.e., the world line of a particle at rest) has $ds^2 = -dt^2$.

But all observers at rest in the frame of the galaxies must agree on the definition of time. This means that the time basis vector $\vec{e}_0$ is perpendicular to the space basis vectors $\vec{e}_i$ in coordinates fixed to the frame of the galaxies, and hence that $g_{0i} = 0$. The metric therefore simplifies to

$$ds^2 = -dt^2 + g_{ij}\,dx^i\,dx^j \quad .$$

We now use the assumption of isotropy. If the Universe is isotropic at time t_0, then isotropy at time t_1 implies that

$$g_{ij}(t = t_1) = f(t_1, t_0)\,g_{ij}(t = t_0)$$

where the function $f(t_1, t_0)$ represents a scaling with time which is the **same** for all metric coefficients. If this were not true, then we could have a different time-dependence in the x than the y directions, which would lead to anisotropy. The consequence of this is that we can factor out the time-dependence in the metric components, and write

$$g_{ij} = [a(t)]^2 \; \gamma_{ij}(\mathbf{x})$$

where $a(t)$ is some function of time, and is called the **scale factor**. You will recall from Lecture 9 that this scale factor is responsible for the redshift.

Isotropy also requires that the γ_{ij} encode spherical symmetry about the origin of coordinates or any other point in the spacetime. This means that at fixed (r, t) the separation of points must go like $d\theta^2 + \sin^2\theta\, d\phi^2$ where (θ, ϕ) are angular coordinates, and this is multiplied by some distance factor. Further, we know that we can choose these angular coordinates to be perpendicular on the surface of a sphere, so $\vec{e}_\theta$ is perpendicular to $\vec{e}_\phi$, and $g_{\theta\phi} = 0$.

We define the radial coordinate r, so that at fixed time t_0, the separation of two points near angle θ, ϕ separated by $d\theta, d\phi$ is

$$a(t_0)\, r \; \left(d\theta^2 + \sin^2\theta\, d\phi^2\right)^{1/2}$$

and so that the radial coordinate basis vector is perpendicular to the $\vec{e}_\theta$ and $\vec{e}_\phi$ basis vectors. This means that $\vec{e}_\theta$ and $\vec{e}_\phi$ lie in spherical surfaces, while $\vec{e}_r$ is perpendicular to these surfaces, and hence that $g_{r\theta} = g_{r\phi} = 0$.

This has simplified the metric, and brings it to the form

$$ds^2 = -dt^2 + a^2 \left(e^{2G(r)}\, dr^2 + r^2 \left(d\theta^2 + \sin^2\theta\, d\phi^2\right)\right)$$

where I've chosen a convenient form for the only remaining metric component that we need to study,

$$g_{rr} = [a(t)]^2 \; e^{2G(r)}$$

where $G(r)$ is an unknown function of r (and can't depend on the other coordinates for reasons of isotropy).

To make progress we will impose homogeneity: we will require that every point in the Universe sees the same curvature at constant time. That is, we require that the Ricci scalar be constant everywhere at constant time.

From this metric we must now calculate the Ricci scalar. We have non-zero metric components

$$\begin{aligned} g_{tt} &= -1 \\ g_{rr} &= a^2\, e^{2G} \\ g_{\theta\theta} &= a^2\, r^2 \\ g_{\phi\phi} &= a^2\, r^2\, \sin^2\theta \end{aligned}$$

from which we can calculate the Christoffel symbols: the only non-zero ones are

$$\begin{aligned}
\Gamma^t{}_{rr} &= a\dot{a}e^{2G} \\
\Gamma^t{}_{\theta\theta} &= a\dot{a}r^2 \\
\Gamma^t{}_{\phi\phi} &= a\dot{a}r^2\sin^2\theta \\
\Gamma^r{}_{rt} = \Gamma^r{}_{tr} = \Gamma^\theta{}_{\theta t} = \Gamma^\theta{}_{t\theta} = \Gamma^\phi{}_{\phi t} = \Gamma^\phi{}_{t\phi} &= \frac{\dot{a}}{a} \\
\Gamma^r{}_{rr} &= G' \\
\Gamma^r{}_{\theta\theta} &= -re^{-2G} \\
\Gamma^r{}_{\phi\phi} &= -r\sin^2\theta e^{-2G} \\
\Gamma^\theta{}_{\theta r} = \Gamma^\theta{}_{r\theta} = \Gamma^\phi{}_{r\phi} = \Gamma^\phi{}_{\phi r} &= \frac{1}{r} \\
\Gamma^\phi{}_{\theta\phi} = \Gamma^\phi{}_{\phi\theta} &= \frac{\cos\theta}{\sin\theta} \\
\Gamma^\theta{}_{\phi\phi} &= -\sin\theta\,\cos\theta
\end{aligned}$$

where $G' \equiv \frac{dG}{dr}$. We can now calculate the Riemann tensor, in the usual way, from

$$R^\alpha{}_{\beta\mu\nu} = \Gamma^\alpha{}_{\beta\nu,\mu} - \Gamma^\alpha{}_{\beta\mu,\nu} + \Gamma^\alpha{}_{\sigma\mu}\Gamma^\sigma{}_{\beta\nu} - \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\beta\mu}$$

or simply contract to the Ricci tensor

$$R_{\mu\nu} = \Gamma^\alpha{}_{\mu\nu,\alpha} - \Gamma^\alpha{}_{\mu\alpha,\nu} + \Gamma^\alpha{}_{\sigma\alpha}\Gamma^\sigma{}_{\mu\nu} - \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\mu\alpha}$$

which turns out to be relatively easy to calculate since many of the terms are zero. The non-zero terms all lie on the diagonal, and are

$$\begin{aligned}
R_{tt} &= -3\frac{\ddot{a}}{a} \\
R_{rr} &= e^{2G}\left[\left(a\ddot{a}+2\dot{a}^2\right) + \frac{2}{r}G'e^{-2G}\right] \\
R_{\theta\theta} &= r^2\left[\left(a\ddot{a}+2\dot{a}^2\right) + \left(rG'-1\right)\frac{e^{-2G}}{r^2} + \frac{1}{r^2}\right] \\
R_{\phi\phi} &= r^2\sin^2\theta\left[\left(a\ddot{a}+2\dot{a}^2\right) + \left(rG'-1\right)\frac{e^{-2G}}{r^2} + \frac{1}{r^2}\right]
\end{aligned}$$

where the close similarity of the $R_{\theta\theta}$ and $R_{\phi\phi}$ is **not** a coincidence. We can now contract once more to calculate the Ricci scalar, which is

$$R = 6\left(\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2}\right) + \frac{1}{a^2}\,\frac{2}{r^2}\left(1 - e^{-2G}\left(1 - 2rG'\right)\right) \quad .$$

For spatial homogeneity at fixed time, R must be independent of r. This means that the second term on the RHS must be a constant, i.e., that

$$\frac{2}{r^2}\left(1-\left(re^{-2G}\right)'\right)=6k$$

where k is some constant (and I write $6k$ because of the factor 6 on the first term on the RHS). This can be rearranged and integrated pretty easily, the result is

$$e^{-2G}=1-kr^2+\frac{A}{r}$$

where A is some constant. And this then gives us a metric component

$$g_{rr}=a^2\,\left(1-kr^2+\frac{A}{r}\right)^{-1}\quad .$$

What can we do about the constant A? Well, near $r=0$, the term A/r becomes large, which causes $g_{rr}\to 0$. But we know that close to any point (and this includes the ordinary point $r=0$), the metric **must be locally flat**, so that $g_{rr}\neq 0$. This requires $A=0$ (otherwise we're locally in a singularity). And so the metric component becomes

$$g_{rr}=a^2\,\left(1-kr^2\right)^{-1}\quad .$$

and the line element of the so-called Friedmann-Robertson-Walker metric (FRW metric for short) is then

$$ds^2=-dt^2+[a(t)]^2\left(\frac{dr^2}{1-kr^2}+r^2\,d\theta^2+r^2\,\sin^2\theta\,d\phi^2\right)$$

This can be rewritten in a number of ways by redefining the radius coordinate (we have perfect freedom to change coordinates as we wish): one good choice is to take

$$d\chi=\frac{dr}{(1-kr^2)^{1/2}}$$

when the metric looks like

$$ds^2=-dt^2+[a(t)]^2\left(d\chi^2+f(\chi)^2\left(d\theta^2+\sin^2\theta\,d\phi^2\right)\right)$$

for some function $f(\chi)$ which depends on the value of k. $f(\chi)$ is then something to do with the relationship between angular size and proper size of a distant object, while χ looks more like what we conventionally think of as a radius. But remember that χ or r are just coordinates (numbers), and not to be thought of as distances in the Newtonian sense.

11.2. The dynamics of the Universe

Adopt the FRW metric in its raw form,

$$ds^2 = -dt^2 + [a(t)]^2 \left(\frac{dr^2}{1-kr^2} + r^2\, d\theta^2 + r^2\, \sin^2\theta\, d\phi^2 \right)$$

and let's use the field equations of GR to work out the dynamics of the Universe. Using the earlier result for $G(r)$, and the expressions that we had for the Ricci tensor and scalar, we now have updated forms

$$\begin{aligned} R_{tt} &= -3\frac{\ddot{a}}{a} \\ R_{rr} &= \frac{1}{1-kr^2}\left(a\ddot{a} + 2\dot{a}^2 + 2k\right) \\ R_{\theta\theta} &= r^2\left(a\ddot{a} + 2\dot{a}^2 + 2k\right) \\ R_{\phi\phi} &= r^2\sin^2\theta\left(a\ddot{a} + 2\dot{a}^2 + 2k\right) \\ R &= 6\left(\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2}\right) \end{aligned}$$

where the high symmetry of the Ricci tensor caused by the assumptions of homogeneity and isotropy are evident. From this symmetry you'd guess that there are much simpler ways of calculating the Ricci tensor than flogging through all the algebra involving the Christoffel symbols and the Riemann tensor, and you'd be right, but this would involve proving some results that we don't otherwise need. I refer you to advanced books on GR to see what tricks are available for spatially homogeneous metrics.

Returning to our problem, we can write down the Einstein tensor

$$\begin{aligned} G^{tt} &= 3\left(\frac{k}{a^2} + \frac{\dot{a}^2}{a^2}\right) \\ G^{rr} &= -\frac{1-kr^2}{a^2}\left(\frac{k}{a^2} + 2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2}\right) \\ G^{\theta\theta} &= -\frac{1}{a^2r^2}\left(\frac{k}{a^2} + 2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2}\right) \\ G^{\phi\phi} &= -\frac{1}{a^2r^2\sin^2\theta}\left(\frac{k}{a^2} + 2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2}\right) \end{aligned}$$

Now we must consider the stress-energy tensor for the Universe. We are dealing with a "fluid" of material (galaxies, radiation). And by assumption, there is a frame (the rest frame of the galaxies) in which this fluid has no bulk motion and to which we have tied our coordinate system. Therefore if we approximate the material in the Universe as a perfect fluid,

$$T^{\mu\nu} = (\rho + P)u^\mu u^\nu + Pg^{\mu\nu}$$

with $\vec{u} = (1, 0, 0, 0)$ in (t, r, θ, ϕ) coordinates. So I have

$$\begin{aligned} T^{tt} &= \rho \\ T^{rr} &= P\frac{1-kr^2}{a^2} \\ T^{\theta\theta} &= P\frac{1}{a^2 r^2} \\ T^{\phi\phi} &= P\frac{1}{a^2 r^2 \sin^2\theta} \end{aligned}$$

and the field equations

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

are

$$\begin{aligned} \dot{a}^2 + k &= \frac{8\pi}{3}\rho a^2 + \frac{\Lambda}{3}a^2 && tt \\ 2a\ddot{a} + \dot{a}^2 + k &= -8\pi P a^2 + \Lambda a^2 && rr \\ 2a\ddot{a} + \dot{a}^2 + k &= -8\pi P a^2 + \Lambda a^2 && \theta\theta \\ 2a\ddot{a} + \dot{a}^2 + k &= -8\pi P a^2 + \Lambda a^2 && \phi\phi \end{aligned}$$

with the other 12 equations being the identity $0 = 0$. Notice that the three equations derived from the rr, $\theta\theta$, and $\phi\phi$ components are identical. This is a consequence of the spatial isotropy of our metric.

These equations can be tidied up into their standard forms

$$\begin{aligned} \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} &= \frac{8\pi}{3}\rho + \frac{\Lambda}{3} && \text{energy equation} \\ \frac{\ddot{a}}{a} &= -\frac{4\pi}{3}(\rho + 3P) + \frac{\Lambda}{3} && \text{acceleration equation} \end{aligned}$$

which describe the evolution of the scale factor of the Universe. The first equation is something like a statement of energy conservation: $\dot{a}^2$ is like the kinetic energy per unit volume of the Universe's expansion, the term in ρa^2 is like the gravitational potential energy, and the term in Λa^2 is like the compressional energy associated with the vacuum. k can then be interpreted as an energy constant (and you'd expect the properties of the Universe to be different depending on whether the total energy is positive or negative).

The second equation is a statement that the Universe is decelerated by the effect of matter ($\propto (\rho + 3P)$, notice the effect of the pressure here), and accelerated by the tension of the vacuum. If the cosmological constant is small, then the Universe is being decelerated (as long as the pressure of the cosmic fluid isn't large and negative, $P > -\frac{1}{3}\rho$). But there are, perhaps, situations in which the Universe might accelerate.

Of these two equations, both of which describe the change of the scale factor with time, one is redundant — and we usually take this to be the second, which is the more

complicated. This is redundant since it follows from the first equation and the law of conservation of energy. For this metric the law of conservation of energy

$$T^{\mu\nu}{}_{;\nu} = 0$$

is four equations, three of which are the trivial $0 = 0$ identity (for $\mu = r$, θ, or ϕ). The remaining equation is

$$\begin{aligned} T^{t\nu}{}_{;\nu} &= 0 \\ T^{t\nu}{}_{,\nu} + T^{\alpha\nu}\Gamma^{t}{}_{\alpha\nu} + T^{t\alpha}\Gamma^{\nu}{}_{\alpha\nu} &= 0 \\ T^{tt}{}_{,t} + T^{rr}\Gamma^{t}{}_{rr} + T^{\theta\theta}\Gamma^{t}{}_{\theta\theta} + T^{\phi\phi}\Gamma^{t}{}_{\phi\phi} + T^{tt}\Gamma^{\nu}{}_{t\nu} &= 0 \\ \dot{\rho} + 3P\frac{\dot{a}}{a} + 3\rho\frac{\dot{a}}{a} &= 0 \end{aligned}$$

which can be written in the more suggestive form

$$\frac{d}{dt}\left(\rho a^3\right) = -P\,\frac{d}{dt}\left(a^3\right) \quad .$$

How to interpret this? The volume element for an infinitesimal element of the Universe is the volume measured in a local Lorentz frame, and this is shown in Lecture 14 to be

$$d^4\Omega = (-g)^{1/2}\,dt\,dr\,d\theta\,d\phi$$

where g is the determinant of the metric tensor written as a matrix. In the current case, $g = g_{tt}\,g_{rr}\,g_{\theta\theta}\,g_{\phi\phi}$, so

$$d^4\Omega = \frac{a^3\,r^2\,\sin\theta}{(1-kr^2)^{1/2}}\,dt\,dr\,d\theta\,d\phi$$

and the (3-space) volume element at fixed time is

$$dV = a^3\,\frac{r^2\,\sin\theta}{(1-kr^2)^{1/2}}\,dr\,d\theta\,d\phi \quad .$$

The volume of a bit of the Universe at a fixed spatial location (r, θ, ϕ) changes with time $\propto a^3$. The rate of change of ρa^3 is therefore the rate of change of the total energy content of a fluid element, and $P(\dot{a^3})$ is the work done by that fluid element as it expands, $-P\frac{dV}{dt}$. So

$$\frac{d}{dt}\left(\rho a^3\right) = -P\,\frac{d}{dt}\left(a^3\right) \quad .$$

is simply a thermodynamic statement about there being no place to lose energy in the Universe: since the Universe is isotropic, there can be no energy flows (there's nowhere for the energy to flow to!), and so $dE = -PdV$, which is exactly what we have here.

We can recover this equation just as easily by differentiating the Friedmann equation.

$$\frac{d}{dt}\left(\dot{a}^2+k\right)=\frac{d}{dt}\left(\frac{8\pi}{3}a^2\rho+\frac{\Lambda}{3}a^2\right)$$
$$2a\ddot{a}=\frac{8\pi}{3}(\dot{\rho a^2})+\frac{2}{3}\Lambda a\dot{a}$$
$$2a\left(-\frac{4\pi}{3}\left(\rho+3P\right)a+\frac{\Lambda}{3}a\right)=\frac{8\pi}{3}(\dot{\rho}a^2+2\rho a\dot{a})+\frac{2}{3}\Lambda a\dot{a}$$

The Λ terms cancel, leaving

$$-\dot{a}(\rho+3P)a=-\dot{\rho}a^2+2\rho a\dot{a}$$
$$\dot{\rho}a^2+3\rho a\dot{a}=-3Pa\dot{a}$$
$$(\dot{\rho a^3})=-P(\dot{a^3})$$

again. This means that any two equations of this set of three are enough, and we usually choose the simplest. But the full set is available for use, and is

$$\frac{\dot{a}^2}{a^2}+\frac{k}{a^2}=\frac{8\pi}{3}\rho+\frac{\Lambda}{3} \qquad \text{Friedmann equation}$$
$$\frac{\ddot{a}}{a}=-\frac{4\pi}{3}\left(\rho+3P\right)+\frac{\Lambda}{3} \qquad \text{acceleration equation}$$
$$\frac{d}{dt}\left(\rho a^3\right)=-P\frac{d}{dt}\left(a^3\right) \qquad \text{equation of state.}$$

12. Cosmological parameters

12.1. The Hubble constant and deceleration parameter

Let us return to the Friedmann and acceleration equations

$$\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = \frac{8\pi}{3}\rho + \frac{\Lambda}{3}$$
$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\rho + 3P) + \frac{\Lambda}{3}$$

and consider what they say at the present (at $t = t_0$). Inserting this time into these relations we get information about the current rate of expansion of the Universe, $\dot{a}_0$, and the current acceleration of the Universe $\ddot{a}_0$. Define two parameters which measure the sizes of these quantities: we measure the current expansion rate of the Universe by the **Hubble constant**,

$$H_0 = \frac{\dot{a}_0}{a_0}$$

and the current slowing down of the Universe by the deceleration parameter,

$$q_0 = -\frac{\ddot{a}_0 a_0}{\dot{a}_0^2}$$

where, as usual, $a_0 = a(t_0)$, and $\dot{a}_0 = \dot{a}(t_0)$, $\ddot{a}_0 = \ddot{a}(t_0)$.

Using these definitions in the Friedmann and acceleration equations at $t = t_0$,

$$H_0^2 + \frac{k}{a_0^2} = \frac{8\pi}{3}\rho_0 + \frac{\Lambda}{3}$$
$$-q_0 H_0^2 = -\frac{4\pi}{3}(\rho_0 + 3P_0) + \frac{\Lambda}{3} \quad .$$

The Universe is now matter-dominated, so $\rho_0 \gg P_0$, and in the simplest models of the Universe we take $\Lambda = 0$. Taking these limiting cases, we can use the second of the above relations to eliminate the current density ρ_0 in terms of the parameters q_0 and H_0,

$$\frac{4\pi}{3}\rho_0 = q_0 H_0^2$$

and replace the curvature parameter, k, using

$$\frac{k}{a_0^2} = (2q_0 - 1)H_0^2 \quad .$$

Substituting these expressions back into the Friedmann equation,

$$\frac{\dot{a}^2}{a^2} + \frac{a_0^2}{a^2}(2q_0 - 1)H_0^2 = 2q_0 H_0^2 \frac{\rho}{\rho_0}$$

and using the dust equation of state, $\rho \propto a^{-3}$, which corresponds to **conservation of the number of particles in the Universe**, which is clearly consistent with a low temperature (so that particles can't pair-produce),

$$\frac{\rho}{\rho_0} = \left(\frac{a_0}{a}\right)^3$$

this becomes

$$\dot{a}^2 = H_0^2 a_0^2 \left(1 - 2q_0 + 2q_0 \frac{a_0}{a}\right) \quad .$$

What we have done here is to replace the physical parameters k and ρ_0, by the kinematic parameters H_0 and q_0 which (as we will see later) should be amenable to direct observational determination using measurements of "geometrical" properties of the Universe.

We can now integrate the expression for $a(t)$. There are three cases that we need to consider.

12.1.1 Flat Universe, $q_0 = \frac{1}{2}$, $k = 0$

In this case the differential equation for $a(t)$ is as simple as it can be:

$$\dot{a}^2 = H_0^2 a_0^2 \left(\frac{a_0}{a}\right)$$

which can be integrated directly to give the **Einstein – de Sitter solution**,

$$\frac{a(t)}{a_0} = \left(\frac{3}{2} H_0 t\right)^{2/3} \quad .$$

We call this a flat Universe since the spatial part of the metric is simply $dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)$, exactly as it would be in a three-dimensional space with zero curvature.

12.1.2 Closed Universe, $q_0 > \frac{1}{2}$, $k = +1$

Before doing the integration, a technical note (a caution). You will see from what we've derived above that

$$\frac{k}{a_0^2} = (2q_0 - 1)H_0^2$$

so how can I take $k = +1$ while also saying that a_0 is the scale factor at the present time, equal to 1? The answer is that $k = +1$ is enforced by making a scaling of the radius coordinate (as I am free to do in General Relativity). If I do so, then I must either accept a_0 as a parameter that I carry along, or absorb the a_0 scaling into the definition of time

(which appears in H_0). But what we do see from the relation between k and q_0 is that $q_0 > \frac{1}{2}$ implies positive k, so there is no inconsistency of sign in choosing $k = +1$.

Solving the equation for $a(t)$ is possible directly, but it is easier to obtain the solution in parametric form. Define an auxiliary parameter, ϕ, called the *development angle*, by

$$\frac{a(t)}{a_0} = \frac{q_0}{2q_0 - 1}\,(1 - \cos\phi)$$

then the equation for $\phi(t)$ is obtained by substituting this into the differential equation for $a(t)$, and making a simple integration, to find

$$\phi - \sin\phi = \frac{H_0 t}{q_0}\,(2q_0 - 1)^{3/2}$$

where I've chosen the constant of integration so that $a(t) = 0$ at $t = 0$, as usual.

We call this a *closed Universe* for reasons which will become apparent shortly.

12.1.3. Open Universe, $q_0 < \frac{1}{2}$, $k = -1$

Once again it is easier to solve for $a(t)$ by introducing a development angle, ψ. The definition of ψ in terms of $a(t)$ is

$$\frac{a(t)}{a_0} = \frac{q_0}{1 - 2q_0}\,(\cosh\psi - 1)$$

and substituting this into the expression for $\dot{a}$ and making a simple integration, we find that $\psi(t)$ is given by

$$\sinh\psi - \psi = \frac{H_0 t}{q_0}\,(1 - 2q_0)^{3/2}$$

in clear analogy with our earlier result for $q_0 > \frac{1}{2}$.

And this is called an *open Universe*, because of its interpretation.

12.2. Interpretation of the solutions

The first result that we can get from these solutions is the present age of the Universe: the time t_0 now. The timescale that relates to the age is set by the Hubble constant, which is conveniently written

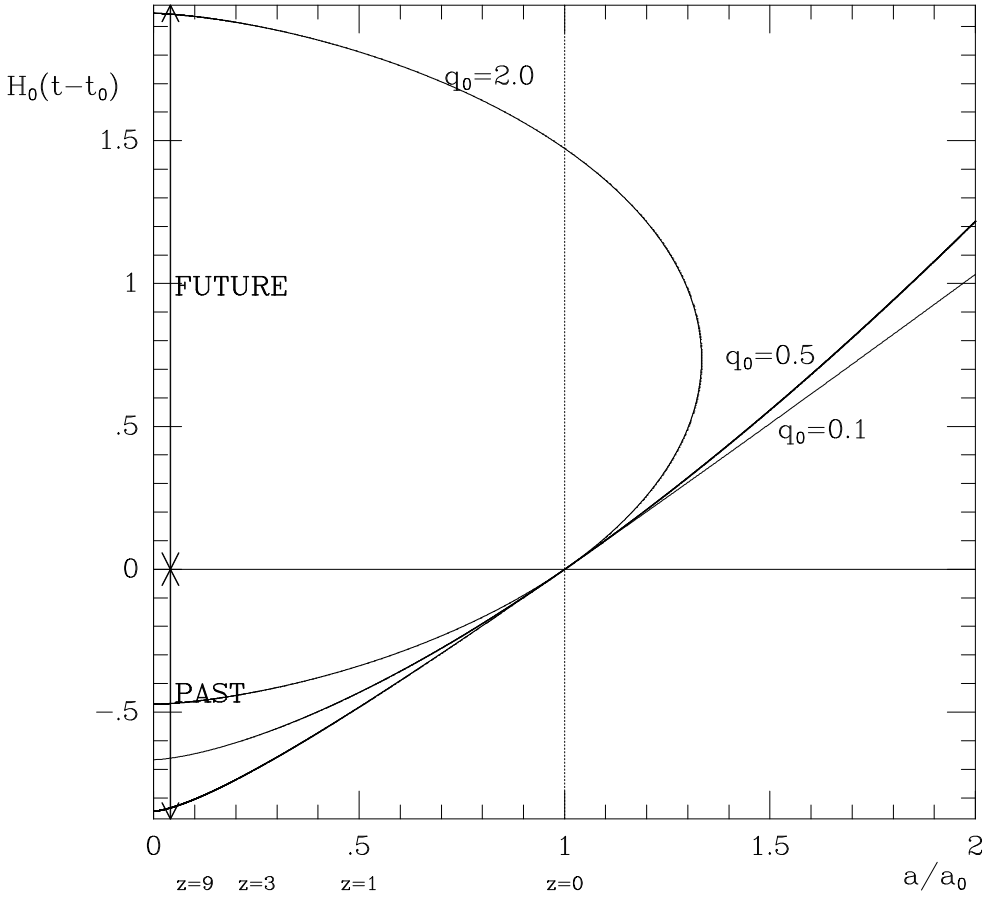
$$H_0 = 100 h_{100}\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$$

where h_{100} is a dimensionless measure of the Hubble constant, and is likely to be $0.5-0.7$. Then the characteristic timescale for the expansion of the Universe is

$$H_0^{-1} = (3.09 \times 10^{17}) h_{100}^{-1}\ \ \mathrm{sec} = (9.78 \times 10^{9}) h_{100}^{-1}\ \ \mathrm{years}.$$

Putting $a = a_0$ in the solution for $a(t)$ for for $k = 0$, we find that the age of the Universe is $t_0 = 6.5h_{100}^{-1}$ years. For $k = +1$ the value of t_0 depends on how rapidly the Universe has been decelerating since the Big Bang — for $q_0 = 2$, the age of the Universe is $4.6h_{100}^{-1}$ years. Finally, for $k = -1$ the age of the Universe again depends on how rapid the deceleration has been, and for $q_0 = 0.1$ the age is $8.3h_{100}^{-1}$ years.

The three solutions are plotted below, where rather than align them at $t = 0$ (which is a little confusing), I've aligned them at $t = t_0$, the present time. It can be seen that the three curves are very similar at the present — the slopes are the same (determined by the value of H_0), so the only difference is in the curvature of the functions (measured by q_0). This close similarity of the curves at the present is what makes it so difficult to tell whether the Universe is open, flat, or closed (what the value of q_0 is). However, the long-time predictions of the functions are very different.



For $k = -1$ or $k = 0$, we see that the solution $a(t)$ increases without limit as $t \to \infty$. At large times, $a \propto t^{2/3}$ if $k = 0$ and $a \propto t$ if $k = -1$. That is, in these two cases the Universe expands for ever, with the rate of expansion, $\dot{a}$, tending to zero if $k = 0$, and tending to a constant value if $k = -1$.

If $k = 0$ we say that we have a *critical-density Universe*: the deceleration of the

Universe caused by its matter content and pressure is just enough that the expansion rate falls to zero eventually, but the matter content is not enough to cause a recollapse.

If $k = -1$, the Universe has insufficient matter to cause the recollapse: at large time there is so little matter to cause deceleration that the Universe expands at constant speed. In fact, you can see that the curvature of the function plotted for $q_0 = 0.1$ is small at all times — the matter content of the Universe causes rather little deceleration for such a small q_0.

What about $k = +1$? Here we see that the Universe expands from zero size to a maximum, and then recollapses. From the parametric solution,

$$\frac{a(t)}{a_0} = \frac{q_0}{2q_0 - 1}\left(1 - \cos\phi\right)$$
$$\phi - \sin\phi = \frac{H_0 t}{q_0}\left(2q_0 - 1\right)^{3/2}$$

we see that $a(t) = 0$ at $\phi = 0$ or 2π, and that this corresponds to times $t = 0$ and

$$t = \frac{2\pi q_0}{H_0}\left(2q_0 - 1\right)^{-3/2}$$

corresponding to the Big Bang and the *Gnab Gib* or Big Crunch. Half-way between these two times there is a phase of maximum expansion, where

$$\frac{a}{a_0} = \frac{2q_0}{2q_0 - 1}$$

so for $q_0 = 2$ (for example), if we are in this type of Universe, the maximum size of the Universe will be $\frac{4}{3}$ times its present-day size.

That is, for the three different models

$k = -1$	$q_0 < \frac{1}{2}$	$t_0 > 6.5h_{100}^{-1}$ Gyr	Universe expands for ever	open
$k = 0$	$q_0 = \frac{1}{2}$	$t_0 = 6.5h_{100}^{-1}$ Gyr	Critical case	flat
$k = +1$	$q_0 > \frac{1}{2}$	$t_0 < 6.5h_{100}^{-1}$ Gyr	Universe will recollapse	closed

The critical case is characterized by the density of the Universe being at its critical value. In the Friedmann equation at $t = t_0$,

$$H_0^2 + \frac{k}{a_0^2} = \frac{8\pi}{3}\rho_0$$

if we continue to take $\Lambda = 0$. For $k = 0$, the density of the Universe now must have the special value

$$\rho_{\rm crit} = \frac{3H_0^2}{8\pi}$$

or, putting back the factors of G and c,

$$\rho_{\rm crit} = \frac{3H_0^2}{8\pi G} = (1.88 \times 10^{-26}) h_{100}^2 \ {\rm kg\,m^{-3}} \quad .$$

If the current density of the Universe exceeds $\rho_{\rm crit}$, the current critical density of the Universe, then the Universe will recollapse.

The ratio of the current density of the Universe to the current critical density of the Universe is an important parameter, Ω_0, the **density parameter**,

$$\Omega_0 = \frac{\rho_0}{\rho_{\rm crit}}$$

and $\Omega_0 > 1$ then corresponds to a closed Universe that will recollapse.

This "physics" parameter, Ω_0, is closely related to the "geometry" parameter, q_0. For a matter-dominated Universe with $\Lambda = 0$, the acceleration equation is

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}\rho$$

which at $t = t_0$ (i.e., the current time) becomes

$$\frac{\ddot{a}_0}{a_0} = -\frac{4\pi}{3}\rho_0 \quad .$$

Eliminate $\ddot{a}_0$ in favour of q_0, then

$$q_0 = \frac{4\pi}{3}\frac{\rho_0}{H_0^2} = \frac{1}{2}\frac{\rho_0}{\rho_{\rm crit}}$$

That is,

$$q_0 = \frac{1}{2}\Omega_0$$

for a matter-dominated Universe with $\Lambda = 0$.

By analogy with Ω_0, we often define an "equivalent density parameter" for Λ. Returning to the Friedmann equation for a matter-dominated Universe, but no longer dropping the cosmological constant, at $t = t_0$

$$\frac{\dot{a}_0^2}{a_0^2} + \frac{k}{a_0^2} = \frac{8\pi}{3}\rho_0 + \frac{\Lambda}{3} \quad .$$

Eliminate $\dot{a}_0$ in favour of H_0, then this can be written

$$H_0^2 + \frac{k}{a_0^2} = H_0^2\Omega_0 + \frac{\Lambda}{3}$$

or

$$\frac{k}{a_0^2} = H_0^2 \left(\Omega_0 + \Omega_\Lambda - 1\right)$$

where

$$\Omega_\Lambda = \frac{\Lambda}{3H_0^2}$$

and clearly $k = 0$ now requires that $\Omega_0 + \Omega_\Lambda = 1$. That is, a *flat Universe* results when the sum of the matter density and the density of the vacuum energy is equal to the critical density.

A few years ago we wouldn't have bothered with this — it was standard to take $\Lambda = 0$. However, cosmological data are now showing some signs that Λ may not be zero: although Λ still appears to be much to small to have any effect in the solar system it may be large enough to affect the Universe as a whole. As I said in an earlier lecture, though, the existence of Λ is still controversial — particle physics arguments tend to suggest that the vacuum energy is large, and $\Omega_\Lambda \gg 1$, but this would say that the Universe could not ever get to more than a few seconds old, which is a little at odds with observation. So many cosmologists still like to assume that it's exactly zero. We shall see within the next five or ten years, as precise data on the structure of the Universe come in.

13. Hubble law and classical cosmological tests

13.1. The Hubble Law

The first indication that the Universe is non-static was the Hubble law, which relates the "velocity" cz of an object to its "distance". How does this law follow from what we've done so far?

It is clear, I think, that the redshift, z, does not necessarily imply a velocity in the usual sense. Galaxies are, after all, at rest in the system of coordinates that we have chosen. However, the spacetime between them is swelling at a rate specified by the scale factor, and this causes galaxies to get a non-Doppler redshift. We can, if we wish, call cz a sort of representative velocity of a galaxy: for small *real* galaxy velocities, v, the redshift contribution made by the Doppler effect is

$$\frac{\Delta\lambda}{\lambda} = \frac{v}{c} = z$$

by the definition of z, and hence small velocities produce small redshifts cz.

For an object at coordinate r_1, what would be the value of the redshift? For simplicity, I'll do this for a flat (Einstein – de Sitter) cosmology, though I could do it equally well in any cosmology.

We relate z and r_1 by knowing that light travels to us on a radial null geodesic. Hence, using the FRW metric

$$ds^2 = -dt^2 + [a(t)]^2 \left(\frac{dr^2}{1-kr^2} + r^2 \left(d\theta^2 + \sin^2\theta \, d\phi^2 \right) \right)$$

with $d\theta = d\phi = 0$, the world line of an incoming light ray has

$$dt = -a(t) \, \left(1 - kr^2\right)^{-\frac{1}{2}} \, dr$$

and so the path $r(t)$ is given by

$$\int_t^{t_0} \frac{dt}{a(t)} = \int_0^r \frac{dr}{(1-kr^2)^{1/2}}$$

where I've taken care to choose the correct sign. Now for an Einstein – de Sitter cosmology, $k = 0$, and

$$\frac{a(t)}{a_0} = \left(\frac{3}{2} \, H_0 \, t \right)^{2/3} \quad .$$

so that the path of the incoming light ray is

$$\int_t^{t_0} \frac{dt}{a_0 \, \left(\frac{3}{2} \, H_0 \, t\right)^{2/3}} = \int_0^r dr$$

which integrates to

$$ra_0 = \left(\frac{2}{3\,H_0}\right)^{2/3} 3\,\left(t_0^{1/3} - t^{1/3}\right) \quad .$$

Suppose we observe a galaxy with redshift z_1. What is its r-coordinate, r_1? Clearly we have observed that galaxy using an incoming light ray, so we can relate the r-coordinate from which the light started to the time from which the light left the galaxy. We have one more piece of information, though: the redshift, which tells us what the time was at which the light left the galaxy, since it is a direct measure of the scale factor of the Universe when the light was emitted compared to the scale factor when the light was detected. That is,

$$(1 + z_1) = \frac{a(t_0)}{a(t_1)}$$

where t_1 was the time at which the light was emitted. Using the expression for $a(t)$,

$$(1 + z_1) = \frac{a(t_0)}{a(t_1)} = \left(\frac{t_0}{t_1}\right)^{2/3}$$

and

$$t_0 = \frac{2}{3H_0} \quad .$$

The result for r_1, the coordinate position of the galaxy seen at redshift z_1, is then

$$a_0 r_1 = \frac{2}{H_0}\left(1 - (1 + z_1)^{-1/2}\right) \quad .$$

The quantity $a_0 r_1$ is the *proper distance* of the galaxy, d_1 — that is, it's the distance that would be measured in a local Lorentz frame that contains both the galaxy and ourselves (taking the galaxy as being near enough to use that the approximation of a local Lorentz frame containing both is OK). So for small d_1 and small z_1, I can expand to get

$$d_1 = \frac{1}{H_0}\left(1 - \left(1 - \frac{1}{2}z_1 + \frac{3}{8}z_1^2 + O(z_1^3)\right)\right)$$

and working to first order in z_1, this becomes

$$H_0 d_1 = z_1 \quad .$$

Reinsert the factor c needed to take this into physical units from scaled units,

$$v_1 = H_0 d_1$$

i.e., the velocity cz_1 of recession of a galaxy is proportional to its distance from us, d_1. The constant of proportionality is the Hubble constant. It was this result, which showed the Universe expanding away from us (plus the Copernican idea that we're not repellent,

it's just that everything in the Universe is expanding uniformly and so moving away from everything else) which led to the idea of the non-infinite, and non-eternal Universe.

13.2. Luminosity distance and the Hubble Diagram

How bright is a source at redshift z? That is to say, what is the flux (the detected energy per unit area per unit time) of radiation from a source of luminosity L (the energy radiated per unit time) that can be detected by an observer?

The answer is

$$F = \frac{L}{4\pi a_0^2 r_1^2 (1+z)^2}$$

where r_1 is the radial coordinate of the source and a_0 is the current value of the scale factor. Deriving this result will give us a good idea about how relativistic arguments proceed.

Suppose the observer has a detector of area A_d, oriented perpendicular to light rays from the source. Let these light rays be detected at time t_0 (the present). If the source is at radial coordinate r_1, then the area of a two-sphere about the source at the time of detection is

$$4\pi r_1^2 a_0^2$$

(recall that we *defined* the radial coordinate so that distances on a two-sphere are $a_0 r_1 \delta\theta$, where $\delta\theta$ is an angle separation, and areas on a two-sphere are $a_0^2 r_1^2 \delta\Omega$, where $\delta\Omega$ is an element of solid angle. The expression above simply integrates over $\delta\Omega$). Therefore the fraction of the emitted energy at the source which is received by the detector is

$$\frac{A_d}{4\pi r_1^2 a_0^2} \quad .$$

But this radiation is redshifted: individual photons emitted at the source at frequency ν are detected at frequency $\nu/(1+z)$, so the energy received from a constant number of photons is decreased by a factor of $(1+z)$.

And also the photons radiated over a time interval Δt_1 at the source are received over a redshifted time interval $\Delta t_0 = \Delta t_1 \, (1+z)$, so less energy per unit is received by a factor $(1+z)$. Therefore the received energy per unit time at the detector is

$$F A_d = L \times \frac{A_d}{4\pi r_1^2 a_0^2} \times \frac{1}{1+z} \times \frac{1}{1+z}$$

where the RHS is the (emitted energy per second) times the (fraction of the sphere covered by the detector) times the (redshifting factor for the received energy) times the (time spread factor). Removing the area of the detector, we get the equation I quoted at the beginning,

$$F = \frac{L}{4\pi a_0^2 r_1^2 (1+z)^2} \quad .$$

It is convenient to write this as

$$F = \frac{L}{4\pi D_L^2}$$

where D_L is the **luminosity distance**, the distance that appears in the relation between flux and luminosity if this relation is written like the usual inverse-square law. With this definition,

$$D_L = a_0\, r_1\, (1+z) \quad .$$

This relationship isn't too useful, unless we know the coordinate r_1 of the emitting source. And we get this in the usual way from an argument about the path of light from the source to the observer. A radial null geodesic connects the emission and reception events at (t_1, r_1) and (t_0, r_0), and so, from the FRW metric

$$-dt^2 + a^2 \frac{dr^2}{1-kr^2} = 0 \quad .$$

Rearranging and choosing the correct sign, and integrating

$$\int_{t_1}^{t_0} \frac{dt}{a(t)} = \int_0^{r_1} \frac{dr}{(1-kr^2)^{1/2}} \quad .$$

It is easier to deal with this equation if we convert from a t integral to an integral over the scale factor a. Clearly

$$da = \dot{a}\, dt$$

and with $a = a_0$ at t_0, $a = a_1 = a_0\,(1+z)^{-1}$ at $t = t_1$,

$$\int_{a_0(1+z)^{-1}}^{a_0} \frac{da}{\dot{a}\, a} = \int_0^{r_1} \frac{dr}{(1-kr^2)^{1/2}}$$

(this sneaky trick is **well worth remembering**). Now we can use the Friedmann equation written in the form

$$\dot{a}^2 = H_0^2 a_0^2 \left(1 - 2q_0 + 2q_0 \frac{a_0}{a}\right) \quad .$$

and replace the curvature parameter, k, using

$$\frac{k}{a_0^2} = (2q_0 - 1) H_0^2 \quad .$$

to get an expressing relating the radial coordinate r_1 to the redshift z. If I also change the scale-factor variable from a to

$$x = \frac{a}{a_0}$$

to make the notation look easier, the integral I'm left with is

$$\frac{1}{a_0 H_0}\int_{(1+z)^{-1}}^{1}\frac{dx}{x}\left(1-2q_0+\frac{2q_0}{x}\right)^{-\frac{1}{2}}=\int_0^{r_1}dr\,\left(1-(2q_0-1)H_0^2a_0^2r^2\right)^{-\frac{1}{2}}\quad .$$

This can be integrated to give

$$a_0 r_1=\frac{zq_0+(q_0-1)\left((1+2q_0z)^{1/2}-1\right)}{H_0q_0^2(1+z)}\quad .$$

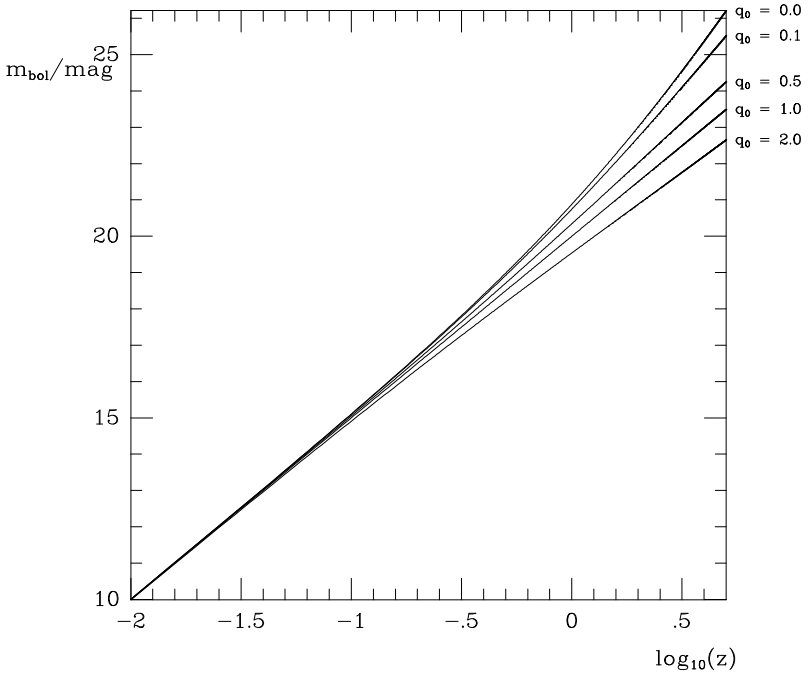
The final result for the luminosity distance is then (reinserting the missing c factor to convert to physical units),

$$D_L=\frac{c}{H_0q_0^2}\left(zq_0+(q_0-1)\left((1+2q_0z)^{1/2}-1\right)\right)\quad .$$

This tells us how to measure H_0 and q_0 (and, by implication, the density and age of the Universe, the curvature of the Universe, and so on). Consider observing "standard candles", galaxies with the same luminosity L_g, at varying redshifts (in fact, we use supernovae in galaxies, galaxy luminosities, Cepheid stars, and so on as distance indicators). Then look at how the flux of such galaxies varies as a function of redshift. If we express the flux in usual optical terms as bolometric magnitude,

$$m_{\rm bol}=m_0-2.5\,\log_{10}F$$

where the constant m_0 defines the zero point of the magnitude scale, and plot $m_{\rm bol}$ against $\log_{10}z$, we get the Hubble diagram, which looks like



At small z,

$$D_L = \frac{cz}{H_0}\left(1 + \frac{1}{2}(1-q_0)z + O(z^2)\right)$$

so

$$m_{\rm bol} = m_0 - 2.5\,\log_{10}\left(\frac{L_g H_0^2}{4\pi c^2}\right) + 5\log_{10}\left(z\left(1 + \frac{1}{2}(1-q_0)z + O(z^2)\right)\right)$$

and the slope of the $m_{\rm bol}/\log_{10}(z)$ graph at small z is 5 for any choice of H_0 or q_0. Deviations from a straight line on such a graph correspond to different values of q_0. And the value of H_0 can be determined from the level of the curve at some z.

However there are pitfalls:

(1) we need to know the absolute value of L_g to get H_0, but we usually don't — we don't have a fully-calibrated galaxy in the laboratory. However we can do a pretty good job with Cepheids.

(2) we need to know that L_g is standard, but in fact when we look at this curve we're looking to very different parts of the Universe (potentially a long way back in time), so it could be that galaxies are changing in properties with redshift (for example because of time-varying star formation rates).

(3) the redshifts have to be representative of the expansion of the Universe, and not of the peculiar motions of galaxies — their random motions add a Doppler shift to the cosmological redshift

$$(1+z) = (1+z_{\rm cosmological})(1+z_{\rm peculiar})$$

which can cause the redshifts of nearby objects to be very non-representative of the cosmological redshift.

As a result, the practical difficulties of using this method are immense. At low redshift, and using Cepheids, this is probably the best way of trying to get at H_0. But as a method of getting at q_0 it is bad.

There are a few complications, too. It's not easy to measure the bolometric magnitude (the integrated light output of an object): normally we measure only the light in a particular waveband, such as the optical V band. Since the emitted light comes from a bluer part of the spectrum, we see different parts of the spectrum of a galaxy (or star, or supernova) as we look at objects of different redshift, and we must correct them all back to a common passband. Such a correction is called a **K-correction**.

Another problem with using fuzzy-edged things like galaxies as distance indicators is that you can't be sure that you're looking at all the galaxy light — the galaxy gradually trails off into the noise, and some of the light is missed. To get around this, one often looks only at the light out to some fixed radius (e.g., 50 kpc) ... but for the faintest objects, which are the most distant, this may mean that the object is not much larger than the point spread function of the telescope being used, and an **aperture correction** may be needed to correct to the same fixed radius of an object.

14. Classical cosmological tests, continued

14.1. The angular diameter distance

An alternative way to attempt to measure the value of q_0 is to look at an object of known linear size as a function of distance from us. That is, we look at the angular sizes of *standard rods*, such as galaxies of a particular type, at different redshift.

How do we calculate the angular size of an object of a given linear size? Suppose a galaxy has a proper diameter (diameter in a local Lorentz frame) d_g ($\ll a(t_1)r_1$), lies at radial coordinate $r = r_1$, and is observed to have redshift z. What is its apparent angular size?

Return to the FRW metric, which tells us what we want to know.

$$ds^2 = -dt^2 + [a(t)]^2 \left(\frac{dr^2}{1-kr^2} + r^2 \left(d\theta^2 + \sin^2\theta \, d\phi^2 \right) \right)$$

Based on this, at a fixed time (so $dt = 0$) and fixed radial coordinate (so $dr = 0$), an element of proper size (ds) can be related to an element of angle ($d\theta$ — we choose to put the object at fixed ϕ, for simplicity) by

$$ds = a(t)\, r \, d\theta$$

so for our galaxy

$$d_1 = a(t_1)\, r_1 \, \theta_g$$

where θ_g is the angular extent of the rod. But light rays from the ends of the rod travel on radial null geodesics to us, situated at the origin of coordinates, and hence the observable angular size of the galaxy is also θ_g, with

$$\theta_g = \frac{d_1}{a(t_1)\, r_1} \quad .$$

We have already shown that for an object at redshift z the radial coordinate is given by

$$a_0 r_1 = \frac{z q_0 + (q_0 - 1)\left((1 + 2q_0 z)^{1/2} - 1 \right)}{H_0 q_0^2 (1+z)}$$

and the scale factor at time t_1 when the light was emitted is related to a_0 by

$$a(t_1) = \frac{a_0}{1+z} \quad .$$

Hence we can write the angular size of the object as

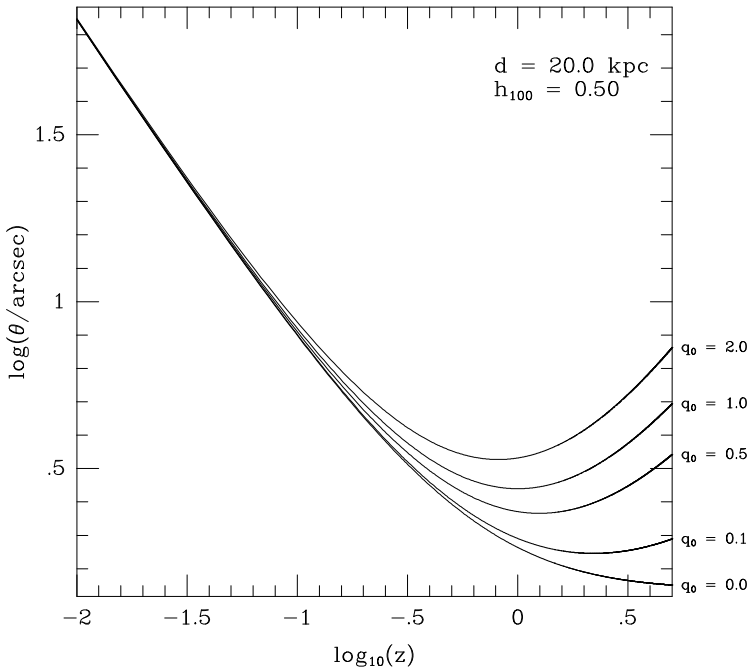
$$\theta_g = \frac{d_g}{D_A}$$

where D_A is the *angular diameter distance*,

$$D_A = \frac{c}{H_0 q_0^2} \frac{\left(zq_0 + (q_0 - 1)\left((1 + 2q_0 z)^{1/2} - 1\right)\right)}{(1+z)^2}$$
$$= \frac{D_L}{(1+z)^2} \quad .$$

Notice that the angular diameter distance is **not** the same as the luminosity distance — the ordinary flat-space ideas about the meaning of distance don't work too well in GR.

But now we can look to see how the angular size of an object of fixed linear size changes with redshift. This is shown in the diagram below, for an object with linear size 20 kpc and $h_{100} = 0.5$.



What is going on here? We see that for $q_0 > 0$, there is a redshift at which the angular size is a minimum, and then the angular size rises for larger z — in other words, the further away an object is, the larger its angular size.

This is a consequence of the curved geometry of the Universe. Think of the meaning of this for a closed space — for example, the apparent angular size of a rod on the surface of a sphere, as seen from the north pole. When the rod is close to the observer, it lies across many lines of longitude, and therefore its angular size (the number of lines of longitude that it crosses) is large. As the rod moves down towards the equator it cuts fewer lines of longitude, until when it reaches the equator it cuts the minimum

number (has the smallest angular size). As we then move the rod towards the south pole, it crosses more lines of longitude again, until when it reaches the south pole it crosses all lines of longitude and has angular size 2π radians.

The analogy isn't exact, however, because we're looking back into a past when the Universe had a different curvature, and much of the shape of the curve is tracking this time-dependence and not the present-day geometry of the Universe. Nevertheless, the minimum angular size should be an observable phenomenon. At $q_0 = 0.5$, the angular size of any object decreases out to $z = 1.25$, and then increases again. And the redshift of minimum angular size depends on the value of q_0 — so we might hope to use "standard rods" in the form of standard galaxies to find this minimum redshift and hence measure q_0. Unfortunately this doesn't work — galaxies evolve too much between $z = 1$ and the present, so we can't treat them as fixed objects. Attempts are still being made to use this technique, though, with other supposedly standard-sized objects.

Note an interesting consequence of the results for D_L and D_A: the surface brightness of an object at redshift z is the flux of that object over its angular size. If we assume that the object is spherical, with luminosity L and radius R, the flux and angular radius of that object are

$$F = \frac{L}{4\pi D_L^2}$$
$$\theta_R = \frac{R}{D_A}$$

and hence the flux per unit solid angle (the surface brightness) is

$$\frac{F}{\pi\theta_R^2} = \frac{L}{4\pi R^2}\frac{D_A^2}{D_L^2} \propto (1+z)^{-4} \quad .$$

That is, the surface brightness of any object drops off as $(1 + z)^4$. This *fading out* of distant objects makes them very hard to see — for example, a galaxy at $z = 2$ has a central brightness that is only about 1 per cent of the central brightness it would have if it was nearby. It might, therefore, vanish under the noise in a detector which is looking for it.

Notice also that this is *different* from the situation in a Minkowski metric, where surface brightness is a relativistic invariant.

14.2. Volume of the Universe

How large is a volume element in a pseudo-Riemannian manifold? By the volume element, I mean the quantity $d^4\Omega$ needed in the 4-D Gauss law,

$$\int_\Omega \nabla . T \, d^4\Omega = \oint_{\partial\Omega} T . d^3\tilde{\Sigma}$$

where $d^3\tilde{\Sigma}$ is the surface element one-form. In a local Lorentz frame patch (that is, in a coordinate system which is tangent to the spacetime at the point we're interested in and which has a flat-spacetime metric), we know that the 4-D element of volume is

$$d^4\Omega = dx^0 dx^1 dx^2 dx^3$$

where the $\{x^\alpha\}$ are coordinates in which $g_{\alpha\beta} = \eta_{\alpha\beta} + O(|\Delta\vec{x}|^2)$ for small shifts $\Delta\vec{x}$ from the tangent point. In another coordinate system, $\{x^{\alpha'}\}$, the Jacobian relates the coordinate expressions of the volume element, as

$$dx^0 dx^1 dx^2 dx^3 = \left[\frac{\partial(x^0 x^1 x^2 x^3)}{\partial(x^{0'} x^{1'} x^{2'} x^{3'})}\right] dx^{0'} dx^{1'} dx^{2'} dx^{3'}$$

where the Jacobian is the usual determinant

$$\frac{\partial(x^0 x^1 x^2 x^3)}{\partial(x^{0'} x^{1'} x^{2'} x^{3'})} = \begin{pmatrix} \frac{dx^0}{dx^{0'}} & \frac{dx^0}{dx^{1'}} & \frac{dx^0}{dx^{2'}} & \frac{dx^0}{dx^{3'}} \\ \frac{dx^1}{dx^{0'}} & \frac{dx^1}{dx^{1'}} & \frac{dx^1}{dx^{2'}} & \frac{dx^1}{dx^{3'}} \\ \frac{dx^2}{dx^{0'}} & \frac{dx^2}{dx^{1'}} & \frac{dx^2}{dx^{2'}} & \frac{dx^2}{dx^{3'}} \\ \frac{dx^3}{dx^{0'}} & \frac{dx^3}{dx^{1'}} & \frac{dx^3}{dx^{2'}} & \frac{dx^3}{dx^{3'}} \end{pmatrix} = \det\Lambda$$

where Λ is the usual transformation matrix,

$$\Lambda^\alpha{}_{\beta'} = \frac{\partial x^\alpha}{\partial x^{\beta'}}$$

But we know that g and η are related because the interval is invariant. This implies that

$$\mathbf{g} = \Lambda\eta\Lambda^{\mathrm{T}}$$

where Λ^{T} is the transpose of Λ, and hence that

$$\begin{aligned} \det\mathbf{g} &= \det\Lambda\ \det\eta\ \det\Lambda \\ &= -(\det\Lambda)^2 \end{aligned}$$

since $\det\Lambda^{\mathrm{T}} = \det\Lambda$, and $\det\eta = -1$. Therefore, if we write $g \equiv \det\mathbf{g}$, where $\mathbf{g}$ is the matrix of the metric coefficients,

$$dx^0 dx^1 dx^2 dx^3 = (-g)^{1/2}\, dx^{0'} dx^{1'} dx^{2'} dx^{3'}$$

That is, in any arbitrary coordinates $\{x^{\alpha'}\}$, the *proper volume element* (that is, the true volume of a space-time 4-volume element, as seen in the tangent flat spacetime at a particular instant) is then

$$(-g)^{1/2} dx^{0'} dx^{1'} dx^{2'} dx^{3'}$$

For example, consider the flat-space metric in polar coordinates, which can be specified by the line element

$$ds^2 = -dt^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \quad .$$

Clearly, the quantity $g = -r^4 \sin^2\theta$, and hence the spacetime volume element is

$$d^4\Omega = r^2 \sin\theta \, dt \, dr \, d\theta \, d\phi \quad .$$

At fixed time, the corresponding three-volume element is the amount which you get by factoring out the thickness of the four-volume in the time direction (dt), and is

$$d^3\Sigma = r^2 \sin\theta \, dr \, d\theta \, d\phi$$

which should be recognisable!

Now let's apply this result to calculate the proper volume of the Universe out to some redshift, z. This volume is

$$V_{\rm pr} = \int_{z=0}^{z} (-g)^{1/2} \, dr \, d\theta \, d\phi$$

For the FRW metric,

$$(-g)^{1/2} = [a(t)]^3 \, \left(1 - kr^2\right)^{-\frac{1}{2}} \, r^2 \, \sin\theta$$

and so the proper volume at the present time is

$$V_{\rm pr,0} = a_0^3 \int_{z=0}^{z} \frac{r^2 \, dr}{\left(1 - kr^2\right)^{1/2}} \, \sin\theta \, d\theta \, d\phi$$

where the integrals are over all θ and ϕ, and over r from the coordinate at which we are located ($r = 0$) to the coordinate corresponding to redshift z. We can do the angular integrals rather simply, to get

$$V_{\rm pr,0} = 4\pi \, a_0^3 \int_{r=0}^{r_1} \frac{r^2 \, dr}{\left(1 - kr^2\right)^{1/2}}$$

and, once again, we have to calculate the relationship between r_1 and z. We've done this before, in the derivation of luminosity distance. The result was (if $\Lambda = 0$)

$$a_0 r_1 = \frac{z q_0 + (q_0 - 1)\left((1 + 2q_0 z)^{1/2} - 1\right)}{H_0 q_0^2 (1 + z)} \quad .$$

We must also substitute for k, using the standard result

$$\frac{k}{a_0^2} = H_0^2 \, (2q_0 - 1)$$

to get the master integration to be performed. The integral isn't too tough: nor is the result terribly pretty for most choices of q_0, so I won't write the answer down. The main character of the result can be seen from the integral and the equation for r_1 — for any

value of q_0, the volume is finite, and remains finite as $z \to \infty$. This occurs because the history of the Universe is finite. If $q_0 > \frac{1}{2}$ we can go further, and say that the proper volume of the Universe remains finite at *all times*, since the recollapse phase limits the amount of the Universe that is ever in causal contact with us. In that condition, note that although the Universe is finite in volume it is unbounded — there are no edges.

We use the dependence of the volume of the Universe on q_0 as a method of measuring q_0, by counting the number of objects that are visible. Since we know how many objects per unit volume there are near us, and how bright they are, we can use the variation of the number of objects of particular brightness with that brightness as a test for how the volume of the Universe changes with z (and hence r). Of course, the calculation is complicated by the need to allow for the differing brightnesses of (say) galaxies, and so we see different fractions of the total galaxy population at different redshifts, and we can't really get at the *proper* volume, but only the volumes on back light cones, but the principle is correct: by measuring the number counts we get to measure the rate of change of the volume of the Universe with r, and hence the value of q_0.

Unfortunately, the properties of most objects in the Universe vary so much with z (i.e., time) that it's almost impossible to do this. However, it ought to be possible using supernovae (SN Ia appear to be very homogeneous objects, but are they really occurring at the same rate now as in the past?) or possibly gravitational lensing. So far no believable and consistent results have emerged.

One further point worth mentioning — the volume of the Universe is particularly strongly affected by Λ, the cosmological constant. So this type of test — counting objects as a function of redshift — should be a good way of measuring Λ, if the problems with source evolution can be solved.

15. The thermal history of the Universe

15.1. Matter and radiation in the Universe

The present-day Universe contains matter, with a dust-like equation of state

$$P = 0 \qquad\qquad \rho_{\rm m} \propto a^{-3}$$

with matter density $\rho_{\rm m}$ and also a small amount of radiation. The energy density in radiation is dominated by the energy of the microwave background radiation (which is everywhere in the Universe). The equation of state of radiation is

$$P = \frac{1}{3}\rho_\gamma$$

and hence using the equation of conservation of energy,

$$\frac{d}{dt}\left(\rho_\gamma a^3\right) = -P\,\frac{d}{dt}\left(a^3\right)$$

we get

$$\rho_\gamma \propto a^{-4} \quad .$$

We can understand this result for radiation very simply. In the Universe at present, photons in the background radiation field interact only very weakly with the matter (the matter is too cold, and there's rather few particles per unit volume). Therefore the number of photons in the background radiation is very nearly conserved, and so the number of photons per unit volume

$$n_\gamma \propto a^{-3} \quad .$$

But in addition, as the Universe expands the energy content of any individual photon is reduced because the photons are redshifted. So

$$\epsilon_\gamma \propto a^{-1}$$

and the energy per unit volume is therefore decreasing as

$$\rho_\gamma \propto a^{-4} \quad .$$

Not coincidentally, this matches the drop-off in surface brightness with redshift.

I've asserted that radiation is dynamically unimportant in the Universe at the present time. Let's investigate further. At present the microwave background radiation, the dominant radiation field in the Universe, is of almost exactly Planckian spectrum and has

$$T_{\rm rad} = 2.728 \pm 0.002 \quad {\rm K} \quad .$$

The energy density in the radiation field is then

$$u_\gamma = a_{\rm R} T_{\rm rad}^4 = 4.2 \times 10^{-14} \quad {\rm J\,m^{-3}}$$

where $a_{\rm R}$ is the radiation constant (7.6×10^{-16} J m^{-3} K^{-4}). The equivalent density in radiation now is therefore

$$\rho_{\gamma\,0} = 4.68 \times 10^{-31} \quad {\rm kg\,m^{-3}} \quad .$$

By contrast, the matter content of the Universe has

$$\begin{aligned} \rho_{\rm m\,0} &= \Omega_0\, \rho_{\rm crit} \\ &= \Omega_0 \left(\frac{3H_0^2}{8\pi G} \right) \\ &= \left(1.88 \times 10^{-26} \right)\, h_{100}^2\, \Omega_0 \quad {\rm kg\,m^{-3}} \end{aligned}$$

so that the present-day ratio of the density in radiation to the density in matter is

$$\frac{\rho_{\gamma\,0}}{\rho_{\rm m\,0}} = (2.5 \times 10^{-5})\, \Omega_0^{-1}\, h_{100}^{-2}$$

which is substantially less than 1. The Universe is, indeed, very matter-dominated at present.

15.2. Equipartition

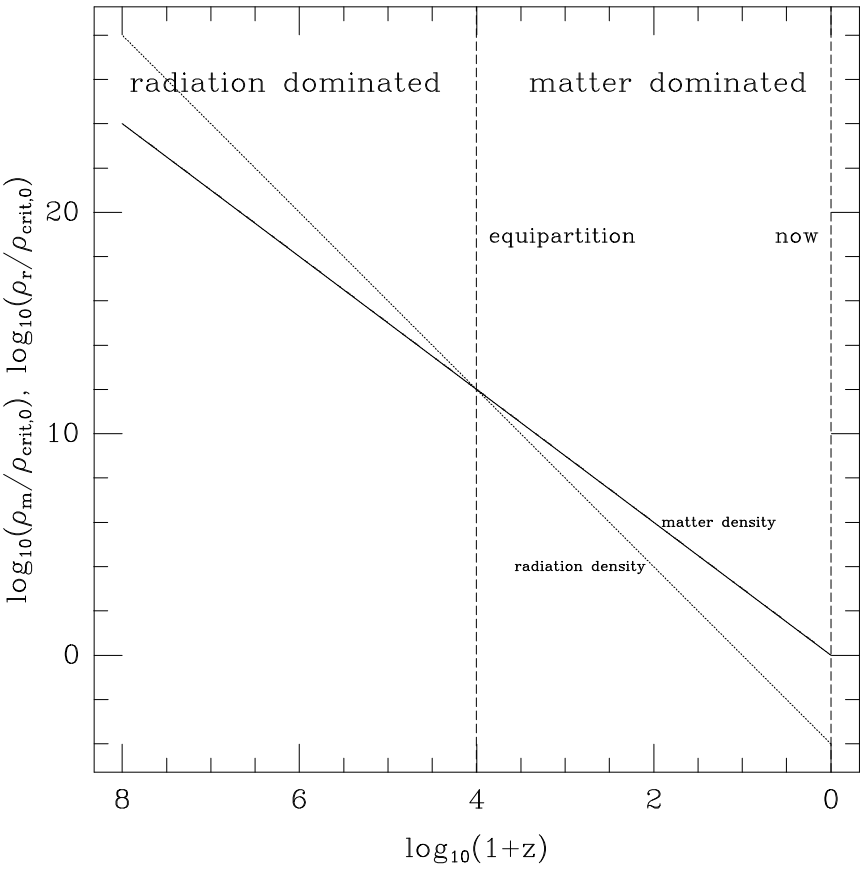
But this was not always the case — we've shown that the matter and radiation densities vary differently with scale factor. At earlier times,

$$\frac{\rho_\gamma}{\rho_{\rm m}} \propto \frac{a_0}{a} = (1+z)$$

so that matter and radiation were of equal density at

$$1 + z_{\rm eq} = \left(4 \times 10^4 \right) \Omega_0\, h_{100}^2$$

which is called the **epoch of equipartition**. At earlier times the Universe was radiation-dominated; at later times the Universe was matter-dominated, as it is at present. This density history can be seen in the plot below, which was calculated for $\Omega_0 = 1$ and $h_{100} = 0.5$, which makes $z_{\rm eq} \approx 10^4$.



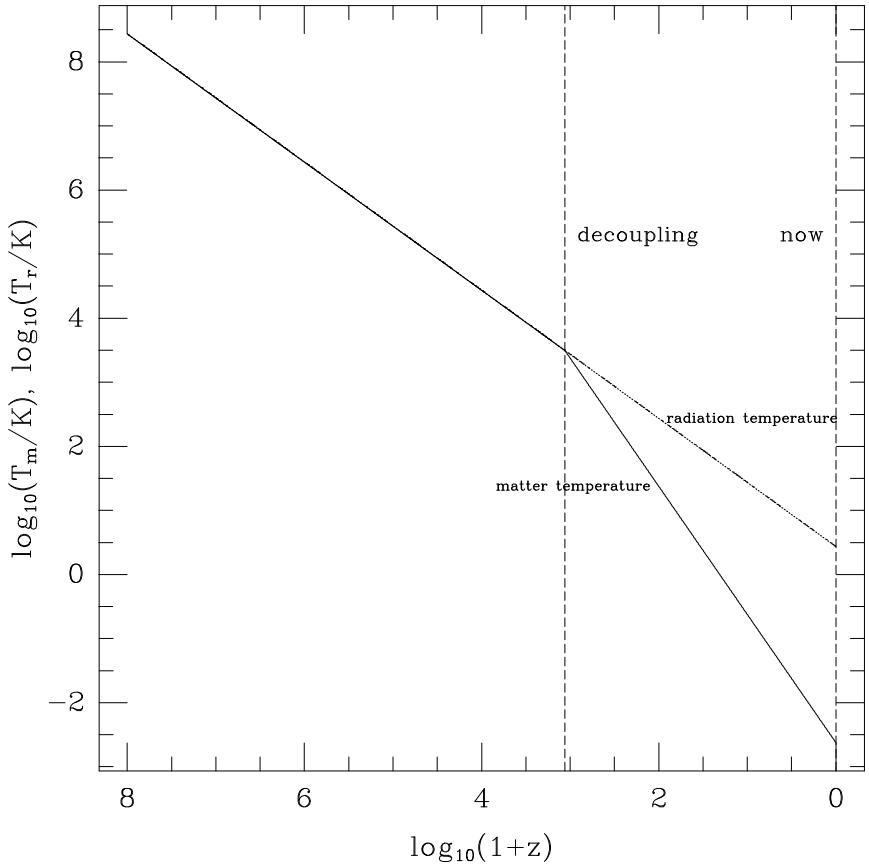
15.3. Decoupling

We might also ask about the temperature of the Universe as a function of time. Since the Universe is homogeneous, there is nowhere for energy flows to go, and hence we expect adiabatic changes in the temperature of matter and radiation that are driven only by the change in (proper) volume of the Universe. Therefore

$$T_{\rm m} \propto \rho_{\rm m}^{\frac{2}{3}} \propto \left(\frac{a_0}{a}\right)^2 \propto (1+z)^2$$
$$T_\gamma \propto \rho_\gamma^{\frac{1}{4}} \propto \frac{a_0}{a} \propto (1+z)$$

so that matter is cooling much faster than radiation at present (and I assumed that the matter had a polytropic index $\gamma = \frac{5}{3}$). As we look back to earlier redshifts, the temperatures of matter and radiation are higher. Matter and radiation interact very little at the present since matter is of low density. In the past the density of matter increases, so the possibility of interaction increases. In addition, the temperature of the matter increases. So there is a particular time in the past, the time of **decoupling**, before which matter was dense enough to interact strongly with radiation (strongly

enough that the interaction time is much less than the expansion time of the Universe). At earlier times, matter and radiation must have had the same temperature. Hence a plot of matter and radiation temperatures against redshift would look like (for $\Omega_0 = 1$, $h_{100} = 0.5$, $\Omega_{\rm B} = 0.1$)



At times before decoupling, the matter and radiation were at the the same temperature, since they strongly interact. As the Universe expands the rate at which radiation is scattered by matter decreases (principally because the density of the scatterers decreases), but the interaction remains fast until the density and temperature drop enough that the scattering length becomes long.

Ignored here is the effect of quasars and young stars, which reheat the intergalactic medium (IGM) at some redshift $z \approx 10$ because of their enormous output of UV radiation. Very quickly at some redshift, which is known to be greater than the redshift of most quasars from the absence of the Gunn-Peterson effect in their spectra, the temperature of the IGM jumps to something like 10^6 K. And this means that most matter in the Universe today is not at the ≈ 10 mK that the calculation above would predict, but rather at X-ray emitting temperatures.

15.4. Recombination and reionization

There is a third important redshift associated with the interaction of matter and radiation — the redshift of **recombination**. Recombination is the time at which the Universe becomes neutral. This is going to be close to the time at which matter and radiation decouple, since neutrality enforces decoupling. The precise relationship between $z_{\rm dec}$ and z_{rec}, however, depends on the rate of expansion of the Universe, and how much baryonic matter the Universe contains.

We can calculate the redshift of reionization using the *Saha equation*, which says that in equilibrium, and at temperature T, the reaction

$$p + e^- \leftrightarrow {}^1{\rm H}$$

produces particle densities

$$\boxed{\frac{n_p n_e}{n_H} = \left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT} \quad .}$$

This equation can be derived from a consideration of the chemical potentials of the species in the recombination equilibrium equation, and involves the ionization potential for hydrogen ($\chi = 13.6$ eV).

Define $z_{\rm rec}$ as the redshift of recombination, and let this be the redshift at which $n_e \approx n_p \approx n_H$, with n_H the number density of neutral hydrogen. The Universe is mostly hydrogen (in other words, I'll ignore the helium content for the moment), so the baryon density at $z_{\rm rec}$ is

$$n_B \approx n_H + n_p = n_{B0}(1 + z_{\rm rec})^3$$

where n_{B0} is the present-day baryon number density, which is (for a pure hydrogen Universe)

$$n_{B0} = \frac{\Omega_B\, \rho_{\rm crit,0}}{m_H} = \Omega_B\, \frac{3\, H_0^2}{8\,\pi\, G\, m_H} = 11.2\, \Omega_B\, h_{100}^2 \ \ {\rm m}^{-3}$$

where Ω_B is the fraction of the critical density which is contained in baryons. For $\Omega_B = 0.1$ and $h_{100} = 0.5$, the current baryon density is 0.28 m^{-3}. We also know that the temperature at recombination is

$$T(z_{\rm rec}) = T_{\rm rad}(1 + z_{\rm rec})$$

since the temperature of radiation in the Universe is simply related to the temperature of the microwave background radiation today. Substituting these into the Saha equation produces

$$\frac{3}{16\pi}\frac{\Omega_B\, H_0^2}{G\, m_H}\,(1 + z_{\rm rec})^3 = \left(\frac{2\,\pi\, m_e\, k_B\, T_{\rm rad}\,(1 + z_{\rm rec})}{h^2}\right)^{3/2} \exp\left(-\frac{\chi}{k_B\, T_{\rm rad}\,(1 + z_{\rm rec})}\right) \quad ,$$

which can conveniently be written

$$(1+z_{\rm rec})^{3/2}\,\exp\left(\frac{A}{1+z_{\rm rec}}\right)=B$$

with

$$A=5.78\times10^4$$
$$B=3.09\times10^{24}\Omega_B\,h_{100}^2$$

Which we can solve to find

$$z_{\rm rec}=1380$$

so that the temperature of the Universe at recombination was about 3770 K. At this time, the timescale of interactions between photons and electrons was

$$t_{\gamma\,e}\approx\frac{1}{n_e\sigma_T c}$$

where $\sigma_T=6.7\times10^{-29}$ m^2 is the Thomson scattering cross-section and $n_e\approx3\times10^8$ m^{-3} (from the value of Ω_B and $z_{\rm rec}$). This gives an interaction time of about 5000 years, which is only about 2 per cent of the age of the Universe at $z_{\rm rec}$ (2.5×10^5 years: all numerical values for $\Omega_B=0.1$, $h_{100}=0.5$, and $\Omega_0=1$). This justifies the idea that the Universe is in thermal equilibrium at the epoch of recombination — photon/electron interactions are so fast that matter and radiation are in excellent thermal contact.

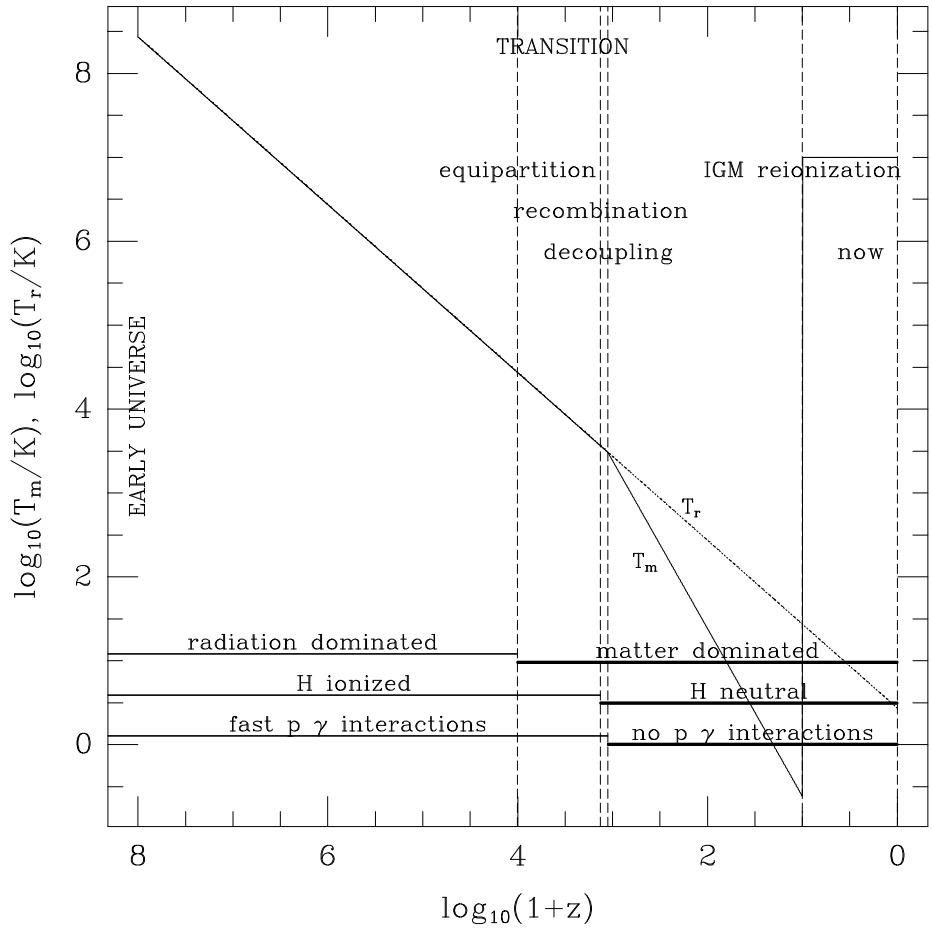
We can also use the Saha equation to see that at z only slightly less than $z_{\rm rec}$ the fractional ionization in the Universe is much lower — because of the exponential factor. Therefore recombination happens quickly, and is swiftly followed by decoupling. Recombination and decoupling define the middle and the end of the phase of the Universe in which *the last scattering of the microwave background radiation* occurred. So we can see that this radiation field is telling us directly about the structure of the Universe at $z\approx1000$.

Finally, there is another interesting thermal phenomenon in the Universe — at the epoch of reionization. We know that from recombination onwards the Universe was mostly neutral. But we also know (from observations of quasar spectra) there there isn't any remaining neutral material in the intergalactic medium (there are no *Gunn-Peterson effect troughs*, where neutral gas has mopped up UV photons). Therefore there must have been a time, the time of re-ionization, at which the Universe became ionized again. The redshift at which this happened is, presumably, the redshift at which the first generation of stars in the Universe formed and started emitting light. This first generation of stars is likely to have been pretty massive, since these stars formed from material with little metal content (see discussions of star formation in the literature, or in earlier courses), and therefore, like the most massive stars today, would have had high temperature, high luminosity, and hence very high ionizing radiation output.

It is calculated that reionization would have been a fast phenomenon. And it must have occurred before most quasars formed, so that quasars see only ionized gas.

Since we see galaxies and quasars to $z \approx 6$, we can assume that $z_{\rm ri}$, the reionization redshift, is 10 or so. And at this time the temperature of the intergalactic medium rises, probably to 30 keV or more — a temperature high enough that this material cools only very slowly. Material which has already condensed out of the IGM, into galaxies, molecular clouds, and so on has a very different history. But that is now a question of the detailed gravitational and fluid-dynamical processes of structure formation, rather than the large-scale relativistic cosmology which is the topic of the present lecture course, and I'll leave it for graduate studies.

Putting all this together, the recent thermal history of diffuse matter in the Universe can be represented by the sketch below (which omits the peculiarities in the temperature of matter after $z_{\rm ri}$ produced by stars, quasars, and other structures).



One final thought. Dark matter is presumably only weakly coupled to matter at present, and even if it has a "dust" equation of state, its current temperature will depend on the redshift at which **it** decoupled from the radiation field. If the dark matter particles are cold, and weakly interact with normal matter and radiation, we would expect that the dark matter decoupling epoch would be much less than $z_{\rm dec}$ for

normal matter, and its temperature now would be low. If, on the other hand, the dark matter has a late decoupling (or may not yet have decoupled — perhaps because the dark particles have very low rest masses, like neutrinos) then the temperature may be close to the temperature of the microwave background radiation. See the discussion of neutrino temperatures in the next lecture.

16. The radiation-dominated era

16.1. The dynamics of the Universe before equipartition

Before equipartition, the Universe was dominated (in energy/density terms) by its radiation content. The Universe was fully-ionized ($z_{\rm eq} > z_{\rm rec}$), matter and radiation were tightly thermally and dynamically coupled ($z_{\rm eq} > z_{\rm dec}$), and the energy density of radiation exceeds the energy density of matter.

Under these circumstances, it's clear that we can no longer use as equation of state for the Universe the *dust* equation of state that we used at $z < z_{\rm eq}$ — during the discussion of the properties of the Universe after the first few hundred thousand years. Instead, a good approximation would be that the density and pressure of the Universe are related by a *relativistic* equation of state, with

$$P = \frac{1}{3}\rho \quad .$$

Note that this doesn't mean that the electrons, protons, helium nuclei (and the few less ionized species) are relativistic — the temperature at $z_{\rm eq}$ is only 10^4 K or so, so that even electrons are non-relativistic), but that the dominant density component of the Universe (the radiation density) has these properties.

With $P = \frac{1}{3}\rho$, the Friedmann equation and the equation of state (energy equation) are

$$\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = \frac{8\pi}{3}\rho + \frac{\Lambda}{3}$$
$$\frac{d}{dt}\left(\rho a^3\right) = -P\,\frac{d}{dt}\left(a^3\right)$$

and the second of these equations reduces to

$$\rho a^4 = \text{constant}.$$

As already discussed, this corresponds to the radiation energy density being proportional to a factor a^{-3} (like the matter density) because of the cosmic expansion, with an additional factor a^{-1} because of the redshifting of the energy of each photon. If we substitute

$$\rho = \rho_{\rm eq}\left(\frac{a_{\rm eq}}{a}\right)^4$$

then the Friedmann equation is

$$\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = \frac{8\pi}{3}\rho_{\rm eq}\left(\frac{a_{\rm eq}}{a}\right)^4 + \frac{\Lambda}{3} \quad .$$

This can be simplified. We know that the Λ term is small, or at most comparable with the present-day value of the density term (from the geometry of the Universe today).

Hence back at $z > 10^3$ the Λ term must have been insignificant, and it can safely be dropped. Similarly, we know that the curvature term (in k) becomes smaller relative to the other terms as we go back in time (this is simply the argument about the "flatness problem" again), and at $z > 10^3$, again it must be small relative to the other terms. Hence in the radiation-dominated phase of the Universe, we can write

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi}{3}\rho_{\rm eq}\left(\frac{a_{\rm eq}}{a}\right)^4$$

to high accuracy. This means that we need to deal *only* with the radiation-dominated Einstein – de Sitter Universe: the closed and open Universe solutions are extremely close in properties to this one.

Solving this equation is simple: the result is

$$a = a_{\rm eq}\left(\frac{32\,\pi\,\rho_{\rm eq}\,t^2}{3}\right)^{1/4}$$

so that $a \propto t^{1/2}$, and $\rho \propto t^{-2}$. That is, the Universe expands more slowly than in the matter-dominated phase (where the scale factor varied as $a \propto t^{2/3}$), but the radiation density of the Universe changes as t^{-2} just as the matter density does after equipartition in an Einstein – de Sitter Universe. In fact we can write the density/time relation more precisely as

$$t = \left(\frac{32\,\pi}{3}\,G\,\rho\right)^{-\frac{1}{2}}$$

(re-inserting the G factor). Note that the assumption that we are dealing with an Einstein – de Sitter Universe factors out the $\rho_{\rm eq}$ factor: all cosmologies that are a reasonable match to the present-day Universe have *very similar* densities at early times ($z > z_{\rm eq}$). Since the density in radiation is related to the temperature by

$$\rho = \frac{a_R\,T^4}{c^2}$$

where $a_R = 7.6 \times 10^{-16}$ J m^{-3} K^{-4} is the radiation constant, I can relate the age of the Universe, t, to the temperature of the Universe, T, by

$$T = \left(\frac{32\,\pi\,G\,a_R}{3c^2}\right)^{-\frac{1}{4}}\,t^{-\frac{1}{2}}$$

which can be written conveniently as

$$(T/{\rm K}) = 1.5 \times 10^{10}\,(t/{\rm sec})^{-\frac{1}{2}} \quad \text{or}$$
$$(k_B T/{\rm MeV}) = 1.3\,(t/{\rm sec})^{-\frac{1}{2}}$$

so that at 1 second after the Big Bang the temperature was about 1.5×10^{10} K (1.3 MeV), and we might expect nuclear reactions to be occurring, while by 1 hour after the Big

Bang, the temperature has dropped to 2.5×10^8 K (20 keV), and the entire Universe resembles the ordinary X-ray emitting gas that is found in clusters of galaxies today.

That is, all the truly relativistic behaviour of the Universe is concentrated in the first few minutes of its life. After approximately an hour, there is no physics other than the ordinary laboratory and astrophysical processes that we observe elsewhere. That is, the Universe is relatively boring from 1 hour after the Big Bang until recombination, and we ought to be able to understand everything that's going on.

Let's check this by putting in values for the density. At one second after the Big Bang, the density of matter in the Universe is about 4×10^8 kg m^{-3}, rather higher than any density that we deal with in laboratory experiments, though substantially less than nuclear density so that the physics of such matter should not be unusual. At one hour, the density has dropped by a factor 1.3×10^7, to only 30 kg m^{-3}, rather less than the density of water.

16.2. e^+e^- annihilation and neutrinos

In fact the result for the density and temperature of the early Universe as a function of time needs some modification, because the details of what particles were present and *relativistic* makes a substantial difference to the early dynamics of the Universe, at the point before $k_BT \approx 2m_ec^2$, which is when and electrons and positrons annihilated.

What were the major constituents of the early Universe? Take as a reference time $t = 0.01$ sec after the Big Bang, when matter was relatively normal and we can be fairly confident about what was going on. Then the basic constituents of the Universe were

baryons	$m_B = 931$ MeV
electrons	$m_e = 511$ keV
photons	$m_\gamma = 0$
neutrinos	$m_\nu = 0$ (assumed)
dark matter	$M_D > 1$ Gev (assumed)

and we expect that both the normal particles and the antiparticles might be present. However at this time, $k_BT \approx 13$ MeV, much less than the rest-mass of a baryon. Thus there is insufficient energy density to create baryon/anti-baryon pairs (and, presumably, any dark matter particles), and so we expect that *almost all anti-baryons will have annihilated with baryons.*

What are the contributions of each of these species to the energy and entropy content of the Universe? Immediately we know that we can ignore the rest-mass equivalent energy content of baryons (since we're well before equipartition), so we need only consider photons, electrons (which can be pair-produced), and neutrinos. The

calculations for the energy densities in photons and neutrinos are similar:

$$u_\gamma = \int_0^\infty \frac{4\pi}{c}\frac{2h\nu^3}{c^2}\left(e^{h\nu/k_BT}-1\right)^{-1}\,d\nu = a_R T^4$$
$$u_\nu = 3\times\int_0^\infty \frac{4\pi}{c}\frac{2h\nu^3}{c^2}\left(e^{h\nu/k_BT}+1\right)^{-1}\,d\nu = \frac{21}{8}a_R T^4$$

where the factor 3 takes account of all three neutrino flavours (the "2" in the equation for neutrinos takes account of anti-neutrinos just as the "2" in the photon expression takes account of the polarizations of photons).

For particles with mass the results are a little more complicated since the chemical potentials need to be taken into account. However, the limiting cases are simple:

$$u_e = \begin{cases} \frac{7}{4}a_R T^4 & k_BT \gg m_ec^2 \\ 0 & k_BT \ll m_ec^2 \end{cases}$$
$$u_B = \begin{cases} \frac{7}{4}a_R T^4 & k_BT \gg m_Bc^2 \\ 0 & k_BT \ll m_Bc^2 \end{cases}$$
$$u_D = \begin{cases} \frac{7}{4}a_R T^4 & k_BT \gg m_Dc^2 \\ 0 & k_BT \ll m_Dc^2 \end{cases}$$

for electrons (and positrons), baryons (and antibaryons), and dark matter particles, respectively.

I can also write results for the entropy per unit volume in the different species,

$$s_\gamma = \frac{4}{3}a_R T^3$$
$$s_\nu = \frac{7}{2}a_R T^3$$
$$s_e = \begin{cases} \frac{7}{3}a_R T^3 & k_BT \gg m_ec^2 \\ 0 & k_BT \ll m_ec^2 \end{cases}$$
$$s_B = \begin{cases} \frac{7}{3}a_R T^3 & k_BT \gg m_bc^2 \\ 0 & k_BT \ll m_bc^2 \end{cases}$$
$$s_D = \begin{cases} \frac{7}{3}a_R T^3 & k_BT \gg m_Dc^2 \\ 0 & k_BT \ll m_Dc^2 \end{cases}$$

once again ignoring the small terms for non-relativistic electrons, baryons, and dark matter particles.

Now, at $t = 0.01$ sec, $k_BT \ll m_bc^2$ for all baryons, for dark matter, and for all leptons except electrons and positrons. Hence the energy budget of the Universe is dominated by the relativistic species (we are well before equipartition, so that the rest-mass energy of dark matter and baryons in the Universe is insignificant), and the

total density of the Universe is

$$\begin{aligned}\rho(T) &= \frac{u_e(T) + u_\nu(T) + u_\gamma(T)}{c^2} \\ &= \left(\frac{7}{4} + \frac{21}{8} + 1\right) \frac{a_R T^4}{c^2} \\ &= \frac{43}{8} a_R T^4 \\ &= g^*_{\rm rel}\, a_R\, T^4\end{aligned}$$

where $g^*_{\rm rel}$ is the statistical weight of all relativistic species in the Universe. We can insert this result into the relationship between time and temperature since the Big Bang

$$t = \left(\frac{32\pi G\rho}{3}\right)^{-\frac{1}{2}}$$

to get an improvement over our previous expression

$$(t/{\rm sec}) = 1.0 \times 10^{20} (T/{\rm K})^{-2}$$

(easier to remember than the other result!). Note the importance of $g^*_{\rm rel}$ here — the dynamics of the early Universe depends on how many species of relativistic particles are present, and the statistics of these species. To obtain the full solution for the evolution of the Universe from $t = 0$ to the present, we would need to know $g^*_{\rm rel}$ at all times, including each species of particle that makes a significant contribution to $\rho(T)$.

Now, the expansion of the Universe is adiabatic (since there are no anisotropies that can drive energy flows), so that the expansion has

$$sa^3 = {\rm constant}$$

and the temperatures of the various constituents of the Universe either change together (if they are tightly coupled by scattering), or independently (if they are weakly coupled). At $t = 0.01$ sec, neutrinos and electrons as well as electrons and photons are strongly coupled — they scatter rapidly compared to the expansion time of the Universe. Therefore the electrons, neutrinos, and photons have the same temperature and I can write (as I did, above)

$$T_e = T_\nu = T_\gamma \equiv T$$

since there is a unique "temperature of the Universe". As the temperature drops, however, the neutrino-electron interactions become progressively slower relative to the expansion of the Universe. That is, there is a time of **neutrino decoupling**, just like the time of electron (and hence matter) decoupling in the later stages of the expansion of the Universe. After decoupling the neutrinos will have a temperature which varies according to

$$\begin{aligned} & s_\nu a^3 = {\rm constant} \\ {\rm i.e.} \qquad & T_\nu \propto a^{-1} \propto (1+z)\ . \end{aligned}$$

For the moment, however, it remains true that $T_e = T_\nu = T_\gamma$, since the electrons, neutrinos, and photons are all relativistic particles (with $k_B T$ greater than their rest masses), so that each independently obeys $s \propto a^{-3}$ and so each has $T \propto a^{-1} \propto (1+z)$.

A major change occurs when T_e ($= T_\gamma = T_\nu$) drops to about 10^{10} K, where $k_B T \approx 2 m_e c^2$, which happens at $t \approx 1$ sec. Now the electrons and positrons annihilate — i.e., the equilibrium

$$e^+ + e^- \rightleftharpoons 2\gamma$$

is driven strongly to the right and favours photon production. The energy released in this annihilation heats the photons *but not the neutrinos* since the neutrinos are no longer strongly interacting with the photons or electrons. The temperature change can be calculated by conservation of entropy

$$s_{e\gamma} a^3|_{\text{before } e^+e^- \text{ annihilation}} = s_{e\gamma} a^3|_{\text{after } e^+e^- \text{ annihilation}}$$
$$s_{\nu} a^3|_{\text{before } e^+e^- \text{ annihilation}} = s_{\nu} a^3|_{\text{after } e^+e^- \text{ annihilation}}$$

Thus if a_1 is the scale factor before annihilation and a_2 is the scale factor after annihilation, $T_{e\gamma 1}$ is the temperature of the electrons (and photons) before annihilation, $T_{\gamma 2}$ is the temperature of the photons after annihilation, and $T_{\nu 1}$ and $T_{\nu 2}$ are the corresponding temperatures of the neutrinos,

$$a_1^3 \left(\frac{4}{3} a_R T_{e\gamma 1}^3 + \frac{7}{3} T_{e\gamma 1}^3 \right) = a_2^3 \left(\frac{4}{3} a_R T_{\gamma 2}^3 \right)$$
$$a_1^3 \left(\frac{7}{2} a_R T_{\nu 1}^3 \right) = a_2^3 \left(\frac{7}{2} a_R T_{\nu 2}^3 \right)$$

Recognising that $T_{e\gamma 1} = T_{\nu 1}$, and dividing,

$$\frac{11}{3} = \frac{4}{3} \left(\frac{T_{\gamma 2}}{T_{\nu 2}} \right)^3$$

so that

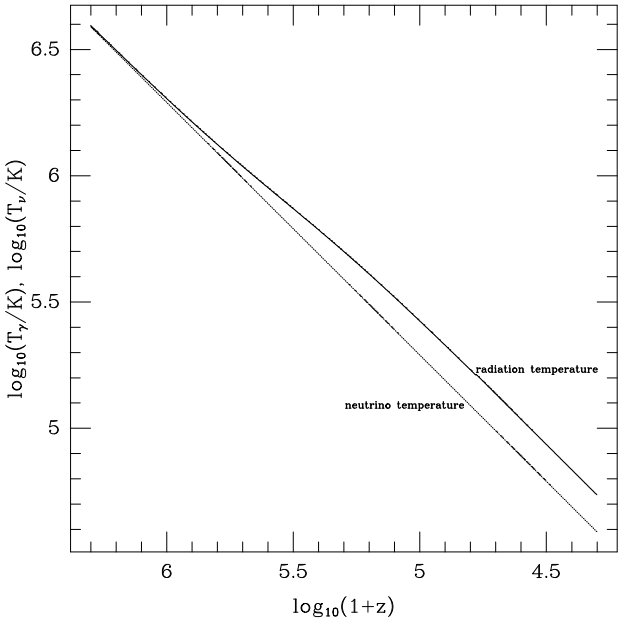
$$T_{\gamma 2} = \left(\frac{11}{4} \right)^{\frac{1}{3}} T_{\nu 2}$$

and hence the annihilation of electrons and positrons raises the temperature of the radiation field to 1.40 times the temperature of the neutrinos. Because the photons and neutrinos remain relativistic from this time to the present (assuming that neutrinos are really massless), their temperatures independently evolve as $(1+z)$, and hence today

$$T_{\nu 0} = \frac{T_{\gamma 0}}{1.40}$$

and the Universe must contain a neutrino background with a temperature of 1.95 K. This neutrino background has essentially no dynamical consequences (its energy density is very low), and appears to be undetectable.

The earliest phases of the thermal history of the Universe therefore are a bit more complicated than the sketch I drew in the last lecture: I must now add the neutrinos as a separate species, and so the temperature history between $t = 0.01$ and 100 sec ($z = 2 \times 10^6$ and 2×10^4) should appear as



17. Nucleosynthesis and baryogenesis

17.1. Nucleosynthesis

An important change in the make-up of the Universe occurs and somewhat after the time of electron/positron annihilation, which marks the end of what's sometimes called the "lepton era", when nucleosynthesis converts some of the elementary particles into composite objects. In many ways the ruling physics here is similar to the physics of recombination (or physics at the end of the quark era, but that we understand rather little about), except that it's not the interaction of electrons and protons to form hydrogen that's important, but the combination of nucleons to produce different types of nuclei. And rather than happening at times a few hundred thousand years after the Big Bang, when the temperature is a few thousand K, nucleosynthesis occurs in the first few minutes of the Big Bang, when the temperature is still 10^9 K or more ($k_B T \gtrsim 100$ keV).

Consider the recombination-like process when neutrons and protons react to become deuterium (and, later, helium). At high enough temperatures ($T \gg 3 \times 10^9$ K), neutrons and protons are kept in some constant ratio by the equilibrium

$$\begin{aligned} n + \nu_e &\rightleftharpoons p + e^- \\ n + e^+ &\rightleftharpoons p + \overline{\nu}_e \\ n &\rightleftharpoons p + e^- + \overline{\nu}_e \quad . \end{aligned}$$

In equilibrium, the chemical potentials are related by

$$\begin{aligned} \mu_n + \mu_{\nu_e} &= \mu_p + \mu_{e^-} \\ \mu_n + \mu_{e^+} &= \mu_p + \mu_{\overline{\nu}_e} \\ \mu_n &= \mu_p + \mu_{e^-} + \mu_{\overline{\nu}_e} \quad . \end{aligned}$$

These three equations are effectively the same, since $\mu_{e^+} = -\mu_{e^-}$ and $\mu_{\nu_e} = -\mu_{\overline{\nu}_e}$. But $\mu \propto \ln(m)$; so $\mu_\nu \ll \mu_e \ll \mu_n$ and

$$\boxed{\mu_n \simeq \mu_p \quad ,}$$

Statistical-physics expressions for the chemical potential relate it to the de Broglie wavelength, λ (which sets the phase space volume element), and the partition function, Z, as

$$\mu = kT \ln\left(\frac{n\lambda^3}{Z}\right)$$

where n is the number of particles per unit volume. Hence the equality of chemical potentials implies a ratio of particle densities of neutrons and protons given by

$$\frac{n_n}{n_p} = \left(\frac{\lambda_p}{\lambda_n}\right)^3 . \frac{Z_n}{Z_p} \quad .$$

But $\lambda_p \sim \lambda_n$, $Z_n = g_n e^{-Q_n/kT}$ and $Z_p = g_p e^{-Q_p/kT}$. The degeneracies $g_n = g_p = 2$ (since both particles have $s = \frac{1}{2}$). Q_n and Q_p are the binding energies (mass deficits) of n and p. Therefore,

$$\frac{n_n}{n_p} = e^{-(Q_n - Q_p)/kT} \quad ,$$

or

$$\boxed{\frac{n_n}{n_p} = e^{-Q/kT} \quad .}$$

where Q is the mass excess of neutrons relative to protons, which is equal to 1.293 MeV.

$Q = kT$ at $T = 1.5 \times 10^{10}$ K. Thus, as long as the neutrons are in good thermal contact with neutrinos (at $T \gg 10^{10}$ K, while the interaction time of neutrons with neutrinos is less than the expansion time of the Universe), the fraction of bosons as neutrons is

$$\frac{n_n}{n_p + n_n} = X_n = \frac{1}{1 + e^{Q/kT}} \quad .$$

In particular,

$$X_n = 0.38$$

at $T = 3 \times 10^{10}$ K.

Now, at $T \lesssim 3 \times 10^{10}$ K, good thermal contact is lost and the reactions tend to fall out of equilibrium, establishing a slightly lower X_n whose value must be obtained from a detailed study of the $n/e/p/\nu_e$ reaction processes. The X_n established is higher than the value of zero predicted by thermal equilibrium, because the Universe is expanding too fast to allow the neutrons to be converted to protons (the expansion time is less than the interaction time between neutrons and positrons or neutrinos). Perhaps an easier way to say this is that the neutron/proton equilibrium can't adjust as fast as is required by the rapidly-falling temperature of the Universe near this time. As a result, the neutron fraction "freezes out" at $X_n \sim 0.16$ (the value appropriate for $T \sim 9 \times 10^9$ K in equilibrium).

But neutrons tend to decay, with a decay time ~ 1000 sec, so X_n is expected to decrease on this timescale.

Helium formation competes with neutron decay to remove free neutrons. The slowest step of ^{4}He production is the formation of deuterium. This is difficult, since ^{2}H

is weakly bound compared to ^{4}He, so although ^{4}He is energetically favored over $n + p$ at $T \lesssim 10^9$ K, the necessary preliminary step of deuterium production limits its formation, and can only start at a lower temperature. We calculate when ^{2}H becomes abundant by looking at the equilibrium of

$$p + n \rightleftharpoons {}^2\mathrm{H} + \gamma$$

(a weak interaction). In equilibrium,

$$\mu_p + \mu_n = \mu_d$$

where I'm using d for the deuteron. Using the expression for the chemical potential in terms of de Broglie wavelength, density, and partition function,

$$kT \ln\left(\frac{n_p\lambda_p^3}{Z_p}\right) + kT \ln\left(\frac{n_n\lambda_n^3}{Z_n}\right) = kT \ln\left(\frac{n_d\lambda_d^3}{Z_d}\right)$$

and so

$$\frac{n_p n_n}{n_d} = \frac{\lambda_d^3}{\lambda_p^3\lambda_n^3} \cdot \frac{Z_p Z_n}{Z_d}$$

$m_d \sim 2m_n$, $m_n \sim m_p$, and so we arrive at

$$\frac{n_p n_n}{n_d} = \left(\frac{\pi m_p kT}{h^2}\right)^{3/2} e^{-I/kT}$$

where $I = 2.2$ MeV is the binding energy of ^{2}H. Solving, we find that $\frac{n_p n_n}{n_d} \sim n_B$, the number of baryons, at $z = 2.9 \times 10^8$, $T = 7.8 \times 10^8$ K, and $n_B = 10^{13}$ m^{-3}. So ^{2}H production (and hence ^{4}He production) only becomes fast at $T \sim 8 \times 10^8$ K (about two minutes after the Big Bang), and somewhat later than ^{4}He production becomes very energetically favorable. By this time some of the neutrons have decayed to protons. At this stage, X_n has dropped from ~ 0.16 to ~ 0.13.

After the "deuterium bottleneck," ^{4}He forms by

$$\mathrm{d} + \mathrm{d} \rightarrow {}^4\mathrm{He} + \gamma$$

$$\underline{\text{or}} \quad \begin{cases} \mathrm{p} + \mathrm{d} \rightarrow {}^3\mathrm{He} + \gamma \\ \mathrm{n} + {}^3\mathrm{He} \rightarrow {}^4\mathrm{He} + \gamma \end{cases}$$

$$\underline{\text{or}} \quad \begin{cases} \mathrm{n} + \mathrm{d} \rightarrow {}^3\mathrm{H} + \gamma \\ \mathrm{p} + {}^3\mathrm{H} \rightarrow {}^4\mathrm{He} + \gamma \end{cases} .$$

These reactions are *fast* and quickly take almost all the deuterium through to ^{4}He, eating up all the neutrons that have not decayed.

The final mass fraction of baryons in helium, Y, can be predicted from the mix of neutrons and protons before nucleosynthesis. It takes two neutrons to produce one ^{4}He nucleus. Therefore,

$$n_{^4\mathrm{He}}(\text{after } ^4\text{He synthesis}) = \frac{1}{2} n_n(\text{before } ^4\text{He synthesis}) \quad ,$$

and hence

$$\begin{aligned} Y &= \frac{4n(^4\mathrm{He})}{n_b} \\ &= \frac{2n_n(\text{before})}{n_n(\text{before}) + n_p(\text{before})} \\ &= 2X_n(\text{before}) \quad , \end{aligned}$$

where $X_n \approx 0.13$ is the fraction of baryons in neutrons surviving to nucleosynthesis. Therefore we expect the Universe to contain a primordial helium mass fraction

$$\boxed{Y \sim 0.26 \quad .}$$

This is consistent with the measured values of ^{4}He abundance, which is roughly constant in many objects (distant galaxies, old stars, HII regions, ...). The constancy of Y and the fit to the theoretical value give us confidence in the reality of a hot Big Bang.

Other elements are also created in the Big Bang: in particular, appreciable amounts of ^{3}He, ^{2}H, and ^{7}Li (that is, appreciable compared to the present observed abundances: all are rare).

We can compare the predicted amounts of, for example, ^{4}He and ^{2}H produced in the Big Bang (as functions of Ω_B, since we do not know Ω_B accurately) with those observed, and try to use this *astrophysical* method to determine Ω_B. (Note that the prediction is only a weak function of Ω_0, since we know that $\Omega \approx 1$ at nucleosynthesis and so all Universes behave similarly).

Problem: Although ^{4}He, once produced, is not destroyed, so that the present Y (corrected for the small contribution from its production in stars) leads directly to the Big-Bang-produced Y, ^{2}H *is* affected by later processes, principally by being destroyed in stars. This correction is probably large, so that the primordial abundance of deuterium is poorly known. If we make the best-guess correction for deuterium destruction, we can try to determine Ω_B.

The sense of the variation of abundance with Ω_0 is that

at higher Ω_B, most of the ^{2}H is burned to ^{4}He
at low Ω_B, little of the ^{2}H is burned to ^{4}He

The observed ^{2}H/^{1}H ($\sim 3 \times 10^{-5}$) is probably a lower limit to the primordial ^{2}H/^{1}H. Thus, ^{2}H/^{1}H really only gives an upper limit to Ω_B. However, after correcting for *astration* (the processing of material in stars), the current best-buy value for Ω_B based on nucleosynthesis (and assuming $\Omega_0 = 1$) is

$$\Omega_{B0} = (0.013 \pm 0.002) h_{100}^{-2}$$

which corresponds to a small fraction of the closure density: most matter in the Universe is presumably dark.

Note that the value of Ω_B deduced from ^{2}H is approximately consistent with the value found from other methods (such as the total virial masses of systems of galaxies), and that a detailed analysis of the abundances of *all* the primordial nucleosynthesized species yields consistent results for Ω_B and Ω_0.

17.2. Photon to baryon ratio, entropy, and matter asymmetry

The photon to baryon ratio is an important parameter of the present-day Universe. This number has been implicit for the last couple of lectures. The present-day baryon number density is

$$n_{\rm B0} = \frac{\Omega_{\rm B0} \rho_{\rm crit,0}}{m_B}$$

where m_B is the mass of a baryon, $m_p \approx m_n \approx m_H$ (to sufficient accuracy at present). Putting in the numbers,

$$n_{\rm B0} = 11.2\, \Omega_{\rm B0}\, h_{100}^2 \quad {\rm m}^{-3}$$

or a pretty good vacuum. $\Omega_{\rm B0}$ is the present-day baryon contribution to the density parameter.

The number of photons per unit volume is given by

$$n_{\gamma 0} = \int_0^\infty \frac{4\pi}{c} \cdot \frac{B(\nu)}{h\nu}\, d\nu$$

where $B(\nu)$ is the black-body spectral intensity (corresponding to the microwave background radiation, which is the dominant radiation field in the Universe)

$$B(\nu) = \frac{2h\nu^3}{c^2} \left(e^{h\nu/k_B T_{\gamma 0}} - 1 \right)^{-1}$$

where $T_{\gamma 0} = 2.728 \pm 0.002$ K. Performing the integral, we find that

$$n_{\gamma 0} = 16\, \pi\, \zeta(3) \left(\frac{k_B T_{\gamma 0}}{hc} \right)^3$$

where $\zeta(x)$ is the Riemann zeta function and $\zeta(3) = 1.202\ldots$, so that

$$n_{\gamma 0} = 4.12 \times 10^8 \quad {\rm m}^{-3}.$$

Thus the number of photons per baryon today is

$$\frac{n_{\gamma 0}}{n_{\rm B0}} = 3.75 \times 10^7 \, \Omega_{B0}^{-1} \, h_{100}^{-2} \gg 1 \quad .$$

This is a large ratio, and we can legitimately be puzzled by it. Why should there be so many more photons than baryons?

We can pose this question another way, by looking at the entropy per baryon in the present-day Universe. The density entropy of the microwave background radiation is

$$s_\gamma = \frac{4}{3} \, a_R \, T_\gamma^3$$

(easily derived from thermodynamics). This corresponds to an entropy per photon of

$$\frac{s_\gamma}{n_\gamma} = \frac{2\pi^4}{45\zeta(3)} k_B \approx 3.6 k_B$$

(a universal constant). Thus the radiation entropy per baryon in the Universe today is proportional to the photon/baryon ratio, and is

$$\frac{s_{\gamma 0}}{n_{B0}} = 1.35 \times 10^8 \, k_B \, \Omega_{B0}^{-1} \, h_{100}^{-2} \quad .$$

This is enormously greater than the entropy that matter has because of its own temperature. Why is the Universe in such a high entropy state today? Why is the disorder so high, but not so high that there is no interesting structure?

The cause of this is, essentially, the baryon asymmetry in the Universe. If n_B and $n_{\overline{B}}$ are the baryon and anti-baryon number densities, then since baryon number is conserved in the Universe (mostly ... see later),

$$\left(n_B - n_{\overline{B}} \right) \, a^3 = \text{constant} \; .$$

In the very early Universe (after the epoch of grand unification), we would expect the number of baryons to be close to the number of antibaryons and the number of photons,

$$n_B \approx n_{\overline{B}} \approx n_\gamma$$

and so the baryon asymmetry

$$\frac{n_B - n_{\overline{B}}}{n_B + n_{\overline{B}}} \approx \frac{n_B - n_{\overline{B}}}{2 n_\gamma} \approx \frac{n_{B0}}{2 n_{\gamma 0}}$$

where n_{B0} and $n_{\gamma 0}$ are the present-day baryon and photon number densities. That is, the large entropy per baryon that we see today, corresponding to the large value of

$$\frac{n_{\gamma 0}}{n_{\rm B0}} = 3.75 \times 10^7 \, \Omega_{B0}^{-1} \, h_{100}^{-2} \gg 1 \quad .$$

corresponds to a *tiny but non-zero* baryon asymmetry created in the early Universe.

This asymmetry is believed to come from a phase of *baryogenesis*, associated with particle-physics processes that involve the non-conservation of baryon number. This must also involve violation of C or CP symmetry (otherwise there would be an equivalent process to cancel out the baryon generation), and must occur out of thermal equilibrium (or there would be no reason to prefer particles over antiparticles and no asymmetry could be generated). This is sometimes referred to as the Sakharov mechanism.

The guess is that some such process is possible under GUTs at $k_B T > 10^{15}$ GeV, that is at a time less than 1 nsec after the Big Bang. Indeed, we can turn the question around and say that a requirement on any GUT is that there should be such a process, otherwise we can't understand where the baryon asymmetry came from, and we would have to put the asymmetry into the initial conditions of the Universe (always an unsatisfactory situation).

There is a residual problem, of course. The fine-tuning needed to get the number of baryons and anti-baryons so nearly equal is strange — when a large asymmetry would have been possible, what process came into action to make only a small asymmetry result from the symmetry-breaking at the end of the GUT phase?

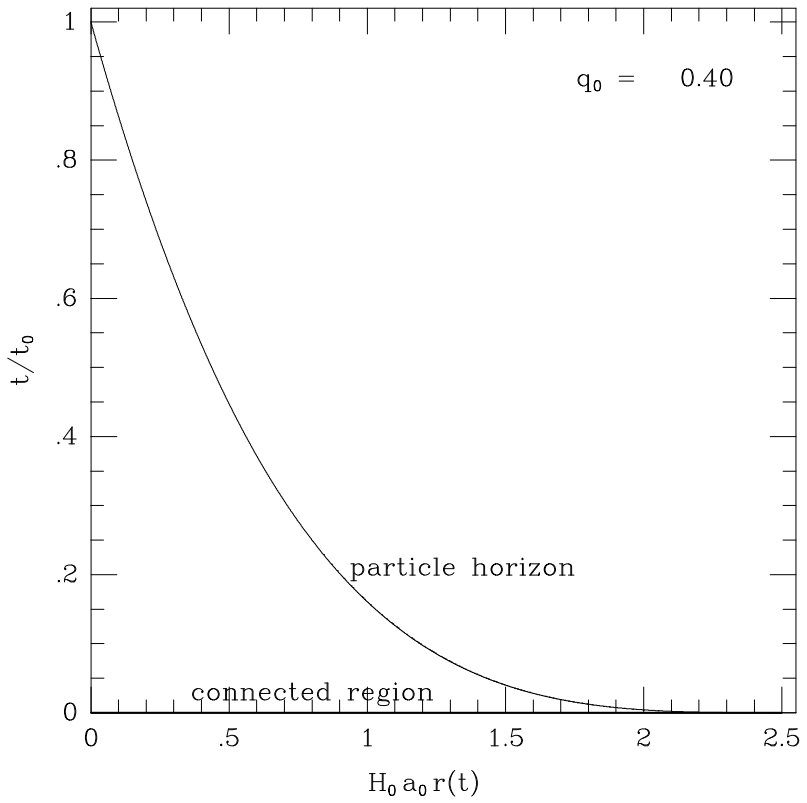
18. Horizons and flatness

18.1. Horizons

Light travels on null geodesics, and so the path of a light ray which moves radially towards an observer at the origin is, as we have used many times,

$$\int_t^{t_0} \frac{dt}{a(t)} = \int_0^r \frac{dr}{(1-kr^2)^{1/2}}$$

which we can solve for $r(t)$. If we look at the world line of such a photon, it appears as in the diagram below (for $q_0 = 0.4$).



What can we see from this diagram? The dominant feature is that there is a domain of $r(t=0)$ which lies within the back light cone, and which therefore can have communicated with us (at $r=0$) by the present time t_0. There is also a region outside the light cone which can never have communicated with us, and which therefore is not causally connected to us. It is also clear that as t_0 increases, a larger and larger coordinate patch of the Universe comes into view.

The region outside the light cone shown is said to lie outside the **particle horizon**. That is, it's in a region of spacetime that is not now causally connected

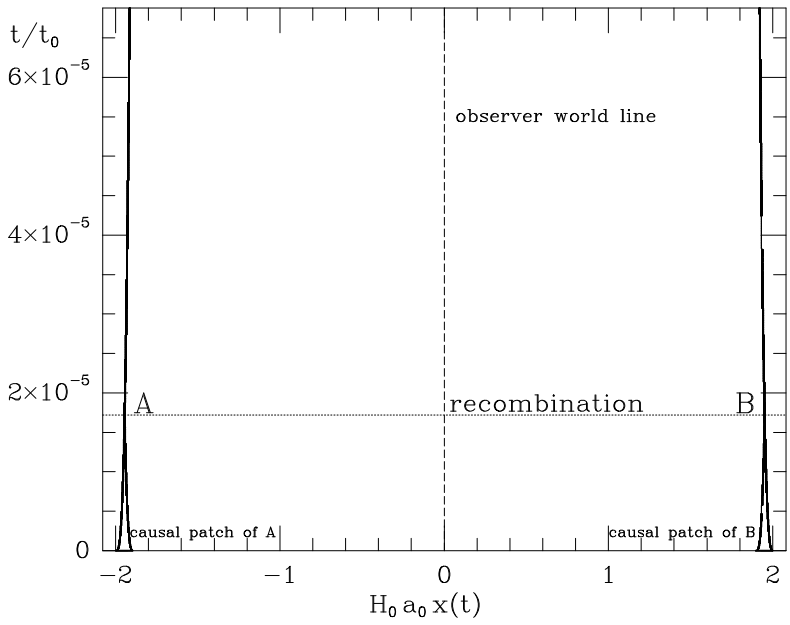
to our current event, but might be so connected in the future. A particle horizon divides the past of an observer into a set of events from which light signals could have been received from a set of event from which no communication is possible.

Contrast with this an **event horizon**. This separates the set of events from which an observer could *ever* receive signals, including at some time in the future, from the set of events from which such communication will never be possible.

Look at the figure again: it highlights the interesting *horizon problem* in cosmology. At present we see the Universe filled with a very uniform radiation field, the microwave background radiation, which originates at $z \approx 1000$. The uniformity of this radiation field is *extreme*, and shows

the Universe is the same over the entire sky at $z = 1000$.

But the back light cones of all points on the sky at $z = 1000$ *do not have a common past*. Let's take a closer look at the very bottom of the first diagram, and consider two locations on the sky, A and B, for which: for example the two points A and B which lie on opposite sides of the sky at $z = 1000$. Then these points have their own particle horizons, which don't overlap — their horizons are very well separated. But points not in causal contact have no reason to exhibit the same physical properties: there's no reason why the Universe should arbitrarily have decided that all locations should be created exactly equal. So why is the microwave background from the directions of A and B similar to a part in 10^5, when there's no reason for the thermodynamic properties of A and B to be the same?



Calculating this for $k = 0$ ($q_0 = \frac{1}{2}$), it's easy to show that the path of a light ray

follows

$$H_0\, a_0\, r(t) = 2\left(1 - \left(\frac{t}{t_0}\right)^{2/3}\right)$$
$$= 2\left(1 - (1+z)^{-\frac{1}{2}}\right)$$

where $t_0 = \frac{2}{3}H_0^{-1}$ is the current age of the Universe. This means that a light-ray from a points with current proper distance (distance today, measured in a local Lorentz frame) of up to

$$a_0 r_{\rm max} = 2\frac{c}{H_0} = 6\, h_{100}^{-1} \quad \text{Gpc}$$

from us could have been in causal contact (this is the proper distance to our horizon). At $z_{\rm rec} = 1000$, points A and B are about $11.6\, h_{100}^{-1}$ Gpc separated in present-day distances, if these are antipodal points on our sky (and the Universe has a trivial topology). But the horizons of these points only extend to a present-day proper radius

$$a_0 r(t_{\rm rec}) = 2H_0^{-1}\,(1+z_{\rm rec})^{-1/2} \approx 0.2\, h_{100}^{-1} \quad \text{Gpc}$$

which is much less than their separation. So, as seen in the diagram, there are *many* independent patches of sky, all of which look very much the same in the microwave background radiation. Indeed, the angular size of these patches is roughly 15°, but the microwave background radiation *does not show much structure on smaller angular scales, where the emitting material at* $z_{\rm rec}$ *should have been very inhomogeneous.* What made the Universe so much more isotropic than it should have been?

18.2. Flatness

Not only is the Universe unexpectedly homogeneous, based on the evidence of the microwave background radiation, but also unexpectedly flat.

Best evidence at the moment says that the density parameter of the Universe (in all constituents: dark matter, baryons, photons, and even vacuum energy) has

$$0.2 \lesssim \Omega_0 \lesssim 2 \quad .$$

This is unexpected: we are within a factor of several of $\Omega_0 = 1$, the critical density Universe. But why is this a problem? Let's look at how Ω depends on time.

The density parameter at time t is the ratio of the density of the Universe to the critical density at that time,

$$\Omega(t) = \frac{\rho(t)}{\rho_{\rm crit}(t)}$$

and both the density in the Universe and the critical density depend on time. At present the Universe is matter-dominated, so

$$\rho = \rho_0 \left(\frac{a_0}{a}\right)^3$$

where ρ_0 is the present matter density (cold dark matter, plus baryons). And the critical density is defined as the density that makes $k = 0$, which from the Friedmann equation is

$$\rho_{\rm crit} = \frac{3H^2}{8\pi} = \frac{3}{8\pi}\frac{\dot{a}^2}{a^2}$$

if we take $\Lambda = 0$. The Friedmann equation, with $\Lambda = 0$, is

$$\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = \frac{8\pi}{3}\rho$$

which I can rewrite in terms of the critical density as

$$\frac{8\pi}{3}\rho_{\rm crit} + \frac{k}{a^2} = \frac{8\pi}{3}\rho$$

or, in terms of the density parameter,

$$\frac{1}{\Omega(t)} = 1 - \frac{3}{8\pi}\frac{k}{\rho a^2} \quad .$$

At $t = t_0$ this still applies, and so

$$\frac{1}{\Omega_0} = 1 - \frac{3}{8\pi}\frac{k}{\rho_0 a_0^2}$$

so, eliminating k,

$$\frac{1 - \frac{1}{\Omega}}{1 - \frac{1}{\Omega_0}} = \frac{\rho_0 a_0^2}{\rho a^2}$$

and so, in a matter-dominated Universe,

$$\frac{1 - \frac{1}{\Omega}}{1 - \frac{1}{\Omega_0}} = \frac{a}{a_0} \quad .$$

This means that as $a \to 0$ (i.e., in the earliest phases of the Universe), for *any* present-day value of Ω_0,

$$\Omega \to 1 \quad .$$

Parenthetically, notice that if $\Omega_0 = 1$, then Ω retains a value of 1 at all times.

But $\Omega(t = 0)$ would be expected to be a parameter of the Big Bang model — it's simply the density of the Universe at the initial point of time, and sets the initial expansion rate. There seems to be no reason why it should be *exactly* one.

And it does need to be one to very high precision. Suppose that at time t_1 close to the time of the Big Bang (say 1 nsec later) the value of $\Omega = 1 + \epsilon_1$, with $\epsilon_1 \ll 1$. Then to first order in ϵ_1, the density parameter today would be

$$\begin{aligned}\Omega_0 &= \left(1 - \frac{a_0}{a_1}\epsilon_1\right)^{-1} \\ &= \left(1 - \left(\frac{t_1}{t_0}\right)^{-\frac{2}{3}}\epsilon_1\right)^{-1}\end{aligned}$$

and so with $t_1 = 1$ nsec $\approx 2 \times 10^{-27}\,\mathrm{t}_0$ (for $h_{100} = 0.5$), we require $\epsilon_1 < 10^{-18}$ if the density today is not to be significantly outside the range in which we know it to lie.

This is a fine-tuning problem. Somehow the early phases of the Universe "knew" that they had to have Ω within a part in 10^{18} of unity (in a number of non-causally connected patches, indeed), so that the Universe today would be anything like it actually is. **Why was the early Universe so incredibly flat?**

18.3. Inflation

The answer to both problems is **inflation**.

It is supposed that some time shortly after the Big Bang, at a temperature $k_{\mathrm{B}}T \approx 10^{14}$ GeV, which by

$$(k_{\mathrm{B}}T/\mathrm{MeV}) \approx 1.3\,(t/\mathrm{sec})^{-\frac{1}{2}}$$

would have happened at time $t \approx 10^{-34}$ sec after the Big Bang, there was a phase transition in the fluid. This phase transition might be associated with the breaking of the symmetry of the strong and weak interactions (or it might have happened earlier, and had some other cause). At that time, the equation of state of the fluid in the Universe would have been

$$\rho = -P$$

since the stress-energy tensor must have been diagonal and proportional to the metric tensor (since we associate the phase of inflation with a *false vacuum*, and so the stress-energy tensor must be proportional to the metric tensor, the only relativistically invariant form). In this case, the equations of motion of the Universe are

$$\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = \frac{8\pi}{3}\rho + \Lambda$$
$$\frac{d}{dt}\left(\rho a^3\right) = -P\,\frac{d}{dt}\left(a^3\right)$$

and with $\rho = -P$, the second equation implies that $\rho =$ constant, ρ_I. This means that ρ is behaving like the cosmological constant, Λ: in the Friedmann equation we can absorb Λ into the definition of ρ_I to obtain

$$\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = \frac{8\pi}{3}\rho_I \quad .$$

In the early Universe the expansion was fast, and the term in k is relatively unimportant (can be demontrated later, if desired — but it's a general result that when the density is high, the curvature term is negligible relative to it. All Universes look flat at early enough times). Making this approximation, the equation of motion becomes

$$\dot{a} = a\left(\frac{8\pi}{3}\rho_I\right)^{1/2}$$

which integrates to the *exponential expansion*

$$a(t) = a_I \exp\left(\left(\frac{8\pi}{3}\rho_I\right)^{1/2} t\right) \quad .$$

It must be emphasized that this is an extremely fast expansion. At time 10^{-34} sec the Universe has not been much decelerated, and is still expanding at about c, so the that proper size of a causally-connected patch of Universe is roughly $ct = 3 \times 10^{-25}$ m.

Now suppose that inflation continues from this time for 1 psec, which is not unreasonable. Then

$$\left(\frac{8\pi}{3}\rho_I\right)^{1/2} t \approx 100$$

is not unreasonable (i.e., it's a guess!), so the exponential factor

$$\exp\left(\left(\frac{8\pi}{3}\rho_I\right)^{1/2} t\right) \approx 10^{43} \quad .$$

Thus by the end of the phase of inflation, the original small causally-connect patch, which had radius 3×10^{-26} m has swollen to a causally-connected patch with radius 3×10^{17} m. Since this is only ~ 1 psec after the Big Bang, $ct \approx 3 \times 10^{-7}$ m. Thus

- a causally-connected patch of Universe that was comfortably smaller than the size of an atomic nucleus has been inflated to a scale $\gg ct$
- even irregularities of a factor 100 in density have therefore been smoothed out by a factor

  $$\left(10^{43}\right)^3 \approx 10^{130}$$

  simply by being stretched. The Universe has therefore been made *very* smooth and uniform. This resolves the horizon problem, since every bit of the Universe that we see was within the original tiny causal patch that was swollen by the inflationary episode.
- and the flatness problem is also solved: in the inflationary phase

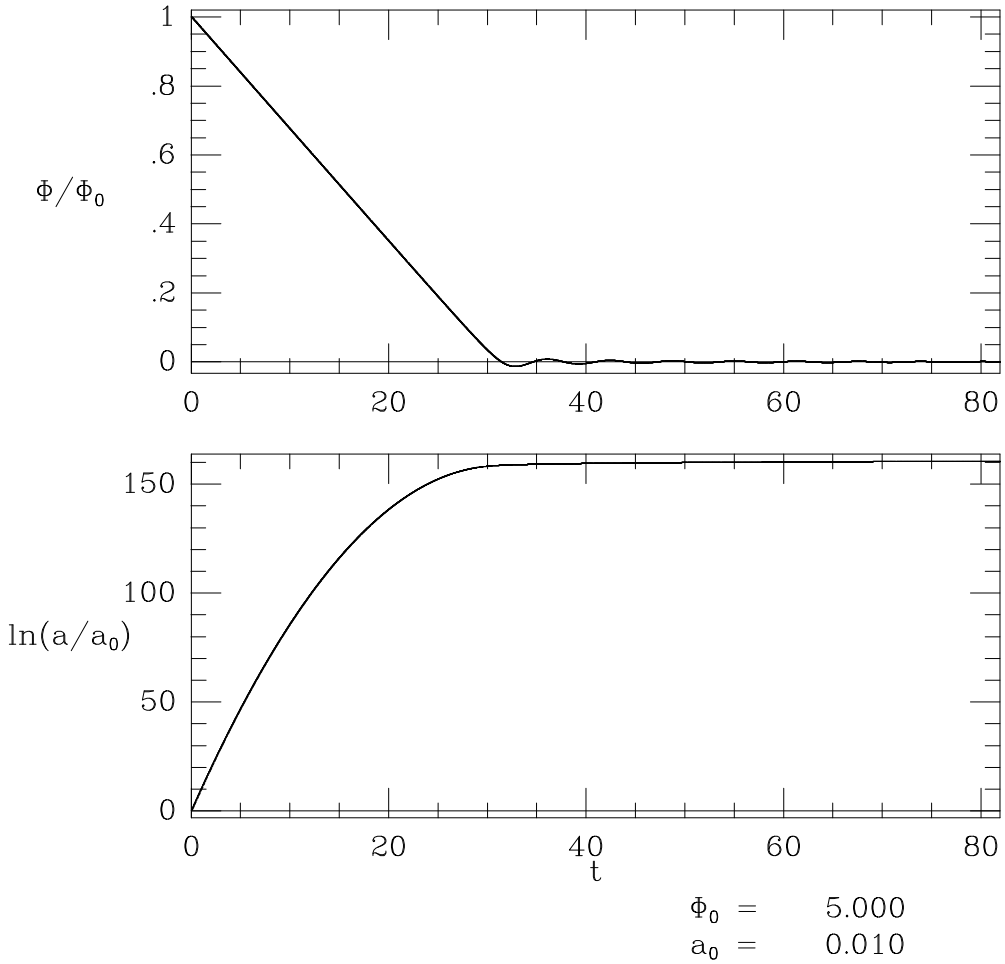
  $$\frac{\dot{a}^2}{a^2} = \frac{8\pi}{3}\rho_I$$

  so that the density parameter

  $$\Omega = \frac{8\pi\rho}{H^2} = \frac{8\pi\rho}{3}\frac{a^2}{\dot{a}^2} = 1 \quad .$$

  That is, the phase of inflation drives the density parameter of the Universe to 1. And since later phases of the expansion (according to the results derived earlier) keep $\Omega = 1$ if $\Omega = 1$ initially, we would expect to find $\Omega = 1$ now.

What does inflation look like? A sample calculation for a particularly simple "inflaton field" (perhaps something to do with some symmetry-breaking at very high energies) is shown below: here Φ is the amplitude of a quantum field, and it can be seen that inflation is driven (the scale factor is expanding very rapidly) even though the dynamics of the inflationary field causes it not to remain constant. For *very many* quantum fields (very many fields in which there's an appreciable energy stored in the field amplitude, rather than the rate of change of the field), a similar sort of solution will occur — a rapid increase in a while the field changes little (or not at all) followed by an oscillatory decay of the field to zero.



What, physically, is happening to solve the flatness problem? Suppose that at the initial state of the Universe there's some density not in the form of the "inflaton field". The rapid expansion takes this density and reduces it exponentially to a very low level (it's expanded away). By contrast, the effective vacuum energy, the inflaton field, doesn't change much as the expansion proceeds once it has reached $\Omega_I = 1$.

So at the end of inflation there's a microscopic amount of non-vacuum like energy,

and a lot of vacuum-like energy. What causes inflation to end? Inflation ends when the inflaton field decays. And this decay is *noisy*, creating a lot of radiation and matter. This *phase of reheating* in the Universe is a major topic of debate — about the mechanism under which the Universe exits from inflation and becomes "normal" again.

But there is agreement about what results — a mixture of hot matter and radiation with $\Omega = 1$. Or perhaps not exactly one, since some energy may end up locked into the Λ-field (with $\Omega_\Lambda < 1$).

And we get one other result for free from this process — we get a mechanism for introducing perturbations (in the form of the noisy pressure and temperature fluctuations caused by the exit from inflation: the ripples in the figure above) which can then start to grow under the influence of gravitation to become modern-day galaxies and clusters of galaxies. That is, the original perturbations which grow to become present-day structures *need not be put in as a boundary condition*, but their properties might be deduced from the physics of the exit from inflation at reheating.