Preprint typeset in JHEP style - HYPER VERSION

General Relativity: An Introduction for Undergraduates*

C.P. Burgess

Department of Physics & Astronomy, McMaster University,
1280 Main Street West, Hamilton, Ontario, Canada, L8S 4M1.
Perimeter Institute for Theoretical Physics,
31 Caroline Street North, Waterloo, Ontario, Canada, N2L 2Y5.

Abstract: These notes present a brief introduction to Einstein's General Theory of Relativity, prepared for the course Physics 3A03.

Contents

---

1. Elements of differential geometry

The essence of general relativity is that gravity is described by the geometry of spacetime, and so this first section pauses to summarize some of the mathematics used to describe non-Euclidean geometries. Before doing so, a brief reminder about Euclidean geometry.

Euclidean geometry

Euclid founded his study of plane (*i.e.* 2-dimensional) geometry on the following five axioms:

1. Any two points can be joined by a straight line.
2. Any straight line segment can be extended indefinitely in a straight line.
3. Given any straight line segment, a circle can be drawn having the segment as radius and one endpoint as center.
4. All right angles are congruent.
5. *Parallel postulate*: If two lines intersect a third in such a way that the sum of the inner angles on one side is less than two right angles, then the two lines inevitably must intersect each other on that side if extended far enough.

All of these seem to be obviously true, given the standard notions of what a point, straight line, circle, right angle and congruence mean. Among the consequences of these axioms are many familiar statements like: the ratio of a circle's circumference, C, to its radius, r, is a universal number: $C/r = 2\pi$; the ratio of a circle's area, A, to the square of its radius is also a universal number $A/r^2 = \pi$; the sum of the interior angles of a triangle sum to 180 degrees, and so on. We are used to taking

these consequences for granted when understanding the relations amongst objects in physical space.

The rest of this section is devoted to describing simple situations where they do not all apply. Once this is done, it becomes an experimental issue whether or not the Euclidean axioms are properties of the space in which we find ourselves situated. The goal of this section is to develop the tools for this, by setting up a precise characterization of these new geometries, and the ways they can differ from Euclidean space.

1.1 Geometry of surfaces

The non-Euclidean geometries that are easiest to visualize are those of two-dimensional surfaces, such as planes, spheres or hyperbolae. These are easy to picture since we can envision these surfaces embedded in 3-dimensional space.

To this end consider the 3-dimensional vector space, $\mathbb{R}_3$, whose vectors, $\mathbf{r}$, describe the distance from an (arbitrary) origin, O, to the various points in space. It is convenient to describe such a vector by its components referred to a 'rectangular' basis of unit vectors, $\{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$, oriented in a fixed but arbitrary direction, so that

$$\begin{aligned} \mathbf{r} &= x\,\mathbf{e}_x + y\,\mathbf{e}_y + z\,\mathbf{e}_z \\ &= x^i\,\mathbf{e}_i\,, \end{aligned} \tag{1.1}$$

where the three coordinates, (x, y, z), each can run from $-\infty$ to ∞.

Some important notation is introduced in the second equality of eq. (1.1), which writes $x^1 = x$, $x^2 = y$, $x^3 = z$, and $\mathbf{e}_1 = \mathbf{e}_x$, $\mathbf{e}_2 = \mathbf{e}_y$ and $\mathbf{e}_3 = \mathbf{e}_z$. There is also an implied sum from 1 to 3 over the repeated index 'i', or any other repeated index taken from the middle of the Latin alphabet for that matter. (Indices taken from the beginning of the Latin alphabet are encountered later, where they run over $a, b = 1, 2$; and indices from the Greek alphabet also come up, and will be summed from $\mu, \nu = 0, 1, 2, 3$.) This rule for summing over repeated indices is called the Einstein summation convention, and in terms of it the dot product of two vectors with components $\mathbf{a} = a^i\,\mathbf{e}_i$ and $\mathbf{b} = b^i\,\mathbf{e}_i$ can be written $\mathbf{a}\cdot\mathbf{b} = \delta_{ij}\,a^i b^j$, where the Kronecker-$\delta$ symbol has the property that $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

We take the distance, $s(\mathbf{r}_1, \mathbf{r}_2)$, between any two points, $\mathbf{r}_1$ and $\mathbf{r}_2$, in $\mathbb{R}_3$ is given in terms of their rectangular coordinates by the usual Pythagorean rule

$$\begin{aligned} s(\mathbf{r}_1, \mathbf{r}_2) &= |\mathbf{r}_1 - \mathbf{r}_2| = \sqrt{(\mathbf{r}_1 - \mathbf{r}_2)\cdot(\mathbf{r}_1 - \mathbf{r}_2)} \\ &= \sqrt{\delta_{ij}(x_1^i - x_2^i)(x_1^j - x_2^j)} \\ &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}\,, \end{aligned} \tag{1.2}$$

where the middle line again uses the Einstein summation convention. This definition has the important property that it does not depend at all on the origin, O, and orientation of the axes, $\mathbf{e}_i = \{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$, that are required to define the coordinates $x^i = \{x, y, z\}$ describing $\mathbf{r}_1$ and $\mathbf{r}_2$.

Curves in Space

Before describing two-dimensional surfaces in $\mathbb{R}_3$, it is worth briefly digressing to describe the simpler case of one-dimensional curves. A curve in $\mathbb{R}_3$ is defined by the locus of points that are swept out as a single parameter varies:

$$\begin{aligned} \mathbf{r}(u) &= x(u)\,\mathbf{e}_x + y(u)\,\mathbf{e}_y + z(u)\,\mathbf{e}_z \\ &= x^i(u)\,\mathbf{e}_i\,. \end{aligned} \tag{1.3}$$

Here the parameter u labels the points on the curve and our interest is usually in component functions $x^i(u) = \{x(u), y(u), z(u)\}$ that are multiply differentiable with respect to u.

For example, straight lines in this picture are described by linear functions, $\mathbf{r}(u) = \mathbf{a} + \mathbf{b}\,u$, where $\mathbf{a}$ and $\mathbf{b}$ are arbitrary constant vectors. When the origin, O, is not on the straight line (*i.e.* $\mathbf{a} \neq 0$) then the origin together with the line define a plane, which is spanned by the vectors $\mathbf{a}$ and $\mathbf{b}$. More generally, a straight line is also given by $\mathbf{r}(u) = \mathbf{a} + \mathbf{b}\,f(u)$, for any function $f(u)$ that satisfies $\mathrm{d}f/\mathrm{d}u \neq 0$, since this simply represents a relabelling of the points along the curve.

By contrast, a curve of the form $\mathbf{r}(u) = \mathbf{c} + \mathbf{a}\cos u + \mathbf{b}\sin u$ traces out a more complicated closed shape, which becomes an ellipse if $\mathbf{a}$ and $\mathbf{b}$ are perpendicular to one another: $\mathbf{a}\cdot\mathbf{b} = a_x b_x + a_y b_y + a_z b_z = 0$. In this case $\mathbf{c}$ specifies the position of the ellipse's centre, and its two semi-major axes are

$$\begin{aligned} a &= |\mathbf{a}| = \sqrt{\mathbf{a}\cdot\mathbf{a}} = \sqrt{a_x^2 + a_y^2 + a_z^2} = \sqrt{\delta_{ij}\,a^i a^j} \\ b &= |\mathbf{b}| = \sqrt{\mathbf{b}\cdot\mathbf{b}} = \sqrt{b_x^2 + b_y^2 + b_z^2} = \sqrt{\delta_{ij}\,b^i b^j}\,. \end{aligned} \tag{1.4}$$

This ellipse is inscribed on the plane spanned by the vectors $\mathbf{a}$ and $\mathbf{b}$, and degenerates into a circle in the special case that $\mathbf{a}$ and $\mathbf{b}$ have the same length: $a = b$.

The family of vectors that lie tangent to a curve $\mathbf{r}(u)$ is found by differentiation,

$$\mathbf{t}(u) = \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}u} = \frac{\mathrm{d}x}{\mathrm{d}u}\,\mathbf{e}_x + \frac{\mathrm{d}y}{\mathrm{d}u}\,\mathbf{e}_y + \frac{\mathrm{d}z}{\mathrm{d}u}\,\mathbf{e}_z = \frac{\mathrm{d}x^i}{\mathrm{d}u}\,\mathbf{e}_i\,, \tag{1.5}$$

and a one-parameter family of unit vectors tangent to the curve is found by normalizing

$$\mathbf{e}_t(u) = \frac{\mathbf{t}(u)}{|\mathbf{t}(u)|}\,, \tag{1.6}$$

so $\mathbf{e}_t \cdot \mathbf{e}_t = 1$ for all u. For a straight line, $\mathbf{r}(u) = \mathbf{a} + \mathbf{b} f(u)$, the tangent

$$\mathbf{t}(u) = \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}u} = \mathbf{b}\,\frac{\mathrm{d}f}{\mathrm{d}u}\,, \tag{1.7}$$

has a constant direction, but a u-dependent length that depends on the precise function $f(u)$ used to parametrize the curve. But for any parametrization the unit tangent vector for a straight line is a constant vector: $\mathbf{e}_t = \mathbf{b}/|\mathbf{b}|$. The basis vectors, $\mathbf{e}_i = \{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ may themselves be regarded as unit tangent vectors to the curves defining the rectangular coordinate axes themselves: that is, $\mathbf{e}_x$ is the unit tangent to the curves along which y and z are constant, and similarly for $\mathbf{e}_y$ and $\mathbf{e}_z$.

The tangent to the elliptical curve centered at the origin, $\mathbf{r}(u) = \mathbf{a}\cos u + \mathbf{b}\sin u$ is given by $\mathbf{t}(u) = -\mathbf{a}\sin u + \mathbf{b}\cos u$, whose direction changes continuously with u, with norm $|\mathbf{t}(u)| = \sqrt{a^2\sin^2 u + b^2\cos^2 u}$ (and we use $\mathbf{a}\cdot\mathbf{b} = 0$). In this case the unit tangent is $\mathbf{e}_t(u) = (-\mathbf{a}\sin u + \mathbf{b}\cos u)/|\mathbf{t}(u)|$. Notice that the inner product between the radius vector and the tangent is $\mathbf{t}(u)\cdot\mathbf{r}(u) = (b^2 - a^2)\sin u\cos u$, which vanishes for all u in the case of a circle, where $b = a$.

Distances along curves

Measures of length and angle play a central role in geometry, and since angle (in radians) is defined in terms of ratios of lengths, the basic problem is how to measure length within curved surfaces. This section describes a first step in this direction: measuring length along curves.

The starting point is eq. (1.2), telling us how distances are measured in $\mathbb{R}_3$. We apply this to find the distance, $\mathrm{d}s$, between two points on a curve, $\mathbf{r}(u)$ and $\mathbf{r}(u+\mathrm{d}u)$, that are infinitesimally far from one another.

$$\mathrm{d}s = |\mathbf{r}(u+\mathrm{d}u) - \mathbf{r}(u)| = \left|\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}u}\right| \mathrm{d}u = \sqrt{\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}u}\cdot\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}u}}\,\mathrm{d}u = \sqrt{\delta_{ij}\,\frac{\mathrm{d}x^i}{\mathrm{d}u}\,\frac{\mathrm{d}x^j}{\mathrm{d}u}}\,\mathrm{d}u\,, \tag{1.8}$$

The arc-length along a finite-sized interval of the curve is then obtained by integration

$$s(u_1, u_2) = \int_{u_1}^{u_2} \mathrm{d}u\,\sqrt{\delta_{ij}\,\frac{\mathrm{d}x^i}{\mathrm{d}u}\,\frac{\mathrm{d}x^j}{\mathrm{d}u}}\,. \tag{1.9}$$

For example, for the circle $\mathbf{r}(u) = a(\mathbf{e}_x\,\cos u + \mathbf{e}_y\,\sin u)$ we have $\mathrm{d}\mathbf{r}/\mathrm{d}u = a(-\mathbf{e}_x\,\sin u + \mathbf{e}_y\,\cos u)$ and so $\mathrm{d}s = a\,\mathrm{d}u$, giving $s(u_1, u_2) = a(u_2 - u_1)$.

Arc-length provides a particularly physical way to parameterize a curve. Once this is done the tangent vector to a curve is automatically a unit vector. To see this consider a generic curve, $\mathbf{r}(u)$, defined using a generic parameter, u. The tangent vector computed using arc-length as a parameter is

$$\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}s} = \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}u}\,\frac{\mathrm{d}u}{\mathrm{d}s} = \frac{\mathbf{t}}{|\mathbf{t}|} = \mathbf{e}_t\,, \tag{1.10}$$

where $\mathbf{t} = \mathrm{d}\mathbf{r}/\mathrm{d}u$ and eq. (1.8) is used to evaluate $\mathrm{d}u/\mathrm{d}s = 1/|\mathbf{t}|$.

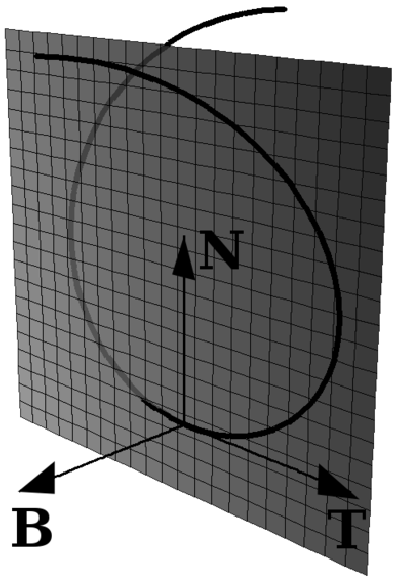


Figure 1: The Frenet-Serret basis vectors and the osculating plane (Wikipedia).

Curvature of curves

In addition to the unit tangent, $\mathbf{e}_t = \mathrm{d}\mathbf{r}/\mathrm{d}s$, there is also a natural family of orthonormal basis vectors that can be defined everywhere along a curve. A unit vector, $\mathbf{n}$, that is always perpendicular to $\mathbf{e}_t$ is found as above by differentiation with respect to arc length: $\mathbf{n}(s) = \mathrm{d}\mathbf{e}_t/\mathrm{d}s$. The fact that this definition gives a vector normal to $\mathbf{e}_t$ can be seen by differentiating the condition $\mathbf{e}_t \cdot \mathbf{e}_t = 1$, as follows:

$$\mathbf{e}_t \cdot \mathbf{n} = \mathbf{e}_t \cdot \frac{\mathrm{d}\mathbf{e}_t}{\mathrm{d}s} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}\Big(\mathbf{e}_t \cdot \mathbf{e}_t\Big) = 0\,. \tag{1.11}$$

The plane spanned by $\mathbf{t}(s)$ and $\mathbf{n}(s)$ at each u is called the *osculating plane* for the curve $\mathbf{r}(s)$. The vectors

$$\mathbf{e}_t(s)\,, \quad \mathbf{e}_n(s) = \frac{\mathbf{n}(s)}{|\mathbf{n}(s)|} \quad \text{and the cross product} \quad \mathbf{e}_b(s) = \mathbf{e}_t(s) \times \mathbf{e}_n(s)\,, \tag{1.12}$$

give an orthonormal triad of vectors at each point along the curve, one of which is always tangent.

Because these vectors form a basis, their derivative along the curve can be expanded in terms of them, leading to:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{e}_t}{\mathrm{d}s} &= \qquad\quad \kappa\,\mathbf{e}_n \\ \frac{\mathrm{d}\mathbf{e}_n}{\mathrm{d}s} &= -\kappa\,\mathbf{e}_t \qquad + \tau\,\mathbf{e}_b \\ \frac{\mathrm{d}\mathbf{e}_b}{\mathrm{d}s} &= \qquad\quad - \tau\,\mathbf{e}_n\,. \end{aligned} \tag{1.13}$$

These expressions are known as the *Frenet-Serret formulae*, and the basis $\mathbf{e}_t$, $\mathbf{e}_n$ and $\mathbf{e}_b$ is called the *Frenet-Serret basis*. The coefficients in this expression give a differential measure of the *curvature*, $\kappa(s)$, and *torsion*, $\tau(s)$, at each point of the curve $\mathbf{r}(s)$.

Exercise 1: Use the definitions of $\mathbf{e}_t$, $\mathbf{e}_n$ and $\mathbf{e}_b$ to prove that only two parameters, κ and τ, are required to label their derivatives as in eqs. (1.13).

Notice that the definitions show that $\kappa = \tau = 0$ for a straight line. Conversely, if κ and τ vanish for all u, then eqs. (1.13) can be integrated twice to show that the corresponding curve, $\mathbf{r}(u)$, is a straight line. Similarly, if τ should vanish for all u (with $\kappa(u)$ arbitrary), then the curve must be confined to the plane that is normal to the constant vector $\mathbf{e}_b$.

> **Exercise 2:** Show that the curvature and torsion of the curve $\mathbf{r}(s) = a[\mathbf{e}_x \cos(s/a) + \mathbf{e}_y \sin(s/a)]$ (a circle of radius a) are constant, with $\kappa = 1/a$ and $\tau = 0$. Repeat for the helical curve $\mathbf{r}(u) = a(\mathbf{e}_x \cos u + \mathbf{e}_y \sin u) + \ell\, u\, \mathbf{e}_z$, keeping in mind that the arc-length in this case satisfies $s = u\sqrt{a^2 + \ell^2}$.

Surfaces in $\mathbb{R}_3$

A two-dimensional surface embedded in $\mathbb{R}_3$ is similarly defined by the locus of points swept out by a two-parameter family,

$$\mathbf{r}(u, v) = x^i(u, v)\, \mathbf{e}_i = x(u, v)\, \mathbf{e}_x + y(u, v)\, \mathbf{e}_y + z(u, v)\, \mathbf{e}_z\,. \tag{1.14}$$

Alternatively, it is sometimes more convenient to define the surface implicitly, rather than explicitly, such as through an algebraic condition of the form $f(\mathbf{r}) = 0$. In this case the expression $\mathbf{r}(u, v)$ can be regarded as being obtained as the solution to this condition. We next provide explicit representations for some simple surfaces, many of which are used as illustrative examples in later sections.

Planar surfaces:

A plane passing through the origin and spanned by two linearly-independent vectors $\mathbf{a}$ and $\mathbf{b}$ is swept out by a surface whose equation has the form

$$\mathbf{r}(u, v) = \mathbf{a}\, u + \mathbf{b}\, v\,, \tag{1.15}$$

with $-\infty < u, v < \infty$. Straight lines can be inscribed inside such a plane, such as $\mathbf{r}(u) = \mathbf{a}\, u$ or $\mathbf{r}(v) = \mathbf{b}\, v$ or $\mathbf{r}(u) = (\mathbf{a} + \mathbf{b})\, u$, as can circles. As is easily verified, the geometry of these circles and straight lines defined for any such a plane satisfies the axioms of Euclidean geometry.

Planes can equally well be specified through a constraint $f(\mathbf{r}) = 0$. For example, the plane $\mathbf{r}(u, v) = \mathbf{e}_x\, u + \mathbf{e}_y\, v$ defined by the x- and y-axes is equally well described as the general solution to the condition $z = 0$, and so $f(\mathbf{r}) := z = \mathbf{e}_z \cdot \mathbf{r}$.

Cylindrical surfaces:

A slightly more interesting example is provided by a cylindrical surface. A representation for a cylinder concentric with the z-axis and having an elliptical profile

aligned with the x- and y-axes would be

$$\mathbf{r}(u,v) = \mathbf{e}_x\, a\cos u + \mathbf{e}_y\, b\sin u + \mathbf{e}_z\, v\,, \tag{1.16}$$

where $0 \le u < 2\pi$ and $-\infty < v < \infty$. The constants a and b define the semi-major axes of the elliptical cross sections taken at fixed z. This elliptical cylinder could equally well be specified by the condition $f(\mathbf{r}) := (x^2/a^2) + (y^2/b^2) - 1 = 0$. It is possible to inscribe straight lines on such a cylinder, but only if they are parallel with the z-axis: for instance $\mathbf{r}(v) = \mathbf{e}_x\, a\cos u_\star + \mathbf{e}_y\, b\sin u_\star + \mathbf{e}_z\, v$, where $u_\star$ is any particular, fixed, value of u.

Spherical surfaces:

The surface of a sphere provides an example of a truly curved surface (in a sense explained in detail below). A representative sphere centered at the origin with radius a can be represented as the surface $f(\mathbf{r}) := x^2 + y^2 + z^2 - a^2 = 0$, or explicitly parameterized using spherical polar coordinates ($u = \theta$ and $v = \phi$) by:

$$\mathbf{r}(u,v) = \mathbf{e}_x\, a\sin u\cos v + \mathbf{e}_y\, a\sin u\sin v + \mathbf{e}_z\, a\cos u\,, \tag{1.17}$$

with $0 < u < \pi$ and $0 \le v < 2\pi$. It is intuitively clear that no straight lines can be inscribed on a sphere.

Inscribed Curves

Given a surface $\mathbf{r}(u,v) = x^i(u,v)\,\mathbf{e}_i$ in $\mathbb{R}_3$, an inscribed curve is a curve, $\mathbf{x}(w) = x^i(w)\,\mathbf{e}_i$, in $\mathbb{R}_3$ whose points also lie within the surface. For instance if the surface is defined by a condition of the form $f(\mathbf{r}) = 0$, then an inscribed curve satisfies $f(\mathbf{x}(w)) = 0$ for all values of its parameter, w. An alternative way of describing an inscribed curve is to specify the curve parameters, $\{u(w), v(w)\}$, that trace out the points along the curve: $\mathbf{r}((u(w), v(w)) = \mathbf{x}(w)$. For instance, the circle $\mathbf{x}(w) = a(\mathbf{e}_x\,\cos w + \mathbf{e}_y\,\sin w)$ is inscribed in the sphere $\mathbf{r}(u,v) = a(\mathbf{e}_x\,\sin u\cos v + \mathbf{e}_y\,\sin u\sin v + \mathbf{e}_z\,\cos u)$, and can be described by the parameter values $\{u(w), v(w)\} = \{\frac{\pi}{2}, w\}$.

The tangent to an inscribed curve can therefore be written either in terms of derivatives of $\mathbf{x}(w)$ or $\mathbf{r}(u,v)$,

$$\mathbf{t} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}w} = \frac{\mathrm{d}}{\mathrm{d}w}\,\mathbf{r}(u(w), v(w)) = \frac{\partial\mathbf{r}}{\partial u}\,\frac{\mathrm{d}u}{\mathrm{d}w} + \frac{\partial\mathbf{r}}{\partial v}\,\frac{\mathrm{d}v}{\mathrm{d}w}\,. \tag{1.18}$$

It is useful to use the Einstein summation convention to combine the above expressions into the more compact notation

$$\mathbf{t} = \frac{\partial\mathbf{r}}{\partial u^a}\,\frac{\mathrm{d}u^a}{\mathrm{d}w} = \frac{\partial x^i}{\partial u^a}\,\frac{\mathrm{d}u^a}{\mathrm{d}w}\,\mathbf{e}_i\,, \tag{1.19}$$

where $a = 1, 2$ with $u^1 = u$ and $u^2 = v$.

A particularly simple family of inscribed curves is obtained by holding fixed either one of the two parameters, u or v, that define the surface itself. Consider for instance a surface defined by the locus of points swept out by a particular parameterization $\mathbf{r}(u, v)$. A family of curves lying in this surface, parameterized by u, is found by setting v to some fixed value $v = v_\star$: $\mathbf{r}(u) = \mathbf{r}(u, v_\star)$. Different values of $v_\star$ produce different members of this family of curves. A second family of curves lying within $\mathbf{r}(u, v)$ is similarly obtained by fixing u at a sequence of values, $u = u_\star$, and letting the variation of v parameterize the curves: $\mathbf{r}(v) = \mathbf{r}(u_\star, v)$. Different choices for $u_\star$ then define different members of this family of curves.

It is possible to use the tangents of inscribed curves to define a pair of linearly independent tangent vectors to any surface that are not necessarily orthogonal. These are given above simply by computing the tangent vector for the inscribed curves along which only one of either u or v varies. The tangent to the curves along which only u varies is given by

$$\mathbf{t}(u) = \frac{\partial \mathbf{r}}{\partial u}(u, v_\star)\,, \tag{1.20}$$

and a family of unit vectors tangent to these curves are then given by $\mathbf{e}_u = \mathbf{t}(u)/|\mathbf{t}(u)|$. The tangents to the curves along which only v varies are similarly given by

$$\hat{\mathbf{t}}(v) = \frac{\partial \mathbf{r}}{\partial v}(u_\star, v)\,, \tag{1.21}$$

and the unit tangent becomes $\mathbf{e}_v = \hat{\mathbf{t}}(v)/|\hat{\mathbf{t}}(v)|$.

Again using the notation $u^a = \{u^1, u^2\} = \{u, v\}$, these may be written

$$\mathbf{t}_a = \frac{\partial \mathbf{r}}{\partial u^a} = \frac{\partial x^i}{\partial u^a}\,\mathbf{e}_i\,, \tag{1.22}$$

where $\mathbf{t}_1 = \mathbf{t}$ while $\mathbf{t}_2 = \hat{\mathbf{t}}$. The span of the normalized vectors $\mathbf{e}_a(u, v)$ define the *tangent plane* to the surface at the point labelled by (u, v).

A normal vector defined everywhere on the surface $\mathbf{r}(u, v)$ may then be constructed using the two families of tangent vectors defined above, $\mathbf{e}_u$ and $\mathbf{e}_v$, by taking the cross product: $\mathbf{e}_n(u, v) = \mathbf{e}_u(u, v) \times \mathbf{e}_v(u, v)$. This defines a basis of vectors that is adapted to the surface at every point.

Notice that if the surface is specified by a constraint, $f(\mathbf{r}) = 0$, then an alternative way to identify this normal direction is by taking the gradient of f:

$$\mathbf{n} = \nabla f = \mathbf{e}_x \left(\frac{\partial f}{\partial x}\right) + \mathbf{e}_y \left(\frac{\partial f}{\partial y}\right) + \mathbf{e}_z \left(\frac{\partial f}{\partial z}\right)\,, \tag{1.23}$$

because the following argument shows this vector is orthogonal to the tangent vectors. The argument relies on the observation that if $\mathbf{r}(u, v)$ is a parametrization of the

surface defined by $f(\mathbf{r}) = 0$, then what this means is $f(\mathbf{r}(u,v)) = 0$ for all u and v. Differentiating this last expression with respect to u or v, and using the chain rule, then implies $\nabla f \cdot (\partial \mathbf{r}/\partial u) = \nabla f \cdot (\partial \mathbf{r}/\partial u) = 0$, or

$$\mathbf{t}_a \cdot \nabla f = \frac{\partial \mathbf{r}}{\partial u^a} \cdot \nabla f = \frac{\partial x^i}{\partial u^a}\, \partial_i f = \frac{\partial x}{\partial u^a}\,\frac{\partial f}{\partial x} + \frac{\partial y}{\partial u^a}\,\frac{\partial f}{\partial y} + \frac{\partial z}{\partial u^a}\,\frac{\partial f}{\partial z} = 0\,, \tag{1.24}$$

which states that ∇f is perpendicular to both the tangent vectors, $\mathbf{t}_a = \{\mathbf{t}, \hat{\mathbf{t}}\}$. Eq. (1.24) introduces the notation

$$\partial_i := \frac{\partial}{\partial x^i}\,, \tag{1.25}$$

and (for practice) is rewritten several ways to emphasize the Einstein summation convention.

Distances along surfaces

Distances along a surface are similarly measured along a curve inscribed in this surface, and in general the distance between two points depends on the details of which curve is used to link these points, just as is also true for points in $\mathbb{R}_3$.

In $\mathbb{R}_3$ when one speaks of the distance between two points without referring to the curve involved, what is meant is the distance along the straight line that connects the two points. Since a straight line cannot in general be inscribed into a generic curved surface it is clear that the same definition cannot generically be used to define a distance between points in a generic surface.

An exception to this is when the two points of interest are infinitesimally separated on the surface: $\mathbf{r}(u,v)$ and $\mathbf{r}(u+\mathrm{d}u, v+\mathrm{d}v)$, since in this case the straight-line curve that connects them is arbitrarily close to an inscribed arc lying on the surface. In this case the distance between the points becomes

$$\mathrm{d}s = |\mathbf{r}(u,v) - \mathbf{r}(u+\mathrm{d}u, v+\mathrm{d}v)| = \left|\frac{\partial \mathbf{r}}{\partial u^a}\,\mathrm{d}u^a\right| = \sqrt{\delta_{ij}\,\frac{\partial x^i}{\partial u^a}\frac{\partial x^j}{\partial u^b}\,\mathrm{d}u^a\,\mathrm{d}u^b}\,. \tag{1.26}$$

The last version of this equation, using the Einstein summation convention, is most commonly written without the ugly square root:

$$\mathrm{d}s^2 = \gamma_{ab}(u,v)\,\mathrm{d}u^a \mathrm{d}u^b\,, \tag{1.27}$$

where the right-hand side defines what was historically called the surface's *first fundamental quadratic form* — or its *induced metric* in more modern parlance — with

$$\gamma_{ab} = \delta_{ij}\,\frac{\partial x^i}{\partial u^a}\frac{\partial x^j}{\partial u^b}\,. \tag{1.28}$$

A central point of the geometry of surfaces is that any *intrinsic* property of the surface — that is, involving only distances and angles associated to inscribed curves on the surface — can be expressed in terms of $\gamma_{ab}(u,v)$ and its derivatives.

Exercise 3: Show that the induced metric for the plane given by $\mathbf{r}(u,v) = \mathbf{e}_x\, u + \mathbf{e}_y\, v$ in $\mathbb{R}_3$ is

$$\gamma_{ab} = \delta_{ab} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{1.29}$$

where δ_{ab} is the Kronecker δ-function, defined just above eq. (1.2). Repeat the calculation for the cylinder $\mathbf{r}(u,v) = a(\mathbf{e}_x\, \cos u + \mathbf{e}_y\, \sin u) + \mathbf{e}_z v$ to show that

$$\gamma_{ab} = \begin{pmatrix} \gamma_{uu} & \gamma_{uv} \\ \gamma_{vu} & \gamma_{vv} \end{pmatrix} = \begin{pmatrix} a^2 & 0 \\ 0 & 1 \end{pmatrix}. \tag{1.30}$$

Finally repeat for the sphere $\mathbf{r}(u,v) = a(\mathbf{e}_x\, \sin u \cos v + \mathbf{e}_y\, \sin u \sin v + \mathbf{e}_z\, \cos u)$, to show

$$\gamma_{ab} = \begin{pmatrix} \gamma_{uu} & \gamma_{uv} \\ \gamma_{vu} & \gamma_{vv} \end{pmatrix} = \begin{pmatrix} a^2 & 0 \\ 0 & a^2 \sin^2 u \end{pmatrix}. \tag{1.31}$$

The arc-length along any inscribed curve running between points A and B may now be found by integrating eqs. (1.27) in the form:

$$s(A,B) = \int_{w_A}^{w_B} \mathrm{d}w\, \frac{\mathrm{d}s}{\mathrm{d}w} = \int_{w_A}^{w_B} \mathrm{d}w\, \sqrt{\gamma_{ab}(w)\, \frac{\mathrm{d}u^a}{\mathrm{d}w}\, \frac{\mathrm{d}u^b}{\mathrm{d}w}}\,, \tag{1.32}$$

where $\gamma_{ab}(w) = \gamma_{ab}(u(w), v(w))$.

Angles between inscribed curves

The angle, θ, between two inscribed curves that intersect at a point P can also be computed using γ_{ab} evaluated at P.

To see this suppose the curves $\mathbf{x}_1(s)$ and $\mathbf{x}_2(s)$ inscribed in the surface $\mathbf{r}(u,v)$ intersect at the point labelled by $(u,v) = (u_\star, v_\star)$. The angle between these curves may be defined as the angle between their tangent vectors, evaluated at P:

$$\mathbf{t}_1 = \frac{\mathrm{d}\mathbf{x}_1}{\mathrm{d}s} = \frac{\partial \mathbf{r}}{\partial u^a}\, \frac{\mathrm{d}u_1^a}{\mathrm{d}s}\,, \tag{1.33}$$

where $u_1^a(s) = \{u_1(s), v_1(s)\}$ describes the parameters which describe the locus of points on the surface through which the inscribed curve $\mathbf{x}_1(s) = \mathbf{r}(u_1(s), v_1(s))$ passes. An identical expression also holds for $\mathbf{t}_2 = \mathrm{d}\mathbf{x}_2/\mathrm{d}s$ and $u_2^a(s) = \{u_2(s), v_2(s)\}$. Clearly the norm of the tangent vector evaluated at P is therefore given by

$$|\mathbf{t}_1|^2 = \frac{\mathrm{d}\mathbf{x}_1}{\mathrm{d}s} \cdot \frac{\mathrm{d}\mathbf{x}_1}{\mathrm{d}s} = \frac{\partial \mathbf{r}}{\partial u^a}\, \frac{\partial \mathbf{r}}{\partial u^b}\, \frac{\mathrm{d}u_1^a}{\mathrm{d}s}\, \frac{\mathrm{d}u_1^b}{\mathrm{d}s} = \gamma_{ab}(u_\star, v_\star)\, \frac{\mathrm{d}u_1^a}{\mathrm{d}s}\, \frac{\mathrm{d}u_1^b}{\mathrm{d}s}\,, \tag{1.34}$$

and similarly for $|\mathbf{t}_2|^2$.

Using $\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$, where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$, we have

$$\cos\theta = \frac{\mathbf{t}_1\cdot\mathbf{t}_2}{|\mathbf{t}_1||\mathbf{t}_2|} = \frac{1}{|\mathbf{t}_1||\mathbf{t}_2|}\left(\frac{\partial\mathbf{r}}{\partial u^a}\cdot\frac{\partial\mathbf{r}}{\partial u^b}\right)\frac{\mathrm{d}u_1^a}{\mathrm{d}s}\frac{\mathrm{d}u_2^b}{\mathrm{d}s} = \frac{\gamma_{ab}(u_\star,u_\star)}{|\mathbf{t}_1||\mathbf{t}_2|}\frac{\mathrm{d}u_1^a}{\mathrm{d}s}\frac{\mathrm{d}u_2^b}{\mathrm{d}s}\,. \tag{1.35}$$

Combining eq. (1.35) with eq. (1.34) applied to both $|\mathbf{t}_1|$ and $|\mathbf{t}_2|$ then shows that θ can be determined purely in terms of $\gamma_{ab}(u_\star,v_\star)$ and the quantities $\mathrm{d}u_1^a/\mathrm{d}s$ and $\mathrm{d}u_2^a/\mathrm{d}s$, all of which would be accessible to an observer trapped to live on the surface.

Geodesics

Although straight lines cannot in general be defined for curves inscribed on a general surface in $\mathbb{R}_3$, there is a natural definition of what is the straightest line possible given the surface's curvature. This definition starts from the observation that a straight line connecting two points in $\mathbb{R}_3$ gives the curve along which the distance between these points is minimized.

This suggests identifying those curves on a given surface that minimize the distance between points, and letting these stand in for straight lines from the point of view of the intrinsic geometry of the surface. Such curves are called geodesics, and are readily computed once the induced metric, $\gamma_{ab}(u,v)$, is everywhere known. The explicit calculation of these curves is left to a subsequent section.

Curvature of surfaces

We have seen that it is always possible to define the curvature, $\kappa(s)$, for a curve, $\mathbf{x}(s)$, by using the Frenet-Serret basis for $\mathbf{x}(s)$ as above, whose derivatives along the curve satisfy the Frenet-Serret formulae, eqs. (1.13). In particular, the first formula, $\mathrm{d}\mathbf{e}_t/\mathrm{d}s = \kappa(s)\,\mathbf{e}_n$, gives κ in terms of the magnitude, $|\mathrm{d}\mathbf{e}_t/\mathrm{d}s|$, of the rate of change of the curve's unit tangent. However, because the $\mathbf{e}_n$ direction need not be specially correlated with the tangent or normal to the surface in which $\mathbf{x}(s)$ is inscribed, this definition of curvature need have little to do with the properties of the surface.

To obtain a measure of the surface's curvature it is therefore useful to focus on a specific family of inscribed curves, $\mathbf{x}(s)$, defined by the intersection of the surface with any of the planes that contain the surface's normal vector, $\mathbf{n}$ (see fig. 2). Because they are defined by construction to lie within a plane, such inscribed curves have vanishing torsion, $\tau(s) = 0$. Furthermore, because the osculation plane spanned by $\mathbf{e}_t$ and $\mathbf{e}_n = \mathrm{d}\mathbf{e}_t/\mathrm{d}s$ contains $\mathbf{n}$, and because the surface's normal, $\mathbf{n}$, is necessarily orthogonal to the tangent of any inscribed curve, it follows that $\mathrm{d}\mathbf{e}_t/\mathrm{d}s$ must be parallel (or antiparallel) to the normal direction, $\mathbf{n}$.

The curvature, $\kappa(s)$, defined using the Frenet-Serret formulae, eqs. (1.13), for such a curve is called a *normal curvature*, $\kappa_n(s)$, of the surface at the point $\mathbf{x}(s)$.

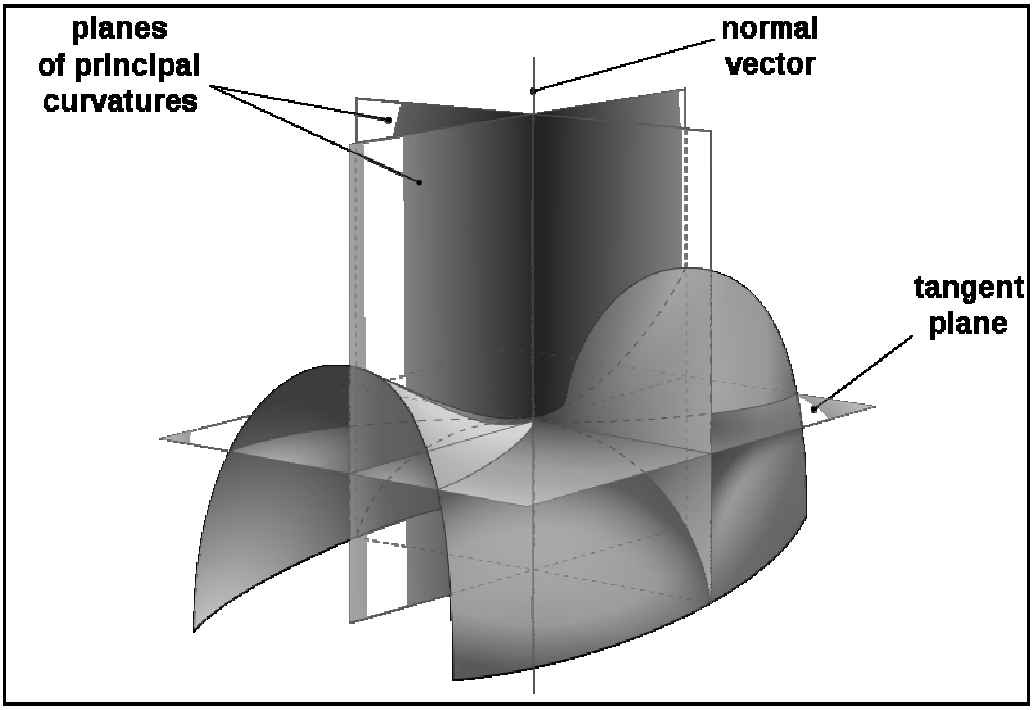


Figure 2: Illustration of several planes whose intersection with a surface define the curves whose curvature is a normal curvature (Wikipedia).

These are not unique, since they depend on the direction of the plane containing $\mathbf{n}$ that is used in the construction. The surface's *principal curvatures*, κ_1 and κ_2, are defined at each point as the maximum and minimum values taken by the normal curvatures as the direction of this plane is varied.

The surface's *mean curvature*, H, and *Gaussian curvature*, K, are then defined as the arithmetic and geometric means of κ_1 and κ_2:

$$H = \frac{1}{2}(\kappa_1 + \kappa_2) \quad \text{and} \quad K = \kappa_1 \kappa_2 \,. \tag{1.36}$$

Although it is not clear from its definition, Gauss' *Theorema Egregium* states that the Gaussian curvature can be determined purely in terms of lengths and angles measured within the surface — that is, in terms of the induced metric γ_{ab} and its derivatives — and so is a property intrinsic to the surface itself (as opposed to an *extrinsic* property that depends on how the surface is embedded into the external $\mathbb{R}_3$).

> **Exercise 4:** Show that the principal curvatures for the plane $\mathbf{r}(u,v) = \mathbf{e}_x\, u + \mathbf{e}_y\, v$ are $\kappa_1 = \kappa_2 = 0$. Repeat for the cylinder $\mathbf{r}(u,v) = a(\mathbf{e}_x \cos u + \mathbf{e}_y \sin u) + \mathbf{e}_z v$ to show they are $\kappa_1 = 0$ and $\kappa_2 = 1/a$. Finally, show that for the sphere $\mathbf{r}(u,v) = a(\mathbf{e}_x \sin u \cos v + \mathbf{e}_y \sin u \sin v + \mathbf{e}_z \cos u)$, the principal curvatures are equal and positive: $\kappa_1 = \kappa_2 = 1/a$.

Changing the parametrization

Notice that the discussion so far did not need to provide any details about the kinds of parameters, (u, v), used to specify the surface. Before generalizing the above discussion to more general spaces, it is worth first digressing briefly about how quantities change as the coordinates used to describe them change.

Contravariant vectors

When describing a surface, $\mathbf{r}(u, v)$, we saw that the inscribed curves along which the parameters $u^a = \{u, v\}$ vary could be used to provide a natural basis, $\mathbf{t}_a = \partial\mathbf{r}/\partial u^a$, for the surface's tangent plane. Because this forms a basis, it can be used to define the components of *any* vector at all that is tangent to the surface:

$$\mathbf{c} = c^a\,\mathbf{t}_a = c^u\,\mathbf{t}_u + c^v\,\mathbf{t}_v\,, \tag{1.37}$$

Suppose we now change our parametrization of the surface, defining new parameters $u^{a'}(u, v) = \{u'(u, v), v'(u, v)\}$ that provide equally good labels for points on the surface: $\mathbf{r}(u, v) = \mathbf{r}(u'(u, v), v'(u, v))$. Provided that the new parameters are really independent of one another (more about this below), the tangents to these new parameter curves define a new basis, $\mathbf{t}_{a'} = \partial\mathbf{r}/\partial u^{a'}$, of the same tangent plane, in terms of which the same vector $\mathbf{c}$ has the expansion

$$\mathbf{c} = c^{a'}\,\mathbf{t}_{a'} = c^{u'}\,\mathbf{t}_{u'} + c^{v'}\,\mathbf{t}_{v'}\,. \tag{1.38}$$

To obtain the relation between the coefficients $c^{a'}$ and c^a we relate the two sets of tangent bases to one another, using the chain rule:

$$\mathbf{t}_a = \frac{\partial\mathbf{r}}{\partial u^a} = \frac{\partial u^{a'}}{\partial u^a}\,\frac{\partial\mathbf{r}}{\partial u^{a'}} = \frac{\partial u^{a'}}{\partial u^a}\,\mathbf{t}_{a'}\,, \tag{1.39}$$

and so

$$\mathbf{c} = c^a\,\mathbf{t}_a = c^a\,\frac{\partial u^{a'}}{\partial u^a}\,\mathbf{t}_{a'} \tag{1.40}$$

which implies

$$c^{a'} = c^a\,\frac{\partial u^{a'}}{\partial u^a}\,. \tag{1.41}$$

Components c^a that transform in this way under a change of parametrization are called *contravariant* components, and $\mathbf{c}$ would be called a contravariant vector.

An earlier-mentioned proviso to this discussion was the requirement that the new coordinates be independent of one another and so provide a faithful parametrization of the surface. Eq. (1.39) provides a local criterion for when this is so, since it is equivalent to asking when the new pair of tangent vectors are linearly independent of

one another (as is required if they are to form a basis). Since eq. (1.39) can equally well be written in matrix notation as

$$\mathbf{t} = J\,\mathbf{t}'\,, \tag{1.42}$$

with

$$\mathbf{t} = \begin{pmatrix} \mathbf{t}_u \\ \mathbf{t}_v \end{pmatrix}\,, \quad \mathbf{t}' = \begin{pmatrix} \mathbf{t}_{u'} \\ \mathbf{t}_{v'} \end{pmatrix} \quad \text{and} \quad J = \begin{pmatrix} \partial u'/\partial u & \partial v'/\partial u \\ \partial u'/\partial v & \partial v'/\partial v \end{pmatrix}\,, \tag{1.43}$$

the new basis is linearly independent if and only if the matrix J is invertible, or equivalently if its determinant, $\mathcal{J} = \det J$ — the *Jacobian* of the transformation $(u, v) \to (u', v')$ — is nonzero: $\mathcal{J} \neq 0$.

Covariant Vectors

There is an alternative way of using parameters on a surface to describe vectors that are tangent to the surface. Instead of defining a set of basis vectors that are tangent to lines along which one parameter varies, $\mathbf{t}_a = \partial\mathbf{r}/\partial u^a = (\partial x^i/\partial u^a)\,\mathbf{e}_i$, one can instead define a basis of vectors, $\mathbf{s}^a$, using the normals to the surfaces along which one of the parameters is held constant. That is we ask the basis $\mathbf{s}^a$ to satisfy the defining condition

$$\mathbf{s}^a \cdot \mathbf{t}_b = \delta^a{}_b\,, \tag{1.44}$$

where, as before, the Kronecker symbol satisfies $\delta^a{}_b = 1$ if $a = b$ and vanishes otherwise. Such a basis is often called a basis *dual* to the basis of tangent vectors.

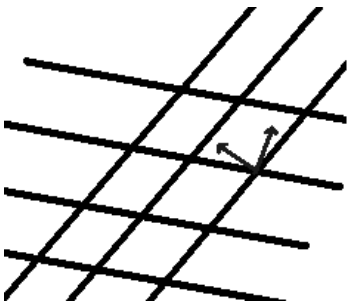


Figure 3: An example where normals to coordinate surfaces (small arrows) are not equivalent to tangents to coordinate directions (lines).

Although these two definitions give the same bases in many of the simple coordinates commonly used (like rectangular or polar coordinates), they need not always do so. An example of an 'oblique' set of coordinates, where these two definitions would not agree, is given by the parameters on a plane defined by $\mathbf{r}(u, v) = \mathbf{a}\,u + \mathbf{b}\,v$, when the vectors $\mathbf{a}$ and $\mathbf{b}$ are not orthogonal to one another. A cartoon of these coordinates is given in figure 3. In general curvilinear coordinates both bases are possible, and most importantly, transform differently when the surface is reparameterized $(u, v) \to (u'v')$.

Since the $\mathbf{s}^a$ form a basis for the surface's tangent plane, a general vector, $\mathbf{c}$, tangent to the surface can be expanded

$$\mathbf{c} = c_b\,\mathbf{s}^b\,, \tag{1.45}$$

and the coefficients in this expansion are given by

$$\mathbf{c} \cdot \mathbf{t}_a = c_b \, \mathbf{s}^b \cdot \mathbf{t}_a = c_b \, \delta^b{}_a = c_a \, . \tag{1.46}$$

These components change if the parameters used to label the surface are changed, $(u, v) \to (u'(u, v), v'(u, v))$, but in a different way than did the components c^a arising when $\mathbf{c}$ is expanded directly in terms of the $\mathbf{t}_a$'s. Using eq. (1.39) with eq. (1.46) gives

$$c_{a'} = \mathbf{c} \cdot \mathbf{t}_{a'} = \frac{\partial u^b}{\partial u^{a'}} \, \mathbf{c} \cdot \mathbf{t}_b = \frac{\partial u^b}{\partial u^{a'}} \, c_b \, , \tag{1.47}$$

where we use that the partial derivatives $\partial u^b / \partial u^{a'}$ make up the elements of the matrix J^{-1} that is inverse to the matrix J whose elements are $\partial u^{a'} / \partial u^b$. Equivalently, using the Einstein summation convention we use the identity

$$\frac{\partial u^b}{\partial u^{a'}} \frac{\partial u^{a'}}{\partial u^c} = \delta^b{}_c \quad \text{and its partner} \quad \frac{\partial u^{b'}}{\partial u^a} \frac{\partial u^a}{\partial u^{c'}} = \delta^{b'}{}_{c'} \, . \tag{1.48}$$

Coefficients that transform as in eq. (1.47) are said to transform as *covariant* components.

Tensors

Since the first fundamental form, γ_{ab}, is so central to the geometry on a surface, it is worth knowing how it transforms when the parameters labelling the surface are changed. Keeping in mind its definition in terms of the distance, $\mathrm{d}s$, along the surface, eq. (1.27), and recognizing that a physical quantity like $\mathrm{d}s$ must be parameter independent shows that if $(u, v) \to (u', v')$, then the chain rule implies

$$\mathrm{d}s^2 = \gamma_{ab} \, \mathrm{d}u^a \mathrm{d}u^b = \gamma_{ab} \frac{\partial u^a}{\partial u^{c'}} \frac{\partial u^b}{\partial u^{d'}} \, \mathrm{d}u^{c'} \mathrm{d}u^{d'} \, , \tag{1.49}$$

which shows

$$\gamma_{c'd'}(u', v') = \gamma_{ab}(u(u', v'), v(u', v')) \frac{\partial u^a}{\partial u^{c'}} \frac{\partial u^b}{\partial u^{d'}} \, . \tag{1.50}$$

Since this looks like two copies of the transformation rule, eq. (1.47), the quantity γ_{ab} is said to transform like a covariant *tensor* of rank 2.

More generally, if something having many indices transforms under a change of parameters like

$$T^{a'_1 .. a'_k}{}_{b'_1 .. b'_\ell}(u', v') = T^{c_1 .. c_k}{}_{d_1 .. d_\ell}(u(u', v'), v(u', v')) \frac{\partial u^{a'_1}}{\partial u^{c_1}} \cdots \frac{\partial u^{a'_k}}{\partial u^{c_k}} \frac{\partial u^{d_1}}{\partial u^{b'_1}} \cdots \frac{\partial u^{d_\ell}}{\partial u^{b'_\ell}} \, , \tag{1.51}$$

is called a tensor of covariant rank ℓ and contravariant rank k. The special case of something which has no indices, such as an inner product between two vectors tangent to a surface,

$$\mathbf{m} \cdot \mathbf{n} = (m^a \, \mathbf{t}_a) \cdot (n^b \, \mathbf{t}_b) = m^a n^b \frac{\partial \mathbf{r}}{\partial u^a} \cdot \frac{\partial \mathbf{r}}{\partial u^b} = \gamma_{ab} \, m^a n^b \, , \tag{1.52}$$

transforms according to

$$\gamma_{a'b'}\, m^{a'} n^{b'} = \left(\gamma_{cd}\, \frac{\partial u^c}{\partial u^{a'}}\, \frac{\partial u^d}{\partial u^{b'}}\right) \left(m^e\, \frac{\partial u^{a'}}{\partial u^e}\right) \left(n^f\, \frac{\partial u^{b'}}{\partial u^f}\right) = \gamma_{ef}\, m^e n^f\,, \tag{1.53}$$

(which uses eq. (1.48) twice) and is called a scalar.

The reason tensors like this are important is that physical laws cannot depend on our arbitrary choice of how we parameterize a surface. A physical statement — like $\mathbf{F} = m\,\mathbf{a}$, say — directly relates physical objects, like vectors, distances or inner products. And although the components of each quantity like $\mathbf{F}$, m and $\mathbf{a}$ can individually change when different parameters or bases are used, it is always true that both sides of the equality transform in precisely the same way. Thus it is important for Newton's Law that $\mathbf{F}$ and the product of m times $\mathbf{a}$ both transform as vectors.

We similarly demand on curved space that any reasonable physical law must have the form $A = B$, where both sides of the equation are tensors of precisely the same type. This ensures that once we know the components ${A^{a_1..a_k}}_{b_1..b_\ell}$ and ${B^{a_1..a_k}}_{b_1..b_\ell}$ are equal in a particular basis, it is *automatic* that they will also be equal in any other basis we should choose to examine.

1.2 General curved space

We are now ready to kick away the crutch of embedding surfaces into flat $\mathbb{R}_3$ and formulate directly what non-Euclidean geometry might look like in three (or more) dimensions. The key in doing so is to focus on those relations derived above for surfaces that do not make any reference at all to how the surface is situated within its embedding space.

Tensors and curvilinear coordinates

We start by choosing an arbitrary set of coordinates, x^i, to label the points in three dimensions, without requiring that these coordinates be the usual rectangular $\{x, y, z\}$. For instance we could instead use spherical polar coordinates $x^i = \{r, \theta, \phi\}$, or any other choice of coordinates which happens to suit our purposes.[1]

Just as for surfaces we can also define curves 'inscribed' within our space by specifying how the coordinates vary along the curve: $x^i(u) = \{x^1(u), x^2(u), x^3(u)\}$. At each point P in our three-dimensional space we can define a *tangent space*, T_P —

[1] As a technical point, it is not necessary that any one choice of coordinates describe *all* of the points in space. It is sufficient to have a collection of coordinate choices which cover the entire space once taken together, with sufficient overlap between pairs of coordinate patches to allow the results of measurements to be translated from one set of coordinates to another.

i.e. a generalized tangent plane — comprising the vector space spanned by all of the tangents at P to the curves that pass through P.

A choice of coordinates provides a natural basis for describing vectors that lie within the tangent space at each point. This can be taken to be defined by the vectors $\mathbf{t}_i$ that are tangent to the curves along which only one of the coordinates varies. Notice that this basis need not be normalized or mutually orthogonal, although it must be linearly independent and complete.

In terms of the basis $\mathbf{t}_i$, the tangent, $\mathbf{t}$, to any other curve defined by $x^i(w)$ has components

$$\mathbf{t} = \frac{\mathrm{d}x^i}{\mathrm{d}w}\,\mathbf{t}_i\,. \tag{1.54}$$

These components define a contravariant vector, in the sense that if we change coordinates from x^i to $x^{i'}$, the components $\mathbf{t}$ in the new basis vectors, $\mathbf{t}_{i'}$, are given by

$$\frac{\mathrm{d}x^{i'}}{\mathrm{d}w} = \frac{\partial x^{i'}}{\partial x^j}\,\frac{\mathrm{d}x^j}{\mathrm{d}w}\,. \tag{1.55}$$

Such a coordinate transformation is only well-defined if the matrix whose entries are $\partial x^{i'}/\partial x^j$ is invertible.

A contravariant tensor, $\mathbf{T}$, having rank p is similarly defined to have components involving p indices, that transform under a coordinate change according to

$$T^{i'_1..i'_p}(x') = T^{j_1..j_p}(x(x'))\,\frac{\partial x^{i'_1}}{\partial x^{j_1}}\cdots\frac{\partial x^{i'_p}}{\partial x^{j_p}}\,. \tag{1.56}$$

Metrics

We now come to the central concept. The essence of the geometry is determined by specifying a notion of distance between points within the space. This is done by giving the *metric*, $g_{ij}(x) = g_{ji}(x)$, which is a symmetric three-by-three positive-definite matrix whose entries are a function of position. $g_{ij}(x)$ is defined to give the distance between two infinitesimally displaced points, situated at x^i and $x^i + \mathrm{d}x^i$, as

$$\mathrm{d}s = \sqrt{g_{ij}(x)\,\mathrm{d}x^i\,\mathrm{d}x^j}\,. \tag{1.57}$$

The square root is always real because g_{ij} is positive definite, and $\mathrm{d}s = 0$ only occurs if $\mathrm{d}x^i = 0$. This last equation is more commonly written without the square root as

$$\mathrm{d}s^2 = g_{ij}(x)\,\mathrm{d}x^i\,\mathrm{d}x^j\,. \tag{1.58}$$

Besides providing a notion of distance, the metric provides a natural way to define the angle between two curves. This is done by using the metric to define an inner product between the tangent vectors of the two curves at their point of intersection. That is, suppose the curves $x^i_1(u)$ and $x^i_2(v)$ both pass through the point P, then

their tangent vectors, $\mathbf{m}_1$ and $\mathbf{m}_2$ respectively have components $\mathrm{d}x_1^i/\mathrm{d}u$ and $\mathrm{d}x_2^i/\mathrm{d}v$. Guided by eq. (1.35), we can then define the intersection angle between the two curves as

$$\cos\theta = \frac{\mathbf{m}_1 \cdot \mathbf{m}_2}{\sqrt{(\mathbf{m}_1 \cdot \mathbf{m}_2)(\mathbf{m}_2 \cdot \mathbf{m}_2)}}\,, \tag{1.59}$$

evaluated at P, where the inner product is defined in terms of the vector components as

$$\mathbf{a} \cdot \mathbf{b} = g_{ij}\, a^i\, b^j\,. \tag{1.60}$$

Notice that in particular the inner product of the tangents of the two curves $x_1^i(u)$ and $x_2^i(v)$ becomes

$$\mathbf{m}_1 \cdot \mathbf{m}_2 = g_{ij}\, \frac{\mathrm{d}x_1^i}{\mathrm{d}u}\, \frac{\mathrm{d}x_2^j}{\mathrm{d}v}\,, \tag{1.61}$$

and so in particular, for basis vectors defined as tangents to the coordinate lines themselves we have

$$\mathbf{t}_i \cdot \mathbf{t}_j = g_{ij}\,. \tag{1.62}$$

Having a notion of angles also means we know what it means for vectors to be orthogonal: $\mathbf{a} \cdot \mathbf{b} = 0$. This then allows the definition of the second natural basis for vectors, $\mathbf{s}^i$, in terms of the normals to the surfaces on which one coordinate is held fixed. Such a dual basis must satisfy $\mathbf{s}^i \cdot \mathbf{t}_j = \delta^i{}_j$, and so if a vector $\mathbf{m}$ is expanded $\mathbf{m} = m^i\, \mathbf{t}_i$ and $\mathbf{m} = m_i\, \mathbf{s}^i$, then the components are given by

$$m_i = \mathbf{m} \cdot \mathbf{t}_i = (m^j\, \mathbf{t}_j) \cdot \mathbf{t}_i = m^j\, g_{ij}\,. \tag{1.63}$$

It is convenient to define the quantities g^{ij} as the components of the matrix that is inverse to the matrix whose components are g_{ij}. Such a matrix always exists because the fact that g_{ij} is positive definite excludes the possibility of it having a zero eigenvector, and so not having an inverse. With this definition we have

$$g^{ij}\, g_{jk} = \delta^i{}_k\,, \tag{1.64}$$

and so multiplying eq. (1.63) by g^{ik} (including the implied sum over i, from the Einstein summation convention) gives

$$m^k = g^{ki}\, m_i\,. \tag{1.65}$$

Notice that its definition, together with the invariance of the distance element, $\mathrm{d}s$, implies that under a coordinate transformation g_{ij} transforms as

$$g_{i'j'} = g_{kl}\, \frac{\partial x^k}{\partial x^{i'}}\, \frac{\partial x^l}{\partial x^{j'}}\,, \tag{1.66}$$

what is called the transformation of a covariant tensor of rank 2. Similarly the *covariant* components, m_i, of a vector $\mathbf{m}$ transform as (compare with eq. (1.56))

$$m_{i'} = m_j \, \frac{\partial x^j}{\partial x^{i'}} \,, \tag{1.67}$$

which is a covariant tensor of rank 1, or one-form. The transformation properties of covariant tensors of higher rank can be similarly defined.

Geodesics

Returning to the main line of development, following the example of curves on a surface, we now define a *geodesic* as the curve that minimizes the distance between two points. Such curves are the natural generalization of the straight lines of Euclidean geometry.

To determine the local equations that govern geodesics, we must first find an expression for the distance between two points, A and B, that is to be minimized. If this distance is measured along a curve, $x^i(u)$, that connects them, the distance may be found by integrating the infinitesimal definition, eq. (1.57), in the form

$$s_{AB} = \int_{u_A}^{u_B} \mathrm{d}u \, \frac{\mathrm{d}s}{\mathrm{d}u} = \int_{u_A}^{u_B} \mathrm{d}u \, \sqrt{g_{ij}(x(u)) \, \frac{\mathrm{d}x^i}{\mathrm{d}u} \, \frac{\mathrm{d}x^j}{\mathrm{d}u}} = \int_{u_A}^{u_B} \mathrm{d}u \, \sqrt{g_{ij}(x(u)) \, \dot{x}^i \, \dot{x}^j} \,. \tag{1.68}$$

This introduces the simplifying notation $\dot{x}^i := \mathrm{d}x^i/\mathrm{d}u$.

If $x^i(u)$ were the curve of minimum length, then the quantity s_{AB} should be stationary with respect to small changes, $x^i(u) \to x^i(u) + \delta x^i(u)$, to the curve, at least to first order in $\delta x^i(u)$. Such variations must vanish in the same way that small changes to a function, $f(x)$, vanish to linear order in δx if they are evaluated at a function's minimum, $x = x_m$: $f(x_m + \delta x) - f(x_m) \simeq f'(x_m) \, \delta x = 0$. The conceptual difference here is that the length s_{AB} is a *functional* that depends on the shape of the entire curve, $x^i(u)$, and not simply on its value at a single point, like A or B.

To see what it means for s_{AB} to be stationary, let us write it as $s_{AB}[x(u)]$, to emphasize that it depends on the shape of the curve $x^i(u)$ in addition to the endpoints. We then evaluate the difference, $\delta s_{AB} = s_{AB}[x(u)+\delta x(u)] - s_{AB}[x(u)]$, and expand the result out to linear order in $\delta x^i(u)$, using $g_{ij}(x(u) + \delta x(u)) \simeq g_{ij}(x(u)) + \delta x^k(u) \, \partial_k g_{ij}(x(u))$ in eq. (1.68) to find

$$\begin{aligned} \delta s_{AB} &= \frac{1}{2} \int_{u_A}^{u_B} \mathrm{d}u \, \left(\frac{\delta x^k \, \partial_k g_{ij} \, \dot{x}^i \dot{x}^j + g_{ij} \, \delta \dot{x}^i \, \dot{x}^j + g_{ij} \, \dot{x}^i \delta \dot{x}^j}{\sqrt{g_{ij} \, \dot{x}^i \, \dot{x}^j}} \right) \\ &= \left[\frac{g_{ij} \, \dot{x}^i \delta x^j}{\dot{s}} \right]_A^B - \int_{u_A}^{u_B} \mathrm{d}u \, \left(\frac{g_{ij} \, \delta x^j}{\dot{s}} \right) \left[\ddot{x}^i + \Gamma^i_{kl} \dot{x}^k \dot{x}^l - \left(\frac{\ddot{s}}{\dot{s}} \right) \dot{x}^i \right] . \end{aligned} \tag{1.69}$$

This uses the notation $\dot{s} = \mathrm{d}s/\mathrm{d}u = \sqrt{g_{ij}\,\dot{x}^i\dot{x}^j}$ and $\ddot{s} = \mathrm{d}^2s/\mathrm{d}u^2$ and defines

$$\Gamma^i_{jk} = \Gamma^i_{kj} := \frac{1}{2}\,g^{il}\Big(\partial_j g_{kl} + \partial_k g_{jl} - \partial_l g_{jk}\Big)\,, \tag{1.70}$$

which is a useful quantity known as the *Christoffel symbol of the second kind*.[2] Finally, the last equality in eqs. (1.69) also arranges there to be no derivatives of δx^i by performing an integration by parts, using the identity

$$\frac{g_{ij}\,\dot{x}^i\delta\dot{x}^j}{\dot{s}} = \frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{g_{ij}\,\dot{x}^i\delta x^j}{\dot{s}}\right] - \delta x^j\,\frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{g_{ij}\,\dot{x}^i}{\dot{s}}\right]\,. \tag{1.71}$$

Exercise 5: Explicitly verify both equalities in eq. (1.69).

Now if $x^i(u)$ is a geodesic connecting A and B then s_{AB} must be minimized for all paths that connect A to B, so we must demand $\delta s_{AB} = 0$ for *any* choice for $\delta x^i(u)$ that satisfies $\delta x^i(A) = \delta x^i(B) = 0$. Since this last condition ensures $[g_{ij}\,\dot{x}^i\delta x^j/\dot{s}]^B_A = 0$, we ask what $x^i(u)$ must satisfy in order to ensure the vanishing of the integral in the last line of eq. (1.69).

But now comes the main point: because $\delta x^i(u)$ is arbitrary, we can choose it to vanish for all u apart from being positive in an arbitrarily narrow interval immediately surrounding some point $u = u_\star$. This insures that the integral receives contributions only from the integrand at $u_\star$, leading to the conclusion that the integrand must therefore vanish at this point. But since we can choose $\delta x^i(u)$ to peak about an arbitrary value of $u_\star$ and s_{AB} must be stationary with respect to *all* such variations, we can conclude that this integrand must vanish for all u when evaluated for any geodesic. But since g_{ij} is positive definite this is only possible if the square bracket vanishes, leading to the following *geodesic equation*:

$$\frac{D\dot{x}^i}{\mathrm{d}u} := \ddot{x}^i + \Gamma^i_{jk}\,\dot{x}^j\dot{x}^k = \left(\frac{\ddot{s}}{\dot{s}}\right)\,\dot{x}^i\,. \tag{1.72}$$

Exercise 6: Use the transformation properties,

$$g_{ij} = g_{i'j'}\,\frac{\partial x^{i'}}{\partial x^i}\,\frac{\partial x^{j'}}{\partial x^j} \quad\text{and}\quad g^{ij} = g^{i'j'}\,\frac{\partial x^i}{\partial x^{i'}}\,\frac{\partial x^j}{\partial x^{j'}}\,, \tag{1.73}$$

under the coordinate transformation $x^i \to x^{i'}$ to derive the transformation law

$$\Gamma^i_{jk} = \Gamma^{i'}_{j'k'}\,\frac{\partial x^i}{\partial x^{i'}}\,\frac{\partial x^{j'}}{\partial x^j}\,\frac{\partial x^{k'}}{\partial x^k} + \frac{\partial^2 x^{i'}}{\partial x^j\partial x^k}\,\frac{\partial x^i}{\partial x^{i'}}\,, \tag{1.74}$$

and show thereby that the Christoffel symbols are not tensors. Similarly show that although $\dot{x}^i$ transforms as a contravariant vector, $\ddot{x}^i$ does not. Finally, show that the sum $\ddot{x}^i + \Gamma^i_{jk}\,\dot{x}^j\,\dot{x}^k$ *does* transform as a contravariant vector, ensuring that if it vanishes in one set of coordinates, it must also vanish in all others.

[2] The *Christoffel symbol of the first kind* is $[i,jk] := g_{il}\Gamma^l_{jk} = \frac{1}{2}(\partial_j g_{ik} + \partial_k g_{ij} - \partial_i g_{jk})$.

The special case where $\ddot{s} = 0$ (and so $u = as + b$ for constants a and b) is called an *affinely-parameterized* geodesic, which satisfies

$$\ddot{x}^i + \Gamma^i_{jk}\, \dot{x}^j \dot{x}^k = 0\,. \tag{1.75}$$

Exercise 7: Use the explicit form computed earlier for the metric on a 2-sphere of radius a, $\mathrm{d}s^2 = a^2(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2)$ in spherical polar coordinates, to show that the only nonzero Christoffel symbols, Γ^a_{bc}, in these coordinates are:

$$\Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta \quad \text{and} \quad \Gamma^\phi_{\theta\phi} = \Gamma^\phi_{\phi\theta} = \cot\theta\,. \tag{1.76}$$

Use this to show that the equations for an affinely parameterized geodesic, $\{\theta(s), \phi(s)\}$, on a sphere are

$$\begin{aligned} \frac{\mathrm{d}^2\theta}{\mathrm{d}s^2} - \sin\theta\cos\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2 &= 0 \\ \text{and} \quad \frac{\mathrm{d}^2\phi}{\mathrm{d}s^2} + 2\,\cot\theta\left(\frac{\mathrm{d}\theta}{\mathrm{d}s}\right)\left(\frac{\mathrm{d}\phi}{\mathrm{d}s}\right) &= 0\,. \end{aligned} \tag{1.77}$$

Show that the solutions to these equations are great circles. (HINT: It will simplify your life to choose your coordinates so that the two points connected by your geodesic are chosen to lie on the sphere's equator.)

Curvature

Since the metric, g_{ij}, can take different forms in different coordinate systems, transforming as eq. (1.66), when confronted with a complicated metric it is important to know how much of the complication comes from complications in the underlying geometry and how much simply arises due to the use of a complicated coordinate system. For instance, the two following metrics describe the same physical distance relation,

$$\begin{aligned} \mathrm{d}s^2 &= \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \\ \mathrm{d}s^2 &= \mathrm{d}x^2 + x^2\,\mathrm{d}y^2 + x^2\sin^2 y\,\mathrm{d}z^2\,, \end{aligned} \tag{1.78}$$

but simply do so with different coordinate choices (rectangular and spherical coordinates, respectively). Given an arbitrary metric,

$$\begin{aligned} \mathrm{d}s^2 = {} & f(x,y,z)\,\mathrm{d}x^2 + g(x,y,z)\,\mathrm{d}y^2 + h(x,y,z)\,\mathrm{d}z^2 \\ & + 2j(x,y,z)\,\mathrm{d}x\,\mathrm{d}y + 2k(x,y,z)\,\mathrm{d}x\,\mathrm{d}z + 2l(x,y,z)\,\mathrm{d}y\,\mathrm{d}z\,, \end{aligned} \tag{1.79}$$

it is useful to have a criterion for deciding when a coordinate transformation exists, $x = x(u,v,w)$, $y = y(u,v,w)$ and $z = z(u,v,w)$, that can put this into a simple form like $\mathrm{d}s^2 = \mathrm{d}u^2 + \mathrm{d}v^2 + \mathrm{d}w^2$, for which $g_{ij} = \delta_{ij}$.

In fact, at first sight it is tempting to conclude that it is *always* possible to perform such a transformation. After all, g_{ij} can be regarded as defining the components of a real symmetric matrix, $\mathbf{g}$, and the transformation rule, eq. (1.66), can be regarded as a similarity transformation,

$$\mathbf{g}' = \mathbf{S}\,\mathbf{g}\,\mathbf{S}^T\,, \tag{1.80}$$

where the superscript 'T' denotes transpose, and the matrix $\mathbf{S}$ has components $\partial x^i/\partial x^{j'}$. But any real symmetric matrix can always be made into the unit matrix with an appropriate choice of $\mathbf{S}$, since it can first be diagonalized using an orthogonal matrix, and then its diagonal elements can be rescaled to unity.

Although the above argument does show that it is always possible to choose coordinates so that $g_{ij} = \delta_{ij}$ *at any one point*, it does not follow that this can be done for an entire region of points at the same time (using the same coordinates). To see why, suppose that the required matrix, $\mathbf{S}(x)$, is found, that when used in eq. (1.80) ensures $g_{i'j'} = \delta_{i'j'}$. This can only be accomplished by a coordinate transformation if there exist coordinates $x^i(x')$ such that

$$\frac{\partial x^i}{\partial x^{j'}} = S_{j'}{}^i\,. \tag{1.81}$$

But there can be *integrability conditions* that can obstruct being able to integrate these equations to find the required $x^i(x')$. For instance, if a solution is to exist it must satisfy $\partial^2 x^i/\partial x^{j'}\partial x^{k'} = \partial^2 x^i/\partial x^{k'}\partial x^{j'}$, so no solution is possible if it should happen that $\partial_{k'} S_{j'}{}^i \neq \partial_{j'} S_{k'}{}^i$.

It turns out that the freedom to change coordinates is sufficient to arrange that both $g_{ij} = \delta_{ij}$ at any particular point, P, *and* that $\Gamma^i_{jk} = 0$ at the same point. Such coordinates are called *Gaussian normal coordinates* at P. Although this can be arranged at any particular point, it cannot in general be arranged *simultaneously* at all points in an open region around a given point.

Flat space

If there exist a set of coordinates for which $g_{ij} = \delta_{ij}$ within a entire region, $\mathcal{R}$, (such as is possible for a 2D plane in $\mathbb{R}_3$, say) then this region is said to be *flat*. A necessary and sufficient condition for this to be possible is that the following tensor:

$$R^i{}_{jkl} = \partial_k \Gamma^i_{jl} - \partial_l \Gamma^i_{jk} + \Gamma^i_{km}\Gamma^m_{jl} - \Gamma^i_{lm}\Gamma^m_{jk}\,. \tag{1.82}$$

must vanish, $R^i{}_{jkl} = 0$, everywhere in $\mathcal{R}$. The tensor $R^i{}_{jkl}$ is called the *Riemann curvature tensor*. (For a proof of this see, for example, the text by Weinberg listed in the bibliography.)

Exercise 8: Use the transformation properties for Γ^i_{jk} derived in Exercise 6 to show that $R^i{}_{jkl}$ transforms as a tensor, ensuring that it suffices to show that the Riemann tensor vanishes in one coordinate system to conclude that it must vanish in them all.

Exercise 9: Use its definition, eq. (1.82), to prove the following symmetry properties of $R_{ijkl} := g_{im} R^m{}_{jkl}$:

$$R_{ijkl} = R_{klij} = -R_{jikl} = -R_{ijlk}\,, \tag{1.83}$$

and

$$R_{ijkl} + R_{iklj} + R_{iljk} = 0\,. \tag{1.84}$$

It is a theorem that $R^i{}_{jkl}$ is the unique tensor that can be constructed only using the metric and its first and second derivatives at a point. Two related tensors can be built from the Riemann tensor by taking traces using the metric. These are the *Ricci tensor*, R_{ij}, and the *Ricci scalar*, R, defined by

$$R_{ij} := R^k{}_{ikj} \quad \text{and} \quad R := g^{ij}\, R_{ij} = g^{ij}\, R^k{}_{ikj}\,. \tag{1.85}$$

Exercise 10: Use the Christoffel symbols computed in exercise 1.2 to compute explicitly the Riemann tensor for a 2-sphere in spherical polar coordinates. Show in this way that its only nonzero component (up to symmetries) is

$$R^\theta{}_{\phi\theta\phi} = \sin^2\theta\,, \tag{1.86}$$

and so $R_{ijkl} = (g_{ik}g_{jl} - g_{il}g_{jk})/a^2$, while $R_{ij} = (1/a^2)\, g_{ij}$ and $R = 2/a^2 = 2K$, where K is the Gaussian curvature.

2. Special Relativity and flat spacetime

Once it is recognized that space can be curved its geometrical properties fall into the domain of experiments, that can ask whether it is curved and how this curvature might manifest itself physically. And if spacetime geometry is a physical quantity, one might also seek the physical laws that govern its properties. General Relativity is the result to which such a search leads.

As a first step towards making the connection between gravity and a physical theory of geometry, it is important to realize that it is not just the geometry of three-dimensional space that is in play; rather it is the geometry of four-dimensional *spacetime*, defined as the union of all possible events in space for all times. Four

coordinates — three spatial coordinates, x^i with $i = 1, 2, 3$, as well as time, $x^0 = t$ — are required to specify positions of events in spacetime. These are collectively denoted by x^μ with Greek indices like μ, ν running from 0 to 3: $x^\mu = \{x^0, x^i\} = \{t, x^1, x^2, x^3\}$.

Within such a picture, point particles can be regarded as sweeping out *world lines*, $x^\mu(u)$, through spacetime as time evolves. For instance, if we use time, t, itself to parameterize such a world line, then a particle that sits motionless at the fixed position $\mathbf{r} = \mathbf{a}$ (or $x^i = a^i$, for constants a^i) has world line $x^\mu(t) = \{t, a^i\}$. The world line of a particle moving at constant speed $\mathbf{v}$ might similarly be written $x^\mu(t) = \{t, v^i t\}$, while that of a particle executing uniform circular motion in the (x, y) plane would be $x^\mu(t) = \{t, a\cos(\omega t), a\sin(\omega t), 0\}$, and so on.

2.1 Minkowski spacetime

If gravity is to be regarded as the physics of curved spacetime, we might expect that the absence of gravity should be describable as the physics of flat spacetime. This section aims to show that this is true, inasmuch as the non-gravitational physics of special relativity is efficiently expressed in terms of the geometry of flat spacetime.

Inertial observers and the Minkowski metric

From this point of view, special relativity can be regarded as describing the motion of particles in a spacetime that is endowed with a metric, $\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu \mathrm{d}x^\nu$, for which coordinates can be found for which $g_{\nu\lambda}$ is a constant (and so for which the Christoffel symbols and curvature vanish $\Gamma^\mu_{\nu\lambda} = R^\mu{}_{\nu\lambda\rho} = 0$). Observers whose measurements are described by such coordinates are called *inertial observers*, and who are the observers for which the standard postulates of special relativity apply:

1. *Principle of Relativity:* All laws of nature take the same form when written in any inertial frame;

2. *Invariance of the Speed of Light:* All inertial observers measure precisely the same numerical value, $c = 299,792,458$ m/s, for the speed of light in vacuum.

We therefore require these observers to use rectangular coordinates in space, $x^i = \{x, y, z\}$, and to move relative to one another by at most a constant velocity.

> **Exercise 11:** Astronomers detect distant objects in the sky that appear to move faster than light — how is this possible? Consider a very distant object moving towards us at speed v at an angle θ to the line of sight. Suppose the object sends us two light rays that depart at times t and $t+\mathrm{d}t$, and these are received at times t' and $t' + \mathrm{d}t'$ (with all times measured in our rest frame), during which time the object moves a distance $\mathrm{d}x =$

$v \sin\theta \, \mathrm{d}t$ transverse to the line of sight. If the distance to the object when the first signal is emitted is $D = c\,(t' - t)$, show that the distance to the object when the second ray is sent is $D - \mathrm{d}D$ where $\mathrm{d}D = c(\mathrm{d}t - \mathrm{d}t') \simeq v \cos\theta \, \mathrm{d}t$, assuming $v \, \mathrm{d}t \ll D$. Use this to show that the apparent lateral speed of the object is

$$v_{\text{eff}} = \frac{\mathrm{d}x}{\mathrm{d}t'} = \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)\left(\frac{\mathrm{d}t}{\mathrm{d}t'}\right) \simeq \frac{v \sin\theta}{1 - (v/c)\cos\theta}\,, \tag{2.1}$$

which can satisfy $v_{\text{eff}} > c$ if θ is close to zero and v is close to (but smaller than) c.

Because all inertial observers measure the same value for c, it is worth defining our unit of distance to be light-seconds — *i.e.* the distance travelled by light in 1 second — so that $c = 1$ and the speed of any particle moving more slowly than light satisfies $0 \leq v < 1$. (Such units would not be useful if all inertial observers did not agree on the speed of light.) These units are used throughout the rest of these notes, and conversion of subsequent formulae to ordinary units is accomplished by inserting whatever factors of c are required to give the expression the correct dimensions. (*E.g.* for a result like $v = 0.2$ to have the dimensions of m/s, its right-hand-side must really be $0.2\,c$. Similarly, for E an energy, p a momentum and m a mass, $E = p$ becomes $E = p\,c$ and $E = m$ becomes $E = m\,c^2$.)

These observations guide us to choose the form taken for the metric to be one for which all inertial observers agree. This suggests the constant metric agreed on by inertial observers should be chosen to be the *Minkowski* metric, $\eta_{\mu\nu}$, defined by

$$\mathrm{d}s^2 = \eta_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu = -\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2\,,$$

and so for rectangular coordinates $\{x^0, x^1, x^2, x^3\} = \{t, x, y, z\}$, we have

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}. \tag{2.2}$$

Notice that this metric is *not* positive definite, unlike the metrics considered when thinking about the geometry of three-dimensional space. But $\mathrm{d}s^2$ *is* positive and agrees with our notion of distance in flat space if it is restricted to a purely spatial interval, along which $\mathrm{d}t = 0$. (The possibility that $\mathrm{d}s^2$ can be zero or negative is the main reason why the geometry of spacetime differs from that of the geometry of four-dimensional space.) If $\mathrm{d}s^2 > 0$ the interval is called *spacelike*, and will turn out

represent the a spatial distance along the interval for the particular inertial observers who see $\mathrm{d}t = 0$ along the interval.

By contrast, the situation $\mathrm{d}s^2 = 0$ describes the trajectory of a light ray. That is, $\mathrm{d}s = 0$ implies $\mathrm{d}t^2 = \mathrm{d}\ell^2$, where $\mathrm{d}\ell^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$ measures the spatial distance traversed. Clearly any such a trajectory satisfies $\mathrm{d}\ell/\mathrm{d}t = 1$, and so moves at the speed of light (since $c = 1$). The requirement that all inertial observers agree on the interval $\mathrm{d}s^2$ therefore includes as a special case the condition that all such observers agree on the speed of light *in vacuo*. An interval for which $\mathrm{d}s^2 = 0$ is called a *null* interval.

In the situation where $\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}\ell^2 < 0$, the interval corresponds to the world line of a trajectory of a particle moving at less than the speed of light, since $v^2 = (\mathrm{d}\ell/\mathrm{d}t)^2 = 1 + (\mathrm{d}s/\mathrm{d}t)^2 < 1$. In this case it is useful to define $\mathrm{d}\tau = \sqrt{-\mathrm{d}s^2}$, since this represents the proper time elapsed by the observer moving along this trajectory (for whom $\mathrm{d}\ell = 0$). For this reason intervals for which $\mathrm{d}s^2 < 0$ are called *timelike*.

Lorentz transformations

The transformations of special relativity may now be defined as those which do not change the Minkowski metric, eq. (2.2), since all such observers will agree on physical distances and so also agree on physical laws that are expressed in terms of them.

The resulting transformations are given by a combination of translations,

$$x^\mu \to x^\mu + a^\mu \,, \tag{2.3}$$

and linear transformations,

$$x^\mu \to \Lambda^\mu{}_\nu \, x^\nu \,, \tag{2.4}$$

where the constant matrices $\Lambda^\mu{}_\nu$ must satisfy

$$\eta_{\alpha\beta} \Lambda^\alpha{}_\mu \Lambda^\beta{}_\nu = \eta_{\mu\nu} \,. \tag{2.5}$$

The group of transformations defined by eqs. (2.3) through (2.5) is called the *Poincaré group*, while those defined by eqs. (2.4) and (2.5) alone are called the *Lorentz group*, or the group $O(3,1)$.

Spatial rotations provide a special case, for which

$$\Lambda^\mu{}_\nu = \begin{pmatrix} 1 & \\ & M^i{}_j \end{pmatrix} \,, \tag{2.6}$$

where $i, j = 1, 2, 3$ runs over purely spatial directions, and $M^i{}_j$ is an arbitrary 3×3 orthogonal matrix: $\delta_{ij} M^i{}_k M^j{}_l = \delta_{kl}$. The group of all such matrices is called $O(3)$.

For instance, for rotations about the z axis through an angle α we would have

$$(M_z)^i{}_j = \begin{pmatrix} \cos\alpha & \sin\alpha & \\ -\sin\alpha & \cos\alpha & \\ & & 1 \end{pmatrix} . \tag{2.7}$$

A second special case is given by a *boost*, which relates two inertial observers who move at constant speed relative to one another. For instance if the motion is along the x axis, then such a boost is described by

$$(\Lambda_x)^\mu{}_\nu = \begin{pmatrix} \cosh\beta & \sinh\beta & & \\ \sinh\beta & \cosh\beta & & \\ & & 1 & \\ & & & 1 \end{pmatrix} , \tag{2.8}$$

and β is a parameter related (more about which below) to the relative speed of the two observers who are related by the boost. Boosts along the y and z axes are similarly given by

$$(\Lambda_y)^\mu{}_\nu = \begin{pmatrix} \cosh\beta & & \sinh\beta & \\ & 1 & & \\ \sinh\beta & & \cosh\beta & \\ & & & 1 \end{pmatrix} \quad \text{and} \quad (\Lambda_z)^\mu{}_\nu = \begin{pmatrix} \cosh\beta & & & \sinh\beta \\ & 1 & & \\ & & 1 & \\ \sinh\beta & & & \cosh\beta \end{pmatrix} . \tag{2.9}$$

Exercise 12: Verify that the transformations (2.6) and (2.8) satisfy condition (2.5).

2.2 Inertial particle motion

Newton's first law states that a particle does not accelerate in the absence of external forces, and so in special relativity the spacetime trajectory (or world-line) of such an inertial particle (on which no forces act) is given by a straight line,

$$x^\mu(\tau) = a^\mu + v^\mu\, f(\lambda) , \tag{2.10}$$

where a^μ and v^μ are constant 4-vectors and λ is a parameter that labels the points along the curve (and so for which the otherwise arbitrary function satisfies $\mathrm{d}f/\mathrm{d}\lambda > 0$). For later purposes notice that any such a curve satisfies

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\lambda^2} = \frac{\mathrm{d}^2 f}{\mathrm{d}\lambda^2}\, v^\mu = \left(\frac{\mathrm{d}^2 f/\mathrm{d}\lambda^2}{\mathrm{d}f/\mathrm{d}\lambda}\right) \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} , \tag{2.11}$$

and so can be interpreted as a geodesic in flat spacetime (*c.f.* eq. (1.72)).

The interval measured along the trajectory is

$$\mathrm{d}s^2 = \eta_{\mu\nu}\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}\,\mathrm{d}\lambda^2 = (v\cdot v)\left(\frac{\mathrm{d}f}{\mathrm{d}\lambda}\right)^2\mathrm{d}\lambda^2\,, \tag{2.12}$$

so it follows that v^μ must satisfy $v\cdot v = \eta_{\mu\nu}\,v^\mu v^\nu < 0$ for a timelike trajectory, in which case the vector v^μ is also said to be timelike. (By contrast, for motion at the speed of light — such as for a photon — v^μ would instead be null: $v\cdot v = 0$.)

For motion slower than the speed of light we define the proper time, τ, as the distance measured along the trajectory, and so $\mathrm{d}s^2 = -\mathrm{d}\tau^2$, and it is convenient to use $\lambda = \tau$ as the parameter along the curve. In this case $u^\mu := \mathrm{d}x^\mu/\mathrm{d}\tau$ is called the 4-velocity of the trajectory, and eq. (2.12) then implies $u\cdot u = -1$. Writing its components as

$$\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = u^\mu = \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}, \frac{\mathrm{d}x}{\mathrm{d}\tau}, \frac{\mathrm{d}y}{\mathrm{d}\tau}, \frac{\mathrm{d}z}{\mathrm{d}\tau}\right) \tag{2.13}$$

$$= \frac{\mathrm{d}t}{\mathrm{d}\tau}\left(1, \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, \frac{\mathrm{d}z}{\mathrm{d}t}\right)\,, \tag{2.14}$$

the condition $u\cdot u = -1$ implies $\mathrm{d}t/\mathrm{d}\tau$ satisfies $(\mathrm{d}t/\mathrm{d}\tau)^2(1-\mathbf{v}^2) = 1$, where the velocity 3-vector, $\mathbf{v}$, is defined to have components $v^i = \mathrm{d}x^i/\mathrm{d}t$. We read off from this the *time dilation* that relates the proper time τ to the time t of the observer with respect to which the trajectory has velocity $\mathbf{v}$:

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} := \gamma = \frac{1}{\sqrt{1-\mathbf{v}^2}}\,. \tag{2.15}$$

where the condition $\mathrm{d}t/\mathrm{d}\tau > 0$ fixes the sign of the square root used in this expression.

We may now relate the parameter β appearing in a Lorentz boost to the speed, v, of the inertial observers involved, and thereby verify that eq. (2.8) describes a standard Lorentz transformation familiar from special relativity. To this end, suppose ${\Lambda^\mu}_\nu$ is the Lorentz boost which transforms from the frame of an observer at rest (and so whose 4-velocity is $u^\mu = (1,0,0,0)$) to the frame of an inertial observer moving with speed v along the x axis (and so whose 4-velocity is $u^\mu = (\gamma, \gamma v, 0, 0)$). Requiring that the Lorentz transformation of eq. (2.8) is the one that relates these two 4-velocities gives the parameter β in terms of the speed, v. That is, if

$$\begin{pmatrix}\gamma\\ \gamma v\\ 0\\ 0\end{pmatrix} = \begin{pmatrix}\cosh\beta & \sinh\beta & & \\ \sinh\beta & \cosh\beta & & \\ & & 1 & \\ & & & 1\end{pmatrix}\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\,, \tag{2.16}$$

then $\cosh\beta = \gamma$ and $\sinh\beta = \gamma v$, and so $\tanh\beta = v$. Notice that the definition $\gamma = (1-v^2)^{-1/2}$ is then equivalent to the identity $\cosh^2\beta - \sinh^2\beta = 1$. β is sometimes called the *rapidity* of the moving particle.

Exercise 13: Prove the identity $\Lambda_x(\beta_1)\Lambda_x(\beta_2) = \Lambda_x(\beta_1 + \beta_2)$ for the composition of two boosts along the x axis, as in eq. (2.8), and use this to show that the inverse of the matrix $\Lambda_x(\beta)$ is $\Lambda_x^{-1}(\beta) = \Lambda_x(-\beta)$. Use your result with the relation $v/c = \tanh\beta$ to derive the relativistic law for adding velocities: if $\beta = \beta_1 + \beta_2$ then

$$v = \frac{v_1 + v_2}{1 + v_1 v_2/c^2}\,. \tag{2.17}$$

Using this connection between β and v in the relation between the coordinates in these two frames, $x^{\mu'} = \Lambda^{\mu'}{}_{\nu} x^{\nu}$, or

$$\begin{pmatrix} t' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \cosh\beta & \sinh\beta & & \\ \sinh\beta & \cosh\beta & & \\ & & 1 & \\ & & & 1 \end{pmatrix} \begin{pmatrix} t \\ x \\ y \\ z \end{pmatrix}, \tag{2.18}$$

leads (temporarily replacing the factors of c) to the familiar expressions

$$t' = \frac{t + vx/c^2}{\sqrt{1 - v^2/c^2}}\,, \qquad x' = \frac{x + vt}{\sqrt{1 - v^2/c^2}}\,, \tag{2.19}$$

together with $y' = y$ and $z' = z$. The fact that these expressions imply that events sharing a common value for t are not the same as those sharing a common value for t' — *i.e.* the *relativity of simultaneity* — that makes it much more efficient to think in terms of spacetime, rather than space and time separately.

Exercise 14: Calculate the relation between the coordinates $\{t', x', y', z'\}$ and $\{t, x, y, z\}$ obtained by first performing a boost in the x direction with speed v followed by a boost in the y direction with speed u.

Lorentz tensors

Physical quantities in different inertial frames in special relativity transform as tensors with respect to Lorentz transformations

$$T^{\mu'_1..\mu'_p}{}_{\lambda'_1..\lambda'_q} = T^{\nu_1..\nu_p}{}_{\rho_1..\rho_q} \left(\Lambda^{\mu'_1}{}_{\nu_1} \cdots \Lambda^{\mu'_p}{}_{\nu_p}\right) \left(\Lambda^{\rho_1}{}_{\lambda'_1} \cdots \Lambda^{\rho_q}{}_{\lambda'_q}\right)\,. \tag{2.20}$$

As a result the Principle or Relativity is automatically satisfied if physical laws having the schematic form of tensor = tensor, since the tensor transformation rule ensures that if the law is true for any one frame, it must be true for them all.

For instance, the instantaneous 4-momentum of a particle having rest-mass m moving along a trajectory $x^\mu(\tau)$ transforms as a 4-velocity, that is defined in terms of the 4-velocity, eq. (2.13), by

$$p^\mu = m\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = m\,u^\mu\,. \tag{2.21}$$

The components of p^μ define the particle's instantaneous energy, $E = p^0$, and 3-momentum, p^i, and so (using the components for u^μ found earlier):

$$p^0 = E = m\,\gamma = \frac{m}{\sqrt{1-v^2}} \qquad \text{and} \qquad p^i = m\,\gamma\,v^i = \frac{mv^i}{\sqrt{1-v^2}}\,. \tag{2.22}$$

Notice that the condition $\eta_{\mu\nu}(\mathrm{d}x^\mu/\mathrm{d}\tau)(\mathrm{d}x^\nu/\mathrm{d}\tau) = -1$ implies $\eta_{\mu\nu}p^\mu p^\nu = p_\mu p^\mu = -m^2$, which is equivalent to the relativistic energy-momentum relation

$$E^2 = \mathbf{p}^2 + m^2\,. \tag{2.23}$$

To describe photons we take the limit $m \to 0$ and $\mathrm{d}\tau \to 0$, so that p^μ remains fixed and well-defined. (The velocity $\mathrm{d}x^\mu/\mathrm{d}\lambda$ is also well-defined, although it is no longer possible to choose proper time, τ, as the parameter along the world line.) The resulting 4-momentum satisfies $\eta_{\mu\nu}p^\mu p^\nu = p_\mu p^\mu = 0$, and so $E = |\mathbf{p}|$.

As an example of the utility of knowing that quantities like p^μ and u^μ transform as 4-vectors under Lorentz transformations, consider the following proof that

$$E = -u_\mu\,p^\mu = -\eta_{\mu\nu}\,u^\mu\,p^\nu\,, \tag{2.24}$$

gives the energy of a particle with 4-momentum p^μ as seen by an observer with 4-velocity u^μ. The proof starts by showing (by direct evaluation) that the result is trivially true in the simple special case where the observer is at rest, in which case $u^\mu = \{1,0,0,0\}$. To obtain the result for a general observer it suffices to recognize that the 4-vector transformation properties of u^μ and p^ν ensure that the quantity $u_\mu\,p^\mu$ is Lorentz invariant. That is, if $x^{\mu'} = \Lambda^{\mu'}{}_\nu x^\nu$ is the Lorentz transformation that takes us to the observer's rest frame, then $p^{\nu'} = \Lambda^{\nu'}{}_\lambda p^\lambda$ and $u^{\mu'} = \Lambda^{\mu'}{}_\rho u^\rho$, and so

$$\eta_{\mu'\nu'}u^{\mu'}p^{\nu'} = \eta_{\mu'\nu'}\,\Lambda^{\mu'}{}_\rho u^\rho\,\Lambda^{\nu'}{}_\lambda p^\lambda = \eta_{\rho\lambda}u^\rho p^\lambda\,, \tag{2.25}$$

which uses eq. (2.5). This ensures that all inertial observers must obtain the same thing for $u_\mu p^\mu$, and so it suffices to show that $E = -u_\mu\,p^\mu$ in the observer's rest frame to conclude it must be true for any frame.

2.3 Non-inertial motion

The geometry of flat space captures equally well the relativistic kinematics of particles that are not moving at constant speed.

Accelerated particles

For instance, consider an arbitrary trajectory, $x^\mu(\tau)$, that does not describe motion at constant velocity, such as the following trajectory describing a particle that accelerates along the x axis from rest at $x = 0$, until its speed reaches $v = v_{\max}$ at which point it then decelerates back to rest a distance ℓ away and then returns to $x = 0$, again at rest:

$$x^\mu(t) = \Big\{t, x(t), y(t), z(t)\Big\} = \left\{t, \ell \sin^2\left(\frac{v_{\max} t}{\ell}\right), 0, 0\right\} . \tag{2.26}$$

Here the inertial observer's time, t, is used to label points on the curve, with $0 \leq t \leq T = \pi\ell/v_{\max}$ describing the entire round trip. The turning point at $x = \ell$ is achieved at $t = \frac{1}{2}T$, and because the instantaneous particle speed seen by the inertial observer is

$$v(t) = \frac{\mathrm{d}x}{\mathrm{d}t} = v_{\max} \sin\left(\frac{2v_{\max} t}{\ell}\right) , \tag{2.27}$$

the maximum speed on the outbound leg takes place at $t = \frac{1}{4}T$.

The proper time measured by a clock riding with the particle along such a trajectory is

$$\mathrm{d}\tau^2 = -\mathrm{d}s^2 = -\eta_{\mu\nu}\, \mathrm{d}x^\mu(t)\mathrm{d}x^\nu(t) = \left[1 - v^2\,(t)\right] \mathrm{d}t^2 , \tag{2.28}$$

and so the 4-velocity and 4-acceleration become

$$\begin{aligned} u^\mu &= \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = \frac{\mathrm{d}t}{\mathrm{d}\tau}\,\frac{\mathrm{d}x^\mu}{\mathrm{d}t} = \frac{1}{\sqrt{1 - v^2\,(t)}}\Big\{1, v\,(t)\,, 0, 0\Big\} \\ \text{and} \quad a^\mu &:= \frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} = \frac{\mathrm{d}t}{\mathrm{d}\tau}\,\frac{\mathrm{d}u^\mu}{\mathrm{d}t} = \frac{\mathrm{d}v/\mathrm{d}t}{[1 - v^2(t)]^2}\Big\{v(t), 1, 0, 0\Big\} , \end{aligned} \tag{2.29}$$

with

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{2v_{\max}^2}{\ell}\, \cos\left(\frac{2v_{\max} t}{\ell}\right) . \tag{2.30}$$

In relativistic Newtonian mechanics the force responsible for this motion is described by a 4-vector, $F^\mu = m\, a^\mu$. Notice that all inertial observers must agree on the *proper acceleration* given by the Lorentz-invariant definition

$$a^2 := \eta_{\mu\nu}\, a^\mu a^\nu = a_\mu a^\mu = \frac{1}{[1 - v^2(t)]^3}\, \left(\frac{\mathrm{d}v}{\mathrm{d}t}\right)^2 . \tag{2.31}$$

Exercise 15: Compute the proper time, 4-velocity, 4-momentum and 4-acceleration for the following trajectories: (a) constant proper acceleration along the z axis, $x^\mu(u) = \{\ell \sinh(\alpha u), 0, 0, \ell \cosh(\alpha u)\}$, and ($b$) uniform circular motion in the x-y plane, $x^\mu(u) = \{t, d\cos(\omega t), d\sin(\omega t), 0\}$. What is the physical interpretation of the parameters ℓ, α, d and ω used in these trajectories?

Exercise 16: Suppose a family of light rays having frequency ω is sent parallel to the x-y plane at an angle θ to the x axis, and so has 4-momentum $k^\mu = \{\hbar\omega_\star, \hbar\omega_\star \cos\theta, \hbar\omega_\star \sin\theta, 0\}$. Show that this satisfies $k_\mu k^\mu = 0$, as it must if it is tangent to the trajectory of a light ray. Use the relation $E = \hbar\omega$ and $E = -\eta_{\mu\nu}\, u^\mu k^\nu$ to evaluate the frequency of the photons that is measured by observers moving along the accelerated trajectories in the previous exercise (Exercise 15).

Twin 'paradox'

The *Twin 'Paradox'* compares the time elapsed for two identical clocks (or twins), one of which travels along an accelerated trajectory as described above, while the other remains at rest at $x = 0$. The time elapsed for the motionless clock is simply the difference in t between the events when the two clocks separate and rejoin, and so is $\Delta t = t_f - t_i = \pi\ell/v_{\max} = T$, while the time elapsed by the moving clock is found by integrating eq. (2.28):

$$\Delta\tau = \int_0^T \mathrm{d}t\, \sqrt{1 - v^2(t)} = \int_0^T \mathrm{d}t\, \sqrt{1 - v_{\max}^2 \sin^2\left(\frac{2\pi t}{T}\right)} = \frac{2\, E(v_{\max})}{\pi}\,, \qquad (2.32)$$

where $E(v)$ denotes the *Elliptic-E function*, defined by

$$E(v) := \int_0^1 \mathrm{d}x\, \sqrt{\frac{1 - vx^2}{1 - x^2}}\,, \qquad (2.33)$$

and so which satisfies $E(1) = 1$. The result for $\Delta\tau/\Delta t$ — the elapsed proper time for the moving twin as a fraction of the time elapsed for the twin at rest — is given as a function of $v_{\max}/c$ in Fig. 4.

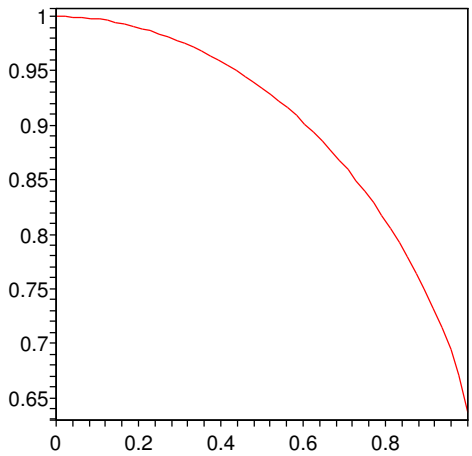


Figure 4: The ratio $\Delta\tau/\Delta t$ of time elapsed for the moving and stationary twins as a function of the moving twin's maximum speed.

The 'paradox' is that the moving twin sees less time pass, but this is not really a paradox at all since there is no reason why clocks on inertial and accelerated trajectories must agree with one another. Indeed, the trajectory of the clock at rest is a geodesic for the Minkowski metric $\eta_{\mu\nu}$ – it is after all a straight line and this metric is constant (and so flat). But the negative sign in the time part of the Minkowski metric ensures that time-like geodesics describe the *maximum* distance between two points in spacetime (whereas, by contrast, geodesics in space give the minimum distance between two points). So

we are guaranteed that all other accelerating clocks also record an elapsed time that is smaller than the one of the clock at rest.

> **Exercise 17:** Imagine two clocks that both perform uniform circular motion of radius a in the x-y plane, but in opposite directions: $x^\mu(u) = \{t, a\cos(\omega t), \pm a\sin(\omega t), 0\}$. Suppose these clocks are synchronized to agree when they are coincident at $x = a$ at $t = 0$. How much time elapses until the next time the clocks are at $x = a$, as seen by each clock as well as by the inertial observer whose time is labelled by t?

Noninertial observers

In special relativity the laws of nature are simpler as seen by inertial observers, whose rectangular positions and times are related by Lorentz transformations $x^{\mu'} = \Lambda^{\mu'}{}_\nu x^\nu$, but look different for observers who do not move at constant speeds relative to inertial observers. This section computes an example of this, and shows in the process that Newton's first law of motion in a non-inertial frame can nonetheless still be regarded as stating that particles move along geodesics in the absence of external forces.

To see how this works, consider the particular case of an observer experiencing uniform circular motion in the x-y plane who uses coordinates, $x^\mu = \{t, x, y, z\}$, in terms of which an inertial observer's coordinates, $x^{\mu'} = \{t, x', y', z\}$, can be written

$$x' = x\cos(\Omega\, t) - y\sin(\Omega\, t) \quad \text{and} \quad y' = x\sin(\Omega\, t) + y\cos(\Omega\, t)\,. \tag{2.34}$$

Ω here represents the angular velocity of the uniform circular motion. Using this relation we have

$$\begin{aligned} \mathrm{d}x' &= \mathrm{d}x\,\cos(\Omega\, t) - \mathrm{d}y\,\sin(\Omega\, t) - \Big[x\sin(\Omega\, t) + y\cos(\Omega\, t)\Big]\Omega\,\mathrm{d}t \\ \mathrm{d}y' &= \mathrm{d}x\,\sin(\Omega\, t) + \mathrm{d}y\cos(\Omega t) + \Big[x\cos(\Omega\, t) - y\sin(\Omega\, t)\Big]\Omega\,\mathrm{d}t\,, \end{aligned} \tag{2.35}$$

so the Lorentz-invariant element of distance becomes

$$\begin{aligned} \mathrm{d}s^2 &= -\mathrm{d}t^2 + \mathrm{d}x'^2 + \mathrm{d}y'^2 + \mathrm{d}z^2 \\ &= \Big[-1 + (x^2+y^2)\Omega^2\Big]\mathrm{d}t^2 + 2(x\,\mathrm{d}y - y\,\mathrm{d}x)\,\Omega\,\mathrm{d}t + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2\,, \end{aligned} \tag{2.36}$$

corresponding to the metric

$$\begin{pmatrix} g_{tt} & g_{tx} & g_{ty} & g_{tz} \\ g_{xt} & g_{xx} & g_{xy} & g_{xz} \\ g_{yt} & g_{yx} & g_{yy} & g_{yz} \\ g_{zt} & g_{zx} & g_{zy} & g_{zz} \end{pmatrix} = \begin{pmatrix} -1 + r^2\,\Omega^2 & -y\Omega & x\Omega & 0 \\ -y\Omega & 1 & 0 & 0 \\ x\Omega & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.37}$$

where $r^2 = x^2 + y^2$.

For this non-inertial observer, a particle whose position is fixed in space defines a world line along which only t varies, so $\mathrm{d}x = \mathrm{d}y = \mathrm{d}z = 0$ (corresponding to a particle executing uniform circular motion from the point of view of the inertial observer). Proper time along such a trajectory as measured with the non-inertial observer's metric is $\mathrm{d}\tau^2 = -\mathrm{d}s^2 = -g_{tt}\,\mathrm{d}t^2 = (1 - r^2\Omega^2)\,\mathrm{d}t^2$, in agreement with the inertial observer's result (given that the inertial observer attributes a speed $v = r\Omega$ due to the uniform circular motion).

In the absence of forces the inertial observer would say that particle trajectories are straight lines: $x^{\mu'} = x_0^{\mu'} + u^{\mu'}\,\tau$ for constant $x_0^{\mu'}$ and $u^{\mu'}$; or $\mathrm{d}^2x^{\mu'}/\mathrm{d}\tau^2 = 0$. These same trajectories do *not* have the same form for the non-inertial observer, since they do not correspond to $x^\mu = x_0^\mu + u^\mu\,\tau$ or $\mathrm{d}^2x^\mu/\mathrm{d}\tau^2 = 0$.

But recall that $\mathrm{d}^2x^{\mu'}/\mathrm{d}\tau^2 = 0$ is the equation for a geodesic for the metric $g_{\mu'\nu'} = \eta_{\mu'\nu'}$, and that the condition for a geodesic can be written for a general metric by

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\nu\lambda}\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\,\frac{\mathrm{d}x^\lambda}{\mathrm{d}\tau} = 0\,, \tag{2.38}$$

which is a form that is equally valid in *any* coordinate system. We should therefore expect that this equation describes motion in the absence of forces as seen by our non-inertial, uniformly rotating observer. But then what is the significance of the Christoffel symbols, $\Gamma^\mu_{\nu\lambda}$, in the non-inertial frame?

To find out we compute the nonzero components of $\Gamma^\mu_{\nu\lambda}$, recalling the definition of the Christoffel symbols

$$\Gamma^\mu_{\nu\lambda} = \frac{1}{2}\,g^{\mu\rho}\Big(\partial_\nu g_{\lambda\rho} + \partial_\lambda g_{\nu\rho} - \partial_\rho g_{\nu\lambda}\Big)\,. \tag{2.39}$$

For the metric of interest the only nonzero metric derivatives are $\partial_x g_{tt} = 2\,x\,\Omega^2$, $\partial_y g_{tt} = 2\,y\,\Omega^2$, $\partial_y g_{tx} = \partial_y g_{xt} = -\Omega$, and $\partial_x g_{ty} = \partial_x g_{yt} = \Omega$, and the inverse metric is

$$\begin{pmatrix} g^{tt} & g^{tx} & g^{ty} & g^{tz} \\ g^{xt} & g^{xx} & g^{xy} & g^{xz} \\ g^{yt} & g^{yx} & g^{yy} & g^{yz} \\ g^{zt} & g^{zx} & g^{zy} & g^{zz} \end{pmatrix} = \begin{pmatrix} -1 & -y\,\Omega & x\,\Omega & 0 \\ -y\,\Omega & 1 - y^2\,\Omega^2 & xy\,\Omega^2 & 0 \\ x\,\Omega & xy\,\Omega^2 & 1 - x^2\,\Omega^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\,, \tag{2.40}$$

and so the nonzero Christoffel symbols (of the second kind) turn out to be

$$\Gamma^x_{tt} = -x\,\Omega^2\,, \quad \Gamma^y_{tt} = -y\,\Omega^2\,, \quad \Gamma^x_{yt} = \Gamma^x_{ty} = -\Omega \quad \text{and} \quad \Gamma^y_{xt} = \Gamma^y_{tx} = \Omega\,. \tag{2.41}$$

With these expressions the equations for a geodesic become

$$\frac{\mathrm{d}^2t}{\mathrm{d}\tau^2} = \frac{\mathrm{d}^2z}{\mathrm{d}\tau^2} = 0 \tag{2.42}$$

and

$$\frac{\mathrm{d}^2 x}{\mathrm{d}\tau^2} - x\,\Omega^2 \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 - 2\,\Omega \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}y}{\mathrm{d}\tau}\right) = 0$$
$$\frac{\mathrm{d}^2 y}{\mathrm{d}\tau^2} - y\,\Omega^2 \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 + 2\,\Omega \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x}{\mathrm{d}\tau}\right) = 0\,. \tag{2.43}$$

The first two of these may be integrated to give

$$z = z_0 + \gamma\, v_z\, \tau = z_0 + v_z\, t \quad \text{and} \quad t = \gamma\, \tau\,, \tag{2.44}$$

where z_0, v_z and γ are constants. Using these in the second two equations, and changing variables using $\mathrm{d}/\mathrm{d}\tau = \gamma(\mathrm{d}/\mathrm{d}t)$, then gives

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} - x\,\Omega^2 - 2\,\Omega \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right) = 0$$
$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} - y\,\Omega^2 + 2\,\Omega \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right) = 0\,. \tag{2.45}$$

Defining the angular momentum vector by $\mathbf{w} := \Omega\,\mathbf{e}_z$, these equations can be written in vector form as

$$\frac{\mathrm{d}^2 \mathbf{r}}{\mathrm{d}t^2} + \mathbf{w} \times (\mathbf{w} \times \mathbf{r}) + 2\,\mathbf{w} \times \mathbf{v} = 0\,, \tag{2.46}$$

where $\mathbf{r} := x\,\mathbf{e}_x + y\,\mathbf{e}_y + z\,\mathbf{e}_z$ and $\mathbf{v} := \mathrm{d}\mathbf{r}/\mathrm{d}t$.

Eqs. (2.45) have as their solutions

$$x = x'(t)\cos(\Omega\, t) + y'(t)\sin(\Omega\, t)$$
$$\text{and} \quad y = -x'(t)\sin(\Omega\, t) + y'(t)\cos(\Omega\, t)\,. \tag{2.47}$$

where $x'(t) = x'_0 + v_x\, t$ and $y'(t) = y_0 + v_y\, t$. The condition $g_{\mu\nu}(\mathrm{d}x^\mu/\mathrm{d}\tau)(\mathrm{d}x^\nu/\mathrm{d}\tau) = -1$ then implies (as usual) $\gamma = (1 - v^2)^{-1/2}$ where $v^2 = v_x^2 + v_y^2 + v_z^2$.

We see that the Christoffel symbols provide precisely the 'fictitious forces' that are required in order to ensure that the geodesics are straight lines, expressed in the non-inertial coordinates. And experience with classical physics allows these fictitious forces to be recognized as old friends; with the $\mathbf{w} \times (\mathbf{w} \times \mathbf{r})$ term of eq. (2.46) representing the *centrifugal force* and the velocity-dependent $\mathbf{w} \times \mathbf{v}$ term giving the *coriolis force* associated with a rotating reference frame. The fact that $\Gamma^\mu_{\nu\lambda}$ do not transform as the components of a tensor is consistent with the fact that these fictitious forces can vanish in some frames — *e.g.* inertial ones — even if they do not in others.

Exercise 18: Show that distances measured by non-inertial observers with coordinates $x^\mu = \{\xi, \chi, y, z\}$ defined by $t = \chi \sinh(a\xi)$ and $x = \chi \cosh(a\xi)$ (with $\chi > 0$) are given by the *Rindler metric*

$$\mathrm{d}s^2 = -a^2\chi^2 \mathrm{d}\xi^2 + \mathrm{d}\chi^2 + \mathrm{d}y^2 + \mathrm{d}z^2\,. \tag{2.48}$$

Show that observers whose world-lines are the curves along which only ξ varies undergo constant proper acceleration with invariant magnitude $\eta_{\mu\nu}(\mathrm{d}u^\mu/\mathrm{d}\tau)(\mathrm{d}u^\nu/\mathrm{d}\tau) = 1/\chi^2$. Show that the only nonzero Christoffel symbols for this metric are $\Gamma^\chi_{\xi\xi} = a^2\chi$ and $\Gamma^\xi_{\chi\xi} = \Gamma^\xi_{\xi\chi} = 1/\chi$, and so show that geodesics satisfy the equations $\mathrm{d}^2y/\mathrm{d}\tau^2 = \mathrm{d}^2z/\mathrm{d}\tau^2 = 0$ and

$$\frac{\mathrm{d}^2\xi}{\mathrm{d}\tau^2} + \frac{2}{\chi}\frac{\mathrm{d}\chi}{\mathrm{d}\tau}\frac{\mathrm{d}\xi}{\mathrm{d}\tau} = 0 \quad \text{and} \quad \frac{\mathrm{d}^2\chi}{\mathrm{d}\tau^2} + a^2\chi\left(\frac{\mathrm{d}\xi}{\mathrm{d}\tau}\right)^2 = 0\,. \tag{2.49}$$

Use these, and the identity $\mathrm{d}\chi/\mathrm{d}\xi = \dot\chi/\dot\xi$ (where over-dots denote $\mathrm{d}/\mathrm{d}\tau$), to show that if ξ parameterizes the geodesics, then $\chi(\xi)$ satisfies

$$\frac{\mathrm{d}^2\chi}{\mathrm{d}\xi^2} = \frac{\ddot\chi}{\dot\xi^2} - \frac{\dot\chi\,\ddot\xi}{\dot\xi^3} = -a^2\chi + \frac{2}{\chi}\left(\frac{\mathrm{d}\chi}{\mathrm{d}\xi}\right)^2\,, \tag{2.50}$$

revealing the fictitious forces required to describe inertial motion in this accelerated frame. Show that the curves $\chi(\xi) = \ell\, e^{\pm a\xi}$ solve this equation.

2.4 Conserved quantities

A special role is played in physics by conserved quantities like electric charge, energy and momentum, since these are all conserved and they all act as sources for known forces of nature. As we shall see, energy and momentum are sources for gravity in much the same way as electric charges and currents source electromagnetism. In order to motivate how energy and momentum density is formulated within a relativistic theory — as will be required in order to state in later sections how they act as sources for gravity — it is convenient first to recall how other conserved quantities, like electric charge density, are formulated.

Electric Current

If there is an observer who sees a nonzero density of electric charge, $\sigma(x,t)$, then anyone else who moves relative to this observer must see a nonzero electric current density, $\mathbf{j}(x,t)$, in addition to seeing a charge density which is different due to the Lorentz contraction of space in the direction of motion, and due to the change in the relative motion of the moving charges. It follows that σ and $\mathbf{j}$ must transform into one another under Lorentz transformations, and it turns out that they transform as a 4-vector with components:

$$j^\mu = \begin{pmatrix} j^0 = \sigma \\ j^i \end{pmatrix}\,, \tag{2.51}$$

where j^i represent the 3 spatial components of the current density vector, $\mathbf{j}$. Being a 4-vector means that it transforms under a Lorentz transformation as

$$j^{\mu'} = \Lambda^\mu{}_\nu\, j^\nu\,, \tag{2.52}$$

and so in the specific case of a boost between inertial observers moving at relative speed $\mathbf{v}$, *c.f.* eqs. (2.8) and (2.19), this becomes

$$\sigma' = j^{0'} = \frac{\sigma + \mathbf{v}\cdot\mathbf{j}/c^2}{\sqrt{1-v^2/c^2}}\,, \qquad \mathbf{j}' = \frac{\mathbf{j} + \mathbf{v}\sigma}{\sqrt{1-v^2/c^2}}\,, \tag{2.53}$$

Conservation of electric charge may be expressed in terms of this 4-vector in a manifestly Lorentz-invariant way, as

$$\partial_\mu j^\mu = \frac{\partial j^0}{\partial t} + \nabla\cdot\mathbf{j} = 0\,. \tag{2.54}$$

Since this is a scalar, if any observer finds the right-hand-side to vanish, then all inertial observers must also find it to vanish. That this equation expresses local charge conservation may be seen by integrating it over a volume V having boundary ∂V, and using Gauss' theorem

$$0 = \int_V \left[\frac{\partial j^0}{\partial t} + \nabla\cdot\mathbf{j}\right] \mathrm{d}^3x = \frac{\mathrm{d}}{\mathrm{d}t}\int_V \sigma\,\mathrm{d}^3x + \int_{\partial V} \mathbf{n}\cdot\mathbf{j}\,\mathrm{d}^2S\,, \tag{2.55}$$

where d^2S denotes an infinitesimal area element of the surface, whose outward-pointing normal vector is $\mathbf{n}$. Written this way it is clear that charge is conserved, inasmuch as the rate of change of the total charge in any volume V is equal to the net flux of charge carried by the current through the boundaries of V.

Electromagnetism

Since charges and currents are sources for electric, $\mathbf{E}$, and magnetic, $\mathbf{B}$, fields, these must similarly transform into one another under Lorentz transformations. It turns out that these six quantities transform as the components of an antisymmetric tensor, $F_{\mu\nu} = -F_{\nu\mu}$, according to

$$\begin{pmatrix} F_{00} & F_{01} & F_{02} & F_{03} \\ F_{10} & F_{11} & F_{12} & F_{13} \\ F_{20} & F_{21} & F_{22} & F_{23} \\ F_{30} & F_{31} & F_{32} & F_{33} \end{pmatrix} = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix}, \tag{2.56}$$

which labels the inertial coordinate in the usual way, $x^\mu = \{x^0, x^1, x^2, x^3\} = \{t, x, y, z\}$.

> **Exercise 19:** Use the transformation properties under Lorentz transformations of a covariant tensor of rank 2 to compute how the components of electric and magnetic fields, $\mathbf{E}$ and $\mathbf{B}$, are related for observers who move relative to one another with constant speed v along the x-axis.

There are two types of fundamental laws in electromagnetism. One of these expresses the forces felt by charges in the presence of electric and magnetic fields, and states that a point charge of magnitude q moving with velocity $\mathbf{v}$ experiences a *Lorentz force* of magnitude

$$\mathbf{F} = q\Big(\mathbf{E} + \mathbf{v} \times \mathbf{B}\Big)\,. \tag{2.57}$$

The second type of law in electromagnetism relates the properties of the electric and magnetic fields to the distribution of charges and currents that source them. These may be summarized as *Maxwell's equations*:

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0\,, \qquad \nabla \cdot \mathbf{B} = 0 \tag{2.58}$$

and

$$\nabla \times \mathbf{B} - \frac{\partial \mathbf{E}}{\partial t} = \mathbf{j}\,, \qquad \nabla \cdot \mathbf{E} = \sigma\,. \tag{2.59}$$

Since all inertial observers must agree on the laws of electromagnetism, it should be possible to formulate these in terms of Lorentz tensors like $F_{\mu\nu}$ and j^μ. Indeed, the two source-free Maxwell equation, eqs. (2.58), can be written as the combined tensor equation

$$\partial_\mu F_{\nu\lambda} + \partial_\nu F_{\lambda\mu} + \partial_\lambda F_{\mu\nu} = 0\,, \tag{2.60}$$

and the two Maxwell equations with sources, eqs. (2.59), similarly can be written

$$\partial_\nu F^{\mu\nu} = j^\mu\,. \tag{2.61}$$

Notice that the antisymmetry $F^{\mu\nu} = -F^{\nu\mu}$ implies $\partial_\mu \partial_\nu F^{\mu\nu}$ vanishes identically, showing that eq. (2.61) would be inconsistent if charge were not conserved, $\partial_\mu j^\mu \neq 0$.

The Lorentz force, eq. (2.57), can also be grouped into a force 4-vector,

$$F_\mu = q F_{\mu\nu} u^\nu\,, \tag{2.62}$$

where u^ν denotes the 4-velocity of the point charge.

Finally, the source-free Maxwell equations, eqs. (2.58), are often solved by writing the fields $\mathbf{E}$ and $\mathbf{B}$ in terms of an electric and magnetic potential, Φ and $\mathbf{A}$, with

$$\mathbf{B} = \nabla \times \mathbf{A} \qquad \text{and} \qquad \mathbf{E} = -\nabla\Phi - \frac{\partial \mathbf{A}}{\partial t}\,, \tag{2.63}$$

and these two equations can be grouped into the single tensor equation

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu\,, \tag{2.64}$$

with the *gauge potential 4-vector* defined by $A^\mu = \{A^0, A^i\} = \{\Phi, A^i\}$.

Exercise 20: Verify that eqs. (2.57), (2.58), (2.59) and (2.63) follow from eqs. (2.62), (2.60), (2.61) and (2.64), together with the definitions of $F_{\mu\nu}$, A_μ and j^μ.

Stress Energy

As the example of electric charge shows, we expect to be able to associate a current 4-vector with each conserved quantity. Since energy is conserved we might therefore naively expect the energy density, ρ, to be combined under Lorentz transformations with an energy flux, $\mathbf{s}$, into a 4-vector $s^\mu = \{s^0, s^i\} = \{\rho, s^i\}$. What makes this expectation naive is the fact that the total energy, $E = \int \rho \, \mathrm{d}^3x$, unlike the total electric charge, $Q = \int \sigma \, \mathrm{d}^3x$, is not itself Lorentz invariant, because it combines with linear momentum, $\mathbf{p}$, into the energy-momentum 4-vector, $p^\mu = \{E, p^i\}$.

The proper statement is instead that the energy density, ρ, energy flux, s^j, momentum density, π^i, and momentum flux (or stress), t^{ij}, all combine under Lorentz transformations into a *Stress-Energy tensor*, $T^{\mu\nu}$, where

$$T^{\mu\nu} = \begin{pmatrix} \rho & s^j \\ \pi^i & t^{ij} \end{pmatrix} . \tag{2.65}$$

In terms of this tensor energy and momentum conservation are expressed by the condition

$$\partial_\nu T^{\mu\nu} = 0 \, , \tag{2.66}$$

since this states that the total change of energy and momentum within any volume V is equal to the net flux of energy and momentum current through the boundaries of V:

$$\partial_\nu T^{0\nu} = 0 \Rightarrow \frac{\mathrm{d}E}{\mathrm{d}t} = \int_V \frac{\partial \rho}{\partial t} \, \mathrm{d}^3V = -\int_{\partial V} s^j \, n_j \, \mathrm{d}^2S \tag{2.67}$$

$$\partial_\nu T^{i\nu} = 0 \Rightarrow \frac{\mathrm{d}P^i}{\mathrm{d}t} = \int_V \frac{\partial \pi^i}{\partial t} \, \mathrm{d}^3V = -\int_{\partial V} t^{ij} \, n_j \, \mathrm{d}^2S \, . \tag{2.68}$$

Furthermore, because of the equivalence between mass and energy in relativity, there is no difference between an energy flux, s^i, and a momentum density, π^i: that is, energy on the move (*i.e.* a flux of energy) carries momentum, and so is equivalent to a density of momentum. Since it is also true that the internal stress tensor can always also be chosen to be symmetric, $t^{ij} = t^{ji}$, the total stress-energy tensor can also be taken to be symmetric: $T^{\mu\nu} = T^{\nu\mu}$.

Examples

At this point it is useful to have explicit forms for the conserved current and stress energy for some simple systems.

Massive point particles:

The simplest system for whom the stress energy can be explicitly written down is for a point particle. A point particle is completely characterized by its world line,

$x^\mu(\tau)$, as well as the value of any conserved physical quantities it might have, such as its rest-mass, m, or its electric charge, q.

The contribution of a massive charged particle to the conserved current is easiest to evaluate in its rest frame, where it is motionless and so contributes no current at all, $\mathbf{j} = 0$, and its contribution to the charge *density* is

$$j^0(\mathbf{r}, t) = q\,\delta^3(\mathbf{r} - \mathbf{y}(t)) \qquad \text{(rest frame)}\,, \tag{2.69}$$

where $\mathbf{y}(t)$ is the particle's spatial trajectory. Here $\delta^3(\mathbf{r}) = \delta(x)\delta(y)\delta(z)$ denotes the 3-dimensional *Dirac delta function*, which can be regarded as a limiting case as $\lambda \to 0$ of the function $(C/\lambda^3)\,e^{-\mathbf{r}^2/\lambda^2}$, with the constant C chosen to ensure that $\int \mathrm{d}^3x\ \delta^3(\mathbf{r}) = 1$. The result is a quantity that is infinitely peaked about zero argument, but with a normalization that diverges in such a way as to ensure constant area under the curve. It has the property that

$$\int \mathrm{d}^3x\ f(\mathbf{r})\,\delta(\mathbf{r} - \mathbf{y}) = f(\mathbf{y})\,, \tag{2.70}$$

for arbitrary smooth functions, f.

The result for j^μ in a general frame is then found simply by identifying a 4-vector that agrees with this result in the rest frame. Any such a 4-vector must be unique, since if two 4-vectors agree in one frame they must agree in them all. The result is

$$j^\mu(x) = q\,u^\mu\,\delta^3(x - y(\tau))\,, \tag{2.71}$$

where $x^\mu = y^\mu(\tau)$ gives the components of the particle's world-line, for which the 4-velocity, $u^\mu(y(\tau)) = \mathrm{d}y^\mu/\mathrm{d}\tau$, satisfies (as usual) $u_\mu u^\mu = -1$. The delta function therefore gives zero contribution except at the particle's world line. Using the components $u^\mu = \gamma(1, \mathbf{v})$, with $\gamma = (1 - v^2)^{-1/2}$, this gives

$$\begin{aligned} j^0 &= q\gamma\,\delta^3(x - y(\tau)) \\ \text{and}\quad \mathbf{j} &= q\gamma\,\mathbf{v}\,\delta^3(x - y(\tau))\,. \end{aligned} \tag{2.72}$$

The stress energy for such a particle is found using the same arguments. In the rest frame there is no internal stress or energy flow, so the only nonzero component is the energy density,

$$T^{00} = m\,\delta^3(\mathbf{r} - \mathbf{x}(\tau)) \qquad \text{(rest frame)}\,. \tag{2.73}$$

The unique result for the tensor $T^{\mu\nu}$ in a general reference frame is then given by

$$T^{\mu\nu}(x) = m\,u^\mu\,u^\nu\,\delta^3(x - y(\tau))\,, \tag{2.74}$$

which has components

$$
\begin{aligned}
T^{00} &= m\gamma^2\,\delta^3(x-y(\tau)) \\
T^{0i} &= m\gamma^2\,v^i\,\delta^3(x-y(\tau)) \\
\text{and}\quad T^{ij} &= m\gamma^2\,v^i v^j\,\delta^3(x-y(\tau))\,.
\end{aligned}
\tag{2.75}
$$

Dust

A more common energy source for gravitational problems is a macroscopic collection of a great many – N, say – individual particles. If these particles do not interact with one another their total current and stress energy is just the sum of the contribution of each, summed over all particles present:

$$
j^\mu = \sum_{k=1}^{N} q_k\,u_k^\mu\,\delta^3(x-x_k(\tau)) \quad \text{and} \quad T^{\mu\nu} = \sum_{k=1}^{N} m_k\,u_k^\mu\,u_k^\nu\,\delta^3(x-x_k(\tau))\,. \tag{2.76}
$$

Most commonly, when the gravitational properties of such a system are of interest it is over distances, L, that are much larger than the typical inter-particle spacing, a: $L \gg a$. (Examples where this will prove to be true include the gravitational field within a star or cloud of interstellar gas — for which the particles are gas molecules or atoms — or for the overall shape of the universe as a whole — for which the particles might be entire galaxies).

In this case only the average properties of the distribution of particles is relevant, and it is unnecessary to carry around information concerning the position of each separate particle. This can be made precise by identifying a region whose size, d, is much larger than the typical inter-particle distance scale, yet still much smaller than the scale, L, of gravitational interest: $L \gg d \gg a$. When such a region exists, because $d \gg a$ it contains a large number of particles, and so has the property that the statistical fluctuations (due to the exchange of individual particles with the surrounding regions, say) about the mean of the energy and charge are very small. However, because $d \ll L$ these mean properties can be well approximated as being constant over each such region, although they can vary slowly from region to region.

In this case we can define the average frame of rest for a given region in terms of the region's average 4-velocity

$$
U^\mu(x) = \mathcal{C}\sum_{k=1}^{\mathcal{N}} u_k^\mu\,, \tag{2.77}
$$

where $\mathcal{C}$ is chosen to ensure that U^μ is normalized, $U_\mu U^\mu = -1$, and $\mathcal{N}$ denotes the number of particles in R. The x-dependence of U^μ emphasizes that the precise

average rest frame can vary slowly from region to region. The average rest frame for R is the frame for which the spatial components vanish: $U^i = \gamma v^i = 0$.

With this definition, the mean charge density, σ, and energy density, ρ, can be defined for the average rest frame by

$$j^\mu(x) := \sigma(x)\, U^\mu(x) \quad \text{and} \quad T^{\mu\nu}(x) := \rho(x)\, U^\mu(x)\, U^\nu(x)\,. \tag{2.78}$$

In the limit where the particles all move non-relativistically these satisfy $\sigma(x) \simeq q\, n(x)$ and $\rho(x) \simeq m\, n(x)$, where $n(x) = \mathcal{N}(x)/\mathcal{V}(x)$ is the macroscopic particle density that reproduces $\int_{\mathcal{V}} \mathrm{d}^3x\; n(x) = \mathcal{N}$ as would the microscopic result, $n_{\rm micro}(x) = \sum_k \delta^3(x - y_k(\tau))$, when integrated over any spatial region of volume $\mathcal{V}$ containing $\mathcal{N}$ particles. This is a less useful description for relativistic particles, since for these the possibility of particle-antiparticle production and annihilation implies the total number of particles is never strictly conserved.

A fluid made up of noninteracting massive particles of this type is known as 'dust', inasmuch as it represents a special case of a more general fluid for which the pressure and viscosity terms are negligible.

Perfect fluids

The similarly simple but more realistic system for which the stress energy can be explicitly written down is for a *perfect fluid*; defined as a system for which the average, macroscopic conserved quantities are functions only of the local average fluid 4-velocity, $U^\mu(x)$, and the local metric, $\eta_{\mu\nu}$ (and not also their derivatives, say).[3]

Under this assumption the conserved current describing the conservation of their total number is given by

$$j^\mu = \sigma\, U^\mu = \begin{pmatrix} \gamma\, \sigma \\ \gamma\, \sigma\, v^i \end{pmatrix}\,, \tag{2.79}$$

where the local rest-frame charge density, $\sigma(x) = -U_\mu\, j^\mu$ has properties (like dependence on temperature or other macroscopic variables) that depend on the details of the microscopic properties of the particles involved. Regardless of these details, conservation of the underlying charge requires $j^\mu(x)$ must satisfy the conservation condition: $\partial_\mu j^\mu = 0$.

Similarly, the most general symmetric tensor depending only on $\eta_{\mu\nu}$ and $U^\mu(x)$ (but not its derivatives) is

$$T^{\mu\nu} = (\rho + p)\, U^\mu\, U^\nu + p\, \eta^{\mu\nu} = \begin{pmatrix} \gamma^2(\rho + p\, v^2) & \gamma^2(\rho + p)\, v^j \\ \gamma^2(\rho + p)\, v^i & \gamma^2(\rho + p)\, v^i v^j + p\, \delta^{ij} \end{pmatrix}\,. \tag{2.80}$$

[3]Inclusion of a dependence on derivatives into the macroscopic currents is what introduces transport coefficients, like conductivities and viscosities, into the discussion.

The interpretation of the coefficient functions $\rho(x)$ is found by going to the rest frame, which reveals $\rho = T^{00}|_{\text{rest}}$ is the rest-frame energy density. Similarly, in the rest frame $T^{ij}|_{\text{rest}} = p\,\delta^{ij}$. But conservation of momentum for a region V within the fluid, eq. (2.68), then reads

$$\frac{\mathrm{d}P^i}{\mathrm{d}t} = \int_V \frac{\partial T^{i0}}{\partial t}\,\mathrm{d}^3V = -\int_{\partial V} T^{ij}\,n_j\,\mathrm{d}^2S = -\int_{\partial V} p\,n^i\,\mathrm{d}^2S\,, \tag{2.81}$$

which uses $\partial_\nu T^{\mu\nu} = 0$, together with Stoke's theorem in the form $\int_V \partial_j T^{ij}\,\mathrm{d}^3V = \int_{\partial V} n_j T^{ij}\,\mathrm{d}^2S$. This shows that each surface element exerts an inward-directed force of magnitude p, along the line defined by the surface element's normal, $\mathbf{n}$. Consequently p can be interpreted as the fluid's *pressure*. In general the detailed properties of both $\rho(x)$ and $p(x)$ can depend on what kind of particles are involved in the fluid, and is often characterized by an *equation of state*, of the form $p = p(\rho, T)$, where T is the fluid's local rest-frame temperature.

3. Weak gravitational fields

We are now in a position to begin making the connection between gravitation and the geometry of spacetime. To this end it is first worth pausing to formulate Newtonian gravity in an explicitly field-theoretic manner.

3.1 Newtonian gravity

In the first encounter with Newtonian gravitation, one is normally taught that the gravitational force acting on a point mass m_1 situated at a position $\mathbf{r}_1$ due to the presence of another point mass, m_2, situated at $\mathbf{r}_2$, is

$$\mathbf{F}_{12} = \frac{G m_1 m_2}{|\mathbf{r}_2 - \mathbf{r}_1|^2}\,\mathbf{e}_{12}\,, \tag{3.1}$$

where $\mathbf{e}_{12} = (\mathbf{r}_2 - \mathbf{r}_1)/|\mathbf{r}_2 - \mathbf{r}_1|$ is the unit vector pointing from particle 1 to particle 2, and $G = 6.673(10) \times 10^{-11}$ N m^2/kg^2 is a universal constant known as *Newton's constant of gravitation*. The force due to a more complicated distribution of masses is then found by summing eq. (3.1) over all of the particles that are present.

The principle of equivalence

Using eq. (3.1) in Newton's 2nd Law of motion gives the acceleration of particle number 1:

$$\mathbf{a}_1 = \frac{\mathbf{F}_{12}}{m_1} = \frac{G m_2}{|\mathbf{r}_2 - \mathbf{r}_1|^2}\,\mathbf{e}_{12}\,, \tag{3.2}$$

and similarly for particle 2. This has the remarkable property of being completely independent of the value of m_1. This property, which assumes that a particle's

inertial mass appearing in Newton's second law — $\mathbf{F} = m\,\mathbf{a}$ — is the same as its *gravitational mass*, appearing in eq. (3.1). As applied to a constant gravitational field, such as arises to good approximation at the Earth's surface, this implies the well-known fact that all objects near the Earth's surface accelerate towards it with a universal acceleration,

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \mathbf{g}\,, \tag{3.3}$$

with magnitude $g = GM_\oplus/R_\oplus^2 \simeq 9.8$ m/s^2, regardless of how massive they are.[4]

The best present test of the mass-independence of eq. (3.2) come from precision measurements of the distance to the Moon that became possible once laser reflectors were left on its surface by astronauts in the late 1960s. These show that the difference between the Moon and the Earth's average acceleration towards the Sun is [1]

$$\frac{\Delta a}{a} = \frac{|a_E - a_M|}{\frac{1}{2}(a_E + a_M)} = (-1 \pm 1.4) \times 10^{-13}\,, \tag{3.4}$$

which is consistent with zero with a precision of one part in 10^{13}.

Measurements such as these provide the experimental cornerstone for understanding gravity theoretically, since they provide guidance about how to modify Newton's theory to be consistent with relativity. In particular, the great accuracy with which a falling particle's acceleration is known to be independent of its mass suggests it be elevated to a principle whose validity is not restricted to its being a consequence of Newton's Laws.

The resulting principle is called the *Principle of Equivalence* because it makes a constant gravitational force, as in eq. (3.3), appear very much like the fictitious centrifugal and coriolis forces encountered earlier in eq. (2.46), since both produce accelerations that are completely independent of the moving particle's mass). A constant gravitational force would in this sense be equivalent to the fictitious constant force associated with being in a non-inertial frame undergoing constant acceleration. Conversely, it is the observers in a *freely falling frame* that are the inertial observers that experience Newton's 2nd Law of motion in a constant gravitational field (as is graphically experienced by astronauts who appear to float freely within their spacecraft, as they all move in orbit around the Earth).

The gravitational field

A more useful way to think about Newtonian gravity for the purposes of generalizing to relativity is in terms of fields, rather than forces. To this end one defines the *gravitational potential*, $\Phi(\mathbf{r}, t)$, throughout all space, whose strength is determined

[4]...provided all non-gravitational complications, like air resistance, are negligible.

by the field equation

$$\nabla^2\Phi = 4\pi G\,\mu\,, \tag{3.5}$$

where $\nabla^2 = \partial_x^2 + \partial_y^2 + \partial_z^2$ and $\mu(\mathbf{r},t)$ denotes the local density of mass, per unit volume. Once $\Phi(\mathbf{r},t)$ is determined by solving this equation, the gravitational force acting on any mass, m, located at a point, $\mathbf{r}$, is found using the relation

$$\mathbf{F} = -m\,\nabla\Phi(\mathbf{r},t)\,. \tag{3.6}$$

To see that eqs. (3.5) and (3.6) reproduce eq. (3.1) one first solves eq. (3.5) using $\mu(\mathbf{r},t) = m_2\,\delta^3(\mathbf{r}-\mathbf{r}_2(t))$ to determine the gravitational potential set up by a point mass, m_2, situated at position $\mathbf{r} = \mathbf{r}_2(t)$. The solution that vanishes at spatial infinity is

$$\Phi(\mathbf{r},t) = -\frac{Gm_2}{|\mathbf{r}-\mathbf{r}_2(t)|}\,, \tag{3.7}$$

and so applying eq. (3.6) to this for a point mass, $m = m_1$, situated at $\mathbf{r} = \mathbf{r}_1$ then gives eq. (3.1).

Because Newton's law of gravity is a conservative force, it can be derived from a potential energy. For a collection of N otherwise noninteracting particles moving under their mutual gravitation the total conserved energy in Newtonian physics is (up to an infinite, but position-independent, constant)

$$E = \frac{1}{2}\sum_{k=1}^{N} m_k\Big[v_k^2 + \Phi(\mathbf{r}_k)\Big]\,, \tag{3.8}$$

where, as usual, $v_k^2 = \mathbf{v}_k\cdot\mathbf{v}_k$. For the special case of two particles, this may be written

$$E = \frac{MV^2}{2} + \frac{m_{\rm red}\,v^2}{2} - \frac{Gm_1m_2}{|\mathbf{r}|}\,, \tag{3.9}$$

where $M = m_1+m_2$ is the total mass, V is the magnitude of the velocity of the *center of mass*, $\mathbf{V} = \mathrm{d}\mathbf{R}/\mathrm{d}t$ with $\mathbf{R} = (m_1\mathbf{r}_1 + m_2\mathbf{r}_2)/M$. The quantity $m_{\rm red} = m_1m_2/M$ defines the *reduced mass* and $\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$ is the relative position of the two particles, whose velocity $\mathbf{v} = \mathrm{d}\mathbf{r}/\mathrm{d}t$ has magnitude v. The relative position and center of mass position are convenient variables because they separately evolve under the equations of motion

$$\frac{\mathrm{d}^2\mathbf{R}}{\mathrm{d}t^2} = 0 \quad\text{and}\quad \frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = -\frac{GM}{|\mathbf{r}|^2}\,\mathbf{e}_r\,, \tag{3.10}$$

where $\mathbf{e}_r = \mathbf{r}/|\mathbf{r}|$ is the unit vector parallel to $\mathbf{r}$.

The solutions to these equations describe both bound orbits and unbound scattering solutions, and an important property of the bound orbits is that their internal

kinetic and potential energies are similar in size. Consequently, the non-relativistic approximation (required for such a Newtonian analysis) is valid if

$$v^2 \simeq \frac{GM}{r} \ll 1\,. \tag{3.11}$$

Putting back the factors of c, the criterion for the validity of a Newtonian description becomes $v^2/c^2 \simeq GM/rc^2 \ll 1$. The size of GM/Rc^2 at the surface of the Sun and Earth is listed in the following Table, and show why a non-relativistic Newtonian approximation works so well for applications in the Solar System.

	M (kg)	R (m)	$GM/R\,c^2$
Earth ($\oplus$)	5.97×10^{24}	6.38×10^{6}	6.95×10^{-10}
Sun ($\odot$)	1.99×10^{30}	6.96×10^{8}	2.12×10^{-6}

Table 1:

The size of non-Newtonian effects near the surface of the Earth and Sun.

Exercise 21: For a bound (elliptical) orbit of a particle in the gravitational field of a large central mass M, use Newton's Law in the form $m\mathbf{a} = -(GMm/r^2)\,\mathbf{e}_r$ (where $\mathbf{e}_r$ is the outward pointing unit vector in the radial direction) to prove $\langle \mathbf{v}^2 \rangle = \langle GM/r \rangle$, where $\langle \cdots \rangle := (1/T) \int \mathrm{d}t (\cdots)$ denotes the time-average of a given quantity over one orbit (where T is the orbital period). Use this to compute the ratio of the average kinetic and potential energy of the particle, $K = \frac{1}{2} m\mathbf{v}^2$ and $U = -GMm/r$, over an orbit, and show that $\langle K \rangle = -\frac{1}{2} \langle U \rangle$.

Consistency with relativity

There are several ways to see that the above Newtonian story is inconsistent with special relativity. One is to notice that eq. (3.7) depends on time, t, only through the specification of the instantaneous position, $\mathbf{r}_2(t)$, of the source particle. This means that the force exerted on other particles, eq. (3.6), changes instantaneously as the source particle changes its position. Information about the source's position therefore travels faster than light to simultaneously tell all other particles that they should fall towards the source particle's new position.

But special relativity states that what is simultaneous for one inertial observer is not simultaneous for all others, and so this same force rule cannot possibly hold for all such observers. This violates the Principle of Relativity. This problem is related to relativity's proscription against things moving faster than light, which the Newtonian force law also violates.

This particular problem arises because eq. (3.5) treats space differently from time, and a naive way to fix it would be to replace the *Laplacian* operator, ∇^2, appearing

in this equation by the Lorentz-invariant *d'Alembertian* operator, $\Box = \nabla^2 - \partial_t^2$, to get the following guess (called *Nördström gravity*):

$$\Box\,\Phi = \eta^{\mu\nu}\,\partial_\mu\partial_\nu\Phi = \partial^\mu\partial_\mu\Phi = \left(-\frac{1}{c^2}\,\frac{\partial^2\Phi}{\partial t^2} + \nabla^2\Phi\right) = 4\pi G\,\mu(x)\,. \tag{3.12}$$

A fully relativistic theory would also have to identify a Lorentz-invariant notion of mass density, like perhaps the rest-frame energy density, $\rho(x)/c^2$, to use on the right-hand side.

This kind of proposal has the nice feature that changes to the forces seen by other masses do not change instantaneously, with the news of changes in source position instead being carried by waves in the field Φ (analogous to electromagnetic waves in the electromagnetic field) that travel at the speed of light. It must be rejected as a successful theory of gravity, however, because its predictions contradict a number of experimental facts, including tests like eq. (3.4), or predictions for the gravitational bending of light rays (to be discussed below).

A clue as to how to proceed comes from recognizing that the famous equation $E = m\,c^2$ implies energy and mass are equivalent to one another in relativity, and so we should be seeking an equation like eq. (3.12), but with the entire conserved stress energy on the right-hand side:

$$\Box\,h_{\mu\nu} = \frac{4\pi G}{c^2}\,T_{\mu\nu}\,, \tag{3.13}$$

in much the same way as it is the entire conserved 4-current, j^μ, that appears on the right-hand side in Maxwell's equations, eqs. (2.61). This indicates we should seek some sort of symmetric tensor field, $h_{\mu\nu}$, to describe gravity, rather than a single scalar field like Φ. Einstein's insight was to see that it is the metric tensor, $g_{\mu\nu}$, that is the field we seek, although eq. (3.13) is only in this case an approximation to the right field equations for gravity.

3.2 Gravity as geometry

To make the case that it is the spacetime metric, $g_{\mu\nu}(x)$, that describes gravity we next investigate in detail spacetime geometry in a spherically symmetric system, such as should apply outside of a spherically symmetric matter distribution like the Sun or Earth.

Spherically Symmetric Geometries

The first step is to identify what restrictions the metric must satisfy in order to be spherically symmetric. For the present purposes, take spherical symmetry to mean the existence of a symmetry acting on the three spatial coordinates, x^i, of the form

given in eq. (2.6): $x^i \to M^i{}_j\, x^j$, where M is an orthogonal matrix ($\delta_{ij}\, M^i{}_k M^j{}_l = \delta_{kl}$). This is to be a symmetry in the sense that it leaves the metric completely unchanged.

The implications for the metric can be found by constructing the most general invariant quadratic line element, $\mathrm{d}s^2$, that can be built from the vectors x^i and $\mathrm{d}x^i$, and from the scalars, t and $\mathrm{d}t$. Given the three rotationally invariant combinations

$$\delta_{ij}\, x^i x^j = \mathbf{x}\cdot\mathbf{x} \equiv r^2\,, \quad \delta_{ij}\, x^i \mathrm{d}x^j = \mathbf{x}\cdot\mathrm{d}\mathbf{x}\,, \quad \delta_{ij}\, \mathrm{d}x^i \mathrm{d}x^j = \mathrm{d}\mathbf{x}\cdot\mathrm{d}\mathbf{x}\,, \tag{3.14}$$

the most general invariant form is

$$\mathrm{d}s^2 = -A\,\mathrm{d}t^2 + B\,\mathrm{d}t(\mathbf{x}\cdot\mathrm{d}\mathbf{x}) + C\,(\mathbf{x}\cdot\mathrm{d}\mathbf{x})^2 + D\,\mathrm{d}\mathbf{x}\cdot\mathrm{d}\mathbf{x}\,, \tag{3.15}$$

where the coefficients $A = A(r,t)$ through $D = D(r,t)$ are arbitrary functions of the invariants r and t.

Given the dependence on r, it is convenient to work in polar coordinates, (r,θ,ϕ), defined as usual by $x^1 = r\sin\theta\cos\phi$, $x^2 = r\sin\theta\sin\phi$ and $x^3 = r\cos\theta$, in which case

$$\mathbf{x}\cdot\mathrm{d}\mathbf{x} = r\,\mathrm{d}r \quad \text{and} \quad \mathrm{d}\mathbf{x}\cdot\mathrm{d}\mathbf{x} = \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,. \tag{3.16}$$

In these coordinates the most general invariant line element is then

$$\mathrm{d}s^2 = -\tilde{A}\,\mathrm{d}t^2 + \tilde{B}\,\mathrm{d}t\mathrm{d}r + \tilde{C}\,\mathrm{d}r^2 + \tilde{D}\,r^2\,(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,, \tag{3.17}$$

where $\tilde{A} = A$, $\tilde{B} = rB$, $\tilde{C} = r^2 C + D$ and $\tilde{D} = D$.

We are still free to redefine the invariant coordinates r and t to further simplify the form of this metric. A convenient choice is to redefine $r \to \hat{r} = r\tilde{D}^{1/2}$, which is possible provided $\tilde{D} \geq 0$. This ensures the last term of eq. (3.17) becomes $\hat{r}^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$. Physically, this means that $\hat{r}$ plays the role usually associated with 'radius', because the sphere obtained by varying θ and ϕ at fixed $\hat{r}$ and t has area $4\pi\hat{r}^2$, and circumference $2\pi\hat{r}$ when these are computed using the proper length $\mathrm{d}s$. Although this choice also mixes up the coefficients of $\mathrm{d}t^2$, $\mathrm{d}t\mathrm{d}r$ and $\mathrm{d}r^2$ in $\mathrm{d}s^2$, this can be absorbed into appropriate redefinitions of the unknown coefficients $\tilde{A}$ through $\tilde{C}$, leaving

$$\mathrm{d}s^2 = -\hat{A}\,\mathrm{d}t^2 + \hat{B}\,\mathrm{d}t\mathrm{d}\hat{r} + \hat{C}\,\mathrm{d}\hat{r}^2 + \hat{r}^2\,(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,. \tag{3.18}$$

Finally, we may remove the cross term $\mathrm{d}t\mathrm{d}r$ by redefining the time coordinate to $t = F(\hat{t},\hat{r})$, for which $\mathrm{d}t = \mathrm{d}\hat{t}\,\partial_{\hat{t}}F + \mathrm{d}\hat{r}\,\partial_{\hat{r}}F$. This makes the cross term in $\mathrm{d}s^2$ become $[-2\hat{A}\partial_{\hat{r}}F + \hat{B}]\partial_{\hat{t}}F\,\mathrm{d}\hat{t}\mathrm{d}\hat{r}$, which can be eliminated by choosing $F(\hat{r})$ as a solution to the linear partial differential equation $-2\hat{A}\partial_{\hat{r}}F + \hat{B} = 0$.

Once this has been done we have the most general form possible for a spherically symmetric metric. Dropping the ' ˆ ' everywhere, it is:

$$\mathrm{d}s^2 = -e^{2a(r,t)}\,\mathrm{d}t^2 + e^{2b(r,t)}\,\mathrm{d}r^2 + r^2\,(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,, \tag{3.19}$$

where the remaining unknown coefficient functions are written as exponentials in order to simplify some expressions that come later.

The coordinates used to put the metric into the form eq. (3.19) are called *Schwarzschild* coordinates, and are defined by the condition that it is r^2 that pre-multiplies the angular terms. An alternative definition of coordinates can instead be defined by the condition that the metric has the alternative *isotropic* form

$$\begin{aligned} \mathrm{d}s^2 &= -e^{2\tilde{a}(\varrho,t)}\,\mathrm{d}t^2 + e^{2\tilde{b}(\varrho,t)}\left[\mathrm{d}\varrho^2 + \varrho^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right] \\ &= -e^{2\tilde{a}(\varrho,t)}\,\mathrm{d}t^2 + e^{2\tilde{b}(\varrho,t)}\left[\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2\right], \end{aligned} \tag{3.20}$$

whose convenience relies on the metric within the square brackets being the metric of flat 3-dimensional space.

Weak Gravitational Fields

To describe weak gravitational fields outside of a spherical source we further suppose that these functions are close to those for a flat geometry (written in spherical coordinates): $\mathrm{d}s^2 \simeq -\mathrm{d}t^2 + \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$. That is, if we write

$$e^{2a(r,t)} := 1 + 2\Phi(r,t) \quad \text{and} \quad e^{2b(r,t)} := 1 + 2\Psi(r,t)\,, \tag{3.21}$$

then the Newtonian limit should correspond to the case where the functions Φ and Ψ are small: $\Phi, \Psi \ll 1$. More precisely, because the Newtonian description of two-body bound orbits implies $v^2 \simeq GM/r$ (*c.f.* the discussion around eq. (3.11)), where v is the relative speed and $M = m_1 + m_2$, we assume $\Phi \simeq \Psi \simeq \mathcal{O}(v^2) \simeq \mathcal{O}(GM/r) \ll 1$.

Geodesics for slowly moving particles

To see what the physical implications of such a metric might be, we must know how it affects the trajectories of particles. To this end — inspired by the example of flat spacetime — we make the additional assumption that in the absence of all non-gravitational forces particles simply follow the geodesics of the metric.

This implies that particle motion maximizes the proper time

$$\begin{aligned} \tau_{AB} &= \int_A^B \mathrm{d}t\,\sqrt{(1+2\Phi) - (1+2\Psi)\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 - r^2\left[\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 - \sin^2\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2\right]} \\ &\approx (t_B - t_A) + \frac{1}{2}\int_A^B \mathrm{d}t\,\left[2\Phi - \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 - \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 - \left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2\right], \end{aligned} \tag{3.22}$$

where we use t as the parameter along the curve. The approximate equality: (a) expands the square root, keeping terms only up to $\mathcal{O}(v^2)$ — and so in particular neglects the product $\Psi(\mathrm{d}r/\mathrm{d}t)^2 \simeq \mathcal{O}(v^4)$; and (b) changes to rectangular coordinates:

$(r, \theta, \phi) \to (x, y, z)$. Assuming Φ is independent of t, and asking eq. (3.22) to be stationary with respect to small variations in the trajectories $\mathbf{r}(t)$ then leads to the following geodesic equation:

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} + \nabla\Phi = 0\,, \tag{3.23}$$

which may be recognized as Newton's equations for particles interacting with gravity provided we regard Φ as being the Newtonian gravitational potential. In particular this implies

$$\Phi \simeq -\frac{GM}{r} + \mathcal{O}\left[\left(\frac{GM}{r}\right)^2\right]\,, \tag{3.24}$$

at a radial position, r, above a weakly-gravitating, spherically symmetric source.

This shows that everything we know about orbits in Newtonian physics can be captured by the postulate that gravity is associated with the curvature of spacetime, with Newton's first law modified to state that particles travel along geodesics in the absence of non-gravitational forces. What is particularly noteworthy is that within this framework the equivalence principle arises automatically: because gravity is associated with motion through a geometry, the acceleration experienced by a moving particle is independent of its mass (for precisely the same reason that the same is true for fictitious forces like the coriolis force).

Exercise 22: Starting from the metric $\mathrm{d}s^2 = -(1+2\Phi)\,\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$, and assuming Φ is small enough that higher powers like Φ^2 can be neglected, show that the only nonzero Christoffel symbols are $\Gamma^t_{tt} = \partial_t\Phi$, $\Gamma^i_{tt} = \partial_i\Phi$ and $\Gamma^t_{ti} = \Gamma^t_{it} = \partial_i\Phi$. Use these results in the geodesic equation to show that geodesics, $x^\mu(w)$, satisfy

$$\ddot{t} + \dot{t}^2\,\partial_t\Phi + 2\,\dot{t}\dot{x}^k\partial_k\Phi = 0 \quad \text{and} \quad \ddot{x}^i + \delta^{ij}\partial_j\Phi\,\dot{t}^2 = 0\,, \tag{3.25}$$

where over-dots denote $\mathrm{d}/\mathrm{d}w$. Use these, with the identity $\mathrm{d}x^i/\mathrm{d}t = \dot{x}^i/\dot{t}$, to derive

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} - \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\,\partial_t\Phi - 2\,\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\cdot\nabla\Phi\right) + \nabla\Phi = 0\,, \tag{3.26}$$

and so also eq. (3.23) in the non-relativistic limit where $\partial_t\Phi = 0$ and products like $\partial_k\Phi\,(\mathrm{d}x^k/\mathrm{d}t)$ can be neglected. (As usual, $\mathbf{r}$ here denotes the vector $x^i\mathbf{e}_i$, where $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_z$ are the unit vectors in the three Cartesian coordinate directions.)

Gravitational redshift

Since the trajectories of Newtonian gravity are reproduced as the geodesics of a metric that depends on the gravitational potential, and since the geodesics are defined as

the curves that maximize the proper time between two events, it must be true that gravitational fields cause time to run differently for observers sitting within them.

To see this quantitatively, consider the world-lines of an observer who hovers (perhaps using rockets) at a fixed distance above a gravitating source: $x^\mu(\tau) = \{t(\tau), r_\star, \theta_\star, \phi_\star\}$, where $(r_\star, \theta_\star, \phi_\star)$ labels the fixed spatial position of the observer. Any such an observer's 4-velocity is given by $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau = \{\mathrm{d}t/\mathrm{d}\tau, 0, 0, 0\}$, where $\mathrm{d}t/\mathrm{d}\tau$ can be computed in terms of the gravitational potential by using the condition $g_{\mu\nu}\, u^\mu u^\nu = g_{tt}(\mathrm{d}t/\mathrm{d}\tau)^2 = -1$, and so using $g_{tt} = -(1+2\Phi)$ we find

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = \sqrt{-g_{tt}} = \sqrt{1+2\Phi} \simeq 1 + \Phi + \mathcal{O}(\Phi^2)\,. \tag{3.27}$$

and so, to linear order in Φ, the difference between the rates of two clocks at different radii, $r_{_A}$ and $r_{_B}$, becomes

$$\left(\frac{\mathrm{d}\tau}{\mathrm{d}t}\right)_{_B} - \left(\frac{\mathrm{d}\tau}{\mathrm{d}t}\right)_{_A} \simeq \Phi(r_{_B}) - \Phi(r_{_A})\,. \tag{3.28}$$

As expected, this states that clocks run at different speeds when situated in a gravitational field.

> **Exercise 23:** For a constant gravitational field pointed along the z axis the Newtonian potential can be written as $\Phi = gz$, where g is the universal acceleration experienced by falling objects. Eq. (3.28) states that two clocks separated by a height $h = \Delta z$ run with rates that differ by an amount gh, with the higher of the two clocks running faster. Verify that this result also follows from special relativity and the principle of equivalence by considering two observers who accelerate in the positive z direction along the trajectories $z_{_A}(t) = \frac{1}{2}\, g\, t^2$ and $z_{_B}(t) = h + \frac{1}{2}\, g\, t^2$ in the absence of a gravitational field, by comparing the times of departure and arrival of two light rays sent from observer A to observer B.

As applied to observers outside of a spherical, weakly gravitating source, for which $\Phi = -GM/r$ these become

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} \simeq 1 - \frac{GM}{r} + \mathcal{O}\left[\left(\frac{GM}{r}\right)^2\right]\,, \tag{3.29}$$

and so in particular $\mathrm{d}\tau = \mathrm{d}t$ for clocks that are infinitely far away ($r \to \infty$). This provides the physical interpretation for the coordinate t, which is seen as the time measured by an infinitely distant observer. Eq. (3.29) then shows how time runs more and more slowly the closer one hovers over the gravitating source. In particular,

reinstating the factors of c, motionless clocks at the top of a building on the surface of the Earth run faster than those on the ground floor by an amount

$$\Delta\left(\frac{\mathrm{d}\tau}{\mathrm{d}t}\right) \simeq \frac{GM_\oplus h}{R_\oplus^2 c^2} \simeq \frac{gh}{c^2} \simeq 1.1\times 10^{-15}\left(\frac{h}{10\,\mathrm{m}}\right)\,, \tag{3.30}$$

for a building of height $h \ll R_\oplus$. Here $g = GM_\oplus/R_\oplus^2 \simeq 9.8$ m/sec^2 denotes the acceleration due to gravity at the Earth's surface. This difference in the clock's rate accumulates over time, adding up to a difference of 9.5×10^{-11} sec (about a tenth of a nanosecond) every day between clocks situated on the two floors. Time differences this large can be measured using accurate atomic clocks, verifying the prediction of eq. (3.30).

Closely related to the slowing of time in a gravitational field is the red-shifting of light as it climbs out of a gravitational potential well (or its blue-shifting as it falls in). Although this is described in more detail below, once the geodesics describing light propagation are determined, the main result also follows from the above discussion of gravitational time dilation. This is possible because of the connection between photon energy and frequency required by quantum mechanics, $E = \hbar\omega$, since frequencies may be directly determined by time measurements (such as measurements of the period $T = 2\pi/\omega$).

Keeping in mind that t measures time as seen by observers at infinity, eq. (3.29) shows that the frequency, $\omega(r)$, of a photon measured by a motionless observer at radius r, differs from the frequency, ω_∞, the same photon would be measured to have at $r\to\infty$ by

$$\frac{\omega(r)}{\omega_\infty} = \frac{T_\infty}{T(r)} = \frac{\mathrm{d}t}{\mathrm{d}\tau} = \frac{1}{\sqrt{1+2\Phi(r)}} \simeq 1 + \frac{GM}{r} + \mathcal{O}\left[\left(\frac{GM}{r}\right)^2\right]\,. \tag{3.31}$$

Physically, the decrease (or red-shift) in frequency seen by observers at successively larger radii corresponds to the photon's energy loss due to its having to climb out of the gravitational potential well. (The only difference between photons and massive particles climbing out of such a well is that for photons this energy loss does *not* imply a corresponding reduction of speed.)

3.3 Relativistic effects in the Solar System

It is very useful to explore the implications of weak gravity in more detail since this is the regime of real interest for most applications in near-Earth orbit, or within the Solar System. But it is also useful to go beyond the strict Newtonian limit since many measurements are sufficiently sensitive to detect the deviations between relativistic gravity and Newton's laws. We do so here in a way that is reasonably model independent, by not restricting to the specific metric that we shall later find

is predicted by Einstein's field equations. The utility of being this general is that it allows a quantitative statement as to the accuracy with which observations support the predictions of General Relativity.

Parameterized Post-Newtonian (PPN) Approximation

Our starting point is the metric, eq. (3.19), which we assume also to be static (*i.e.* t independent), and so write as

$$\begin{aligned} \mathrm{d}s^2 &= -e^{2a(r)}\,\mathrm{d}t^2 + e^{2b(r)}\,\mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right) \\ &= -\Big[1+2\Phi(r)\Big]\,\mathrm{d}t^2 + \Big[1+2\Psi(r)\Big]\,\mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right). \end{aligned} \tag{3.32}$$

Unlike in previous sections we do not stop at the Newtonian approximation for Φ and Ψ and instead write

$$\Phi(r) = -\frac{GM}{r} + (\beta-\gamma)\left(\frac{GM}{r}\right)^2 + \cdots\,, \quad \text{and} \quad \Psi(r) = \gamma\left(\frac{GM}{r}\right) + \cdots\,, \tag{3.33}$$

where β and γ are dimensionless quantities that will differ for different theories of gravity. As we shall see in detail below, in General Relativity the exact spherical solution to Einstein's equations gives

$$e^{2a(r)} = 1 + 2\Phi(r) = e^{-2b(r)} = \Big[1+2\Psi(r)\Big]^{-1} = 1 - \frac{2\,GM}{r}\,, \tag{3.34}$$

and so predicts

$$\beta = \gamma = 1 \qquad \text{(General Relativity)}\ . \tag{3.35}$$

Most of the experimental tests of General Relativity can be summarized as constraints on the range for β and γ that are allowed by observations, some of the most important of which are described in the next sections.

General properties of geodesics

Since the observational tests all involve the motion of particles or light rays within the geometry, the first step is to identify and solve the geodesic equations. We start with some general properties of geodesics for any geometry, before specializing to the spherically symmetric case.

The equation of motion which defines the trajectory, $x^\mu(\tau)$, of a freely-falling particle is the geodesic equation

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\nu\lambda}[x(\tau)]\left(\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\lambda}{\mathrm{d}\tau}\right) = 0\,, \tag{3.36}$$

where for time-like geodesics τ is the proper time measured along the trajectory. There are several first integrals of these equations that may be obtained on general grounds.

To find the first integral of this type, take the inner product of eq. (3.36) with the velocity 4-vector, $\mathrm{d}x^\mu/\mathrm{d}\tau$, and use eq. (2.39) to simplify the result:

$$\begin{aligned} 0 &= g_{\mu\nu}\left(\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\right)\left[\frac{\mathrm{d}^2x^\nu}{\mathrm{d}\tau^2}+\Gamma^\nu_{\alpha\beta}\left(\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau}\right)\right] \\ &= g_{\mu\nu}\left(\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}^2x^\nu}{\mathrm{d}\tau^2}\right)+\frac{1}{2}\,\partial_\alpha g_{\mu\beta}\left(\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau}\right) \\ &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\left[g_{\mu\nu}\left(\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\right)\right]. \end{aligned} \tag{3.37}$$

The last line uses that $g_{\mu\nu}[x(\tau)]$ is itself evaluated along the trajectory, and so must be implicitly differentiated.

This shows that the quantity $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$ is a constant along a geodesic (where $\dot{x}^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$) and so in particular its sign does not change. As a result it follows that if a particle initially starts out moving at the local speed of light, $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0$, then this is always true. Similarly, if a particle initially moves more slowly than light, $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu < 0$, then this is also always true.

Another first integral of the geodesic equations is immediate if the metric should happen to have an *isometry*. That is, if there are directions in the geometry along which the metric does not change. Recall that the metric transforms as a tensor, under a coordinate change, so $g_{\mu'\nu'}(x') = g_{\alpha\beta}(x)\,(\partial x^\alpha/\partial x^{\mu'})(\partial x^\beta/\partial x^{\nu'})$. Specializing to an infinitesimal transformation, $x^\mu = x^{\mu'} + \xi^\mu(x')$, gives $\partial x^\mu/\partial x^{\alpha'} \simeq \delta^\mu{}_\alpha + \partial_\alpha\xi^\mu$ and so the transformed metric becomes $g_{\mu'\nu'} \simeq g_{\mu\nu} + \delta g_{\mu\nu}$, with

$$\delta g_{\mu\nu} = \xi^\lambda\partial_\lambda g_{\mu\nu} + \partial_\mu\xi^\lambda g_{\lambda\nu} + \partial_\nu\xi^\lambda g_{\mu\lambda}\,. \tag{3.38}$$

This transformation is called an isometry for those ξ^μ for which eq. (3.38) vanishes, and if such a $\xi^\mu(x)$ exists it is called a *Killing vector field*. In the time-independent and spherically symmetric applications of present interest there are four such directions, corresponding to arbitrary shifts in t, and to the three independent rotations of 3-dimensional space (including in particular constant shifts of ϕ). The simplest Killing vectors are those corresponding respectively to the constant shifts in the coordinates t and ϕ, for which $\xi^\mu_{(t)} = \{1,0,0,0\}$ or $\xi^\mu_{(\phi)} = \{0,0,0,1\}$ (in the coordinates $x^\mu = \{t,r,\theta,\phi\}$), since for these $\partial_\mu\xi^\nu = 0$, and the fact that the metric does not depend on this coordinate implies $\xi^\mu_{(t)}\partial_\mu g_{\nu\lambda} = \partial_t g_{\nu\lambda} = 0$ and $\xi^\mu_{(\phi)}\partial_\mu g_{\nu\lambda} = \partial_\phi g_{\nu\lambda} = 0$.

To see why isometries help integrate the geodesic equations, multiply eq. (3.36)

through by $\xi_\mu = g_{\mu\nu}\,\xi^\nu$, to get

$$\begin{aligned}
0 &= g_{\mu\nu}\,\xi^\mu \left[\frac{\mathrm{d}^2 x^\nu}{\mathrm{d}\tau^2} + \Gamma^\nu_{\alpha\beta}\left(\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau}\right)\right] \\
&= g_{\mu\nu}\,\xi^\mu \left(\frac{\mathrm{d}^2 x^\nu}{\mathrm{d}\tau^2}\right) + \partial_\alpha g_{\mu\beta}\,\xi^\mu \left(\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau}\right) - \frac{1}{2}\,\xi^\mu \partial_\mu g_{\alpha\beta}\left(\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau}\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}\tau}\left[g_{\mu\nu}\,\xi^\mu\left(\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\right)\right] - \frac{1}{2}\,\delta g_{\alpha\beta}\left(\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau}\right). \qquad (3.39)
\end{aligned}$$

Clearly, for any ξ^μ for which $\delta g_{\alpha\beta} = 0$ the geodesic equation implies the quantity

$$Q = g_{\mu\nu}\,\xi^\mu\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\,, \qquad (3.40)$$

is a constant along the geodesic. That is, there is a *conserved quantity* for a geodesic corresponding to each symmetry of the metric.

Geodesics in static spherically symmetric spacetimes

In terms of the metric functions $a(r)$ and $b(r)$ the nonzero components of the Christoffel symbols turn out to be,

$$\begin{aligned}
&\Gamma^r_{tt} = e^{2(a-b)}\,\partial_r a\,, \quad \Gamma^t_{tr} = \partial_r a\,, \quad \Gamma^r_{rr} = \partial_r b \\
&\Gamma^r_{\theta\theta} = -r\,e^{-2b}\,, \quad \Gamma^r_{\phi\phi} = -r\sin^2\theta\,e^{-2b}\,, \quad \Gamma^\theta_{r\theta} = \frac{1}{r} \qquad (3.41) \\
&\Gamma^\phi_{r\phi} = \frac{1}{r}\,, \quad \Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta\,, \quad \Gamma^\phi_{\theta\phi} = \cot\theta\,.
\end{aligned}$$

and so the geodesic equations become

$$\begin{aligned}
\frac{\mathrm{d}^2 t}{\mathrm{d}\tau^2} + 2\,\partial_r a\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right) &= 0 \\
\frac{\mathrm{d}^2\theta}{\mathrm{d}\tau^2} - \sin\theta\cos\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2 + \frac{2}{r}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\right) &= 0 \qquad (3.42) \\
\frac{\mathrm{d}^2\phi}{\mathrm{d}\tau^2} + 2\,\cot\theta\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right) + \frac{2}{r}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right) &= 0\,,
\end{aligned}$$

and

$$\frac{\mathrm{d}^2 r}{\mathrm{d}\tau^2} + e^{2(a-b)}\,\partial_r a\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 + \partial_r b\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 - r e^{-2b}\left[\left(\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\right)^2 + \sin^2\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2\right] = 0\,. \qquad (3.43)$$

One of the above equations can be traded for the first integral corresponding to the condition that $g_{\mu\nu}(\mathrm{d}x^\mu/\mathrm{d}\tau)(\mathrm{d}x^\nu/\mathrm{d}\tau) = -1$ (or zero, for a null geodesic) along the geodesic, which implies

$$-e^{2a}\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 + e^{2b}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + r^2\left[\left(\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\right)^2 + \sin^2\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2\right] = -1 \quad (\text{or } 0)\,. \qquad (3.44)$$

The conserved quantity, $\mathcal{E}$, associated with the symmetry corresponding to shifts in t is found by multiplying the t geodesic equation by $g_{tt} = -e^{2a}$ and integrating, leading to

$$\mathcal{E} = -g_{\mu\nu}\, \xi^{\mu}_{(t)} \frac{\mathrm{d}x^{\nu}}{\mathrm{d}\tau} = e^{2a} \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right) = (1+2\Phi) \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right) , \tag{3.45}$$

being a constant along the geodesic. The corresponding conserved quantity, L, associated with shifting ϕ is similarly found by multiplying the ϕ geodesic equation by $g_{\phi\phi} = r^2 \sin^2\theta$ and integrating, implying that the angular momentum

$$L = g_{\mu\nu}\, \xi^{\mu}_{(\phi)} \frac{\mathrm{d}x^{\nu}}{\mathrm{d}\tau} = r^2 \sin^2\theta \left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right) , \tag{3.46}$$

is also constant along any geodesic.

The resulting equations can be further simplified by using the observation that motion in a spherically symmetric gravitational field lies completely within a plane.[5] This allows us the freedom to choose the orientation of the coordinate axes so that the relevant plane is described by $\theta(\tau) = \frac{\pi}{2}$, for all τ. (Notice that this choice solves the geodesic equation for θ, eq. (3.42), as claimed.) With this simplifying choice, we may use eqs. (3.45) and (3.46) to eliminate $\mathrm{d}t/\mathrm{d}\tau$ and $\mathrm{d}\phi/\mathrm{d}\tau$ from eq. (3.44), leading to the following first-order equation governing the radial motion of a geodesic

$$-\mathcal{E}^2\, e^{-2a} + e^{2b} \left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + \frac{L^2}{r^2} = -\zeta\,, \tag{3.47}$$

where $\zeta = 1$ for a time-like geodesic and $\zeta = 0$ for a null geodesic. Alternatively, this may be written

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + W_{\mathrm{eff}}(r) = 0\,, \tag{3.48}$$

which has the form $E = 0$ for the energy of one-dimensional motion in the presence of an *effective potential*

$$\begin{aligned} W_{\mathrm{eff}}(r) &= \left(\frac{L^2}{r^2} + \zeta\right) e^{-2b(r)} - \mathcal{E}^2\, e^{-2[a(r)+b(r)]} \\ &= \frac{1}{1+2\Psi} \left(\frac{L^2}{r^2} + \zeta - \frac{\mathcal{E}^2}{1+2\Phi}\right) . \end{aligned} \tag{3.49}$$

The advantage of writing the equation in this form is the intuition it provides about the kinds of orbits that are possible (once the functions $a(r)$ and $b(r)$ — or $\Phi(r)$ and $\Psi(r)$ — are specified).

[5] The fact that the motion lies in a plane ultimately can be traced to the existence of the two isometries to do with rotations that do not correspond simply to shifts in ϕ.

Gravitational redshift

We are now in a position to directly verify the earlier expression for the redshift (or energy loss) of a light ray as seen by motionless observers as it climbs away from a gravitational source. To this end suppose a light ray is sent radially outward from an observer at $(r, \theta, \phi) = (r_A, \theta_\star, \phi_\star)$ to another observer at position $(r, \theta, \phi) = (r_B, \theta_\star, \phi_\star)$. To compute the energy of this light ray as seen by these observers we must compute both their 4-velocity, u^μ, and the 4-momentum of the outgoing light ray, p^μ, and evaluate $E = -g_{\mu\nu}\, u^\mu p^\nu$.

The 4-velocity of an observer sitting at fixed spatial position, (r, θ, ϕ), is easiest to compute since it must point purely in the time direction: $u^\mu = \{u^t, 0, 0, 0\}$. The condition $g_{\mu\nu} u^\mu u^\nu = g_{tt}(u^t)^2 = -1$ then implies

$$u^t(r) = \frac{1}{\sqrt{-g_{tt}(r)}} = e^{-a(r)} = \frac{1}{\sqrt{1 + 2\Phi(r)}}\,. \tag{3.50}$$

The trajectory of the light ray, $x^\mu(w)$, is a radially out-going null geodesic for the given metric, for which the equations of the previous section can be applied, specialized to the case of radial motion: $\mathrm{d}\theta/\mathrm{d}\tau = \mathrm{d}\phi/\mathrm{d}\tau = 0$. In particular, the condition $g_{\mu\nu}(\mathrm{d}x^\mu/\mathrm{d}w)(\mathrm{d}x^\nu/\mathrm{d}w) = 0$, eq. (3.44), in this case implies

$$0 = -e^{2a(r)} \left(\frac{\mathrm{d}t}{\mathrm{d}w}\right)^2 + e^{2b(r)} \left(\frac{\mathrm{d}r}{\mathrm{d}w}\right)^2 = -\left(1 + 2\Phi\right) \left(\frac{\mathrm{d}t}{\mathrm{d}w}\right)^2 + \left(1 + 2\Psi\right) \left(\frac{\mathrm{d}r}{\mathrm{d}w}\right)^2, \tag{3.51}$$

and so the trajectory of the light ray satisfies

$$\frac{\mathrm{d}r/\mathrm{d}w}{\mathrm{d}t/\mathrm{d}w} = \frac{\mathrm{d}r}{\mathrm{d}t} = \pm\, e^{a-b} = \pm\sqrt{\frac{1 + 2\Phi}{1 + 2\Psi}}\,, \tag{3.52}$$

where the sign depends on whether the light ray is in-going or out-going. Similarly, eq. (3.45) implies that

$$\mathcal{E} = e^{2a} \left(\frac{\mathrm{d}t}{\mathrm{d}w}\right) = \left(1 + 2\Phi\right) \left(\frac{\mathrm{d}t}{\mathrm{d}w}\right), \tag{3.53}$$

is constant along the outgoing null geodesic. The tangent vector to the light ray's world-line then is

$$\frac{\mathrm{d}x^\mu}{\mathrm{d}w} = \left\{\frac{\mathrm{d}t}{\mathrm{d}w}, \frac{\mathrm{d}r}{\mathrm{d}w}, 0, 0\right\} = \frac{\mathcal{E}}{1 + 2\Phi} \left\{1, \pm\sqrt{\frac{1 + 2\Phi}{1 + 2\Psi}}, 0, 0\right\}, \tag{3.54}$$

in terms of which the photon's 4-momentum may be written $p^\mu(w) = k\, \mathrm{d}x^\mu/\mathrm{d}w$ for some constant k.

We may now compute the energy of the photon seen by the stationary observers at fixed position, which is given by

$$E(r) = -g_{\mu\nu}\, u^\mu p^\nu = -g_{tt}\, u^t p^t = \frac{k\mathcal{E}}{\sqrt{1+2\Phi(r)}}\,. \tag{3.55}$$

In particular, since $\Phi \to 0$ as $r \to \infty$ it follows that $k\mathcal{E} = E_\infty$ can be interpreted as the photon's energy as seen by observers at rest very far from the gravitating source. In this case, the energy seen by observers at rest at general r is

$$\frac{E(r)}{E_\infty} = \frac{1}{\sqrt{1+2\Phi(r)}} \simeq 1 + \frac{GM}{r} + \left(\frac{3}{2} - \beta + \gamma\right)\left(\frac{GM}{r}\right)^2 + \cdots\,, \tag{3.56}$$

which agrees with the result, eq. (3.31), of the previous section.

Deflection of light by the Sun

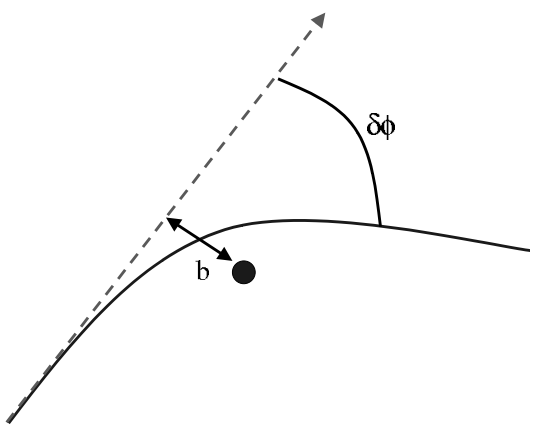


Figure 5: The geometry of light deflection by a gravitating body, showing the impact parameter, b, and deflection angle, $\delta\phi$.

The equation governing the radial motion for a more general light ray in a spherically symmetric gravitational field is eq. (3.48), together with eq. (3.49) specialized to $\zeta = 0$:

$$W_{\text{eff}}(r) = \frac{1}{1+2\Psi}\left(\frac{L^2}{r^2} - \frac{\mathcal{E}^2}{1+2\Phi}\right)\,. \tag{3.57}$$

These describe trajectories that typically escape to infinity, particularly in the weak field limit, since light rays move so swiftly they are difficult to bind into orbits. The point of closest approach to the gravitating source of such a ray corresponds to the place where $\mathrm{d}r/\mathrm{d}\tau = 0$, and so — from eq. (3.48) — occurs at $r = r_\star$, where $W_{\text{eff}}(r_\star) = 0$. That is,

$$\begin{aligned} b^2 := \frac{L^2}{\mathcal{E}^2} &= \frac{r_\star^2}{1+2\Phi(r_\star)} \\ &\simeq r_\star^2 + 2\,GMr_\star + 2(GM)^2(2-\beta+\gamma) + \mathcal{O}\left[\frac{(GM)^3}{r_\star}\right]\,. \end{aligned} \tag{3.58}$$

If $\Phi = 0$ then $r_\star = b$, and since the geodesics are straight lines in this limit, b is revealed as the *impact parameter*: the point of closest approach of the straight line obtained by extrapolating the asymptotic trajectory far from the gravitating source.

The radial coordinate of closest approach for the full trajectory is instead smaller than b, approximately given by

$$r_\star \simeq b - GM + \mathcal{O}\left[\frac{(GM)^2}{b}\right], \tag{3.59}$$

when $b \gg GM$. This is a good approximation within the solar system, since then $b \geq R_\odot$, and Table 1 shows that $GM_\odot/R_\odot \simeq 10^{-6}$.

The spatial shape of the trajectory in space, $r(\phi)$, is found by using eqs. (3.46) and (3.48) to compute $\mathrm{d}r/\mathrm{d}\phi = (\mathrm{d}r/\mathrm{d}\tau)/(\mathrm{d}\phi/\mathrm{d}\tau)$, leading to

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)^2 + \left(\frac{r^2}{L}\right)^2 W_{\text{eff}}(r) = \left(\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)^2 + \frac{r^2}{1+2\Psi}\left[1 - \frac{1}{1+2\Phi}\left(\frac{r^2}{b^2}\right)\right] = 0\,. \tag{3.60}$$

Very far from the gravitating source $\Phi, \Psi \to 0$ and so this reduces to

$$\frac{\mathrm{d}r}{\mathrm{d}\phi} \simeq \pm\left(\frac{r}{b}\right)\sqrt{r^2 - b^2}\,, \tag{3.61}$$

where the sign corresponds to which angular direction the light ray travels relative to the gravitating source. This has as solutions $b = r\cos(\phi - \phi_\star)$ (upper sign) or $b = r\sin(\phi - \phi_\star)$ (lower sign). These are the equations of a straight line, as must be so in the absence of gravity. This form confirms that b is the impact parameter of the asymptotic trajectory.

The measured quantity when a light ray is deflected by a gravitating source is the deflection, $\delta\phi$, between the asymptotic lines defined by the incident and departing rays. This is computed by inverting the expression for $\mathrm{d}r/\mathrm{d}\phi$ to obtain $\mathrm{d}\phi/\mathrm{d}r$, using eq. (3.60), and integrating the result from the initial asymptotically distant region to the final one. Since the scattering is symmetric about the point of closest approach, the total change, $\Delta\phi$, over the whole trajectory is twice the result integrated from $r = r_\star$ to $r = \infty$, leading to

$$\Delta\phi = 2\int_{r_\star}^{\infty} \mathrm{d}r\,\left(\frac{\mathrm{d}\phi}{\mathrm{d}r}\right) = 2b\int_{r_\star}^{\infty}\frac{\mathrm{d}r}{r}\sqrt{\frac{(1+2\Phi)(1+2\Psi)}{r^2 - b^2(1+2\Phi)}}\,. \tag{3.62}$$

Changing variables to $x = r/r_\star$, using the leading approximations $\Phi \simeq -GM/r$, $\Psi \simeq \gamma\, GM/r$, $r_\star \simeq b - GM$, and expanding to linear order in GM/b, this becomes

$$\begin{aligned}\Delta\phi &= 2\int_1^\infty \frac{\mathrm{d}x}{x}\sqrt{\frac{(1+2\Phi)(1+2\Psi)}{(xr_\star/b)^2 - 1 - 2\Phi}} \\ &= 2\int_1^\infty \frac{\mathrm{d}x}{x\sqrt{x^2-1}}\left[1 + \frac{GM}{bx}\left(\gamma + \frac{x^2}{x+1}\right)\right] + \mathcal{O}\left[\left(\frac{GM}{b}\right)^2\right] \\ &= \pi + 2(\gamma+1)\left(\frac{GM}{b}\right) + \mathcal{O}\left[\left(\frac{GM}{b}\right)^2\right].\end{aligned} \tag{3.63}$$

The desired scattering angle subtracts the result in the absence of gravity, $\delta\phi = \Delta\phi - \pi$, and so (restoring factors of c)

$$\delta\phi = \frac{\gamma+1}{2}\left(\frac{4GM}{b\,c^2}\right) + \mathcal{O}\left[\left(\frac{GM}{b\,c^2}\right)^2\right] \quad \text{(radians)}\,. \tag{3.64}$$

In particular, for General Relativity we have $\gamma = 1$, so applying eq. (3.64) to trajectories that just graze the Sun — *i.e.* for which $M = M_\odot$ and $b = R_\odot$ — gives $\delta\phi \simeq 1.75$ seconds of arc. (An arc-second is defined to be 1/3600 of a degree.)

> **Exercise 24:** Compute the deflection angle in Newtonian gravity for a particle whose trajectory is bent by gravity as it passes a second particle, as a function of its impact parameter, b. Specialize the result to the case where the particle's speed is $v = c$ and show that Newton would have predicted a result that is half as large as Einstein's prediction of $\delta\phi \simeq 4GM/b\,c^2$. Here $M = m_1 + m_2$ is the total mass of the two-particle system.

This effect was first observed in 1919, by searching for the deflection of starlight as it passes very close to the Sun during a total solar eclipse. The deflection is then observable as an apparent change in the position of the stars seen near the Sun during the eclipse as compared with their relative positions when the Sun is elsewhere in the sky. Because the light rays are bent towards the Sun, during the eclipse their apparent position as seen from Earth is displaced away from the Sun, by an amount that falls off with their angular separation from the Sun.

Modern measurements instead perform this measurement using very long baseline radio telescopes to observe astrophysical radio sources when these are near the Sun. The use of long baseline interferometry provides much improved angular resolution, as well as the advantage that the Sun is not as brilliant a foreground obstruction in radio wavelengths as it is in visible light. The main complications arise from the presence of a plasma of ionized particles in the solar corona near the Sun, whose presence provides an index of refraction for the radio waves and so can bend their trajectories. Unlike the relativistic effect, the influence of the solar corona is frequency dependent, however, and so can be disentangled by making observations at more than one frequency. The resulting constraint on the PPN parameter γ is

$$\gamma = 1.007 \pm 0.009\,, \tag{3.65}$$

and so agrees well with the prediction $\gamma = 1$ of General Relativity.

Shapiro time delay

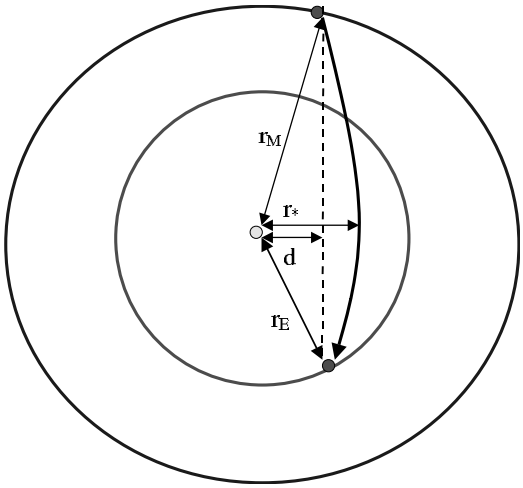


Figure 6: The geometry of time delay measurements, showing the impact parameter, d, point of closest approach, $r_\star$, and the distances to the Earth, $r_E = r_\oplus$ and Mars r_M.

A second observable related to the trajectories of light rays in the presence of gravity is associated with the change in transit time for light rays that travel very close to the solar surface [2]. This can be measured by sending signals to other planets (such as to space probes orbiting Mars or on the Martian surface) and back and measuring the result as a function of the planetary position as it passes through *superior conjunction* (*i.e.* when it is on the opposite side of the Sun from the Earth).

Recall that it takes light about 8 minutes to reach the Earth from the Sun, and it takes a radio signal about 40 minutes to make the round trip across the 740 million km from Earth to Mars at its most distant. Since the Earth's orbital speed is roughly 30 km/s, during this time the Earth only moves through about 70,000 km, largely at right angles to the line of sight to Mars. As a result we can treat the Earth and Mars to be at rest for the purposes of the calculation.

Suppose the instantaneous Sun-Earth distance is denoted $r_\oplus$, and the same for Mars is r_M, and if the radial position of the radio signal's closest approach to the Sun is $r_\star$. In the absence of gravity the time taken for the round-trip passage of a signal from Earth to Mars (see Figure 6) is

$$\Delta t_0 = 2\left(\sqrt{r_M^2 - d^2} + \sqrt{r_\oplus^2 - d^2}\right), \tag{3.66}$$

where d is the distance from the Sun of the nearest point on the straight line connecting Mars to the Earth. Each square root of the form $\sqrt{r^2 - d^2}$ gives the light travel time along the straight-line trajectory to r from $r = d$, and the factor of 2 appears because we seek the round-trip time. Notice that, unlike for the impact parameter b in the calculation for the deflection of light, the quantity d satisfies $d < r_\star$, because the relevant straight-line trajectory is the one passing directly from Earth to Mars, and not the one tangent to the asymptotic light ray at infinite distance.

With gravity present, the radius of closest approach is found by asking where $\mathrm{d}r/\mathrm{d}\tau = 0$ along the geodesic trajectory, leading to eq. (3.58), which states $W_{\text{eff}}(r_\star) = 0$, and so $r_\star = b - GM + \cdots$, where $b = L/\mathcal{E}$.

The time elapsed (as seen by a distant motionless observer) during the radio signal's trip is found by integrating $\mathrm{d}t/\mathrm{d}r = (\mathrm{d}t/\mathrm{d}\tau)/(\mathrm{d}r/\mathrm{d}\tau)$, using eqs. (3.48) and (3.45). That is,

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 + \left(\frac{1+2\Phi}{\mathcal{E}}\right)^2 W_{\mathrm{eff}}(r) = 0\,, \tag{3.67}$$

and so the round-trip time evolved becomes $\Delta t = 2[\mathcal{T}(r_\star, r_M) + \mathcal{T}(r_\star, r_\oplus)]$, with

$$\mathcal{T}(r_\star, r_x) = \int_{r_\star}^{r_x} \mathrm{d}r \, \left(\frac{\mathrm{d}t}{\mathrm{d}r}\right) = \int_{r_\star}^{r_x} \mathrm{d}r \sqrt{\frac{1+2\Psi}{1+2\Phi}} \, \left[1 - \frac{b^2}{r^2}\,(1+2\Phi)\right]^{-1/2} . \tag{3.68}$$

Writing $r = x\, r_\star$ and expanding to leading order in $GM/r_\star$ then gives

$$\begin{aligned}
\mathcal{T}(r_\star, r_x) &= r_\star \int_1^{r_x/r_\star} \mathrm{d}x \sqrt{\frac{1+2\Psi}{1+2\Phi}} \, \left[1 - \frac{b^2}{x^2 r_\star^2}\,(1+2\Phi)\right]^{-1/2} \\
&= r_\star \int_1^{r_x/r_\star} \frac{\mathrm{d}x}{\sqrt{x^2-1}} \left[x + \frac{GM}{r_\star}\left(1+\gamma+\frac{1}{x+1}\right)\right] + \mathcal{O}\left[\left(\frac{GM}{r_\star}\right)^2\right] \\
&\simeq \sqrt{r_x^2 - r_\star^2} + GM\left\{(1+\gamma)\cosh^{-1}\left(\frac{r_x}{r_\star}\right) + \tanh\left[\frac{1}{2}\cosh^{-1}\left(\frac{r_x}{r_\star}\right)\right]\right\} .
\end{aligned} \tag{3.69}$$

These expressions may be simplified using $\cosh^{-1} x = \ln\left(x + \sqrt{x^2-1}\right)$ and $\tanh\left(\frac{1}{2}\,z\right) = (e^z - 1)/(e^z + 1)$ and so

$$\tanh\left(\frac{1}{2}\cosh^{-1} x\right) = \frac{x - 1 + \sqrt{x^2-1}}{x + 1 + \sqrt{x^2-1}} = \sqrt{\frac{x-1}{x+1}}\,, \tag{3.70}$$

to get (with c's re-instated) $\Delta t = 2[\mathcal{T}(r_\star, r_M) + \mathcal{T}(r_\star, r_\oplus)]$, with

$$c\mathcal{T}(r_\star, r_x) \simeq \sqrt{r_x^2 - r_\star^2} + \frac{GM}{c^2}\left[(1+\gamma)\ln\left(\frac{r_x + \sqrt{r_x^2 - r_\star^2}}{r_\star}\right) + \sqrt{\frac{r_x - r_\star}{r_x + r_\star}}\right] , \tag{3.71}$$

up to terms of order $(GM)^2/r_\star c^4$.

In the applications of interest to the solar system this may be simplified using $r_x \gg r_\star$ to drop all terms suppressed by $(r_\star/r_x)^2$, whose accuracy is controlled by $(R_\odot/r_\oplus)^2 \simeq 2\times 10^{-5}$, an amount about 10 times larger than $GM_\odot/R_\odot$. In this case the total time delay becomes

$$\Delta t \simeq \Delta t_0 + \left(\frac{1+\gamma}{2}\right)\frac{4\,GM}{c^3}\ln\left(\frac{4\,r_M r_\oplus}{r_\star^2}\right) . \tag{3.72}$$

This neglects the product $(r_\star/r_\oplus)^2(GM/r_\star)$, which means that the difference between d and $r_\star$ can be neglected in the first term, allowing it to be written as the transit time, Δt_0, found in the absence of gravity, eq. (3.66).

The size of this effect for signals sent to Mars during superior conjunction is about 250 μsec out of a total round-trip travel time of about 40 minutes. Although this represents only a change of one part in 10^7, it can be measured precisely due to the great stability of atomic clocks, which can be accurate to a part in 10^{12}. The orbits of the planets are also known to sufficient precision to make their positions known to an accuracy of about a kilometre, meaning that the timing effect is also not swamped by the distance uncertainty. The biggest measurement errors are associated with the effects of propagation through the ions of the solar corona, as was the case for measurements of the solar deflection of light. The resulting precision obtained for the PPN parameter γ from the Viking Mars Mission is [3]

$$\gamma = 1.000 \pm 0.002\,. \tag{3.73}$$

More recent measurements of the same effect for signals sent to the Cassini probe at Saturn have improved this accuracy to [4]

$$\gamma - 1 = (-1.3 \pm 5.2) \times 10^{-5}\,, \tag{3.74}$$

again in good agreement with the prediction $\gamma = 1$ of General Relativity.

Orbital precession

Another classic test of General Relativity within the solar system concerns the orbits of planets and satellites rather than the motion of light rays. In this case the relevant equations are those for a time-like geodesic, rather than a null one, and so the radial dependence is given by eqs. (3.48) and (3.49), with $\zeta = 1$ rather than zero, and so

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + W_{\text{eff}}(r) = 0\,, \tag{3.75}$$

with

$$W_{\text{eff}}(r) = \frac{1}{1+2\Psi}\left(\frac{L^2}{r^2} + 1 - \frac{\mathcal{E}^2}{1+2\Phi}\right)\,, \tag{3.76}$$

with conservation of energy and momentum given by eqs. (3.45) and (3.46),

$$\mathcal{E} = (1+2\Phi)\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right) \quad \text{and} \quad L = r^2\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)\,. \tag{3.77}$$

The Newtonian Limit

It is useful to have in mind the properties of the Newtonian orbits before investigating their relativistic corrections.

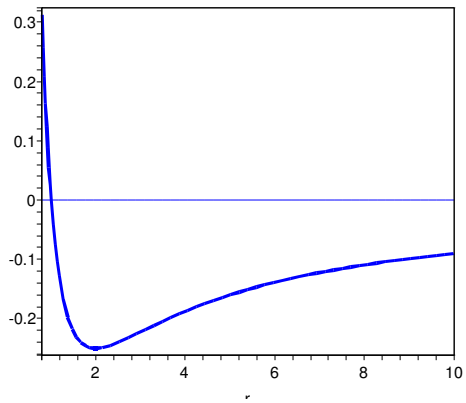


Figure 7: A plot of the Newtonian effective potential against r.

Recall that for orbits, the Newtonian limit of these equations corresponds to $\Phi = -GM/r = \mathcal{O}(v^2)$, and $\mathrm{d}t/\mathrm{d}\tau = 1 + \mathcal{O}(v^2)$ and so $\mathcal{E} = 1 + \varepsilon$ with $\varepsilon = \mathcal{O}(v^2)$. In this case $\Psi = \mathcal{O}(v^2)$ only contributes at $\mathcal{O}(v^4)$ and so can be completely neglected, leaving eq. (3.75) in the familiar form from the Newtonian Kepler problem,

$$\frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + \frac{L^2}{2\,r^2} + \Phi(r) = \varepsilon\,. \tag{3.78}$$

In particular we see that $L = r^2(\mathrm{d}\phi/\mathrm{d}\tau)$ is the usual specific angular momentum, while ε plays the role of the total Newtonian energy. The effective potential appearing here, $V_{\text{eff}}(r) = (L^2/2r^2) - GM/r$, is plotted in Fig. 7, which displays the divergence $V_{\text{eff}} \to +\infty$ as $r \to 0$ when $L \neq 0$, thereby showing how angular momentum excludes an orbiting particle from approaching too close to $r \to 0$. For $r \to \infty$ the limit instead is $V_{\text{eff}} \to 0$ from below, showing that orbits with $\varepsilon \geq 0$ escape to infinity while those with $\varepsilon < 0$ describe bound orbits.

The bound orbits are confined to lie within a finite range of radii, $r_- \leq r \leq r_+$, whose endpoints are determined by the conditions $\mathrm{d}r/\mathrm{d}\tau = 0$. Eq. (3.78) allows these to be determined in terms of the conserved quantities L and ε, since they must be roots of

$$\frac{L^2}{2\,r^2} - \frac{GM}{r} = \varepsilon\,. \tag{3.79}$$

The smaller of the two roots, r_-, corresponds to the point of closest approach to the Sun, and is called its *perihelion*. *Aphelion*[6] defines the point on the orbit furthest from the Sun, given by the larger of the two roots, $r = r_+$. Solving eq. (3.79) gives the explicit expressions

$$\frac{1}{r_\pm} = \frac{GM \mp \sqrt{(GM)^2 - 2L^2|\varepsilon|}}{L^2}\,, \tag{3.80}$$

or, equivalently,

$$r_\pm = \frac{GM \pm \sqrt{(GM)^2 - 2L^2|\varepsilon|}}{2|\varepsilon|}\,. \tag{3.81}$$

[6]For orbits about the Earth the corresponding points are instead called *perigee* and *apogee*, and for orbits about other stars the terms are *periastron* and *apastron*.

The explicit shape, $r(\phi)$, of the bound orbits in the Newtonian case is found by combining eqs. (3.78) and (3.77) to obtain

$$\frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)^2 = \frac{1}{2}\left(\frac{\mathrm{d}r/\mathrm{d}\tau}{\mathrm{d}\phi/\mathrm{d}\tau}\right)^2 = \left(\frac{r^2}{L}\right)^2\left(\varepsilon + \frac{GM}{r} - \frac{L^2}{2\,r^2}\right). \tag{3.82}$$

This can be explicitly integrated by changing variables to $u = 1/r$, giving solutions $u = A + B\cos\phi$ where A and B are constants. These describe bound orbits that are ellipses, with the constants A and B related to their semi-major axis a and eccentricity $0 \le e < 1$. In terms of a and e:

$$r(\phi) = \frac{a(1-e^2)}{1 + e\cos\phi}\,. \tag{3.83}$$

This shows that the points closest to and furthest from the Sun are given by $r_\pm = a(1 \pm e)$. Comparing this with the expressions for $r_\pm$ in terms of L and ε allows these conserved quantities to be given in terms of a and e by

$$L^2 = GMa(1-e^2) \quad \text{and} \quad \varepsilon = -\frac{GM}{2a}\,, \tag{3.84}$$

and so $L^2/(2\,|\varepsilon|) = a^2(1-e^2) = r_+ r_-$ and $r_+ + r_- = 2\,a = GM/|\varepsilon|$.

There are two different ways to define the *period* of the orbit, both of which happen to give the same result in the Newtonian limit. One definition, P_r, is defined in terms of the radial motion as the time taken to move between successive perihelia. This can be found by recognizing that $\mathrm{d}t/\mathrm{d}\tau \simeq 1$ in the Newtonian limit, and integrating eq. (3.78)

$$\begin{aligned} P_r &= 2\int_{r_-}^{r_+} \mathrm{d}r\left(\frac{\mathrm{d}t}{\mathrm{d}r}\right) = 2\int_{r_-}^{r_+} \mathrm{d}r\left[2\,\varepsilon + \frac{2\,GM}{r} - \frac{L^2}{r^2}\right]^{-1/2} \\ &= \frac{2}{\sqrt{2\,|\varepsilon|}}\int_{r_-}^{r_+} \frac{r\mathrm{d}r}{\sqrt{(r_+ - r)(r - r_-)}} = \frac{\pi(r_+ + r_-)}{\sqrt{2\,|\varepsilon|}}\,, \end{aligned} \tag{3.85}$$

and so

$$\left(\frac{P_r}{2\pi}\right)^2 = \frac{(GM)^2}{(2\,|\varepsilon|)^3} = \frac{a^3}{GM}\,, \tag{3.86}$$

in agreement with Newton's modification of *Kepler's Third Law*.

A second way to define the orbital period is in terms of the angular motion, as the time, P_ϕ, required to sweep out 2π radians:

$$\begin{aligned} P_\phi &= \int_0^{2\pi} \mathrm{d}\phi\left(\frac{\mathrm{d}t}{\mathrm{d}\phi}\right) = \int_0^{2\pi} \mathrm{d}\phi\left(\frac{r^2}{L}\right) = \frac{2\,a^2(1-e^2)^2}{L}\int_0^{\pi} \frac{\mathrm{d}\phi}{(1+e\cos\phi)^2} \\ &= \frac{2\,a^2(1-e^2)^2}{L}\left[\frac{\pi}{(1-e^2)^{3/2}}\right] = P_r\,. \end{aligned} \tag{3.87}$$

Because these two notions of period agree with one another, the Newtonian orbit passes through precisely the same points every time ϕ cycles through 2π radians, and so is said to be *closed*.

More generally this is not the case in relativistic systems, and any mismatch $P_r \neq P_\phi$ implies the orbit *precesses*, with successive perihelions occurring at different angular positions, displaced by the *perihelion shift*, $\delta\phi_{\rm prec} := \Delta\phi - 2\pi$, with

$$\Delta\phi := 2\int_{r_-}^{r_+} \mathrm{d}r \left(\frac{\mathrm{d}\phi}{\mathrm{d}r}\right) . \tag{3.88}$$

For the Newtonian orbits $\delta\phi_{\rm prec} = 0$, because

$$\begin{aligned}\Delta\phi &= L\int_{r_-}^{r_+} \frac{\mathrm{d}r}{r\sqrt{2\,GMr - 2\,|\varepsilon| r^2 - L^2}} = \frac{L}{\sqrt{2\,|\varepsilon|}}\int_{r_-}^{r_+} \frac{\mathrm{d}r}{r\sqrt{(r_+ - r)(r - r_-)}} \\ &= \sqrt{a^2(1-e^2)}\left(\frac{2\pi}{\sqrt{r_+ r_-}}\right) = 2\pi\,,\end{aligned} \tag{3.89}$$

as expected, since $P_r = P_\phi$.

Relativistic Precession

We may now see how the leading relativistic corrections change these Newtonian results. The main observable effect from the point of view of testing General Relativity is the violation of the relation $P_r = P_\phi$ that relativistic effects induce, leading to a nonzero prediction for the orbital precession angle, $\delta\phi_{\rm prec}$.

To this end we recompute eq. (3.88) by going back to the full expressions, eqs. (3.75) and (3.77), for the orbital shape, $r(\phi)$. These give

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\phi}\right)^2 = \frac{1}{r^4}\left(\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)^2 = -\frac{1}{L^2(1+2\Psi)}\left(\frac{L^2}{r^2} + 1 - \frac{\mathcal{E}^2}{1+2\Phi}\right) , \tag{3.90}$$

where $u := 1/r$. Expanding this out to next-to-leading order in powers of $GM/r = GMu$ and $\mathcal{E} = 1 + \varepsilon$ gives

$$\begin{aligned}\left(\frac{\mathrm{d}u}{\mathrm{d}\phi}\right)^2 \simeq \frac{1}{L^2}\Big[&\left(-L^2u^2 + 2\,GMu + 2\,\varepsilon\right)(1 - 2\gamma\,GMu) \\ &+2(2+\gamma-\beta)(GMu)^2 + \varepsilon^2 + 4\,\varepsilon GMu\Big] .\end{aligned} \tag{3.91}$$

The relativistic correction terms have several effects. First, they change the position of the zeroes of the right-hand side of eq. (3.91), to $u_\pm = u_{0\pm} + \delta u_\pm$, where

$$\begin{aligned}\delta u_\pm &\simeq \frac{(2+\gamma-\beta)(GMu_{0\pm})^2 + \frac{1}{2}\,\varepsilon^2 + 2\,\varepsilon\,(GMu_{0\pm})}{L^2 u_{0\pm} - GM} \\ &= \frac{(2+\gamma-\beta)(GMu_{0\pm})^2 + \frac{1}{2}\,\varepsilon^2 + 2\,\varepsilon\,(GMu_{0\pm})}{\pm\sqrt{(GM)^2 + 2L^2\varepsilon}} \\ &\simeq \pm\left(\frac{GM}{a^2 e}\right)\left[\frac{2+\gamma-\beta}{(1\pm e)^2} + \frac{1}{8} - \frac{1}{1\pm e}\right] ,\end{aligned} \tag{3.92}$$

in which the second line uses eq. (3.80) to simplify the denominator, and the third line expresses L, ε and $u_{0\pm}$ in terms of a and e using the equations for the Newtonian orbits.

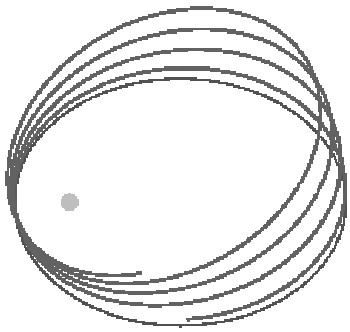


Figure 8: The precession of an elliptical orbit, such as is caused by deviations from the inverse-square force law.

The angle $\Delta\phi$ then becomes

$$\begin{aligned}\Delta\phi &= 2\,L\int_{u_+}^{u_-}\frac{\mathrm{d}u}{\sqrt{Au^3+Bu^2+Cu+D}} \\ &\simeq 2\pi - L\int_{u_+}^{u_-}\mathrm{d}u\,\left[\frac{\delta A\,u^3+\delta B\,u^2+\delta C\,u+\delta D}{(B_0u^2+C_0u+D_0)^{3/2}}\right],\end{aligned} \tag{3.93}$$

where

$$B_0 = -L^2\,,\quad C_0 = 2\,GM \quad\text{and}\quad D_0 = 2\,\varepsilon\,, \tag{3.94}$$

while

$$\begin{aligned}\delta A = 2\gamma GML^2\,,\quad \delta B = 2(2-\gamma-\beta)(GM)^2\,,\\ \delta C = 4(1-\gamma)\,\varepsilon\,GM \quad\text{and}\quad \delta D = \varepsilon^2\,.\end{aligned} \tag{3.95}$$

The integral in the second line of eq. (3.93) is subtle to evaluate because it diverges as $u \to u_{0\pm}$. Although $\delta u_+ > 0$ and $\delta u_- < 0$, so the range of integration does not include $u_{0\pm}$, it is nonetheless true that this near-divergence complicates the expansion of the integral in powers of GM/a. Such an evaluation gives (restoring factors of c)

$$\delta\phi_{\text{prec}} = \Delta\phi - 2\pi = \left(\frac{2+2\gamma-\beta}{3}\right)\frac{6\pi}{(1-e^2)}\left(\frac{GM}{a\,c^2}\right). \tag{3.96}$$

Exercise 25: Verify that eq. (3.96) follows from eq. (3.93), as claimed.

Astronomy has a long history of precise observations of planetary orbits, and most orbits are observed to precess. However there are several complication to be addressed before these can be compared with the prediction, eq. (3.96). First of all, Newton's Law only predicts strictly elliptical orbits for a planet orbiting the Sun in the absence of the gravitational pull of all of the other planets, and in the approximation that the Sun is perfectly spherical. Deviations from these two idealizations perturb the orbits, typically causing them to precess. The calculated contribution of these more mundane perturbations must be subtracted from any observed precession before any relativistic effects can be identified.

Deviations of this type from the predictions of Newtonian mechanics were identified very early, and were historically used to predict the existence of some of the outer planets before their actual discovery. By the turn of the 20th century all such

planetary effects had been accounted for, and only one observation remained in disagreement with predictions: a small anomalous precession in the orbit of Mercury. This is measured to precess — relative to the vernal equinox (*i.e.* the place in the sky where the Sun crosses the celestial equator in the spring as seen from the Earth) — by a very small amount: 5599.7 arc-seconds per century. For comparison, the amount expected within Newtonian gravity is given in the first three rows of the following table, which sum to the Newtonian prediction of 5557.0 arc-sec/century.

Source	Amount (arcsec/century)
Earth's spin precession	5025.6
Other planets	531.4
Solar oblateness	0.03
Relativity	42.98 ± 0.04
Total	5600.0

The difference between the observations and the Newtonian prediction, 43 arc-sec/century, is larger than the theoretical and observational errors, and its interpretation remained a puzzle, until the discovery of General Relativity. Remarkably, the contribution of eq. (3.96) for $\beta = \gamma = 1$ is precisely the amount required to bring theory into agreement with observations. This was one of the clinchers for Einstein and others in the early days of General Relativity. Given the bounds on γ coming from the deflection of light and the Shapiro time delay, the agreement of predictions with the orbit of Mercury gives the following limit on β:

$$\beta = 1.000 \pm 0.003\,. \tag{3.97}$$

There is an analogous relativistic precession of the orbits of other planets, and some asteroids, and although the orbits of the remaining innermost planets are so close to circular that their precession is hard to measure, all extant observations agree well with the predictions. The comparison for the innermost planets and the asteroid Icarus is given in the following table [5].

Object	GR prediction (arcsec/century)	Observation (arcsec/century)
Mercury	43.0	43.1 ± 0.05
Venus	8.6	8.4 ± 4.8
Earth	3.8	5.0 ± 1.2
Icarus	10.3	9.8 ± 0.8

4. Field equations for curved space

The content of general relativity has been summarized (by John Wheeler) as the statements that "Spacetime tells Matter how to move" and "Matter tells Spacetime how to curve".

The previous section has explored the implication of generalizing Newton's First Law to the assumption that particles move on geodesics in the absence of any non-gravitational forces. This is how spacetime tells matter to move. The remainder of this section describes the field equations, which is how matter makes spacetime curve. These equations are necessary for predicting which metric should be relevant to describe the gravitational field in any given situation.

4.1 Gravity as curvature

The first step towards formulating the field equations is to identify how they should depend on the metric. To this end we seek a quantity that expresses precisely what is different about a gravitating geometry. Whatever this quantity is, it should be a tensor so that whatever the distinction is, all observers will agree on it (much as they all agree on what it means to be a geodesic).

Freely falling observers

The principle of equivalence states that a freely-falling observer in a gravitational field finds the local laws of physics are the same as those given in special relativity. These observers are those whose coordinates are such that $g_{\mu\nu} = \eta_{\mu\nu}$ and $\Gamma^{\mu}_{\nu\lambda} = 0$ at the relevant point, and so geodesics correspond to the condition $\mathrm{d}^2x^\mu/\mathrm{d}\tau^2 = 0$. Mathematically, it is always possible to find such an observer at any point, and the coordinates of these observers are called *Gaussian normal coordinates*.

In general it is *not* possible to find a similar class of observers simultaneously for all of the points throughout an entire region of spacetime, and according to Einstein the failure to be able to do so is the signature of the existence of a gravitational field. We therefore seek a tensor that can be used to distinguish a metric that describes a gravitational field, from one which is simply Minkowski space written in a bizarre set of coordinates.

Since the issue is whether or not $\Gamma^{\mu}_{\nu\lambda}$ can be made to vanish throughout an entire region, even though this is always possible at a given point, the obstruction is to do with the ability to choose coordinates that set *derivatives*, $\partial_\rho\Gamma^{\mu}_{\nu\lambda}$, to zero at a given point, as well $\Gamma^{\mu}_{\nu\lambda}$ itself. We therefore expect the tensor which expresses the obstruction to involve derivatives of the Christoffel symbols, and so second derivatives of the metric.

The tensor that provides the obstruction to making the Christoffel symbols vanish throughout a region is the natural generalization to spacetime of the curvature, encountered in earlier sections when describing the differential geometry of space. That is, the existence of observers for which $\Gamma^{\mu}_{\nu\lambda} = 0$ throughout some region can be shown to be equivalent to the vanishing of the Riemann curvature tensor, $R^{\mu}{}_{\nu\lambda\rho}$, throughout the same region, where

$$R^{\mu}{}_{\nu\lambda\rho} = \partial_\lambda \Gamma^{\mu}_{\nu\rho} + \Gamma^{\mu}_{\lambda\sigma}\Gamma^{\sigma}_{\nu\rho} - (\lambda \leftrightarrow \rho)\,. \tag{4.1}$$

Recalling that the Christoffel symbols are defined by

$$\Gamma^{\mu}_{\nu\lambda} = \frac{1}{2}\, g^{\mu\rho}\Big(\partial_\nu g_{\lambda\rho} + \partial_\lambda g_{\nu\rho} - \partial_\rho g_{\nu\lambda}\Big)\,, \tag{4.2}$$

it is clear that the Riemann tensor involves second derivatives of the metric tensor.

Because $R^{\mu}{}_{\nu\lambda\rho}$ transforms as a tensor, if it vanishes in any set of coordinates, it must also vanish for all others. This means that although the laws of nature can be made into those of special relativity simply by transforming to an appropriate freely-falling frame), this does not mean that all the effects of gravity are removed in such a frame. This cannot be true, since the curvature tensor, $R^{\mu}{}_{\nu\lambda\rho}$, cannot be similarly removed simply by performing a coordinate transformation. Einstein's point with the principle of equivalence was not that gravity is purely a fictitious frame-dependent thing, but rather that it is the *tidal* forces of gravity that are present for all observers, and it is the curvature of spacetime that encodes these tidal effects.

4.2 Einstein's field equations

We may now state the field equation that expresses how sources of mass and energy give rise to gravitational fields, that generalizes the Newtonian field equation for the gravitational potential, Φ:

$$\nabla^2\Phi = 4\pi G\,\mu\,, \tag{4.3}$$

where μ is the local mass density. We've seen that the Newtonian potential, Φ, is naturally expressed as a component of the metric, $g_{\mu\nu}$, and since eq. (4.3) involves second derivatives of Φ it is natural to seek a generalization with the curvature tensor appearing on the left-hand side.

Einstein proposed that the spacetime curvature tensor, $R^{\mu}{}_{\nu\lambda\rho}$, is related to the local distribution, $T_{\mu\nu}$, of stress-energy by the following field equations:

$$R_{\mu\nu} - \frac{1}{2}R\, g_{\mu\nu} = \frac{8\pi G}{c^2}\, T_{\mu\nu}\,, \tag{4.4}$$

where $T_{\mu\nu}$ is the stress-energy tensor that describes the conserved energy and momentum of matter, $R_{\mu\nu} = R^{\lambda}{}_{\mu\lambda\nu}$ is the spacetime's *Ricci tensor* and $R = g^{\mu\nu}R_{\mu\nu}$ is its *Ricci scalar*. The left-hand-side of this equation is the most general one which satisfies the following three conditions:

1. It transforms as a symmetric tensor (as does $T_{\mu\nu}$);

2. It involves exactly two derivatives of $g_{\mu\nu}$ (which is the relativistic generalization of the Newtonian potential Φ, because $g_{tt} \approx -1 - 2\Phi$ in the non-relativistic limit); and

3. It is covariantly conserved inasmuch as: $\nabla^\mu \left(R_{\mu\nu} - \frac{1}{2} R\, g_{\mu\nu} \right) = 0$.

In the above ∇_μ denotes the *covariant derivative*, defined so that $\nabla^\mu T_{\mu\ldots} = g^{\mu\nu} \nabla_\mu T_{\nu\ldots}$, and

$$\nabla_\mu T^{\alpha_1\ldots}{}_{\beta_1\ldots} = \partial_\mu T^{\alpha_1\ldots}{}_{\beta_1\ldots} + \Gamma^{\alpha_1}_{\mu\rho} T^{\rho\ldots}{}_{\beta_1\ldots} + \cdots - \Gamma^{\rho}_{\mu\beta_1} T^{\alpha_1\ldots}{}_{\rho\ldots} - \cdots . \tag{4.5}$$

It is defined in this way in order to have the following properties: $\nabla_\mu T_{\nu\ldots\lambda}$ transforms as a tensor under coordinate changes if $T_{\nu\ldots\lambda}$ does, and for a freely-falling observer (for whom $\Gamma^\mu_{\nu\lambda} = 0$ at a particular point) it reduces to a regular partial derivative: $\partial_\mu T_{\nu\ldots\lambda}$. Given these properties, the third condition listed above is motivated by the generalization to curved space,

$$\nabla_\mu T^{\mu\nu} = \partial_\mu T^{\mu\nu} + \Gamma^\mu_{\mu\alpha} T^{\alpha\nu} + \Gamma^\nu_{\mu\alpha} T^{\mu\alpha} = 0\,, \tag{4.6}$$

of the conservation of stress-energy, eq. (2.66). Notice that because it is a tensor equation, if $\nabla_\mu T^{\mu\nu}$ vanishes for any observer it must vanish for all observers. But energy conservation requires $\nabla_\mu T^{\mu\nu} = 0$ because eq. (4.6) reduces to eq. (2.66) for a freely falling observer, for whom $\Gamma^\mu_{\nu\lambda}$ vanishes at a particular point.

Two comments are in order about Requirement 2, that the left-hand side involve only two derivatives:

1. Requirement 2 should not be regarded as being fundamental. Rather, keeping in mind that our observational knowledge of gravity is largely confined to comparatively weak gravitational fields, it should be regarded as the leading contribution in an expansion of the left-hand side in powers of the curvature. As such it expresses our ignorance about strong curvatures, and we should expect any inferences drawn from General Relativity to be suspect when the curvatures become sufficiently large. How large? This is not known, but we should beware whenever any dimensionless measure of curvature (like $Gg^{\mu\nu}R_{\mu\nu}$ or $G^2 R_{\mu\nu\lambda\rho} R^{\mu\nu\lambda\rho}$) should become large.

2. Requirement 2 states that the left-hand side should contain precisely two derivatives of the metric, but if this equation is to be regarded as being a derivative expansion one should really keep all terms having *up to* two derivatives. In fact there is one possible term involving no derivatives at all, and this

should be expected to dominate if derivatives are small. Including this term revises eq. (4.4) to

$$R_{\mu\nu} - \frac{1}{2} R\, g_{\mu\nu} + \lambda\, g_{\mu\nu} = 8\pi G\, T_{\mu\nu}\,, \tag{4.7}$$

where the constant λ is known as the 'cosmological term'. At present there is evidence from cosmology that λ is actually nonzero, but very small compared with the contribution of the right-hand side of eq. (4.7) in all applications apart from cosmology. For simplicity we ignore this term in the following sections, but return to it in the later discussion of cosmology.

Taking the trace of eq. (4.4) implies $R = -8\pi G\, T$, where $T = g^{\mu\nu} T_{\mu\nu}$ is the trace of the stress tensor. Using this in eq. (4.4) gives the Einstein equations in their *trace-reversed* form:

$$R_{\mu\nu} = 8\pi G \left(T_{\mu\nu} - \frac{1}{2} T g_{\mu\nu} \right). \tag{4.8}$$

In particular, a *vacuum spacetime* is one for which no matter is present, and so $T_{\mu\nu} = 0$. Eq. (4.8) implies any such spacetime is Ricci flat: $R_{\mu\nu} = 0$.

Exercise 26: Use the definitions to compute the Ricci scalar for an n-dimensional space whose metric is $g_{\mu\nu} = e^{2\phi}\, \eta_{\mu\nu}$, where $\phi(x)$ is a scalar function and $\eta_{\mu\nu}$ is the usual flat Minkowski metric. Show that it is given by

$$R = -2(n-1)\partial^2\phi - (n-1)(n-2)(\partial\phi)^2\,, \tag{4.9}$$

where $\partial^2\phi := \eta^{\mu\nu}\partial_\mu\partial_\nu\phi$ and $(\partial\phi)^2 := \eta^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi$.

4.3 Rotationally invariant solutions

This section now derives some of the solutions to Einstein's equations which describe the geometries outside of symmetric gravitating sources, such as stars, planets or black holes.

Birkhoff's Theorem: Spherical symmetry implies static

Consider first the geometry outside of a spherical distribution of matter. It is assumed that there is no matter outside of the distribution, and so $T_{\mu\nu} = 0$ in the region of interest. The goal of this section is to identify the most general solution to the vacuum Einstein equations which is spherically symmetric. We do so *without* making the additional assumption of time-independence.

We saw earlier that it is always possible to choose coordinates in a spherically symmetric geometry so that the metric takes the form of eq. (3.19). The metric cannot be simplified further using only symmetries and coordinate choices, so the

functions $a(r,t)$ and $b(r,t)$ must be determined by solving Einstein's field equations for the vacuum: $R_{\mu\nu} = 0$. To this end the next step is to specialize Einstein's equations to the special case of the metric given in eq. (3.19).

Plugging into the definitions the nonzero components of the Christoffel symbols become:

$$\begin{aligned}
&\Gamma^t_{tt} = \partial_t a\,, \quad \Gamma^t_{tr} = \partial_r a\,, \quad \Gamma^t_{rr} = e^{2(b-a)}\,\partial_t b \\
&\Gamma^r_{tt} = e^{2(a-b)}\,\partial_r a\,, \quad \Gamma^r_{tr} = \partial_t b\,, \quad \Gamma^r_{rr} = \partial_r b \\
&\Gamma^r_{\theta\theta} = -r\,e^{-2b}\,, \quad \Gamma^r_{\phi\phi} = -r\sin^2\theta\,e^{-2b}\,, \quad \Gamma^\theta_{r\theta} = \frac{1}{r} \\
&\Gamma^\phi_{r\phi} = \frac{1}{r}\,, \quad \Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta\,, \quad \Gamma^\phi_{\theta\phi} = \cot\theta\,.
\end{aligned} \tag{4.10}$$

Exercise 27: Verify that eqs. (4.10) follow from a direct application of the definition of $\Gamma^\mu_{\nu\lambda}$ to the metric of eq. (3.19), as claimed.

Using these components of the Christoffel symbols in the definition of the Riemann tensor then leads to the following nonzero components:

$$\begin{aligned}
&R^t{}_{rtr} = e^{2(b-a)}\Big[\partial_t^2 b + (\partial_t b)^2 - \partial_t a\,\partial_t b\Big] - \partial_r^2 a - (\partial_r a)^2 + \partial_r a\,\partial_r b\,, \\
&R^t{}_{\theta t\theta} = -r\,e^{-2b}\partial_r a\,, \quad R^t{}_{\phi t\phi} = -r\,e^{-2b}\sin^2\theta\,\partial_r a\,, \\
&R^t{}_{\theta r\theta} = -r\,e^{-2a}\partial_t b\,, \quad R^t{}_{\phi r\phi} = -r\,e^{-2a}\sin^2\theta\,\partial_t b\,, \\
&R^r{}_{\theta r\theta} = r\,e^{-2b}\partial_r b\,, \quad R^r{}_{\phi r\phi} = r\,e^{-2b}\sin^2\theta\,\partial_r b\,, \\
&R^\theta{}_{\phi\theta\phi} = (1-e^{-2b})\sin^2\theta\,.
\end{aligned} \tag{4.11}$$

Finally, taking the trace of this to obtain the Ricci tensor leads to

$$\begin{aligned}
&R_{tt} = \partial_t^2 b + (\partial_t b)^2 - \partial_t a\,\partial_t b + e^{2(a-b)}\left[\partial_r^2 a + (\partial_r a)^2 - \partial_r a\,\partial_r b + \frac{2\partial_r a}{r}\right] \\
&R_{rr} = -\partial_r^2 a - (\partial_r a)^2 + \partial_r a\,\partial_r b + \frac{2\partial_r b}{r} + e^{2(b-a)}\Big[\partial_t^2 b + (\partial_t b)^2 - \partial_t a\,\partial_t b\Big] \\
&R_{tr} = \frac{2\partial_t b}{r}\,, \quad R_{\theta\theta} = 1 + e^{-2b}\Big[r(\partial_r b - \partial_r a) - 1\Big]\,, \quad R_{\phi\phi} = R_{\theta\theta}\sin^2\theta\,.
\end{aligned} \tag{4.12}$$

Exercise 28: Verify that eqs. (4.11) and (4.12) follow from a direct application of the definitions, using the components of $\Gamma^\mu_{\nu\lambda}$ given in eq. (4.10), as claimed.

The goal is to use the five equations found by setting $R_{\mu\nu} = 0$ to solve for the two unknown functions $a(r,t)$ and $b(r,t)$. Although this seems like it should be an over-determined problem (too many equations for the number of unknowns), it is not for two reasons. The first reason is the spherical symmetry of the problem (which

is also what reduced the metric to two independent functions). For example the conditions $R_{\theta\theta} = 0$ and $R_{\phi\phi} = 0$ are not independent conditions, and this is a generic consequence of spherical symmetry. However, the remaining four equations still do not over-determine a and b because of the *Bianchi identity*, $\nabla^\mu \left(R_{\mu\nu} - \frac{1}{2} R\, g_{\mu\nu}\right) = 0$, implies that they are not all independent.

Birkhoff's Theorem

The simplest equation to solve is $R_{tr} = 0$, which implies $b = b(r)$ is t-independent. Differentiating $R_{\theta\theta} = 0$ with respect to t and using $\partial_t b = 0$, then implies the further condition $\partial_t \partial_r a = 0$, whose general solution is $a(r,t) = f(r) + g(t)$, for arbitrary functions f and g. This makes the time component of the metric become $-e^{2a}\mathrm{d}t^2 = -e^{2f(r)}[e^{g(t)}\mathrm{d}t]^2$, which shows that the function $g(t)$ can be removed by redefining the t coordinate from t to t', with $\mathrm{d}t' = e^{g(t)}\mathrm{d}t$. Once this has been done it follows that the remaining metric functions are independent of time: $a = a(r)$ and $b = b(r)$:

$$\mathrm{d}s^2 = -e^{2a(r)}\mathrm{d}t^2 + e^{2b(r)}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,. \tag{4.13}$$

This result is important, so it has a name: *Birkhoff's theorem.* It states that the assumption of spherical symmetry is sufficient in itself to ensure that the geometry is also time-independent. A metric like eq. 4.13, for which coordinates exist for which all components of $g_{\mu\nu}$ are independent of t and there are no terms linear[7] in $\mathrm{d}t$, is called *static.* If the metric can only be made t-independent in coordinates for which $\mathrm{d}t\,\mathrm{d}x^i$ cross-terms exist, then the metric is instead called *stationary.*

The Schwarzschild Solution

Given the t-independence of a and b, the components of the Ricci tensor simplify to

$$\begin{aligned} R_{tt} &= e^{2(a-b)}\left[\partial_r^2 a + (\partial_r a)^2 - \partial_r a\,\partial_r b + \frac{2\partial_r a}{r}\right] \\ R_{rr} &= -\partial_r^2 a - (\partial_r a)^2 + \partial_r a\,\partial_r b + \frac{2\partial_r b}{r} \\ R_{\theta\theta} &= 1 + e^{-2b}\Big[r(\partial_r b - \partial_r a) - 1\Big]\,, \quad R_{\phi\phi} = R_{\theta\theta}\sin^2\theta\,. \end{aligned} \tag{4.14}$$

A simple equation is obtained by taking the combination $R_{tt}e^{2(b-a)} + R_{rr} = 0$, which gives

$$\frac{2}{r}\Big(\partial_r a + \partial_r b\Big) = 0\,. \tag{4.15}$$

This implies $a + b = k$, where k is an r-independent constant. The constant k can be set to zero without loss of generality simply by rescaling the time coordinate $t \to e^{-k}t$, leaving the result $a(r) = -b(r)$. Using this in $R_{\theta\theta} = 0$ implies

$$e^{2a}\Big(2r\partial_r a + 1\Big) = \partial_r\Big(re^{2a}\Big) = 1\,, \tag{4.16}$$

[7] More precisely, for which the vector in the time direction is 'hypersurface orthogonal'.

whose solution is

$$e^{2a} = 1 - \frac{r_s}{r}\,, \tag{4.17}$$

where the integration constant, r_s, has dimensions of length, and is called the *Schwarzschild radius*. As is easily checked, no further information is obtained from setting to zero any of the other components of $R_{\mu\nu}$ given in eq. (4.14).

The value of the integration constant, r_s, can be found by examining the large-r limit, for which the metric approaches the metric for flat space (written in polar coordinates): $\mathrm{d}s^2 \to -\mathrm{d}t^2 + \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$. A metric having this property is said to be *asymptotically flat*. Since the flatness of the metric at large r implies the gravitational field is weak there, the Newtonian limit applies and so $g_{tt} \approx -1 - 2\Phi$, where $\Phi = -GM/r$ is the Newtonian potential for a spherical source having mass[8] M. Comparing this with the large-r limit of $g_{tt} = -e^{2a} = -1 + r_s/r$ gives (re-introducing the factors of c),

$$r_s = \frac{2\,GM}{c^2}\,. \tag{4.18}$$

The final result is the Schwarzschild geometry

$$\mathrm{d}s^2 = -\left(1 - \frac{r_s}{r}\right)\mathrm{d}t^2 + \left(1 - \frac{r_s}{r}\right)^{-1}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,, \tag{4.19}$$

whose weak-field limit ($r \gg r_s$) is obtained by expanding in powers of r_s/r, and gives the Parameterized Post-Newtonian form, $e^{2a} = -[1 - 2\,GM/r + (\beta-\gamma)(GM/r)^2 + \cdots]$ and $e^b = 1 + \gamma(GM/r) + \cdots$, with $\beta = \gamma = 1$, as was discussed in earlier sections.

Notice that r_s is very small for ordinary astrophysical objects. For instance using the solar mass, $M_\odot = 2 \times 10^{33}$ g, leads to $r_s = 3$ km. For such objects the geometry of eq. (4.19) becomes inappropriate once one reaches the 'edge' of the sun, $r = R_\odot = 700,000$ km, inside of which $T_{\mu\nu}$ no longer vanishes. Because of this the entire exterior of the star is effectively in the weak-field limit $r > R_\odot \gg r_s$.

5. Compact stars and Black Holes

This section explores some of the physical consequences of the spherically symmetric solutions obtained in the previous section, going beyond the limit of weak gravitational fields considered earlier.

Geodesics

Given the metric, the motion of freely falling observers can be found by integrating the geodesic equations, eq. (3.36). This relies on having explicit expressions for the

[8] There are a number of definitions of mass in GR, and defining M in this way is equivalent to using that of Arnowitt, Deser and Misner (ADM) in this case.

Christoffel symbols for the Schwarzschild geometry, which are given by specializing eqs. (4.10) to the case $e^{-2b} = e^{2a} = 1 - r_s/r$:

$$\begin{aligned}
&\Gamma^t_{tr} = -\Gamma^r_{rr} = \frac{r_s}{2r(r-r_s)} \qquad \Gamma^r_{tt} = \frac{r_s(r-r_s)}{2r^3} \\
&\Gamma^r_{\theta\theta} = -(r-r_s)\,, \quad \Gamma^r_{\phi\phi} = -(r-r_s)\sin^2\theta \\
&\Gamma^\theta_{r\theta} = \Gamma^\phi_{r\phi} = \frac{1}{r}\,, \quad \Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta\,, \quad \Gamma^\phi_{\theta\phi} = \cot\theta\,.
\end{aligned} \tag{5.1}$$

Using these gives the geodesics as solutions, $x^\mu(\tau) = [t(\tau), r(\tau), \theta(\tau), \phi(\tau)]$, to the following equations:

$$\begin{aligned}
\frac{\mathrm{d}^2 t}{\mathrm{d}\tau^2} + \left[\frac{r_s}{r(r-r_s)}\right]\frac{\mathrm{d}r}{\mathrm{d}\tau}\frac{\mathrm{d}t}{\mathrm{d}\tau} &= 0 \\
\frac{\mathrm{d}^2 r}{\mathrm{d}\tau^2} + \left[\frac{r_s(r-r_s)}{2r^3}\right]\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 - \left[\frac{r_s}{2r(r-r_s)}\right]\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 & \\
-(r-r_s)\left[\left(\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\right)^2 + \sin^2\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2\right] &= 0 \\
\frac{\mathrm{d}^2\theta}{\mathrm{d}\tau^2} + \frac{2}{r}\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\frac{\mathrm{d}r}{\mathrm{d}\tau} - \sin\theta\cos\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2 &= 0 \\
\frac{\mathrm{d}^2\phi}{\mathrm{d}\tau^2} + \frac{2}{r}\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\frac{\mathrm{d}r}{\mathrm{d}\tau} + 2\cot\theta\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\frac{\mathrm{d}\phi}{\mathrm{d}\tau} &= 0\,.
\end{aligned} \tag{5.2}$$

5.1 Orbits

As discussed earlier, solving these equations themselves is in general a mess. However because of the symmetries of the geometry there are a number of conservation laws, which help obtain solutions. Spherical symmetry ensures the conservation of angular momentum, and the conservation of the direction of angular momentum requires the trajectory to be restricted to a plane in space. We are free to choose our coordinates so that this plane corresponds to $\theta = \pi/2$, and it is clear that $\theta(\tau) = \pi/2$ is indeed a solution to the third of eqs. (5.2). Using this result, the conservation of the magnitude of angular momentum can then be seen by multiplying the last of eqs. (5.2) by r^2, to give $(\mathrm{d}/\mathrm{d}\tau)[r^2(\mathrm{d}\phi/\mathrm{d}\tau)] = 0$. This leads to the first integral

$$r^2\frac{\mathrm{d}\phi}{\mathrm{d}\tau} = L \tag{5.3}$$

where L is a constant.

Time-translation invariance similarly leads to energy conservation, whose form is found by multiplying the first of eqs. (5.2) by $(1 - r_s/r)$, to get $(\mathrm{d}/\mathrm{d}\tau)[(1 - r_s/r)(\mathrm{d}t/\mathrm{d}\tau)] = 0$. Integrating then gives the first integral

$$\left(1 - \frac{r_s}{r}\right)\frac{\mathrm{d}t}{\mathrm{d}\tau} = \mathcal{E}\,, \tag{5.4}$$

where $\mathcal{E}$ is a constant. Furthermore, eq. (3.37) shows that it is also always true that

$$\begin{aligned}\zeta &= -g_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} \\ &= \left(1-\frac{r_s}{r}\right)\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 - \left(1-\frac{r_s}{r}\right)^{-1}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 \\ &\qquad -r^2\left[\left(\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\right)^2 + \sin^2\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2\right], \end{aligned} \tag{5.5}$$

is also conserved along any geodesic. For timelike geodesics we usually choose τ to be proper time along the trajectory, in which case $\zeta = 1$. For null geodesics describing the propagation of light we must instead choose $\zeta = 0$. This last equation may be simplified by using the three conservation laws given above, allowing the derivatives $\mathrm{d}t/\mathrm{d}\tau$, $\mathrm{d}\theta/\mathrm{d}\tau$ and $\mathrm{d}\phi/\mathrm{d}\tau$ to be eliminated in favour of the constants L and $\mathcal{E}$, giving the following first-order equation to be solved for $\mathrm{d}r/\mathrm{d}\tau$:

$$\zeta = \left(1-\frac{r_s}{r}\right)^{-1}\mathcal{E}^2 - \left(1-\frac{r_s}{r}\right)^{-1}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 - \frac{L^2}{r^2}\,. \tag{5.6}$$

In principle one solves this equation for $r(\tau)$, and after plugging the result into eqs. (5.3) and (5.4), integrates these to obtain $\phi(\tau)$ and $t(\tau)$.

This last equation can be put into a form with which one can become emotionally involved, by multiplying through by $\frac{1}{2}(1 - r_s/r)$:

$$\frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + V(r) = E\,, \tag{5.7}$$

where

$$\begin{aligned} V(r) &= \frac{1}{2}\left(1-\frac{2GM}{r}\right)\left(\frac{L^2}{r^2}+\zeta\right) \\ &= \left[\frac{L^2}{2r^2} - \frac{\zeta GM}{r}\right] - \frac{L^2 GM}{r^3} + \frac{\zeta}{2} \\ \text{and}\quad E &= \frac{\mathcal{E}^2}{2}\,. \end{aligned} \tag{5.8}$$

What is attractive about eq. (5.7) is that it has the form of the energy equation in elementary mechanics for one-dimensional motion in a potential, $V(r)$, for a particle having energy $\mathcal{E}$. This is attractive because there is considerable intuition about the properties of the solutions based on the shape of the potential.

Orbits of massive particles

Consider first the timelike geodesics which describe the world-lines of massive particles moving slower than the speed of light, corresponding to the choice $\zeta = 1$ in the

above expressions. It is useful to contrast the relativistic result with what happens for orbits in the Newtonian limit. To this end notice that the effective potential governing the radial motion of orbits in the Newtonian limit is given by the square bracket in the second equality for $V(r)$: that is $V_c(r) = (L^2/2r^2) - GM/r$.

To infer the qualitative properties of orbits in the Newtonian limit notice that $V_c(r) \to +\infty$ as $r \to 0$ and $V_c(r) \to 0$ from below as $r \to \infty$. This implies $V_c(r)$ must have a minimum for some intermediate value, $r = r_c$, which differentiation shows lies at $r = r_c \equiv L^2/GM$. Furthermore, r is time independent at this minimum provided that the 'energy' satisfies $E = V_c(r_c)$, and so $r = r_c$ gives the position of the circular orbits for a given L. Since this is a minimum of V_c, circular orbits are stable for any L and orbits which start near $r = r_c$ will oscillate about this point. The period of this radial oscillation is given by $\omega_r^2 = V_c''(r_c)$, and so $\omega_r = (GM)^2/L^3$. On the other hand, for circular orbits the angular frequency of the orbit's angular motion is given by $\omega_\phi = (\mathrm{d}\phi/\mathrm{d}\tau) = L/r_c^2 = (GM)^2/L^3$. This result $\omega_r = \omega_\phi$ is related to these orbits being ellipses having a fixed orientation in space, since the time between successive closest approaches (perihelia) is the same as the time taken to circumnavigate the orbit once.

How does all this change in the relativistic case? In this case $V(r) \to -\infty$ as $r \to 0$ and $V(r) \to \frac{1}{2}$ from below as $r \to \infty$. Differentiating V shows that $V'(r)$ vanishes when

$$r = r_{c\pm} \equiv \frac{L}{r_s}\Big[L \pm \Big(L^2 - 3\,r_s^2\Big)^{1/2}\Big]\,. \tag{5.9}$$

We see from this that if $L < \sqrt{3}\,r_s$ then $V(r)$ has no real minima or maxima, and so no circular orbits are possible at all. Orbits then come in two classes: those coming in from infinity, which have $E \geq \frac{1}{2}$, or $\mathcal{E}^2 \geq 1$; and those which cannot escape from the gravitational source, having $E < \frac{1}{2}$, or $\mathcal{E}^2 < 1$. In both cases, once r begins to decrease it necessarily reaches $r = 0$ (and so at some point either reaches $r = r_s$ or crashes into the source's surface).

If, on the other hand, $L > \sqrt{3}\,r_s$, then $V(r)$ has a local minimum at $r = r_{c+}$ and a maximum at $r = r_{c-}$. This shows that stable orbits occur at $r = r_{c+}$, and the radius of these orbits grows as L does. The smallest stable orbit occurs when $L = \sqrt{3}\,r_s$, and occurs at $r_{\min} = 3\,r_s = 6GM$. On the other hand, for $L \gg r_s$ the radius of the stable orbit becomes $r_{c+} \to L^2/GM$, which agrees with the Newtonian result (as we should expect because $GM/r_c = (GM)^2/L^2 = (r_s/2L)^2 \ll 1$). Orbits which start near such circular orbits will oscillate about this radius, with frequency

$$\omega_r^2 = V''(r_c) = \frac{3L^2}{r_c^4} - \frac{2GM}{r_c^3} - \frac{12L^2GM}{r_c^5}\,. \tag{5.10}$$

Since this does not agree with the frequency of the angular motion, defined by $\omega_\phi = (\mathrm{d}\phi/\mathrm{d}\tau) = L/r_c^2$, the motion describes the precessing ellipses seen in earlier lectures.

Exercise 29: Compute ω_ϕ for a stable circular orbit as a function of L and GM and compare it to the frequency, ω_r, of small radial oscillations about the same circular orbit, computed using eq. (5.10). From these calculate the precession angle, $\delta\phi_{\rm prec}$, that accumulates per period for nearly circular elliptical orbits in Schwarzschild spacetime. Does your result agree with the small-eccentricity limit of the post-Newtonian result found earlier in eq. (3.96)?

Circular orbits are also possible at $r = r_{c-}$, which *decreases* with increasing L. However because this is a maximum of $V(r)$ these orbits are unstable, and small perturbations from them cause the trajectory to veer into the source or to escape out to infinity. In particular, the outermost of these circular orbits occurs for the smallest possible L, corresponding to $r_{c-} \to 6\,GM$ as $L \to \sqrt{3}\,r_s$ (which coincides with r_{c+} in this limit). The smallest possible unstable circular orbit instead occurs as $L \to \infty$, which corresponds to $r_{c-} \to \frac{3}{2}\,r_s = 3\,GM$.

Exercise 30: Show that circular orbits in Schwarzschild spacetime exactly satisfy Kepler's 3rd Law: $\Omega^2 = GM/r^3$, where $\Omega = \mathrm{d}\phi/\mathrm{d}t = (\mathrm{d}\phi/\mathrm{d}\tau)/(\mathrm{d}t/\mathrm{d}\tau)$.

Orbits of light rays

The trajectories for massless particles (like photons, gravitons and possibly some neutrinos) are found in an identical fashion, using instead $\zeta = 0$, as appropriate for null geodesics. In this case the potential $V(r)$ degenerates to

$$V(r) = \frac{L^2}{2r^2}\left(1 - \frac{2\,GM}{r}\right)\,, \tag{5.11}$$

for which V' vanishes at the L-independent value $r = r_c \equiv \frac{3}{2}\,r_s = 3\,GM$. Since $V(r) \to -\infty$ as $r \to 0$ and $V(r) \to 0^+$ as $r \to \infty$, V has a maximum at $r = r_c$. This shows that there is only one possible circular orbit for a light ray, and this is unstable — occurring at $r = 3\,GM$. Furthermore, since $V(r_c) = \frac{1}{6}[L/(3\,GM)]^2$, photons which approach from infinity do not get closer than $r = 3\,GM$ provided their 'energy' satisfies $E < V(r_c)$, or $|\mathcal{E}| < L/(3\sqrt{3}\,GM)$. Trajectories having $|\mathcal{E}|$ larger than this necessarily reach $r < \frac{3}{2}\,r_s$.

5.2 Radial geodesics

We have seen that orbits exist for which test particles can move to arbitrarily small r, and this means that we may have to take seriously the potential singularities of the metric at $r = 0$ and $r = r_s$. (More about these singularities in the next section.)

Since for any given $\mathcal{E}$ the orbits which penetrate to small r have small L it is useful to study in more detail radially-directed geodesics corresponding to particles which fall directly in (or climb directly out) of the gravitational potential. For simplicity it is also useful to follow the fastest-moving particles, and so specialize to null geodesics.

If we focus on the shape of these geodesics in the (r, t) plane, it is convenient not to separately find $r(\tau)$ and $t(\tau)$, and to instead directly use the condition

$$\mathrm{d}s^2 = 0 = -(1 - r_s/r)\,\mathrm{d}t^2 + (1 - r_s/r)^{-1}\mathrm{d}r^2\,, \tag{5.12}$$

to get

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \pm\left(1 - \frac{r_s}{r}\right)\,. \tag{5.13}$$

This integrates to give the curves $r_*(r) = \pm t$, where the upper (lower) sign corresponds to outward-going (in-falling) geodesics. Since the *tortoise coordinate*, r_*, defined by

$$r_* = r + r_s \ln\left|\frac{r}{r_s} - 1\right|\,, \tag{5.14}$$

approaches r for large r, these trajectories get closer and closer to the flat-space geodesics, $r = \pm t$, as $r \to \infty$. Notice also that $r_* \to -\infty$ as $r \to r_s$.

Suppose we now examine what happens to an in-falling light ray, for which $r_* = -t$. At asymptotically late times, $t \to \infty$, r_* approaches $-\infty$ and so r asymptotically approaches r_s from above. Even though we found that orbits are not energetically precluded from reaching $r = 0$, the above result makes it seem as if an infinite amount of time is required to reach the Schwarzschild radius. And this is indeed true, although it is important in relativity to specify more precisely whose time the coordinate t keeps track of.

Imagine therefore filling spacetime with observers who hover at a fixed radius and angle in the Schwarzschild gravitational field. (Since these are not geodesics, these observers would have to use rockets to accelerate and keep from falling in the ambient gravitational field.) Only the coordinate t varies along the world-line of such an observer, but the proper time as measured by one of these observers is given by

$$\mathrm{d}\tau^2 = -\mathrm{d}s^2 = \left(1 - \frac{r_s}{r}\right)\mathrm{d}t^2\,, \tag{5.15}$$

and so $\mathrm{d}\tau = (1 - r_s/r)^{1/2}\,\mathrm{d}t$. In general this differs from $\mathrm{d}t$ because of the gravitational redshift associated with each observer's position, and so t represents the time measured only by the asymptotic observer at $r \to \infty$.

The result above therefore shows that as seen by an observer at infinity, an in-falling light ray takes an infinite amount of time to reach $r = r_s$. It does *not* show that this takes an infinite amount of time as measured by the in-falling observers

themselves. This can be determined by returning to our geodesic expression, eq. (5.7), in the case $L = 0$:

$$\frac{\mathrm{d}r}{\mathrm{d}\tau} = \pm \left[\mathcal{E}^2 - \zeta\left(1 - \frac{r_s}{r}\right)\right]^{1/2} . \tag{5.16}$$

For in-falling null geodesics we choose $\zeta = 0$, and so $r = r_0 - \mathcal{E}(\tau - \tau_0)$, showing that $r = r_s$ is reached in a finite parameter interval along the null geodesic. A similar conclusion can be drawn for in-falling timelike geodesics, for which $\zeta = 1$. This is most simply done by choosing the special case $\mathcal{E} = 1$, for which $\mathrm{d}r/\mathrm{d}\tau = -\sqrt{r_s/r}$, and so $r \propto \tau^{2/3}$.

We conclude that in-falling observers pass $r = r_s$ in a finite amount of their own time. Paradoxically, this is not inconsistent with the infinite amount of time taken as seen by the observer from infinity. To understand this suppose that the in-falling astronaut were to send regularly spaced signals out to the observer at infinity during the trip. Because of the gravitational redshift, these signals arrive at infinity spaced further apart than they were on their emission, with this redshift becoming infinite as the astronaut reaches $r = r_s$.

5.3 Singularities of the solution

Because coordinates can be chosen arbitrarily in General Relativity, it is always important to check that they mean what they are assumed to mean. This is usually done by using the metric to compute physical distances, such as when we chose the radial coordinate earlier to be the radius or area of the spheres at fixed r and t. However, because the metric itself is only found after solving the field equations, it may be that the coordinates do not end up having all of the properties they were assumed to have when they were chosen. For this reason it is always important to check the properties of the metric which results, to see what it implies about the properties of various coordinate surfaces.

The first thing to check is that the metric is well-defined: *i.e.* that its components are finite and the metric is invertible. (Invertibility is important because if $g_{\mu\nu}$ is not invertible then the infinitesimal coordinate displacements, $\mathrm{d}x^\mu$, are not linearly independent and so do not span all of the possible directions in the space.) Inspection of the Schwarzschild metric, eq. (4.19), shows that there are two places which might be problems: $r = 0$ and $r = r_s$. Clearly neither of these is of real interest for most astrophysical objects, for which the solution does not apply down to such small radii.

Curvature singularity: $r = 0$

The geometry near $r = 0$ is counter-intuitive because for all $0 < r < r_s$ it is g_{rr} which is negative, while g_{tt} is positive. This means that in this region it is r, *and not* t, that is the time coordinate!

$r = 0$ seems problematic, because the components of the metric and curvature tensors all diverge at this point. This need not represent a physical problem in itself, however, because the components of a tensor are different in different coordinate systems, and it could just be that our coordinates are poorly chosen near $r = 0$. For instance, even starting with the flat metric $\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$ can lead to divergent metric components after performing the coordinate change $x = 1/(w-3)$, since $\mathrm{d}x = -\mathrm{d}w/(w-3)^2$ implies there are metric components which diverge as $w \to 3$ or vanish as $w \to \infty$. In this case this is a sign that the coordinate transformation $x \to w$ is singular (because x or w diverges).

Is this what is happening for the Schwarzschild solution at $r = 0$? If so there would exist an inspired change of coordinates which would remove the singularities in the metric and curvature as $r \to 0$. A sufficient condition for such a change of coordinates *not* to exist is if there is a scalar quantity which diverges, since a scalar takes the same value in all coordinate systems. Examples of scalars might be R, $R_{\mu\nu}R^{\mu\nu}$ or the eigenvalues, λ_a, of the matrix $R^\mu{}_\nu$. (These last are scalars because the covariance of the eigenvalue equation $R^\mu{}_\nu v_a^\nu = \lambda_a v_a^\mu$ requires λ_a to be a scalar. Notice the same would not be true for the eigenvalues of the matrix $R_{\mu\nu}$!) Unfortunately, none of those listed helps for the Schwarzschild geometry, because this satisfies $R_{\mu\nu} = 0$ by construction. However, there is a scalar that diverges at $r = 0$:

$$R_{\mu\nu\lambda\rho}R^{\mu\nu\lambda\rho} = \frac{12\, r_s^2}{r^6}\,, \tag{5.17}$$

and this shows that $r = 0$ really is a *curvature singularity*, and not merely a *coordinate singularity*.

Given that we believe Einstein's equations are likely to be weak-curvature expansions of something more fundamental, we should be wary of taking too seriously the properties of the Schwarzschild solution very near $r = 0$.

Coordinate singularity: $r = r_s = 2\,GM$

What about the singularity seen in eq. (4.19) as $r \to r_s$? Is this also a curvature singularity? It is the purpose of this section to argue that this is a coordinate singularity, which merely expresses the breakdown of the Schwarzschild coordinates for $r \leq r_s$. An indication that this is possible comes from the fact that nothing particular seems to happen to in-falling observers as they reach r_s.

This suggests dropping the coordinates r and t and instead trying coordinates which are adapted to in-falling and out-going radial light rays. To this end consider the new coordinates u and v, defined in terms of the tortoise coordinate, r_*, of eq. (5.14):

$$u = t - r_* \qquad\qquad v = t + r_*\,. \tag{5.18}$$

Radially in-falling light rays are described in these coordinates by constant v, while radially out-going light rays travel along the lines of constant u.

The idea is to trade t, the Schwarzschild time variable, for either u or v. For instance, *Eddington-Finkelstein* coordinates are defined by using the coordinates (v, r), in terms of which the Schwarzschild metric becomes

$$\mathrm{d}s^2 = -\left(1 - \frac{r_s}{r}\right)\mathrm{d}v^2 + 2\mathrm{d}v\mathrm{d}r + r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2)\,, \tag{5.19}$$

and so $g_{vv} = -(1 - r_s/r)$, $g_{rv} = g_{vr} = 1$ and $g_{rr} = 0$. It is clear that none of the metric components diverge anymore as $r \to r_s$, although some do pass through zero there. However, zero values for metric elements are not in themselves a problem — after all, there were plenty of zeros in the diagonal Schwarzschild metric for $r \neq r_s$ — provided that the metric remains invertible. However the determinant of the above metric is $g \equiv \det g_{\mu\nu} = -r^4 \sin^2\theta$, which is independent of GM and so takes the same value as it does in flat space. This shows that the vanishing of $\det g$ when r or $\sin\theta$ vanish is just a reflection of the breakdown of spherical polar coordinates at $r = 0$, or at $\theta = 0$ and $\theta = \pi$ for nonzero r.

5.4 Black Holes and Event Horizons

Since the metric can be made nonsingular at $r = r_s$ simply by performing a change of coordinates, there is nothing intrinsically singular about the Schwarzschild geometry at $r = r_s$. It is just that the Schwarzschild coordinates, (t, r, θ, ϕ), themselves break down at this point. Now that we have coordinates that do not break down, what then is the interpretation of the surface $r = r_s$?

To see this consider again the trajectories of in-falling and out-going light rays. In Eddington-Finkelstein coordinates the condition $\mathrm{d}s = 0$ implies these satisfy

$$\mathrm{d}s^2 = 0 = \mathrm{d}v\left[2\mathrm{d}r - \left(1 - \frac{r_s}{r}\right)\mathrm{d}v\right]\,, \tag{5.20}$$

and so $\mathrm{d}v/\mathrm{d}r = 0$ for in-falling light rays, and $\mathrm{d}v/\mathrm{d}r = 2(1 - r_s/r)^{-1}$ for out-going light rays. Notice that for the outgoing rays, this means that $\mathrm{d}r/\mathrm{d}v$ is positive only if $r > r_s$, but $\mathrm{d}r/\mathrm{d}v < 0$ when $r < r_s$. This shows that r always decreases when $r < r_s$, *even for the outgoing light ray*! At $r = r_s$, the outgoing ray satisfies $\mathrm{d}r/\mathrm{d}v = 0$, and so the ray simply 'hovers' at $r = r_s$. That is to say, the surface $r = r_s$ is a *null surface*, spanned by null geodesics.

These arguments show that the surface $r = r_s$ serves as the point of no return, inasmuch as no light signal emitted at $r < r_s$ can escape to $r > r_s$. The same is also true for timelike geodesics, as might have been expected since these particles necessarily move more slowly than do light rays. The existence of such a surface also makes sense from the following point of view. If the escape speed, v_{esc}, were

computed as a function of r using Newtonian physics, it would be defined as that speed that gets the object to infinity with precisely zero kinetic energy, and so would satisfy

$$\frac{mv_{esc}^2}{2} - \frac{GMm}{r} = 0\,, \tag{5.21}$$

and so $v_{esc}^2 = 2\,GM/r$. The radius at which $v_{esc} = c$ would then be $r_s = 2\,GM/c^2$, in agreement with the Schwarzschild radius. (There is no reason why the numerical factors of the Newtonian calculation should agree exactly with the relativistic calculation, but it is nonetheless a happy accident that they do.) What is new to special relativity is the proscription of motion with $v > c$, which completely precludes the ability for anything to escape from $r < r_s$.

A surface such as this, which divides spacetime into regions between which signals cannot be sent due to the speed of light being the maximum speed, is called an *event horizon*. The surface $r = r_s$ is an event horizon for the Schwarzschild geometry. The region with $r < r_s$ is called a *black hole*, since it is something from which nothing, not even light, can classically escape.

Validity of the approximations

If gravitational effects are so dramatic as to divide spacetime into two regions like this, one might ask whether the curvatures are too large to trust our use of Einstein's equations to predict them. It is worth keeping in mind when doing so that we have three scales in the problem, namely G, M and r, and so there are two independent dimensionless ratios which we can form from them. These are GM/r and G/r^2 (in our units with $\hbar = c = 1$). It turns out that each of these controls a different kind of approximation.

- *Relativistic Effects:* We have already seen that the first of these, $2GM/rc^2 \sim r_s/r$ (once the factors of c are restored), controls the importance of relativistic effects, and the fact that this is $O(1)$ when $r \sim r_s$ shows that relativistic effects are crucial to understanding the properties of the event horizon.

- *Quantum effects:* Our treatment of gravity has been purely classical, and it turns out that the relative size of quantum corrections to our treatment are of order G/r^2 — or $\hbar G/c^3 r^2$ once $\hbar$ and c are restored (notice the tell-tale $\hbar$, the signature of quantum effects). The classical approximation is typically a good one provided that this ratio is small. Since the unit of length — the Planck length — associated with G is very small, $\ell_p = (\hbar G/c^3)^{1/2} \sim 10^{-33}$ cm, the condition $G/r^2 = (\ell_p/r)^2 \ll 1$ is not a very strong restriction for any r of astrophysical interest!

- *Weak curvature:* Recall that Einstein's equations are motivated as being the weak-curvature approximation to some possibly more fundamental theory, and so corrections to these equations might be expected to arise that are of order GR, where R is any invariant notion of the local curvature. (For example, one might think of GR being the square root of the invariant $G^2 R_{\mu\nu\lambda\rho} R^{\mu\nu\lambda\rho}$ for the Schwarzschild geometry.) But inspection of the Riemann tensor for Schwarzschild shows that in order of magnitude the components of $R_{\mu\nu\lambda\rho}$ are of order GM/r^3 and so $GR \sim G^2 M/r^3 \sim (G/r^2)(r_s/r)$, which shows how curvature corrections to Einstein's equations are related to the size of quantum corrections.

The above arguments indicate that the effects of quantum gravity near the event horizon, when $r = r_s$, should be of order

$$\delta_s = \frac{\hbar G}{c^3 r_s^2} = \frac{\ell_p^2}{r_s^2} = \frac{\hbar c}{4GM^2} = \frac{M_p^2}{4M^2}\,, \tag{5.22}$$

and this gets smaller the larger M is. The quantity $M_p = \sqrt{\hbar c/G} = 2.18 \times 10^{-8}$ kg is called the Planck mass, and its size shows that $\delta_s \ll 1$ for the black holes of astrophysics, for which $M > M_\odot \simeq 1.99 \times 10^{30}$ kg. Given that the interpretation of astrophysical objects as black holes is based purely on the classical predictions of general relativity, one might have worried that this interpretation might be undermined by unknown quantum gravity effects. The fact that $\delta_s \ll 1$ for such black holes shows that this worry is likely to be groundless.

Quantum effects would be important, however, for very light black holes, such as if their mass were as small as that of an elementary particle like a proton, whose mass is $m_p \simeq 1.67 \times 10^{-27}$ kg. We should not trust any classical inferences about the gravitational field of a proton at radii as small as its Schwarzschild radius, and so have no reason to believe these should behave gravitationally as classical black holes.

5.5 Quantum effects near Black Holes

What about black holes with masses in between these two extremes? For a black hole with $M \simeq 10^{-3}$ kg (*i.e.* 1 gram) we have $\delta_s \simeq 10^{-10}$, ensuring that it is massive enough that the classical approximation would be very good, even at the Schwarzschild radius. On the other hand, although quantum effects are small for such a black hole, they need not be completely negligible. Are there any novel quantum phenomena that might arise?

Particle Production

The first quantum property of a black hole that can arise in this way as a small quantum correction in a controlled semiclassical approximation was found in the 1970 s by Stephen Hawking. He discovered that black holes need not be strictly black, since quantum effects can make them radiate elementary particles.

The effect he discovered for black holes is a special case of a more general quantum phenomenon: the spontaneous production of particles by an external field. This kind of effect had been predicted theoretically decades earlier by Julian Schwinger, who predicted that a sufficiently strong electric field would create electrons and positrons out of the vacuum.

It is instructive to see how particle production like this works energetically. Because of the randomness of quantum mechanics the vacuum of empty space is better imagined as a frothing soup of particles and antiparticles that are forever trying to emerge as real particles. (In quantum mechanics, whatever is not forbidden is compulsory.) They normally cannot emerge, however, because their appearance is forbidden by conservation laws. For instance, electrons cannot emerge from the vacuum alone without violating conservation of electric charge, since each electron carries charge $q = -e$, where $e = 1.60 \times 10^{-19}$ Coulomb. But since positrons carry the opposite charge, charge conservation cannot forbid the joint emergence of an electron-positron pair. But it is energy conservation that keeps such pairs from emerging all the time from the vacuum around us, because such an emergence would require the production of sufficient energy to account for their masses, $E = 2mc^2$. Although there is a sense that the Uncertainty Principle allows quantum fluctuations to violate energy conservation, they can only do so very briefly and in the long term energy conservation is inviolate.

The situation changes in the presence of an electric field, $\boldsymbol{E}$, because the energy of a pair of oppositely charged particles is a function of their separation. Such particles can lower their energy by separating because their opposite charges make them feel forces in opposite directions due to the electric field. It is the work done by these forces that lowers their energy, and if their total energy (including their mass) can be lowered to zero in this way then energy conservation can no longer forbid their being produced spontaneously from the vacuum. The energy (including the rest mass) of an electron-positron pair (held at rest) a distance x apart in a constant electric field turns out to be

$$E = 2mc^2 - e|\boldsymbol{E}|x\,, \tag{5.23}$$

and so this can vanish (just like for the vacuum), once $x > 2mc^2/e|\boldsymbol{E}|$.

Using the quantum probability of having the electrons emerge a distance x apart from the vacuum, $p(x) \sim e^{-2mcx/\hbar}$, implies the probability for producing electron-

positron pairs by an electric field is given by

$$p \sim \exp\left(-\frac{4m^2c^3}{e|\boldsymbol{E}|\hbar}\right) . \tag{5.24}$$

Notice that the exponential dependence makes this probability extremely small unless $e|\boldsymbol{E}| \gtrsim 4m^2c^3/\hbar$, which is why electrons don't pop out of the vacuum all the time in the presence of the stray electric fields that arise in day-to-day life. The kinds of fields that are required can exist very near very heavy nuclei (having more protons than the heaviest naturally occuring nuclei), once all of their screening electrons have been stripped off.

Hawking Radiation

Hawking's observation was that a similar phenomenon can happen in the gravitational field produced by a black hole. As particles and antiparticles pop in and out of the fermenting froth of the vacuum near the Schwarzschild radius, $r = r_s$, one member of a pair can fall into the black hole and so be unable to recombine with its erstwhile partner. And the energy that is released by having this member fall into the hole can be sufficient to carry its surviving partner far enough away from the black hole that it can escape. The resulting prediction is that a black hole should emit a constant stream of elementary particles, now called the *Hawking radiation.*

To see why sufficient energy is liberated, consider a particle having 4-momentum $p^\mu = mv^\mu$, where m is the particle rest-mass and $v^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$ is its 4-velocity, moving along a radial trajectory, $r = r(t)$, in a Schwarzschild geometry. Since $v \cdot v = g_{\mu\nu}v^\mu v^\nu = -(1 - r_s/r)(\mathrm{d}t/\mathrm{d}\tau)^2 + (1 - r_s/r)^{-1}(\mathrm{d}r/\mathrm{d}\tau)^2 = -1$, we have

$$v^\mu = \begin{pmatrix} \gamma \\ \gamma\, v \\ 0 \\ 0 \end{pmatrix} , \tag{5.25}$$

where $v = \mathrm{d}r/\mathrm{d}t$ and

$$\gamma = \left[\left(1 - \frac{r_s}{r}\right) - \frac{v^2}{(1 - r_s/r)}\right]^{-1/2} . \tag{5.26}$$

Notice that the requirement that $\mathrm{d}t/\mathrm{d}\tau = \gamma$ be real requires $v^2 \leq (1 - r_s/r)^2$, which approaches zero as $r \to r_s$. This limit arises because v is defined using the asymptotic time t, and reflects the breakdown of this coordinate near $r = r_s$ due to the infinite redshift that exists between this coordinate and the proper time for freely-falling observers in this limit.

Recall that the quantity that is conserved along the trajectory of a particle as it falls in (or climbs out) of the black hole is

$$\mathcal{E} = -g_{tt}\,\frac{\mathrm{d}t}{\mathrm{d}\tau} = \left(1 - \frac{r_s}{r}\right)\frac{p^t}{m} = \gamma\left(1 - \frac{r_s}{r}\right)\,, \tag{5.27}$$

and so is the quantity of interest for deciding whether one of a particle-antiparticle pair can escape to infinity. This is an energy inasmuch as it agrees at $r \to \infty$ with the energy, $E = -u \cdot p = -g_{\mu\nu}u^\mu p^\nu = -g_{tt}\,u^t p^t$, of the particle as seen by a static observer hovering at fixed radius whose 4-velocity, u^μ, is $u^t = [1 - r_s/r]^{-1/2}$ and $u^i = 0$.

Since $\mathcal{E} \to 1$ as $r \to \infty$, the obstacle to having a particle escape to infinity is that $\mathcal{E}$ for the escaping particle must get to unity whereas the sum $\mathcal{E}_1 + \mathcal{E}_2$ for the particle-antiparticle pair starts at zero (same as in the absence of the pair) and is conserved as they move along their respective geodesics. In order to have $\mathcal{E}_1 = 1$ for the particle, say, its partner must be able to tunnel to a region for which $\mathcal{E}_2 = -1$. The remarkable thing is that eq. (5.27) shows that this is possible, provided $r < r_s$ because $\mathcal{E} < 0$ in this region. Furthermore, $\mathcal{E} = -1$ can be reached if r gets close enough to $r = 0$.

Particle production can therefore occur provided the particle-antiparticle pair can tunnel to a separation of order $r \simeq r_s$, since one particle must remain outside the event horizon (in order to escape) while the other must get deep enough inside to ensure that it reaches an area for which $\mathcal{E} \leq -1$. Using the quantum amplitude, $\psi \simeq e^{-mr}$, for the amplitude for a pair of mass m to separate by a distance r leads one to expect a particle production rate that is suppressed by a power of e^{-mr_s}.

It happens that a more precise calculation does give this result, and the distribution of particles that are released in this way closely resembles what would be expected for the radiation from a hot body, $\propto \exp(-m/T_H)$, with the temperature given by

$$k_B T_H = \frac{\hbar c}{4\pi r_s} = \frac{\hbar c^3}{8\pi G M}\,, \tag{5.28}$$

where $k_B = 1.38 \times 10^{-23}$ Joule/Kelvin is Boltzmann's constant, which tells how much energy is associated with a given temperature. T_H is called the black hole's *Hawking temperature*, $T_H \propto 1/r_s$. Numerically, for a solar-mass astrophysical black hole with $M = M_\odot$, this predicts the completely negligible temperature $T_H \simeq 10^{-8}$ Kelvin.

For thermal emission into radiation the surface brightness (energy loss rate per unit area), f, is completely characterized by the temperature, with $f = \alpha_B N_r T^4$, where N_r counts the number of species of particles in the radiation and $\alpha_B = \pi^2 k_B^4/(60\hbar^3 c^2) = 5.67 \times 10^{-8}$ Watts/(metre)^2(Kelvin)^4 is the Stefan-Boltzmann constant. The total rate of energy loss that is produced in this way far from the black

hole whose surface area is $4\pi r_s^2$ is then of order

$$\frac{\mathrm{d}E}{\mathrm{d}t} \simeq -4\pi\alpha_B N_r T_H^4 r_s^2 = -N_r \left(\frac{M_\odot}{M}\right)^2 9.00 \times 10^{-29} \quad \text{Watts}\,. \tag{5.29}$$

Although this is negligible for any astrophysical system, for a black hole with $M = 1$ gram, it is 10^{66} times bigger than for a solar mass, implying a whopping power release of 10^{38} Watts! Since the black hole energy is given by its mass, the above equation can be read as implying $\mathrm{d}M/\mathrm{d}t \propto -M^{-2}$, which can be integrated to infer how M varies with time. The result is a monotonically decreasing function that ultimately reaches zero, describing the black hole's evaporation.

Because the radiation rate grows as M falls, for relatively small black holes the energy loss due to Hawking radiation can be appreciable. And the more energy that is lost, the smaller becomes the mass of the black hole, making the Hawking temperature (and so also the radiation rate) larger. This is the recipe for a runaway evaporation, wherein the radiation becomes faster and faster, ultimately becoming explosive once the black hole mass gets down to the vicinity of the Planck mass, $M_p \simeq 2 \times 10^{-8}$ kg. The time taken for the evaporation of such a black hole turns out to be

$$\tau_{ev} = \frac{5120\pi G^2 M^3}{\hbar c^4} = \left(\frac{M}{M_\odot}\right)^3 6.62 \times 10^{74} \quad \text{seconds}\,. \tag{5.30}$$

This is much larger than the age of the universe (10^{10} years, or 3×10^{17} seconds) in the case of a solar-mass black hole, but is in the ballpark of 10^{-25} seconds for a one-gram black hole.

Hawking radiation is one of the few cases where a quantum effect can be reliably computed in a gravitating environment, and it carries many surprises. It tells us that very small black holes are unlikely to exist, since they are likely to evaporate very quickly and explosively. It also turns out that the similarity between black holes and thermal systems appears to be very deep, with 1/4 of the area of the black hole event horizon (in Planck units) playing the role of its entropy

$$S = \frac{\pi r_s^2}{\ell_p^2} = \frac{4\pi G M^2}{\hbar c}\,, \tag{5.31}$$

called the *Bekenstein-Hawking entropy*. The classical evolution of the black hole then combines precisely with the thermodynamic evolution of any surrounding hot particles to ensure the validity of the three laws of Thermodynamics (including the inevitability of the increase of total entropy), in a deep way that even now remains poorly understood.

5.6 Rotating Black Holes

The Schwarzschild solution described to this point describes the unique gravitational field outside of any spherically symmetric source (including a black hole). But because such a source carries no angular momentum, it cannot describe the gravitational field exterior to a rotating source, or the field external to a black hole formed by the collapse of initially rotating matter.

Rotating black holes are instead described by what is called the Kerr metric,[9] which is axially symmetric rather than spherically symmetric. The Kerr metric can be explicitly written using *Boyer-Lindquist* coordinates, $\{t, r, \theta, \phi\}$, where $0 < \theta < \pi$ and $0 < \phi < 2\pi$ are periodic angular variables (as for spherical polar coordinates), while both t and r can take arbitrarily large values. It is given by

$$\begin{aligned} \mathrm{d}s^2 &= -\left(1 - \frac{2GMr}{\rho^2}\right)\mathrm{d}t^2 - \frac{2GMar\sin^2\theta}{\rho^2}\Big(\mathrm{d}t\,\mathrm{d}\phi + \mathrm{d}\phi\,\mathrm{d}t\Big) \\ &\quad + \frac{\rho^2}{\Delta}\,\mathrm{d}r^2 + \rho^2\mathrm{d}\theta^2 + \frac{\sin^2\theta}{\rho^2}\Big[\big(r^2+a^2\big)^2 - a^2\Delta\,\sin^2\theta\Big]\mathrm{d}\phi^2 \qquad (5.32) \\ &= -\frac{\Delta}{\rho^2}\Big[\mathrm{d}t - a\sin^2\theta\,\mathrm{d}\phi\Big]^2 + \frac{\sin^2\theta}{\rho^2}\Big[(r^2+a^2)\mathrm{d}\phi - a\,\mathrm{d}t\Big]^2 + \frac{\rho^2}{\Delta}\,\mathrm{d}r^2 + \rho^2\mathrm{d}\theta^2\,, \end{aligned}$$

where a and GM are positive real parameters with dimensions of length while $\rho(r,\theta)$ and $\Delta(r)$ are functions, given explicitly by

$$\Delta := r^2 - 2GMr + a^2\,, \qquad (5.33)$$

and

$$\rho^2 := r^2 + a^2\cos^2\theta\,. \qquad (5.34)$$

As is straightforward (but tedious) to verify, the Ricci tensor constructed from this metric vanishes — $R_{\mu\nu} = 0$ — so it satisfies the vacuum Einstein equations. Other useful quantities are the metric's determinant, $g = \det g_{\mu\nu}$, whose square root is

$$\sqrt{-g} = \rho^2\sin\theta = (r^2 + a^2\cos^2\theta)\sin\theta\,, \qquad (5.35)$$

and the inverse metric, which has nonzero components

$$-g^{tt} = \frac{(r^2+a^2)^2 - a^2\Delta\sin^2\theta}{\Delta\rho^2}\,, \qquad g^{t\phi} = \frac{2GMar}{\Delta\rho^2} \qquad (5.36)$$

and

$$g^{rr} = \frac{\Delta}{\rho^2}\,, \qquad g^{\theta\theta} = \frac{1}{\rho^2} \quad \text{and} \quad g^{\phi\phi} = \frac{(r-2GM)r + a^2\cos^2\theta}{\Delta\sin^2\theta\rho^2}\,. \qquad (5.37)$$

[9] Both Schwarzschild and Kerr solutions to Einstein's equations are named after their discoverers.

For $r \gg a$ and $r \gg GM$ these functions become $\rho \simeq r$ and $\Delta \simeq r^2 - 2GMr$ and so the metric becomes

$$\begin{aligned} \mathrm{d}s^2 \simeq -\left(1 - \frac{2GM}{r}\right)\mathrm{d}t^2 - \frac{2GMa\sin^2\theta}{r}\Big(\mathrm{d}t\,\mathrm{d}\phi + \mathrm{d}\phi\,\mathrm{d}t\Big) \\ + \left(1 + \frac{2GM}{r}\right)\mathrm{d}r^2 + r^2\Big(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\Big)\,, \end{aligned} \tag{5.38}$$

up to terms that are subdominant by two powers of $1/r$. This asymptotes to Minkowski space in spherical polar coordinates as $r \to \infty$, showing that this geometry is asymptotically flat at large r.

Keeping terms of order $1/r$ shows $g_{00} \simeq -1 + 2GM/r$ and so the Newtonian potential seen by very distant observers is $\Phi = -GM/r$, as appropriate for an object of mass M (where G, as usual, denotes Newton's gravitational constant). This interpretation of M as the black hole mass is also supported by taking the $a \to 0$ limit for arbitrary r, in which case (5.32) becomes the Schwarzschild metric, with $r_s = 2GM$.

The dependence on θ implies the metric (5.32) has less symmetry than does the Schwarzschild metric, making it not spherically symmetric. It *is* symmetric under the independent constant shifts of the coordinates t and ϕ, however, showing that it is both time-translation invariant — *i.e.* 'stationary' — and invariant under rotations for which θ remains fixed. For the asymptotically flat geometry at large r, shifts of ϕ with fixed θ correspond to rotations about only the z-axis.

As usual there is a conserved angular momentum associated with this rotational invariance, but because the invariance is only about the z-axis, there is only a single conserved quantity, J, instead of a vector's-worth of quantities, $\mathbf{J}$. This conserved angular momentum works out to be related to a by

$$J = Ma\,, \tag{5.39}$$

so the $a \to 0$ limit corresponds to turning off the geometry's angular momentum (in which limit we saw above the geometry becomes Schwarzschild).

The presence of the $\mathrm{d}t\,\mathrm{d}\phi + \mathrm{d}\phi\,\mathrm{d}t$ term implies the Kerr geometry is (unlike the Schwarzschild geometry) *not* 'static' — *i.e.* not invariant under time-reversal, for which $t \to -t$ — even though the geometry *is* stationary.[10] The absence of time-reversal invariance is also what would be expected for nonzero J because time-reversal also changes the sign of J, and indeed the Kerr solution remains invariant under $t \to -t$ if at the same time we take $a \to -a$.

[10] Strictly speaking, a geometry is stationary when it has a time-like Killing vector field, ξ^μ — see the discussion around eq. (3.38) — and it is static if this vector field is 'hypersurface orthogonal', *i.e.* perpendicular to surfaces of constant t.

In the limit $M \to 0$ with a fixed, the metric (5.32) becomes

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \frac{r^2 + a^2\cos^2\theta}{r^2 + a^2}\,\mathrm{d}r^2 + \left(r^2 + a^2\cos^2\theta\right)\mathrm{d}\theta^2 + \left(r^2 + a^2\right)\sin^2\theta\,\mathrm{d}\phi^2\,. \quad (5.40)$$

This is again flat space but written in ellipsoidal coordinates, related to cartesian coordinates by

$$x = \sqrt{r^2 + a^2}\,\sin\theta\cos\phi\,, \quad y = \sqrt{r^2 + a^2}\,\sin\theta\sin\phi\,, \quad z = r\cos\theta\,. \quad (5.41)$$

Surfaces of constant r in these coordinates are ellipsoids that satisfy

$$\frac{x^2 + y^2}{r^2 + a^2} + \frac{z^2}{r^2} = 1\,. \quad (5.42)$$

As $r \to 0$ these ellipsoids degenerate down to a circular disk, $x^2 + y^2 \leq a^2$, at $z = 0$, whose centre corresponds to $\cos\theta = 1$ and whose boundary at $x^2 + y^2 = a^2$ corresponds to $\cos\theta = 0$.

Event horizons and Ergosphere

The Kerr geometry describes the spacetime surrounding a spinning black hole, and it is a black hole inasmuch as there is a region of the spacetime from which it is impossible to escape to spatial infinity. The boundary of this region defines an 'event horizon' through which the flow of test particles is purely a one-way trip. To explore the physically significant surfaces like this, consider various families of observers moving within this spacetime.

The first class of observers to consider are those who simply 'hover' at fixed r, θ and ϕ. These are the observers who remain at rest with stationary observers at infinity, relative to whose clocks the hovering observers experience time-dilation (or redshift). The 4-velocity, u^μ, of any such a hovering observer points purely in the t direction and must be time-like or null, so that $g_{\mu\nu}u^\mu u^\nu = g_{tt}(u^t)^2 \leq 0$. Increments of proper time, $\mathrm{d}\tau$, for such an observer are given by

$$\mathrm{d}\tau^2 = \left(1 - \frac{2GMr}{\rho^2}\right)\mathrm{d}t^2\,. \quad (5.43)$$

Such observers are only possible when $2GMr < \rho^2 = r^2 + a^2\cos^2\theta$ and so

$$r > GM + \sqrt{(GM)^2 - a^2\cos^2\theta}\,, \quad (5.44)$$

and for radii smaller than this all timelike observers must also move in the direction of the black hole's rotation. For the equator (for which $\theta = \frac{\pi}{2}$ and so $\cos\theta = 0$) this amounts to $r > 2GM$ — just like the corresponding condition for Schwarzschild. It occurs for smaller radii than this at higher latitudes, with hovering observers

allowed for $r > r_+ := GM + \sqrt{(GM)^2 - a^2}$ at the poles (for which $\cos\theta = \pm 1$). Eq. (5.44) defines the exterior of the 'ergosphere', defined as the region within which it is impossible to simply hover at fixed r, θ and ϕ.

Consider next a photon that moves in the equatorial plane ($\cos\theta = 0$) initially with no radial velocity. Such a photon instantaneously has a 4-momentum pointing purely in the ϕ and t directions, and so satisfies $g_{tt}\mathrm{d}t^2 + g_{t\phi}(\mathrm{d}t\,\mathrm{d}\phi + \mathrm{d}\phi\,\mathrm{d}t) + g_{\phi\phi}\mathrm{d}\phi^2 = 0$, and so

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = -\frac{g_{t\phi}}{g_{\phi\phi}} \pm \sqrt{\left(\frac{g_{t\phi}}{g_{\phi\phi}}\right)^2 - \frac{g_{tt}}{g_{\phi\phi}}}\,. \tag{5.45}$$

Evaluating this right at the boundary of the ergosphere (which for $\theta = \frac{\pi}{2}$ corresponds to $r = 2GM$) implies $g_{tt} = 0$ and so

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = 0 \quad \text{or} \quad \frac{\mathrm{d}\phi}{\mathrm{d}t} = -2\left(\frac{g_{t\phi}}{g_{\phi\phi}}\right) = \frac{a}{2(GM)^2 + a^2}\,. \tag{5.46}$$

These show that a photon moving in a retrograde sense relative to the black hole rotation has zero transverse speed when at the edge of the ergosphere. A massive particle not moving radially at this radius moves more slowly than a photon and so must be carried along by the rotation within the ergosphere. By contrast, motion in the same sense as the black hole rotation has nonzero speed, suggesting that the edge of the ergosphere is unlikely also to define the event horizon in the equatorial plane.[11]

This disagreement between the position of the event horizon and the boundary of the ergosphere arises because Kerr is stationary but not static. To identify the position of the event horizon consider the trajectory $r(t)$ of a radially out-going light ray. This satisfies $\mathrm{d}s^2 = 0$ and so $r(t)$ must satisfy

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \sqrt{\frac{\Delta}{\rho^2}\left(1 - \frac{2GMr}{\rho^2}\right)}\,. \tag{5.47}$$

The radial position, r, no longer increases with increasing t once the right-hand side of this equation vanishes. This either occurs when $2GMr = \rho^2 = r^2 + a^2\cos^2\theta$ or when $\Delta = r^2 - 2GMr + a^2 = 0$.

The problem at $2GMr \leq \rho^2$ proves to be more about the breakdown of the ability to use the coordinate t to parameterize time along a timelike curve inside the ergosphere. Instead it is radii for which $\Delta(r) = 0$ that turn out to correspond to event horizons for the Kerr metric, corresponding to where $g^{rr} = 1/g_{rr}$ vanishes. This implies the event horizons occur as surfaces of constant r, at the specific values

$$r = r_\pm := GM \pm \sqrt{(GM)^2 - a^2}\,. \tag{5.48}$$

[11] Recall for a Schwarzschild black hole a photon cannot have tangential components to momentum right at the horizon, $r = 2GM$.

External observers only access information from outside the outermost of the two event horizons — *i.e.* the one at $r = r_+$. Precisely as for the Schwarzschild geometry, the apparent singularity of the metric at $r_\pm$ is only an artifact of the breakdown there of the coordinates $\{t, r, \theta, \phi\}$.

Notice that the external horizon becomes the Schwarzschild horizon $r_+ \to 2GM$ as $a \to 0$, and also corresponds to the boundary of the ergosphere (for all θ) in this limit. The ergosphere touches the outer horizon only at the poles, but elsewhere (for all $\cos^2\theta < 1$) is strictly exterior to the outer horizon.

Both the boundary of the ergosphere and the event horizons are only real for all θ if $a \leq GM$, in which case the black-hole angular momentum satisfies the upper bound

$$J = Ma \leq GM^2 \,. \tag{5.49}$$

This is believed to be a physical condition for black holes because geometries with $a > GM$ turn out to have regions of infinite curvature that are not masked by event horizons (what are called 'naked singularities'), that are unstable and are believed to be unphysical.

6. Other astrophysical applications

The universe is a violent place, containing many examples of matter situated in very extreme environments. Many of the most violent of these involve black holes located in galactic centres whose masses are many millions of times the mass of our Sun. These release enormous amounts of energy as material falls into the black hole, in amounts that can only be understood within a relativistic framework.

Furthermore more sophisticated surveying techniques are now mapping out larger and larger regions of the universe, allowing a more detailed understanding of how much matter is out there, where it is, and how it interacts with its surroundings. Since most of this material turns out to be dark, there is a high premium for understanding how it gravitates, since this provides the only observational handle on knowing where it is.

Many of these studies rely heavily on General Relativity, and some are accurate enough to provide precision tests of the theory that are similar in spirit to those performed in the solar system. This section summarizes a few of these.

6.1 Stellar interiors

For an astrophysical object like a star the properties of the event horizon are irrelevant, because the Schwarzschild geometry only applies down to the star's radius, $R_\star$, below which we must re-solve Einstein's equations in the presence of matter,

$T_{\mu\nu} \neq 0$. To illustrate how this works, this section finds this interior geometry using a simple model of the physics of the star. The absence of stable orbits in the Schwarzschild solution too close to $r = r_s$ should make one expect that stars should not be able to stave off gravitational collapse if they become too dense, $R_\star \sim r_s$, and this expectation is borne out in detail in the analysis below.

If the star is spherically symmetric then the arguments made earlier show that it is always possible to choose coordinates so that the metric has the form

$$\mathrm{d}s^2 = -e^{2a}\mathrm{d}t^2 + e^{2b}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2)\,, \tag{6.1}$$

where we may take $a = a(r)$ and $b = b(r)$ if the star's interior is time-independent. The goal is to solve for these functions using the field equations,

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}\,R\,g_{\mu\nu} = 8\pi G T_{\mu\nu}\,, \tag{6.2}$$

given a simple choice for $T_{\mu\nu}$. To this end we require the components of the Einstein tensor, $G_{\mu\nu}$, which can be found using eqs. (4.12):

$$\begin{aligned} G_{tt} &= \frac{e^{2(a-b)}}{r^2}\left(2r\partial_r b - 1 + e^{2b}\right)\,, \quad G_{rr} = \frac{1}{r^2}\left(2r\partial_r a + 1 - e^{2b}\right)\,, \\ G_{\theta\theta} &= r^2 e^{-2b}\left[\partial_r^2 a + (\partial_r a)^2 - \partial_r a\,\partial_r b + \frac{1}{r}(\partial_r a - \partial_r b)\right] \end{aligned} \tag{6.3}$$

and $G_{\phi\phi} = G_{\theta\theta}\sin^2\theta$.

For the stress energy, we take the stellar interior to be a perfect fluid that is characterized by an energy density, ρ, and pressure, p, which are related by some sort of equation of state, $p = p(\rho, S)$, where S is the fluid's entropy. Any such a fluid must have a local rest frame, whose 4-velocity is denoted by $u^\mu(x)$, where as usual $g_{\mu\nu}u^\mu u^\nu = -1$.

To determine the stress tensor for such a fluid, we appeal to the principle of equivalence. First consider the limit of flat space for which we would like $T_{tt} = \rho$ and $T_{ij} = p\,\delta_{ij}$ in the fluid's rest frame (for which $u^\mu = (1,0,0,0)$). This implies $T_{\mu\nu}$, written in terms of u^μ, ρ and p, must be defined by

$$T_{\mu\nu} = (\rho + p)u_\mu u_\nu + p\,g_{\mu\nu}\,, \tag{6.4}$$

where $g_{\mu\nu} = \eta_{\mu\nu}$ in flat space. The principle of equivalence says that this same expression should also hold in the presence of a gravitational field, since it is a generally covariant expression that agrees with the flat-space result of special relativity in the special frame for which $g_{\mu\nu} = \eta_{\mu\nu}$.

Evaluating eq. (6.4) using the metric, eq. (6.1) leads to the following components for $T_{\mu\nu}$:

$$T_{tt} = e^{2a}\rho\,, \quad T_{rr} = e^{2b}p\,, \quad T_{\theta\theta} = r^2 p\,, \tag{6.5}$$

and $T_{\phi\phi} = T_{\theta\theta}\sin^2\theta$. The expression of energy conservation for this metric, $\nabla^\mu T_{\mu\nu} = 0$, then implies

$$(\rho + p)\frac{\mathrm{d}a}{\mathrm{d}r} = -\frac{\mathrm{d}p}{\mathrm{d}r}\,. \tag{6.6}$$

Using eqs. (6.5) in the Einstein equations leads to three independent expressions:

$$\begin{aligned}
\frac{e^{-2b}}{r^2}\left(2r\partial_r b - 1 + e^{2b}\right) &= 8\pi G\rho \quad (tt \text{ equation}) \\
\frac{e^{-2b}}{r^2}\left(2r\partial_r a + 1 - e^{2b}\right) &= 8\pi G p \quad (rr \text{ equation}) \\
r^2 e^{-2b}\left[\partial_r^2 a + (\partial_r a)^2 - \partial_r a\,\partial_r b + \frac{1}{r}(\partial_r a - \partial_r b)\right] &= 8\pi G p \quad (\theta\theta \text{ equation})\,.
\end{aligned} \tag{6.7}$$

Since the (tt) equation does not involve $a(r)$, it can be put into a more physically intuitive form by performing a change of variables from $b(r)$ to

$$m(r) = \frac{r}{2G}\Big(1 - e^{-2b}\Big)\,, \tag{6.8}$$

for which $e^{2b} = [1 - 2Gm(r)/r]^{-1}$. In terms of this variable the (tt) equation becomes

$$\frac{\mathrm{d}m}{\mathrm{d}r} = 4\pi r^2\,\rho\,, \tag{6.9}$$

which integrates to give

$$m(r) = 4\pi\int_0^r \mathrm{d}\hat{r}\;\hat{r}^2\rho(\hat{r})\,. \tag{6.10}$$

If the boundary of the star is taken to be $r = R_\star$, then for $r > R_\star$ the geometry is given by the Schwarzschild metric. Continuity of the metric across $r = R_\star$ then requires the function $m(r)$ must satisfy the boundary condition $m(R_\star) = M$, where M is the mass of the star. That is,

$$M = 4\pi\int_0^{R_\star} \mathrm{d}\hat{r}\;\hat{r}^2\rho(\hat{r})\,. \tag{6.11}$$

This last equation almost (but not quite) says that $m(r)$ is the integral of the energy density out to radius r, and so that M is the integral of this energy density over the entire volume of the star. The qualification 'almost' is required here because the integral of the energy density would really have been weighted by the covariant measure of volume which involves the determinant of the entire spatial metric, $\sqrt{\det g_{ij}} = e^b r^2\sin\theta\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi$, and so the integrated energy is really given by

$$M_{\text{tot}} = 4\pi\int_0^{R_\star} \mathrm{d}\hat{r}\;e^{b(\hat{r})}\hat{r}^2\rho(\hat{r}) = 4\pi\int_0^{R_\star} \mathrm{d}\hat{r}\;\frac{\hat{r}^2\rho(\hat{r})}{[1 - 2Gm(\hat{r})/\hat{r}]^{1/2}} > M\,. \tag{6.12}$$

This shows that $M_{\rm tot}$ is better thought of as the energy the star would have if it were distributed to infinity and so had no gravitational field, making the difference $M_{\rm tot} - M$ the star's gravitational binding energy.

Trading $b(r)$ for $m(r)$ in the (rr) equation then gives the following result for $a(r)$:

$$\frac{\mathrm{d}a}{\mathrm{d}r} = \frac{Gm(r) + 4\pi G r^3 p}{r[r - 2\,Gm(r)]}\,. \tag{6.13}$$

Rather than trying to simplify this using the $(\theta\theta)$ equation, it is simpler instead to use conservation of energy, eq. (7.70), to trade $\mathrm{d}a/\mathrm{d}r$ for $\mathrm{d}p/\mathrm{d}r$, to get

$$\frac{\mathrm{d}p}{\mathrm{d}r} = -\frac{(\rho + p)[Gm(r) + 4\pi G r^3 p]}{r[r - 2\,Gm(r)]}\,. \tag{6.14}$$

This equation, called the *Tolman-Oppenheimer-Volkoff equation*, expresses how the pressure profile in the star's interior must adjust in order to balance the gravitational force required to support the star's outer layers, and provides the condition of hydrostatic equilibrium for the interior of the star. In particular, so long as p and ρ are both positive and $r > 2\,Gm(r)$, eq. (6.14) implies $\mathrm{d}p/\mathrm{d}r < 0$ and so the pressure profile decreases monotonically with radius within the star, taking its maximum value at the star's centre at $r = 0$.

Notice that in the Newtonian limit we may take $p \ll \rho$, since the energy density is dominated by the rest mass of the atoms in the star, as well as $r \gg 2\,Gm(r)$, allowing eq. (6.14) to be approximated by the more familiar form

$$\frac{\mathrm{d}p}{\mathrm{d}r} = -\frac{Gm(r)\rho}{r^2}\,. \tag{6.15}$$

This equation simply states that the pressure gradient adjusts to ensure that the net force acting on any particular fluid element vanishes. To see this, consider a small fluid element that extends from r to $r + \mathrm{d}r$ with cross-sectional area A. Since pressure is force per unit area, the radial component of the net fluid force acting on this element is

$$\mathrm{d}F_p = p(r)A - p(r + \mathrm{d}r)A \simeq -\frac{\mathrm{d}p}{\mathrm{d}r}\,A\,\mathrm{d}r\,. \tag{6.16}$$

Eq. (6.15) simply states that this force must balance the gravitational attraction between the matter in the fluid element (whose mass is $\rho A\,\mathrm{d}r$) and the matter that lies interior to it in the star (whose mass is $m(r)$) and so whose radial component is

$$\mathrm{d}F_g = -\frac{Gm(r)\rho A}{r^2}\,\mathrm{d}r\,. \tag{6.17}$$

Implications for stellar phenomenology

In general, hydrostatic equilibrium relates $\mathrm{d}p/\mathrm{d}r$ to ρ and m (which is itself related to ρ), and this can be integrated to obtain explicit profiles, $p(r)$ and $\rho(r)$, once an

equation of state is given, like $p = p(\rho, S)$ where S is the entropy density of the fluid. For example, for a perfect fluid one might use $p = \kappa\rho T$ where κ is a constant related to the mass per particle of the atoms making up the fluid and T is the local fluid temperature (and so is related to its entropy).

Once such an equation of state is known p can be eliminated (in principle) in terms of ρ, as can m using eq. (6.10), allowing eq. (6.14) to be regarded as an equation involving ρ only. This can be integrated, typically numerically, to give the profile $\rho(r)$ from which the equation of state then gives $p(r)$, while $m(r)$ and $a(r)$ are obtained from eqs. (6.10) and (6.13). Of course this process can become complicated in detail if the changes in pressure and density trigger phase changes in the stellar material, or in the dominant mechanism for energy transfer within the star, but the logic still remains the same in such cases provided one is careful to use the proper new expression relating p and ρ in the relevant areas.

The upshot is that an assumed equation of state leads to a prediction for all three of these profiles that depends on a single integration constant, usually taken to be the value of the energy density at the stellar centre: $\rho_\star = \rho(r = 0)$. What is important is that this means that the two external properties of a star — its mass M and radius $R_\star$ — must be related to one another because both of these can be calculated once $\rho_\star$ is known. The stellar radius is calculable because it may be defined as the radius $r = R_\star$ where $p(R_\star) = 0$. The mass is then found by using eq. (6.10) at $r = R_\star$. Because these two variables are both predicted from the one integration constant one expects to find a relation $M = M(R_\star)$ that relates all stars that share the same equation of state.

The importance of this observation is that both M and $R_\star$ can often be determined by observations. For instance, the mass can often be found by observing how other objects orbit around the given star. Although such orbits exist for a surprisingly large number of stars, since just under half of stars are found in *binary systems* with pairs of (or more than two) stars orbiting one another, in practice the two stars are usually required to eclipse one another (from the Earth's point of view) in order to obtain the stellar mass. This is because it is Kepler's third law that gives the mass in terms of the orbital period and semi-major axis, but the semi-major axis can only be determined if the orientation of the stellar orbit relative to the line of sight is known.

The radius, on the other hand, is more easily observable because it typically controls the star's overall *luminosity*, L, defined as its rate of energy emission. This depends on $R_\star$ because stars emit energy thermally and so do so with a flux — *i.e.* rate per unit surface area — that is characterized purely by their surface temperature: $f = f(T) = \sigma T^4$, where σ is a known constant. Since this temperature can be

measured from the spectrum of radiation the star emits, as can the total luminosity, $L = 4\pi f R_\star^2$, from the total observed brightness of the star (once its distance from the Earth is known), the radius $R_\star$ can be inferred from observations.

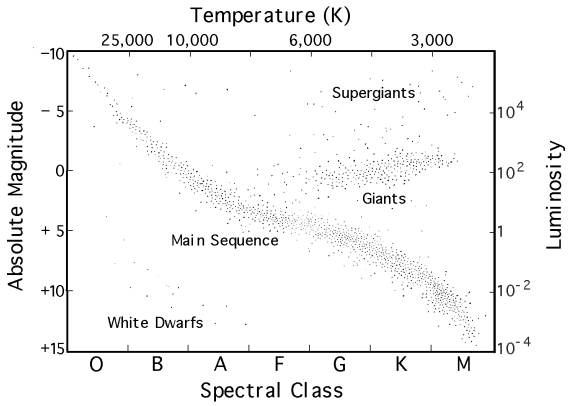


Figure 9: A Hertzsprung-Russell (HR) diagram showing the correlation between stellar luminosity and temperature.

In the event, 90% of stars are dominantly made up of hydrogen, and provide the pressure gradients required to stave off gravitational collapse by fusing hydrogen into helium in their cores. Consequently they share the same equation of state, and so their pressure, density and temperature profiles are all calculable in terms of their central density, $\rho_\star$. They should be expected to fall along a single curve $M(R_\star)$ if their masses and radii are plotted in the $M - R_\star$ plane. Astronomers really test this prediction by instead plotting their luminosity against their temperature, and looking for a correlation between L and T since these are the two quantities that are the most easily observable. And they indeed find that most stars — known as *main sequence* stars — do fall along a curve when plotted in the $L - T$ plane (known as a *Hertzsprung-Russell* diagram).

When mass can be measured it is also observed to be correlated with luminosity when main sequence stars are plotted in the $M - L$ plane. Because the energy source in stars ultimately comes from nuclear reactions, small increases in mass lead to fairly small increases in the central temperature, but this leads to a large change in luminosity. Observationally one finds the strong variation $L \propto M^{3.5}$, with more massive stars being much more luminous.

For ordinary stars the balance between pressure and gravity is perilously achieved, because it relies on the pressures associated with the energy release due to nuclear fusion which becomes possible at the high pressures found in stellar cores. This is perilous because it can only work so long as there is nuclear fuel to burn in this way, and so ends once this fuel is depleted. Furthermore, since the main sequence lifetime is of order $\tau \propto M/L$ the observed mass-luminosity correlation shows that $\tau \propto M^{-2.5}$, and so more massive stars have a much shorter lifetime than do lighter ones.

At some point either a new, more stable, source of pressure must be found to balance gravity if a permanent object is to be formed, or gravity wins – leading to a runaway gravitational collapse.

An incompressible star

To see in more detail what the options are for balancing gravity with various forms of pressure it is instructive to specialize to the very simple case of an incompressible fluid, $\rho = \rho_\star$ for all p. This represents the extreme case where the stellar material resists changing its density regardless of how high the pressures get. It also has the advantage of allowing explicit solutions which illustrate the behaviour in the more general case.

Suppose, then, that we assume the incompressible density profile

$$\rho(r) = \begin{cases} \rho_\star \text{ if } r < R_\star \\ 0 \text{ if } r > R_\star \end{cases}, \tag{6.18}$$

which is characterized by the two parameters $\rho_\star$ and $R_\star$. In this case we may directly integrate to obtain $m(r)$, leading to

$$m(r) = \begin{cases} 4\pi\rho_\star r^3/3 & \text{if} \quad r < R_\star \\ 4\pi\rho_\star R_\star^3/3 = M & \text{if} \quad r > R_\star \end{cases}, \tag{6.19}$$

which last relation allows one to trade $R_\star$ and M as independent parameters. Similarly, the pressure profile found by integrating (6.14) becomes

$$p(r) = \rho_\star \left[\frac{R_\star\sqrt{R_\star - r_s} - \sqrt{R_\star^3 - r_s r^2}}{\sqrt{R_\star^3 - r_s r^2} - 3R_\star\sqrt{R_\star - r_s}} \right] \quad \text{if } r < R_\star\,, \tag{6.20}$$

where, as usual, $r_s = 2GM$. Notice that the pressure goes to zero at $r = R_\star$: $p(R_\star) = 0$, as expected by hydrostatic equilibrium for the stellar surface.

Similarly, integrating eq. (6.13) gives the metric component, $g_{tt} = -e^{2a}$:

$$e^{a(r)} = \frac{3}{2}\left(1 - \frac{r_s}{R_\star}\right)^{1/2} - \frac{1}{2}\left(1 - \frac{r_s r^2}{R_\star^3}\right)^{1/2} \quad \text{if } r < R_\star\,. \tag{6.21}$$

Notice that this implies $e^{2a(R_\star)} = 1 - r_s/R_\star$, as required by continuity with the exterior Schwarzschild solution.

The pressure equation, eq. (6.20), says something really interesting. Recall that it implies the pressure goes to zero at the stellar surface, $p(R_\star) = 0$, and then grows monotonically as one moves into the interior (*i.e.* for decreasing r), as is required by hydrostatic equilibrium. The maximum pressure reached is at the stellar center, and is given by

$$p_{\max} = p(0) = \rho_\star \left[\frac{\sqrt{R_\star - r_s} - \sqrt{R_\star}}{\sqrt{R_\star} - 3\sqrt{R_\star - r_s}} \right]. \tag{6.22}$$

Notice in particular that if we increase M (and so also r_s) for fixed $R_\star$, then $p(0) \to \infty$ once $r_s = \frac{8}{9}R_\star$, or $M_{\max} = \frac{4}{9}(R_\star/G)$. This states that once the star becomes too

dense it is completely impossible to support it against gravitational collapse. A similar conclusion is reached using more realistic equations of state, but for these it is also true that $M_{\max} \leq \frac{4}{9}(R_\star/G)$, a result known as Buchdahl's theorem. This is as one might have expected: an incompressible fluid supports the maximum mass possible.

If M should be larger than $M_{\max}$ for any given equation of state then there is no static solution possible, and the star collapses. It continues to collapse, either until the equation of state modifies so that M becomes smaller than the new $M_{\max}$, or until the entire star[12] falls below $r = r_s$, forming a black hole.

For real astrophysical objects several kinds of equations of state are known to produce objects that are stable against gravitational collapse, each typically working for a specific range of masses. These include planets (for which gravity is balanced by material stresses); white dwarf stars (for which gravity is balanced by electron degeneracy pressure); and neutron stars (for which gravity is balanced by neutron degeneracy pressure). These can remain stable indefinitely, unless additional matter is added to them in such a way as to push them over the limit of stability. (For instance, some supernovae arise when white dwarfs are pushed over their limits in this way as they accrete matter within binary star systems.)

6.2 Gravitational lensing

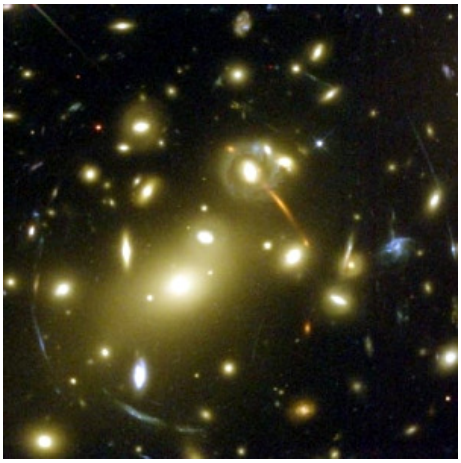


Figure 10: A photograph of gravitational lensing (the arc-like shapes) of distant galaxies by a foreground galaxy cluster.

Since we can now see objects that are very distant in the Universe, we should expect to find a reasonably large number of coincidences with distant galaxies appearing to lie very close to the same line of sight as nearer galaxies in the foreground. Because of this we expect the widespread occurrence of gravitational lensing, wherein light from very distant galaxies is deflected by the gravitational field of a foreground mass. This kind of lensing has in fact been seen many times, such as the *strong lensing* that is shown in fig. 10, where the arcs are lensed images of a distant galaxy distorted by a large cluster of galaxies in the foreground. But other examples of lensing have also been seen, including the *micro-lensing* of stars in our galaxy (and in nearby galaxies) by other stars that pass along the intervening line of sight, and the *weak lensing* that slightly distorts the shape of a great many galaxies across the sky.

[12] In real stellar collapse changes in the equation of state cause much of the initial star to be explosivly ejected, with only a fraction of the initial mass collapsing to form a black hole.

This section describes the basics of such lensing events. One should keep in mind that lensing phenomena are typically not used to test GR, because comparatively little is known about the properties of the foreground masses that are doing the lensing. Because of this it is difficult to have precise predictions with which to compare the observations. What is done instead is to use the observed lensing to infer the distribution of foreground matter, under the assumption that GR provides a good description of the lensing physics. It is arguments such as these that point to the widespread existence throughout the Universe of an unknown form of matter — called *Dark Matter* — whose presence is only inferred from its gravitational effects.

Lensing Basics

The starting point for the story is the basic observation, derived in section 3.3, that General Relativity predicts that light rays passing close to a spherical gravitational source are deflected through an angle

$$\alpha \simeq \frac{4GM}{b} = \frac{2\,r_s}{b}\,, \tag{6.23}$$

where M is the source's mass and b is the impact parameter of the passing light ray.

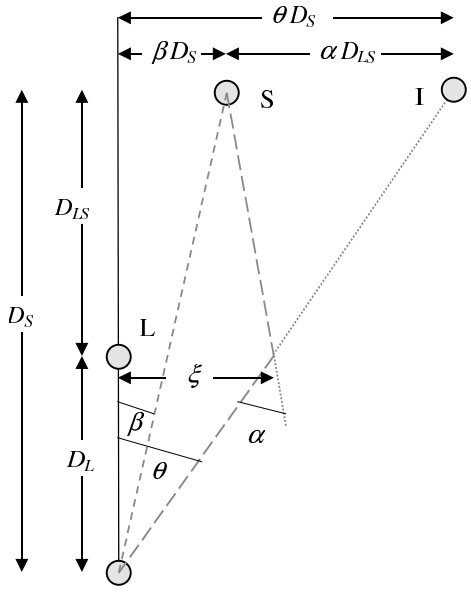


Figure 11: A diagram of the geometry of a lensing event.

The goal is to re-express this angle in terms of those that are more suited to what is actually measured when a lensing event is seen. In particular, rather than knowing the deflection angle, α, it is more useful to know the angular position, θ, of the image, I, relative to the angular position, β, of the source, S, as seen from by the observer, O. The figure, fig. 11, shows how these are related. A great help when using this figure to solve for θ is the early recognition that for most events the distances involved are enormous and the deflection angles are consequently very small. Because of this we can idealize the change of direction of the light ray as being completely localized at a single instant when it passes by the plane of the lens' position in the sky, and we can liberally use the approximation $\sin x \simeq x$ for $x \ll 1$.

Inspection of the top of the figure shows that the angles α, β and θ are related by

$$\theta\, D_S = \beta\, D_S + \alpha\, D_{LS}\,, \tag{6.24}$$

and so dividing through by D_S and eliminating α using eq. (6.23), with $b \simeq \xi \simeq \theta\, D_L$ then gives

$$\theta = \beta + \frac{\theta_E^2}{\theta}\,, \tag{6.25}$$

where the *Einstein angle*, θ_E, is defined in terms of the distances in the problem by

$$\theta_E = \sqrt{\frac{2\, r_s D_{LS}}{D_S D_L}}\,. \tag{6.26}$$

Solving eq. (6.25) for θ gives the desired solutions, $\theta = \theta_\pm$, for the angular positions of the two perceived images (one on each side of the lens in the plane defined by the observer source and lens), with

$$\theta_\pm = \frac{1}{2}\Big(\beta \pm \sqrt{\beta^2 + 4\theta_E^2}\Big)\,. \tag{6.27}$$

In the degenerate situation that the lens lies directly in front of the source — *i.e.* if $\beta = 0$, and so the observer, lens and source do not define a plane — then the observed image would be an *Einstein ring* that surrounds the lens, whose angular radius is $\theta = \theta_E$.

To get an idea of how big this ring is, suppose the source and lens are as distant from each other as the lens is from us, $D_{LS} \simeq D_L := D$ and so $D_S \simeq 2D$. Then if[13] $D \simeq 1$ Mpc and the lens has a mass $M \simeq 10^8\, M_\odot$, its Schwarzschild radius would be $r_s \simeq 3 \times 10^{11}$ m and so $\theta_E \simeq \sqrt{r_s/2D} \simeq 2 \times 10^{-6}$ radians, or 0.5 seconds of arc.

When the source and lens are instead only slightly off-set this ring degenerates into two arcs, much like those seen in fig. 10, and these are relatively easy to recognize. There are also several ways to check that two candidate objects in the sky are really multiple, lensed images of the same source. One is to compare their spectra, which should be identical for two images of the same source because (unlike for lenses in the lab) the bending of light by gravity is the same for all wavelengths. The other is to watch for correlations in any time-dependence in the intensity of the received light, since any fluctuations in the intensity of one must be repeated for the other — possibly after a delay due to any difference in the path length along the two light trajectories. Time-lags of this sort are observed for pairs of gravitationally lensed images, with changes in one image followed by changes in the other, often several weeks later.

But lensing events are not always so easy to identify, since the lenses are often too dark to see and the images needn't be so strongly distorted if the lens and source are not aligned sufficiently closely. Alternatively, for some objects the angle θ_E can be too small to be resolved. It turns out that there are nonetheless sometimes useful ways for searching for lensing that do not rely on directly detecting the independent images of a particular source.

[13] Mpc denotes a *megaparsec*, or 10^6 parsecs, which is a commonly used distance unit in extra-galactic astronomy. A parsec is an astronomical measure of distance, defined to be one AU per arc-second, where an astronomical unit (AU) is the mean Earth-Sun distance. This makes a parsec about 3.262 light years, or 3.086×10^{16} m, which is roughly the distance to the nearest stars. So 1 Mpc $\simeq 3 \times 10^{22}$ m.

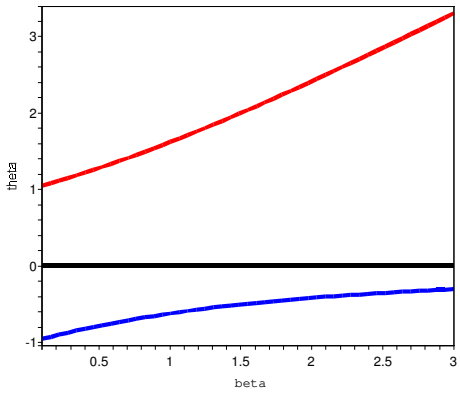


Figure 12: A plot of θ_+ (upper red) and θ_- (lower blue) vs β, in units of θ_E.

Weak Lensing

When looking at a field of view filled with distant galaxies, evidence, even for relatively weak lensing, can be found using statistical methods even if it is hopeless to find multiple images of individual galaxies. This evidence relies on statistically identifying the distortion that lensing produces on a galaxy's shape.

To quantify this distortion imagine describing the sky using angular coordinates that are centered on the position of the foreground object that is responsible for the lensing. In this case, as before, we use θ to describe the 'radial' angle of an image away from the lens, and φ to measure the 'azimuthal' angle of the image transverse to the radial direction θ. Lensing only moves the image of the source away or towards the lens (in the θ direction), with one image inside of and one outside of $\theta = \theta_E$, but does not also change φ.

In terms of these coordinates, suppose a narrow beam of light rays has angular widths $\Delta\theta$ and $\Delta\varphi$ when it leaves the source. Since the source is displaced relative to the lens by the angle β, the spread in $\Delta\theta$ can be interpreted as a spread $\Delta\beta$ in the initial angular position of the beam relative to the lens. Once the beam has been lensed its new angular position relative to the lens is $\theta_\pm$, and although the spread in the beam in the φ direction remains unchanged, in the θ direction the spread becomes

$$\Delta\theta_\pm = \left(\frac{\mathrm{d}\theta_\pm}{\mathrm{d}\beta}\right)\Delta\beta = \frac{1}{2}\left[1 \pm \frac{\beta}{\sqrt{\beta^2 + 4\theta_E^2}}\right]\Delta\beta\,. \tag{6.28}$$

Because of this distortion the images of a galaxy that would have appeared to us as being spherical without the lens, become elliptical in a precisely calculable way. Observationally, the problem is that galaxies are not perfectly spherical, and so the trick is to distinguish the distortions due to lensing from general oddities in galactic shapes. This is where statistics come in, because galaxies are usually randomly oriented in the sky and can come in a fairly random pattern of shapes. But the distortions due to lensing in the part of the sky near a source are preferentially distorted along the direction towards the lens. If one samples a large sample of galaxies in a particular part of the sky and finds a bias for galaxies to be distorted (on average) in a particular direction, this can be interpreted as evidence for lensing by a source that lies in this direction. By repeating this process over and over again

for nearby regions it is possible to provide a map of the foreground mass distribution that is doing the lensing, regardless of whether this distribution is directly visible or not.

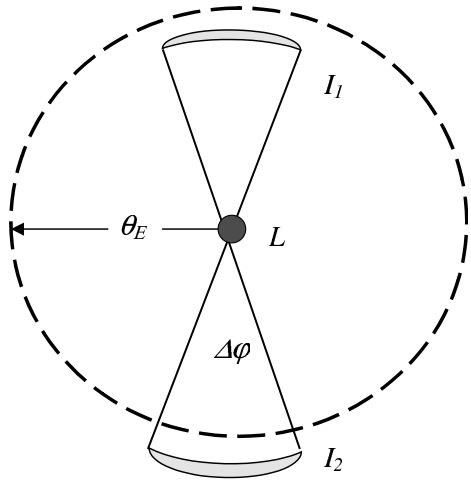


Figure 13: A sketch of the angular distortion of the two lensed images.

Maps of the mass distribution in the universe produced by weak lensing surveys of this type are just now (2008) being performed, and are providing one of the main lines of evidence for the existence of vast amounts of *Dark Matter* throughout the universe (more about which later).

Microlensing

Lensing can also be applied to objects in our own galaxy, and for nearby galaxies (like the *Magellanic Clouds* – which are two small galaxies that orbit our own) since stars, planets and other objects can periodically pass into the line of sight towards other stars. Seeking these kinds of lensing events could allow us to count the number of relatively small and dark objects that may be floating about the galaxy, otherwise unseen.

A major complication in this case is the very small size of the angular deflection, however, since a solar-mass lens situated 10 kpc away (a typical galactic distance) lensing a source that is 10 kpc $\simeq 3 \times 10^{20}$ m) $\theta_E \simeq 2 \times 10^{-9}$ radians, or 5×10^{-4} arc seconds. Angular distances this small are too small to measure from Earth, even if two stars could be found lying this close to the same line of sight.

But even if the two separate images of the source star are not separately visible, taken together they increase the total amount of light received at the earth from the source, compared with what would have arrived in the absence of lensing. Although we do not know how bright the initial source star intrinsically is, we know that within our galaxy stars are moving, with an average speed of roughly 200 km/sec. So although the absolute brightness of the unresolved images cannot be compared with a known initial source brightness, the *change* in brightness of the images as the lens and source move into and out of alignment can be measured.

What does this change of brightness look like? Since a star emits radiation thermally, its surface brightness depends only on its temperature and so its apparent brightness as seen by any given observer is controlled purely by the total fraction of the star's radiation that the observer is able to catch. And this fraction is controlled by the solid angle that the source subtends as seen by the observer. (This is why

the apparent brightness of a star usually falls off with distance, d, from the star like $1/d^2$.) The increase in brightness due to the lensing may therefore be computed by calculating the increased solid angle subtended due to the splitting and distorting of the lensed images.

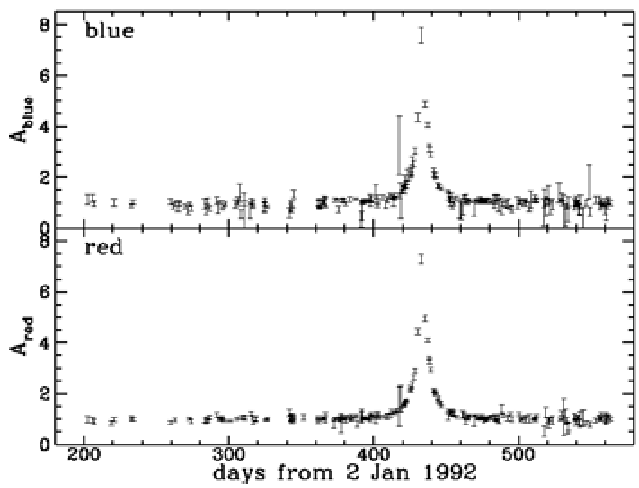


Figure 14: Observational traces of brightness vs day of observation for candidate microlensing events (from the MACHO Collaboration).

Consider then a beam of light coming from the source at $\theta = \beta$ having a small angular width $\Delta\theta = \Delta\beta$ and $\Delta\varphi$. The solid angle, observed from Earth, spanned by this beam at the source is then $\Delta\Omega = \sin\theta\Delta\theta\Delta\varphi \simeq \beta\Delta\beta\Delta\varphi$. Once it has been lensed, we have seen that the beam acquires a new angular position, $\theta = \theta_\pm$, and widths $\Delta\theta_\pm = (\mathrm{d}\theta_\pm/\mathrm{d}\beta)\Delta\beta$ and $\Delta\varphi$. After the lensing the beam subtends the new solid angle $\Delta\Omega_\pm \simeq |\theta_\pm\Delta\theta_\pm\Delta\varphi|$, where the absolute value arises because θ_- is negative. The change in intensity due to the imaging is therefore given by the ratio of these two solid angles summed over the two images:

$$\begin{aligned}\frac{I_{\rm lens}}{I_{\rm source}} &= \sum_{i=\pm}\left|\frac{\theta_i\Delta\theta_i\Delta\varphi}{\beta\Delta\beta\Delta\varphi}\right| = \sum_{i=\pm}\left|\frac{\theta_i}{\beta}\left(\frac{\mathrm{d}\theta_i}{\mathrm{d}\beta}\right)\right| \\ &= \frac{1}{2}\left[\frac{\beta}{\sqrt{\beta^2+4\theta_E^2}} + \frac{\sqrt{\beta^2+4\theta_E^2}}{\beta}\right]. \end{aligned} \tag{6.29}$$

Notice that since $f(x) = \frac{1}{2}(x + 1/x) \geq 1$ (with equality occurring only when $x = 1$) we have $I_{\rm lens} \geq I_{\rm source}$, with equality occurring only if $\theta_E/\beta \to 0$.

The time-dependence enters this intensity because β varies with time as the relative positions of the source and lens change. The maximum change occurs once $\beta \simeq \theta_E$ and so for lens and source a distance of order D away, the time required for a maximal change of intensity can be estimated to be $\tau \simeq \theta_E D/v$, where $v \simeq 200$ km/sec is the typical speed of galactic objects. Taking $D \simeq 10$ kpc and $\theta_E \simeq 2\times10^{-9}$ radians, as above, then gives the estimate $\tau \simeq 0.1$ years, or a few months.

Although it might seem like a million-to-one shot to happen to see a lens and source line up in precisely this way, these kinds of *microlensing* events have been sought by dedicating a telescope to repeatedly photograph large fields of stars over

many nights, and then looking for the few stars whose brightness changes. Such a search inevitably finds various types of variable stars, whose brightness changes for other reasons internal to the star, but these can be identified by seeing how their pattern of variation differs at different wavelengths. Once these are removed, a handful of *bona fide* microlensing events remain, some of which are shown in fig. 14. The frequency of these events is consistent with what is known for small stellar and planetary objects, but is too small to account for the dark matter (whose presence is inferred in all galaxies from measurements of how they rotate).

6.3 Gravitational waves

Waves are a generic consequence of relativistic field theory, and correspond to the fact that information can only travel out through the field at a finite speed (at most, the speed of light), bringing the news to other particles about how their sources have moved. For the special case of General Relativity, since gravity is represented as the geometry of spacetime, gravitational waves are ripples in the fabric of spacetime itself. These are generated when masses are moved relative to one another. These waves are the precise analogs of the electromagnetic waves that are generated by moving electrical charges, and which we know as light, radio waves, x-rays *etc.*, depending on their frequency.

To understand many of the properties of gravitational waves it suffices to consider very small geometrical ripples about flat spacetime, for which the metric has the form

$$g_{\mu\nu}(x) = \eta_{\mu\nu} + h_{\mu\nu}(x)\,, \tag{6.30}$$

where $h_{\mu\nu}$ represents a small deviation that depends on position and time. Calculating the Christoffel symbols and curvature tensor, but dropping all terms that are quadratic and higher in the small quantity $h_{\mu\nu}$ leads a Ricci tensor of the form

$$R_{\mu\nu} = -\frac{1}{2}\,\eta^{\alpha\beta}\Big(\partial_\alpha\partial_\beta\, h_{\mu\nu} - \partial_\mu\partial_\alpha\, h_{\beta\nu} - \partial_\nu\partial_\alpha\, h_{\beta\mu} + \partial_\mu\partial_\nu\, h_{\alpha\beta}\Big) + \mathcal{O}(h^2)\,. \tag{6.31}$$

We can simplify this by using the freedom to change coordinates, in which case a small change, $x^\mu \to x^\mu + \xi^\mu$, leads to

$$h_{\mu\nu} \to h_{\mu\nu} + \eta_{\nu\lambda}\partial_\mu\xi^\lambda + \eta_{\mu\lambda}\partial_\nu\xi^\lambda\,, \tag{6.32}$$

up to quantities that quadratic in ξ^μ. A convenient choice is to use the four independent quantities in ξ^μ to impose the following four independent constraints on $h_{\mu\nu}$:

$$\eta^{\sigma\nu}\partial_\sigma\left(h_{\mu\nu} - \frac{1}{2}\,\eta_{\mu\nu}\eta^{\alpha\beta}h_{\alpha\beta}\right) = 0\,, \tag{6.33}$$

since with this choice the vacuum Einstein equations become, to linear order in h:

$$R_{\mu\nu} = -\frac{1}{2}\,\Box\, h_{\mu\nu} = 0\,, \tag{6.34}$$

where $\Box = \eta^{\alpha\beta}\,\partial_\alpha\partial_\beta$ denotes the d'Alembertian operator (which was introduced in the earlier sections devoted to special relativity).

The significance of this last equation is that it is a *wave equation*, as may be seen by writing it out without the benefit of the Einstein summation convention (and re-introducing the factors of c):

$$\Box\, h_{\mu\nu} = -\frac{1}{c^2}\,\frac{\partial^2 h_{\mu\nu}}{\partial t^2} + \nabla^2 h_{\mu\nu} = 0\,. \tag{6.35}$$

This has as solutions arbitrary linear combinations of plane waves,

$$h_{\mu\nu}(x) = \varepsilon_{\mu\nu}(k)\ \exp\Big[ik_\mu x^\mu\Big]\,, \tag{6.36}$$

where the quantity $\varepsilon_{\mu\nu}$ describes the wave's *polarization* (of which there are two independent forms, more about which below), and the 4-vector k^μ must satisfy

$$k^2 = \eta_{\mu\nu}k^\mu k^\nu = 0\,. \tag{6.37}$$

Writing $k^\mu = \{\omega, \mathbf{k}\}$, eq. (6.37) implies $\mathbf{k} = \omega\hat{\mathbf{k}}$, where $\hat{\mathbf{k}} := \mathbf{k}/|\mathbf{k}|$ is the unit vector normal to the plane of the wave-front. The plane wave then becomes

$$\exp\Big[ik_\mu x^\mu\Big] = \exp\Big[-i\omega t + i\mathbf{k}\cdot\mathbf{x}\Big] = \exp\Big[-i\omega\Big(t - \hat{\mathbf{k}}\cdot\mathbf{x}\Big)\Big]\,. \tag{6.38}$$

General spatial profiles, $h_{\mu\nu}(\mathbf{x})$, are built as linear combinations of the above solutions (*i.e.* by Fourier transformation). Eq. (6.38) implies the waves are functions of the combination $t - x/c$, where $x = \hat{\mathbf{k}}\cdot\mathbf{x}$ and the factors of c are restored. This shows that wave profiles propagate with speed c: both gravitational and electromagnetic waves move at the speed of light.

The two polarizations of gravitational waves correspond to the choices possible for the polarization tensor, $\varepsilon_{\mu\nu}(k)$, which eqs. (6.33) imply satisfy

$$k^\mu\varepsilon_{\mu\nu} - \frac{1}{2}\,k_\nu\,\varepsilon^\mu{}_\mu = 0\,. \tag{6.39}$$

This condition has many more than two solutions, but it is also true that this condition does not completely remove the freedom to change $\varepsilon_{\mu\nu}(k)$ by using coordinate transformations of the form (6.32) with $\xi^\mu(x) = \zeta^\mu e^{ik\cdot x}$, with constant ζ^μ and $k^2 = k_\mu k^\mu = 0$. To see why, notice that under such a transformation we have $\varepsilon_{\mu\nu}(k) \to \tilde{\varepsilon}_{\mu\nu}(k)$ with

$$\tilde{\varepsilon}_{\mu\nu}(k) := \varepsilon_{\mu\nu}(k) + ik_\mu\zeta_\nu + ik_\nu\zeta_\mu\,, \tag{6.40}$$

and so if $\varepsilon_{\mu\nu}(k)$ satisfies (6.39) then so also does $\tilde{\varepsilon}_{\mu\nu}(k)$, since (using $k^2 = 0$)

$$k^\mu \tilde{\varepsilon}_{\mu\nu}(k) - \frac{1}{2}\, k_\nu \tilde{\varepsilon}^\mu{}_\mu = i(k\cdot\zeta)k_\nu - \frac{i}{2} k_\nu\,(2k\cdot\zeta) = 0\,. \tag{6.41}$$

This remaining freedom to redefine coordinates is also removed if $\varepsilon_{\mu\nu}(k)$ is required to also satisfy a second condition: $\ell^\mu \varepsilon_{\mu\nu}(k) = 0$ where ℓ^μ is a second future-pointed null vector, $\ell_\mu \ell^\mu = 0$, chosen such that $k_\mu \ell^\mu = -1$. Contracting (6.39) with ℓ^ν then implies $\varepsilon^\mu{}_\mu = 0$ and so $k^\mu \varepsilon_{\mu\nu} = 0$. For instance, for a wave moving along the positive z-axis with frequency $\omega \neq 0$, we can choose

$$k^\mu = \omega \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \ell^\mu = \frac{1}{2\omega} \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \tag{6.42}$$

and so the most general $\varepsilon_{\mu\nu}(k)$ satisfying $k^\mu \varepsilon_{\mu\nu} = \ell^\mu \varepsilon_{\mu\nu} = \varepsilon^\mu{}_\mu = 0$ is

$$\varepsilon_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \varepsilon_+ & \varepsilon_\times & 0 \\ 0 & \varepsilon_\times & -\varepsilon_+ & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{6.43}$$

where ε_+ and $\varepsilon_\times$ denote the two types of polarizations. Notice the wave is transverse (just like an electromagnetic wave) because these polarizations are only nonzero in the x and y directions for a wave travelling along the z-axis.

If such a wave were to pass by a material it causes nearby particles to move relative to one another in an oscillatory fashion. Because gravity is so weak the induced motion for test particles on Earth is likely to be *extremely* small.

Remarkably, such relative motion was recently observed for the arms of a pair of long laser interferometers, built by the LIGO collaboration with precisely the goal of determining if such waves actually exist in nature. The LIGO interferometers each have arms several kilometers long, and were situated thousands of kilometers away from one another (so that their reactions to any stray environmental effects would not be correlated, unlike for the passage of a gravitational wave). The observed wave had precisely the properties that would have been expected if the wave were emitted by two distant black holes, that initially orbited one another but whose orbits decayed (for reasons described below) until they eventually merged together into a larger, spinning, black hole.

6.4 Binary pulsars

The most precise extra-solar tests of GR come from the study of the orbits of binary pulsars. This section briefly describes what these systems are, and what new features

arise in their study beyond those that are familiar from tests of GR within the solar system.

What are binary pulsars?

A *pulsar* is an astrophysical object that is observed to send regularly repeated bursts of radiation (which could be radio waves, or x-rays *etc.*), whose repetition period ranges from a few seconds to a few milliseconds (see Fig. 15).

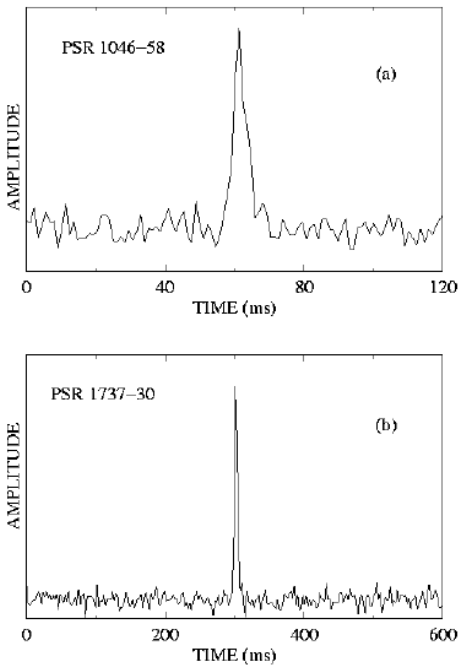


Figure 15: Plots of the spectrum of radiation from two representative pulsars. The pattern shown is repeated over and over again.

Their properties fit what would be expected for a very compact star, called a neutron star, that is rapidly spinning. A neutron star is an exotic beast, with a mass similar to that of the Sun but a radius of only a few kilometres, which is not much larger than its Schwarzschild radius. This small size makes it capable of rotating as quickly as many times a second. Such a star, once rapidly rotating, would tend to set up a large magnetic field which would tend to fire very energetic particles into space along a well-directed beam. Such a beam would rotate with the neutron star, causing a lighthouse-like beam of particles that sweeps around as the neutron star turns. The regular pattern of pulses of radio waves or x-rays seen from the Earth then arises as this lighthouse beam repeatedly sweeps past us.

A *binary* pulsar is a pulsar (*i.e.* a neutron star) that orbits a companion star. This companion can be an ordinary star like the Sun, or possibly even another neutron star. (Stars orbiting one another like this are actually not an unusual occurrence, since just under half the stars visible in the sky orbit a partner in this way.) The fact that the pulsar is in an orbit around another star can be inferred from the shifts that this motion induces in the frequency of the light the pulsar emits, a phenomenon called the Doppler effect.

It takes a pulsar a few days or so to orbit once around its partner, indicating that the pulsar and its companion are closer to one another than Mercury is to our Sun. Together with the compactness of the pulsar itself, this means that the gravitational

fields through which these stars pass are much stronger than those to which we are accustomed in the solar system. What's more, the fact that the pulsar sends out such regularly repeating signals means that we see an exquisitely precise clock in orbit around another star, providing a remarkable chance to measure the nature of space and time in these orbits.

For all of these reasons there are a number of relativistic effects that are comparatively large relative to those seen in the solar system. This allows a potentially greater suite of tests of GR than are possible in the solar system. Some of the relativistic effects that have been seen in these systems are the ones that are also seen in the solar system. These include

- the relativistic precession, or periastron shift, of the pulsar orbits;
- the relativistic slowing of time as counted by the pulsar as it moves in the gravitational field of its companion;
- the Shapiro time delay of the pulsar signals as they pass through the gravitational field of the massive companion.

Orbital Decay

There are also new effects seen in binary pulsar systems, that have not been seen before. Foremost among these is the observed decay of the pulsar orbit, which are very slowly spiralling in towards one another. This orbital decay is observed as an extremely small, slow, secular increase in the orbital period, seen in Fig. 16. Although small, the increase is observable because the pulsars have been watched consistently over a long period of time, in some cases — for the Hulse-Taylor pulsar, for example — for several decades.

Why is this decay a relativistic effect? It is because orbital decay indicates that the pulsar orbit is losing energy. General Relativity predicts such an energy loss, due to the emission of gravitational waves. After a short aside to summarize the properties of gravitational waves, we return to a discussion of their implications for pulsar orbits in more detail.

Gravitational Waves and Orbital Decay

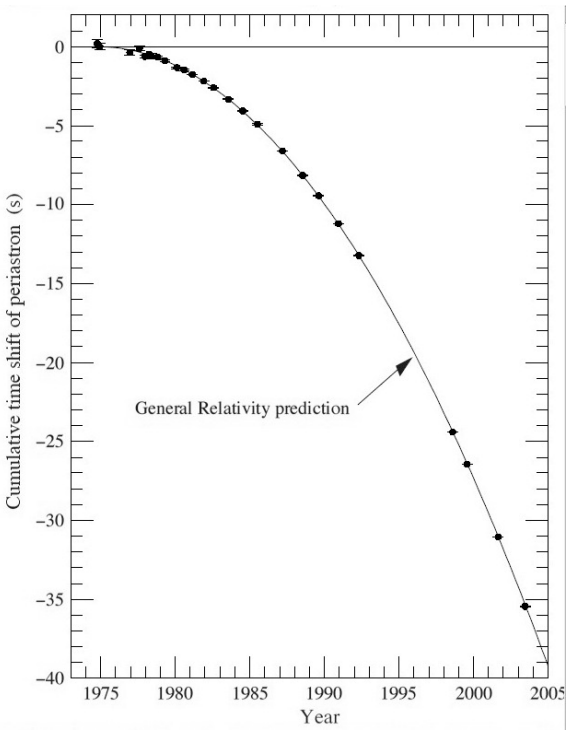


Figure 16: A plot comparing measurements of the rate of decay of the pulsar orbital period as a function of time, with the prediction following from gravitational radiation in GR.

Because the waves are produced by moving masses, much as electromagnetic waves arise from the motion of electric charges, the energy loss rate into gravitational radiation turns out to be proportional to a power of both the total mass, M, of the orbiting system and of its orbital angular frequency, $\Omega = 2\pi/P$:

$$L = \frac{128G}{5c^5} M^2 R^4 \Omega^6 \simeq 2\times 10^{33} \frac{\text{erg}}{\text{sec}} \left[\left(\frac{M}{M_\odot}\right)\left(\frac{1\text{ hour}}{P}\right.\right. \tag{6.44}$$

and the second equality uses Kepler's 3rd Law, $\Omega^2 = GM/R^3$, to trade R for M and Ω (or the orbital period, P). Alternatively, $L \simeq 25\,(GM^2\Omega/R)\,(v/c)^5$, where $v \simeq R\,\Omega$ is of order the orbital speed. This way of writing things shows that the first term represents an emission of an appreciable fraction of the gravitational binding energy per period, while the second factor suppresses the result by 5 powers of v/c. For orbits with a period of $P \simeq 1$ hour $\simeq 1.1 \times 10^{-4}$ year, Kepler's 3rd Law implies a mean orbital radius that is of order $R \simeq (1.1\times 10^{-4})^{2/3}$ AU $\simeq 0.002$ AU, where 1 AU $\simeq 1.5\times 10^{11}$ m. Consequently, $v/c \simeq R\Omega/c \simeq 0.1$.

Equating this to the loss rate of orbital energy and using the properties of Newtonian orbits to relate the energy to the orbital period, give the resulting GR prediction for the period change

$$\frac{\mathrm{d}P}{\mathrm{d}t} = -3 \times 10^{-12} \left[\left(\frac{M}{M_\odot}\right)\left(\frac{1\text{ hour}}{P}\right)\right]^{5/3} . \tag{6.45}$$

Fig. 16 plots the comparison between the prediction of eq. (6.45) and the observed rate of decrease of orbital period for the *Hulse-Taylor* pulsar, PSR B1913+16, which has been closely and continually watched for several decades now.

But in order for this to provide a test of General Relativity it is necessary to know what the masses are for both the pulsar and its companion. How were these

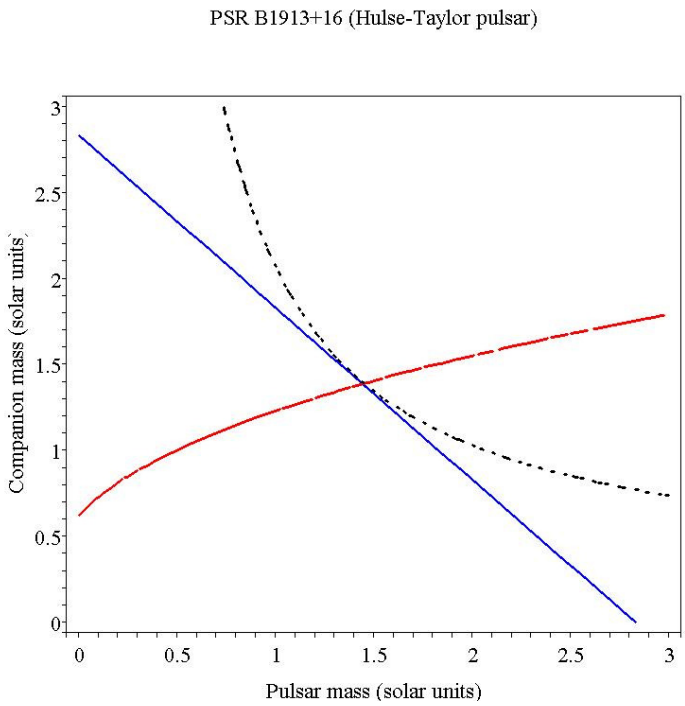


Figure 17: Plot of the prediction for periastron shift (blue), orbital decay rate (dotted black) and relativistic time delay (dashed red) for the Hulse-Taylor binary, PSR 1913+16 in General Relativity, as functions of the mass of the pulsar and its companion in the binary system. If GR is true all three lines should touch at a point (within errors), revealing the masses of the actual bodies involved.

measured in order to make the comparison of fig. 16? Although they cannot be measured directly, since the companion is not in this case visible, progress is possible because the masses also appear in the prediction for the size of the other relativistic effects that are observed for pulsars. The strategy is to use the agreement of these predictions with experiment to infer the masses of the orbiting stars, and then to use these to predict the gravitational radiation rate.

Fig. 17 illustrates this strategy, showing three curves that give the relationship between the pulsar mass and the mass of its companion that follows by requiring the prediction of GR for the precession of the orbit, the slowing down of the pulsar clock, and the orbital decay caused by gravitational radiation, to agree with what is seen for a particular pulsar. If GR provides a correct description of the pulsar system, all three of these curves should touch at a single point, corresponding to the masses of the two bodies in the orbit. The remarkable fact is that they do, and because they do we learn both that GR is working well, and what the masses of the two stars must be. And given these masses the rate of decay evolves in time in precisely the way predicted by GR, as seen in Fig. 16.

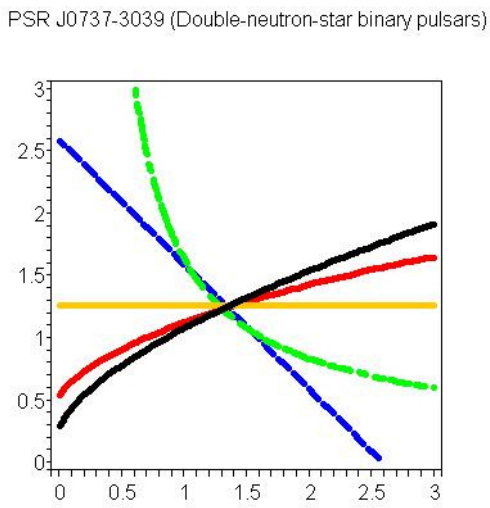


Figure 18: Plot of the prediction for periastron shift (dashed blue), orbital decay rate (dashed green), relativistic time delay (red), Shapiro time delay (orange) and shift (black), for the double-binary system, PSR J0737-3039 in General Relativity, as functions of the mass of the two pulsars. If GR is true all five lines should touch at a point (which they do within the errors).

The Double Binary Pulsar

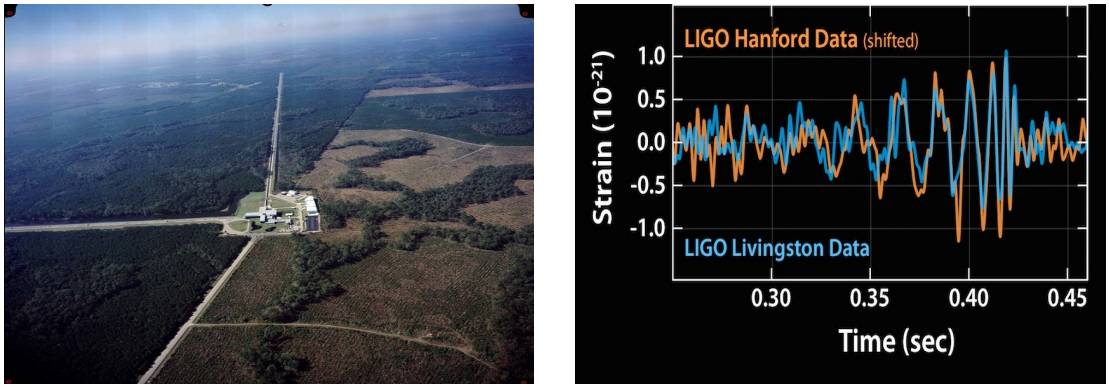
Almost a thousand pulsars have been discovered over the years, and some of the ones found more recently promise to provide new ways to test General Relativity. A particularly promising system is given by the pulsar J0737-3039, which (unusually) consists of a pulsar being orbited by another pulsar. Even better, the pulsars almost eclipse one another (that is, the beam from one passes through the astrophysical detritus that surrounds the other), and so their orbit is inclined so that we see it edge-on from the point of view of the Earth.

This system is something of a holy grail for testing general relativity, since it provides access to more relativistic effects than do other pulsar systems. For example, the near-eclipsing of one pulsar by the other implies that the observed light signal from one pulsar passes very close to the other on its way to the Earth, and so experiences a Shapiro time delay that is observable and may be compared with predictions.

The prediction of General Relativity for the five observable relativistic effects as a function of the two pulsar masses is given in Fig. 18. If GR provides a correct description of the pulsar system, all of the curves should touch at a single point, corresponding to the masses of the two pulsars. Remarkably, again they all do to within the errors, confirming that GR is working well. And just like for the Hulse-Taylor pulsar, the precision of these tests will improve the longer its signals are watched.

Direct Detection of Gravitational Waves

Big things took place in the years since these notes were first written (2008), with the direct detection in 2015 of gravitational waves by the LIGO detector on Earth. The detector used a very large interferometer to detect the passage of gravitational waves. This is done by splitting a laser beam and having it travel back and forth multiple times down the length of two perpendicular arms, before recombining the



(a) No magnetic field

(b) Magnetic field applied

Figure 19: Left panel: Aerial view of one of the two LIGO interferometers. Laser light is split and sent down each arm, multiply reflecting on mirrors at each end before being recombined to mutually interfere. This allows precision comparison of the length of each arm as a function of time, with the passage of a gravitational wave observed as an oscillatory relative change in length. Right panel: The first detected gravitational-wave signal, with the two colours representing the correlated signals seen by to the two LIGO interferometers. (Figure source: LIGO Collaboration).

beam and measuring the resulting interference pattern. The interference pattern is extremely sensitive to the relative length of the interferometer's two arms, and the passage of a gravitational is detected by observing the characteristic oscillation in this relative length caused by the passage of a gravitational wave.

Extreme care is taken when doing so to isolate the interferometer mirrors from their environment, and extremely large interferometer ams were used, in order to allow the extremely small signal to be detected and to be distinguished from other spurious sources of vibration. Coincident detection by two independent interferometers situated more than 1,000 km apart from one another was also required to eliminate any unknown local sources at either site. The detected signal is shown in the right-hand panel of Fig. 19, showing in particular how the signals of the two interferometers indeed correlated with one another.

6.5 Astrophysical Black Holes

There is considerable evidence within astrophysics for the existence of black holes in the universe, and this provides support for the general picture of these objects that is painted by General Relativity even though they do not yet provide precision tests of the theory.

In each case the thrust of the evidence identifies the total mass of an unseen central object by watching how it is orbited by objects that we can see. This is

compared with upper limits to its size, that either come from direct observations of the innermost positions of the orbiting objects, or by considerations related to how fast the object is observed to change its brightness. In many cases there is so much mass crammed into so small a region that there is no known way for it to support itself against collapse into a black hole.

There are two broad classes of black holes that have been reliably identified in this way: stellar-sized black holes; and super-massive black holes in the centers of galaxies. (Intermediate-sized black holes, with masses of order thousands of solar masses, are also believed to exist — perhaps at the centers of globular clusters of stars — but evidence for them is more controversial.)

Stellar-sized black holes

Among the first black hole candidates were those having masses not so different from that of the Sun, as would be expected as the endpoints of the gravitational collapse of a sufficiently massive star. Although the black hole itself is not visible, it can be observed when matter that falls into it radiates. And such infalling matter is particularly likely in situations where the black hole is in an orbit with another ordinary star, since in this case material from the companion star can be siphoned off to continually feed the black hole (as illustrated in fig. 20). As it falls in, this matter can become hot enough to emit x-rays, and many examples of such *x-ray binaries* are known (some of which are among the brightest objects in the sky when viewed in x-rays).

Figure 20: A drawing (courtesy of the European Space Agency and Hubble Space Telescope) of an x-ray binary system.

Sometimes the stellar companion to the black hole is a star that is sufficiently luminous to be directly visible in optical or radio wavelengths. In such cases the black hole dominates the luminosity of the binary pair in x-rays, while its stellar partner is the one that can be seen in the visible spectrum. Among the most famous x-ray binaries that are believed to consist of black holes is Cygnus X-1, the brightest x-ray source in its constellation as seen from Earth. The orbital partner of the x-ray source has been identified to be the super-giant star AGK2 +35 1910 = HDE 226868, which is itself incapable of emitting the x-rays observed from its partner. Both stars are 2 kpc away from us, and move together in an aggregation of stars, indicating a probable common origin.

The light from this star exhibits the characteristic Doppler shifts that are associated with being in an orbit about a massive partner, with an orbital period of 5.6 days. Because the plane of the orbit relative to the sky is unknown it is trickier to determine unambiguously the mass of the partner, but the best estimates lead to a mass of $8.7 \pm 0.8\ M_\odot$. On the other hand, since the x-ray source varies in time with a timescale faster than several times a second, it cannot be larger than a fraction of a light-second across (the best estimates indicate its size is smaller than 10^5 km). The compact object is believed to be a black hole since neutron stars cannot be this massive, and no other object is known that can have this much mass compressed within the allowed size.

Galactic black holes

Enormous black holes, more massive than a million Suns, are believed to reside at the center of most galaxies. When fed by infalling material, these can be among the most luminous objects in the universe.

The Milky Way

Detailed studies of the properties of the galactic center give very good evidence that our own galaxy, the Milky Way, itself contains such a super-massive black hole. This evidence partially comes from the indications that there is a very powerful energy source located near the galactic center, as would be expected if material accretes there onto a black hole. The galactic center is an active energy emitter when viewed in radio and x-ray wavelengths. (Studies with visible light are more difficult because this is obscured by the dust that lies along our line of sight to the galactic center.) Fig. 21 shows an x-ray photograph of our galactic center, showing the presence of a variety of sources.

A more detailed picture of the Milky Way's central object is formed by studying how it affects the motion of stars in its immediate vicinity. The motion of a handful of such stars have been observed continually for 16 years, allowing a detailed reconstruction of their orbits that in some cases includes enough time for them to have completed an entire revolution about the galactic center [6].

The observed orbits are consistent with motion in the presence of a very massive point source, since they are very close to Keplerian. For instance, one of the innermost stars — the star S2 of fig. 22 — moves in an orbit whose eccentricity is $e = 0.88$ and whose semimajor axis subtends an angle (seen from the Earth) of 0.1 seconds of arc, or 4×10^{-7} radians. Since the galactic center is 8.3 kpc away, this corresponds to an orbit whose semimajor axis is 0.01 light years, or about 4 light days.

These orbits indicate that the mass of the central mass is $4.3 \times 10^6\ M_\odot$. On the other hand, the size of the source must be much smaller than the point of closest

approach of the smallest orbit (which turns out to be 17 light hours) because the orbits are consistent with the central object being at a single point. For comparison, the Schwarzschild radius corresponding to a mass of $4.3 \times 10^6\ M_\odot$ is about 1.2×10^7 km, or 43 light seconds. (For reference, our Sun is about a light second across, and the Earth is about 8 light minutes – or 480 light seconds – from the Sun.)

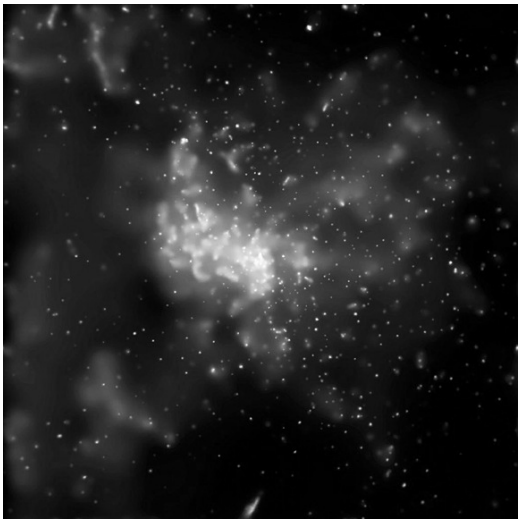


Figure 21: An Chandra satellite x-ray image of the center of our galaxy.

The central source being orbited is believed to be a black hole because there is no other known way to cram this much mass into so small a region, without its being directly visible. If there were a black hole at the galactic center, simulations show that stars would naturally be found orbiting it that are formed as huge gas clouds fall into the black hole.

Active Galaxies

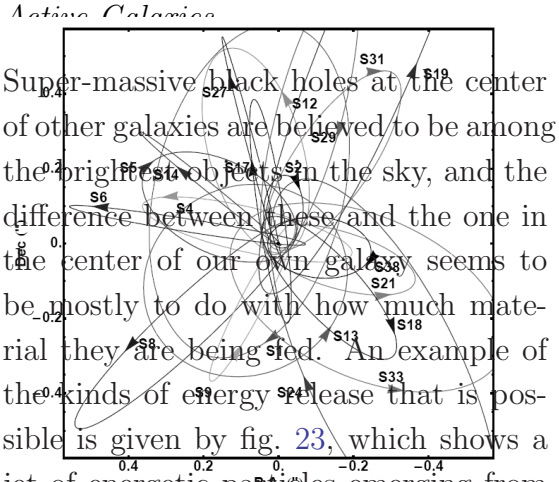


Figure 22: A plot of reconstructed orbits of several stars orbiting the center of our galaxy [6].

Super-massive black holes at the center of other galaxies are believed to be among the brightest objects in the sky, and the difference between these and the one in the center of our own galaxy seems to be mostly to do with how much material they are being fed. An example of the kinds of energy release that is possible is given by fig. 23, which shows a jet of energetic particles emerging from the center of the large elliptical galaxy M87 in the Virgo cluster about 17 Mpc away from us. This jet is more than 5000 light years long, and the apparent speed of the matter being ejected along it has been measured using the Hubble space telescope. This finds the apparent motion to be between 4 to 6 times the speed of light, an illusion that indicates (see exercise 11 in chapter 2) that the jet is moving at relativistic speeds (but slower than light) largely directed towards us, along the line of sight. Other indications of a strong energy source in M87 comes from its strong emissions in x-rays and gamma rays.

The argument that the energy source at the galactic center is a black hole again comes from measurements indicating that an enormous amount of mass resides within

a comparatively small volume. For M87 the mass measurement is made by following the speed with which hot gas orbits the central object in a central disc as a function of the gas' distance from the center. The speed of the orbits is measurable as a net Doppler red-shift on one (receding) side of the central object, and a net blue-shift on the other (approaching) side. These measurements indicate the central object is enormously massive: its mass is 3×10^9 solar masses.

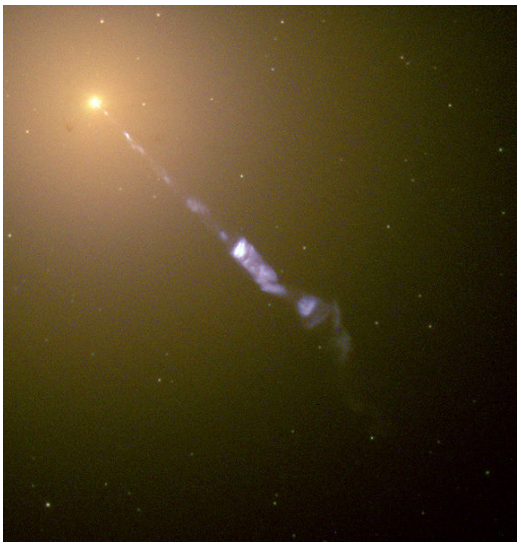


Figure 23: A Hubble Space Telescope photograph of an energetic jet emerging from the center of galaxy M87.

An upper limit to the size within which this mass is compressed comes from the HESS gamma ray telescope which sees variations in the gamma ray flux that occur over timescales of a few days. This indicates that the 3 billion solar masses lie within a region that is a few light days across. For comparison, the Schwarzschild radius of a black hole whose mass is $3 \times 10^9\ M_\odot$ is about 8 light hours (which is larger than the planetary orbits of our solar system). The only known object that can be this massive and yet so small is a black hole.

A second piece of circumstantial evidence for the central object being a black hole is the enormous efficiency – 6 times better than the nuclear fusion that drives stars like the Sun – with which a black hole is able to convert mass into energy. To see why this is so consider the conserved quantity, $\mathcal{E} = -(1 - r_s/r)(\mathrm{d}t/\mathrm{d}\tau)$, of a particle moving in a circular orbit at radius r. The 4-velocity for such a particle is

$$u^\mu = \gamma \begin{pmatrix} 1 \\ 0 \\ 0 \\ \Omega \end{pmatrix}, \tag{6.46}$$

where $\gamma = \mathrm{d}t/\mathrm{d}\tau$ and $\Omega = \mathrm{d}\phi/\mathrm{d}t$. Since $1 = -u \cdot u = \gamma^2[(1 - r_s/r) - r^2\Omega^2]$, we have

$$\gamma = \frac{1}{\sqrt{(1 - r_s/r) - r^2\Omega^2}} = \frac{1}{\sqrt{1 - 3r_s/2r}}, \tag{6.47}$$

where the last equality uses Kepler's 3rd Law, $\Omega^2 = GM/r^3 = r_s/(2r^3)$ — which is exactly satisfied for circular orbits in Schwarzschild spacetime (see exercise 28) — to write $r^2\Omega^2 = r_s/(2r)$.

Using eq. (6.47) in the expression for $\mathcal{E}$ then gives

$$\mathcal{E} = \frac{1 - r_s/r}{\sqrt{1 - 3r_s/2r}}\,, \tag{6.48}$$

which for the innermost circular orbit at $r = 6M = 3r_s$ becomes

$$\mathcal{E} = \frac{2\sqrt{2}}{3} \simeq 0.94\,. \tag{6.49}$$

Since $\mathcal{E} = 1$ for a particle at rest at infinity, $\mathcal{E}$ can be interpreted as the energy per unit rest mass, and eq. (6.49) shows that as much as 6% of the rest mass of a particle can be converted to gravitational binding energy as a particle falls into an orbit close to the black hole. This is ultimately the energy that is released to drive the acceleration of the few particles that escape the black hole by being accelerated out the jet (which emerges along the axis of rotation for the accretion disc that infalling matter forms around the black hole).

By comparison, typical nuclear interactions release the nuclear binding energy, and comparing the 27 MeV released by each fusion of a Helium nucleus from four Hydrogen nuclei shows that this type of fusion releases roughly only 1% of the rest mass available as energy. The energy released from matter infalling into a black hole is therefore expected to be roughly 6 times more abundant than would have been released by using the same amount of matter in some sort of a nuclear reaction.

7. Cosmology

The earlier sections show that once one accepts Einstein's point of view that the right way to describe gravity is as the curvature of spacetime, it becomes possible to relate our local geometry to the distribution of matter in our immediate vicinity. However the same logic also connects geometry to the matter distribution over much larger scales, and in principle should relate the geometry of the Universe as a whole to the average distribution of matter on the largest observable scales.

It is this realization that underlies the science of cosmology, which uses observations of the distribution of matter on very large scales to make inferences about the overall curvature of space and time, and how these change in time. This section provides a brief overview of the Big Bang theory of cosmology, with an emphasis is on the theoretical ideas that pertain to General Relativity.

7.1 Kinematics of an expanding Universe

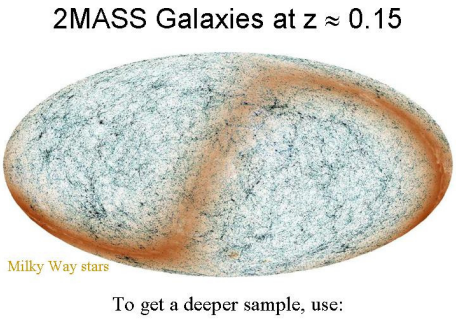


Figure 24: A map of the nearby distribution of galaxies on the sky seen from the earth, as obtained from the 2-Mass galaxy survey. The 'S'-shaped smear is where our view is obscured by the presence of our own galaxy in the foreground.

We start with a section describing the geometry of spacetime on which all of the subsequent sections rely. The key underlying assumption in this section is that the universe is homogeneous and isotropic when seen on the largest distance scales. Here isotropic means that all directions are equivalent as seen by an observer situated at a particular point, and so is equivalent to the spherical symmetry of the geometry about this point. Homogeneity states that the above isotropy holds for an observer located at *any* point. Until relatively recently this assertion about the homogeneity and isotropy of the universe was an assumption, often called the *Cosmological Principle*. More recently it has become possible to put this assertion on an observational footing, based on large-scale surveys of the distribution of matter and radiation within the observed universe. The isotropy of this distribution relative to our own vantage point can be seen in fig. 24, which shows a the results of a representative galaxy survey.

The LFRW Metric

The assumption that the universe is spherically symmetric and homogeneous puts a strong restriction on the form of the universe's overall geometry. We have already seen that spherical symmetry by itself ensures that the metric can always be written in the 'isotropic' form of eq. (3.20):

$$\mathrm{d}s^2 = -e^{2\alpha(\varrho,\tau)}\,\mathrm{d}\tau^2 + e^{2\beta(\varrho,\tau)}\left[\mathrm{d}\varrho^2 + \varrho^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right], \tag{7.1}$$

for some unknown functions, $\alpha(\varrho,\tau)$ and $\beta(\varrho,\tau)$.

These functions are further restricted by the requirement of homogeneity, which says that α must be a function only of the time coordinate, $\alpha = \alpha(\tau)$. This function can then be completely eliminated by redefining the time coordinate, $\tau \to t$, so that $e^{\alpha(\tau)}\mathrm{d}\tau = \mathrm{d}t$.

It is tempting to conclude that homogeneity amounts to translation invariance, and so β must also be independent of ϱ. Although this does provide a homogeneous and isotropic space, it does not produce the most general one. The condition on β is slightly weaker: β must come as a sum, $\beta = f(\tau) + g(\varrho)$. Although β can depend on ϱ, the allowed dependence is very restrictive. Homogeneity turns out to require

that $g(\varrho)$ is such that we can change variables $\varrho \to r$, in a way that allows it to be put into the LeMaitre-Friedmann-Robertson-Walker (LFRW) form:

$$\begin{aligned} \mathrm{d}s^2 &= -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-\kappa r^2/r_0^2} + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2\right] \\ &= -\mathrm{d}t^2 + a^2(t)\left[\mathrm{d}\ell^2 + r^2(\ell)\,\mathrm{d}\theta^2 + r^2(\ell)\sin^2\theta\,\mathrm{d}\phi^2\right], \end{aligned} \tag{7.2}$$

where r_0 is a constant and κ can take one of the following three values: $\kappa = 1, 0, -1$. This is the most general 4D geometry that is consistent with isotropy and homogeneity of its spatial slices, and it is characterized by the one unknown function, $a(t) = e^{f(\tau(t))}$. The content of Einstein's equations will be to relate the shape of the function $a(t)$ to the matter content of the universe.

The coordinate ℓ in this metric is related to r by $\mathrm{d}\ell = \mathrm{d}r/(1-\kappa r^2/r_0^2)^{1/2}$, so if we demand $\ell(r=0) = 0$ then

$$r(\ell) = \begin{cases} r_0\,\sin(\ell/r_0) & \text{if} \quad \kappa = +1 \\ \ell & \text{if} \quad \kappa = 0 \\ r_0\,\sinh(\ell/r_0) & \text{if} \quad \kappa = -1 \end{cases}. \tag{7.3}$$

Notice that the metric, eq. (7.2), is invariant under the following re-scaling of parameters: $a \to a/\lambda$, $r_0 \to \lambda r_0$, provided we also re-scale the coordinate $\ell \to \lambda\ell$. This freedom is often used to choose convenient units, such as by choosing λ to ensure $r_0 = 1$ (if $\kappa \neq 0$), or perhaps to set $a(t_0) = 1$ for some t_0.

The coordinates used all have the following simple physical interpretations.

- t represents the proper time along the time-like trajectories along which ℓ, θ and ϕ are fixed. The range over which t may run is defined by the region over which the function $a(t)$ is neither zero nor infinite.

- ℓ is simply related to the proper distance measured along the radial directions along which t, θ and ϕ are fixed, since this proper distance is given by

$$D(\ell, t) = \ell\, a(t)\,. \tag{7.4}$$

  If $\kappa = 0, -1$ then ℓ takes values in the range $0 < \ell < \infty$, but if $\kappa = +1$ then ℓ is restricted to run over $0 < \ell < \pi r_0$ because $r(\ell)$ vanishes at $\ell = \pi r_0$.

- $0 < \theta < \pi$ and $0 < \phi < 2\pi$ represent the usual angular coordinates of spherical polar coordinates. (Spherical coordinates furnish a convenient description of our view of the universe, with the origin of coordinates representing our vantage point.) The geometry is invariant under the $SO(3)$ rotations of the 2-dimensional spherical surfaces at fixed ℓ and t which these coordinates parameterize.

- $r(\ell)$ is simply related to the arc-length measured along these spherical surfaces of fixed ℓ and t in the sense that a small angular displacement, $\mathrm{d}\theta$, is subtended by a proper arc-length
$$\mathrm{d}s = a(t)\, r(\ell)\, \mathrm{d}\theta\,, \tag{7.5}$$
at a coordinate position ℓ. It follows that the sphere having proper radius $\ell a(t)$ has a proper circumference of $\mathcal{C} = 2\pi\, r(\ell) a(t)$ and its proper area is $\mathcal{A} = 4\pi\, r^2(\ell) a^2(t)$.

The quantities κ and r_0 characterize the curvature of the spatial slices at fixed t, in the following way.

Flat Spatial Curvature

If $\kappa = 0$ then $r(\ell) = \ell$ and the spatial part of the LFRW metric reduces (apart from the overall factor, $a^2(t)$) to the metric of flat 3-dimensional space, written in spherical polar coordinates:

$$\mathrm{d}s_3^2 = \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2)\,, \tag{7.6}$$

as may be seen by performing the standard coordinate transformation

$$x = r\, \sin\theta\, \cos\phi\,, \qquad y = r\, \sin\theta\, \sin\phi\,, \qquad z = r\, \cos\theta \tag{7.7}$$

in the metric of eq. (2.2). In this case the parameter r_0 does not appear in the metric.

Positive Spatial Curvature

When $\kappa = 1$ we have $r(\ell) = r_0 \sin(\ell/r_0)$ and the metric for t fixed describes the geometry of a 3-dimensional sphere whose radius of curvature is r_0. For instance, in this case the circumference of a circle of proper radius $a(t)\,\ell$ is

$$\mathcal{C} = 2\pi a(t)\, r_0\, \sin\left(\frac{\ell}{r_0}\right)\,, \tag{7.8}$$

which is strictly smaller than the corresponding flat result: $\mathcal{C} < 2\pi a(t)\,\ell$.

Furthermore, for fixed t, $\mathcal{C}$ is a monotonically increasing function of ℓ until $\ell = \pi r_0/2$, but beyond this point $\mathcal{C}$ decreases until it vanishes at $\ell = \pi r_0$. The maximum coordinate circumference obtained in this way is $\mathcal{C}_{\max} = 2\pi a(t)\, r_0$.

Notice also that the flat $\kappa = 0$ case is retrieved in the limit of infinite curvature radius: $r_0 \to \infty$.

Negative Spatial Curvature

When $\kappa = -1$ we have $r(\ell) = r_0 \sinh(\ell/r_0)$, which makes the metric for constant t describe the geometry of a 3-dimensional surface of negative constant curvature.

(The surface of a saddle is close to being a 2-dimensional surface having constant negative curvature.) The radius of curvature of this space is r_0. In this case the circumference of a circle of proper radius $a(t)\,\ell$ grows monotonically with ℓ,

$$\mathcal{C} = 2\pi a(t)\, r_0 \,\sinh\left(\frac{\ell}{r_0}\right) , \tag{7.9}$$

and is always larger than the corresponding flat-space result: $\mathcal{C} > 2\pi a(t)\,\ell$.

Again the flat $\kappa = 0$ case is retrieved in the limit of infinite curvature radius: $r_0 \to \infty$.

Particle Motion

For the purposes of cosmology galaxies are particles, and so their trajectories in this spacetime are given, as usual, by solutions to the geodesic equation, eq. (3.36)

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}s^2} + \Gamma^\mu_{\nu\lambda}[x(s)] \left(\frac{\mathrm{d}x^\nu}{\mathrm{d}s}\right)\left(\frac{\mathrm{d}x^\lambda}{\mathrm{d}s}\right) = 0 , \tag{7.10}$$

with the Christoffel symbols, $\Gamma^\mu_{\nu\lambda}$, given by eq. (2.39).

For the LFRW metric the only nonzero Christoffel symbols turn out to be given by

$$\begin{gathered}
\Gamma^t_{\ell\ell} = a\dot{a} , \quad \Gamma^t_{\theta\theta} = a\dot{a}\, r^2 , \quad \Gamma^t_{\phi\phi} = a\dot{a}\, r^2 \,\sin^2\theta , \\
\Gamma^\ell_{t\ell} = \Gamma^\ell_{\ell t} = \Gamma^\theta_{t\theta} = \Gamma^\theta_{\theta t} = \Gamma^\phi_{t\phi} = \Gamma^\phi_{\phi t} = \frac{\dot{a}}{a} , \\
\Gamma^\ell_{\theta\theta} = -rr' , \quad \Gamma^\ell_{\phi\phi} = -rr' \,\sin^2\theta , \quad \Gamma^\theta_{\ell\theta} = \Gamma^\theta_{\theta\ell} = \Gamma^\phi_{\ell\phi} = \Gamma^\phi_{\phi\ell} = \frac{r'}{r} , \\
\Gamma^\theta_{\phi\phi} = -\sin\theta\,\cos\theta , \quad \Gamma^\phi_{\theta\phi} = \cot\theta ,
\end{gathered} \tag{7.11}$$

where the dots denote differentiation with respect to t and the primes represent derivatives with respect to ℓ.

Using these expressions for the Christoffel symbols, the four geodesic equations then become

$$\begin{aligned}
\frac{\mathrm{d}^2 t}{\mathrm{d}s^2} + a\dot{a}\left\{\left(\frac{\mathrm{d}\ell}{\mathrm{d}s}\right)^2 + r^2\left[\left(\frac{\mathrm{d}\theta}{\mathrm{d}s}\right)^2 + \sin^2\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2\right]\right\} &= 0 \\
\frac{\mathrm{d}^2 \ell}{\mathrm{d}s^2} + 2\left(\frac{\dot{a}}{a}\right)\frac{\mathrm{d}\ell}{\mathrm{d}s}\frac{\mathrm{d}t}{\mathrm{d}s} - rr'\left[\left(\frac{\mathrm{d}\theta}{\mathrm{d}s}\right)^2 + \sin^2\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2\right] &= 0 \\
\frac{\mathrm{d}^2 \theta}{\mathrm{d}s^2} + 2\left(\frac{\dot{a}}{a}\right)\frac{\mathrm{d}\theta}{\mathrm{d}s}\frac{\mathrm{d}t}{\mathrm{d}s} + 2\left(\frac{r'}{r}\right)\frac{\mathrm{d}\theta}{\mathrm{d}s}\frac{\mathrm{d}\ell}{\mathrm{d}s} - \sin\theta\cos\theta\left(\frac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2 &= 0 \\
\frac{\mathrm{d}^2 \phi}{\mathrm{d}s^2} + 2\left(\frac{\dot{a}}{a}\right)\frac{\mathrm{d}\phi}{\mathrm{d}s}\frac{\mathrm{d}t}{\mathrm{d}s} + 2\left(\frac{r'}{r}\right)\frac{\mathrm{d}\phi}{\mathrm{d}s}\frac{\mathrm{d}\ell}{\mathrm{d}s} + 2\cot\theta\,\frac{\mathrm{d}\theta}{\mathrm{d}s}\frac{\mathrm{d}\phi}{\mathrm{d}s} &= 0
\end{aligned}$$

Since the metric is rotationally invariant, angular momentum is conserved along these geodesics in precisely the same way as it was for the Schwarzschild metric. That is, the motion is guaranteed to take place entirely within a plane, and we are free to choose our coordinates so that this plane is described by the equator, $\theta = \frac{\pi}{2}$, for all s (which is clearly a solution to the $\mathrm{d}^2\theta/\mathrm{d}s^2$ equation above). Rotational invariance implies that the equation of motion for ϕ may be integrated once, to give (using $\theta = \pi/2$)

$$L = a^2 r^2 \frac{\mathrm{d}\phi}{\mathrm{d}s}\,, \tag{7.12}$$

where L is a constant.

The remaining equations can often be explicitly integrated. When $\dot{a} = 0$ they describe motion at constant speed along the geodesics of the spatial geometry (along straight lines if this geometry is flat: $\kappa = 0$). When $\dot{a} \neq 0$, motion along these geodesics instead tends to damp out under the influence of the $\dot{a}/a$ terms in the equations (called the *Hubble 'friction'* terms). This damping arises because the expansion of the universe extracts energy from the motion. Several special cases are of particular interest.

- *Radial Motion:* If $\mathrm{d}\theta/\mathrm{d}s = \mathrm{d}\phi/\mathrm{d}s = 0$ at one point, then these quantities remain zero along the entire geodesic. This shows that an initially radial motion continues in the radial direction for all times. Radial free fall is described by the equations of motion

$$\frac{\mathrm{d}^2 t}{\mathrm{d}s^2} + a\,\dot{a}\left(\frac{\mathrm{d}\ell}{\mathrm{d}s}\right)^2 = 0 \quad \text{and} \quad \frac{\mathrm{d}^2\ell}{\mathrm{d}s^2} + 2\,\frac{\dot{a}}{a}\left(\frac{\mathrm{d}\ell}{\mathrm{d}s}\right)\left(\frac{\mathrm{d}t}{\mathrm{d}s}\right) = 0\,. \tag{7.13}$$

  These together imply the constancy of the proper distance along the geodesic, $(\mathrm{d}/\mathrm{d}s)[(\mathrm{d}t/\mathrm{d}s)^2 - a^2(\mathrm{d}\ell/\mathrm{d}s)^2] = 0$, as expected on general grounds.

- *Inertial Motion:* If a galaxy is initially at rest — and so $\mathrm{d}\ell/\mathrm{d}s = \mathrm{d}\theta/\mathrm{d}s = \mathrm{d}\phi/\mathrm{d}s = 0$ — then it remains at rest, at fixed coordinate position, for all t. This shows that observers who remain at fixed position ℓ (the analogs of the observers at fixed r for the Schwarzschild metric) move along geodesics (unlike for the fixed-r observers in Schwarzschild).

Hubble Flow and Peculiar Motion

Consider now a particle moving more slowly than light, but for which some force keeps it from moving along a geodesic. This might happen for a galaxy, for instance, if some local density enhancement attracts it. In particular, consider for simplicity a galaxy having coordinates $(t, \ell = \ell(t), \theta = \theta_0, \phi = \phi_0)$, which moves on a purely radial trajectory. The proper distance to this galaxy from, say, the origin is given by

$D(\ell, t) = \ell(t)a(t)$, and so its proper velocity relative to an observer at the origin is

$$V_p = \frac{\mathrm{d}D}{\mathrm{d}t} = \ell\,\frac{\mathrm{d}a}{\mathrm{d}t} + a\,\frac{\mathrm{d}\ell}{\mathrm{d}t} = H\,D + a\,\frac{\mathrm{d}\ell}{\mathrm{d}t}\,, \tag{7.14}$$

where

$$H(t) := \frac{1}{a}\left(\frac{\mathrm{d}a}{\mathrm{d}t}\right). \tag{7.15}$$

The first term of eq. (7.14) describes the galaxy's apparent motion due to the overall universal expansion, and expresses the *Hubble Law*: in the absence of other motions at any given instant all galaxies recede with a proper speed which is proportional to their proper distance. (This law describes the observed overall motion of galaxies very well, as is illustrated in fig. 25.) By contrast, the second term describes *peculiar* velocity,

$$V_{\rm pec} = a\,\frac{\mathrm{d}\ell}{\mathrm{d}t}\,, \tag{7.16}$$

which expresses any deviation from geodesic motion in the overall LFRW metric.

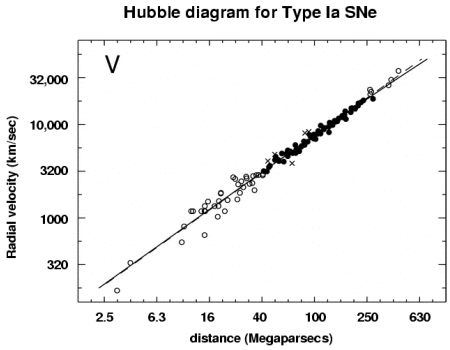


Figure 25: A plot of velocity (redshift) vs (luminosity) distance for a class of bright, distant objects that are used to trace the motions of very distant galaxies (courtesy of Michael Richmond).

Measurements of H at the present epoch, $H_0 = H(t = t_0)$, give $H_0 = 70{\pm}10$ km/sec/Mpc, which for a galaxy 1,000 Mpc distant (using present-day proper distance) would represent an apparent Hubble velocity of $V_H = 70,000$ km/sec, or $V_H/c \sim 0.2$.

If the proper time of an observer riding in this galaxy, τ, is used as the parameter along its trajectory, then (as usual)

$$g_{\mu\nu}\left(\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\right) = -1\,. \tag{7.17}$$

This expression allows the time dilation of observers in the galaxy to be related to the motion just described. Specialized to the radial motion $\ell = \ell(t)$ this last equation reads

$$\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 - a^2\left(\frac{\mathrm{d}\ell}{\mathrm{d}\tau}\right)^2 = \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2\left[1 - V_{\rm pec}^2\right] = 1\,, \tag{7.18}$$

and so the local time dilation is

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \gamma_{\rm pec} = \frac{1}{\sqrt{1 - V_{\rm pec}^2}}\,. \tag{7.19}$$

We see that there is no time dilation in the absence of peculiar motion, so t describes the proper time for all observers who sit at fixed coordinate positions. In the presence of proper motion a time dilation arises, given by the usual special relativistic expression in terms of the peculiar velocity, $V_{\rm pec}$.

Light Rays and Redshift

The trajectories of particles (like photons) moving at the speed of light similarly satisfy

$$g_{\mu\nu}\left(\frac{\mathrm{d}x^\nu}{\mathrm{d}s}\right)\left(\frac{\mathrm{d}x^\nu}{\mathrm{d}s}\right) = 0\,, \tag{7.20}$$

which for radial motion specializes to

$$\frac{\mathrm{d}t}{\mathrm{d}s} = \pm\, a\,\frac{\mathrm{d}\ell}{\mathrm{d}s}\,. \tag{7.21}$$

Consider now a photon which is sent to us (at the origin) along a radial trajectory from a galaxy which is situated at fixed coordinate position $\ell = L$. If we suppose the photon to arrive at our position at $t = 0$ then we may compute its departure time at the emitting galaxy, $t = -T$. Explicitly, the *look-back time*, T, is given by eq. (7.21) to be

$$L = \int_0^T \frac{\mathrm{d}t}{a(t)}\,. \tag{7.22}$$

Imagine now repeating this calculation for a sequence of photons (or for a train of wave crests) which are emitted from the galaxy and are received here. Suppose two consecutive photons are emitted at events which are labelled by the coordinate positions $(-T, L, \theta_0, \phi_0)$ and $(-T+\delta T, L+\delta L, \theta_0, \phi_0)$, with the first of these received at the origin at time $t = 0$ and the second arriving at $(\delta t, \delta\ell, \theta_0, \phi_0)$. The redshift of such a wave train may be found by computing how δt depends on δT, the scale factor, $a(t)$, and the peculiar motions of the emitter and observer.

We know that the trajectories of both photons satisfy eq. (7.21), and so we know

$$L = \int_0^T \frac{\mathrm{d}t}{a(t)} \quad \text{and} \quad (L+\delta L) - \delta\ell = \int_{-\delta t}^{T-\delta T} \frac{\mathrm{d}t}{a(t)}\,. \tag{7.23}$$

Subtracting the first of these from the second, and expanding the result to first order in the small quantities δt, δT δL leads to the following relation

$$\delta L - \delta\ell = \int_{-\delta t}^{T-\delta T} \frac{\mathrm{d}t}{a(t)} - \int_0^T \frac{\mathrm{d}t}{a(t)} \approx \frac{\delta t}{a_0} - \frac{\delta T}{a(T)}\,, \tag{7.24}$$

where $a_0 = a(0)$. Dividing by δT then gives

$$\frac{\delta L}{\delta T} - \frac{\delta\ell}{\delta t}\left(\frac{\delta t}{\delta T}\right) = \frac{1}{a_0}\left(\frac{\delta t}{\delta T}\right) - \frac{1}{a(T)}\,. \tag{7.25}$$

This may now be solved for $\delta t/\delta T$ as a function of a_0, $a(T)$ and the emitter and observer's peculiar velocities, $V_{\rm pec} = a(T)[\delta L/\delta T]$ and $v_{\rm pec} = a_0[\delta\ell/\delta t]$ to give

$$\frac{\delta t}{\delta T} = \frac{a_0}{a(T)}\left(\frac{1+V_{\rm pec}}{1+v_{\rm pec}}\right) . \tag{7.26}$$

The redshift, z, of the light is defined in terms of its wavelength at emission, $\lambda_{\rm em}$, and at observation, $\lambda_{\rm obs}$, by $z = (\lambda_{\rm obs} - \lambda_{\rm em})/\lambda_{\rm em}$ and so

$$\begin{aligned} 1+z = \frac{\lambda_{\rm obs}}{\lambda_{\rm em}} &= \frac{\delta\tau_{\rm obs}}{\delta\tau_{\rm em}} = \frac{\delta t}{\delta T}\left[\frac{1-v_{\rm pec}^2}{1-V_{\rm pec}^2}\right]^{1/2} \\ &= \frac{a_0}{a(T)}\left(\frac{1+V_{\rm pec}}{1+v_{\rm pec}}\right)\left[\frac{1-v_{\rm pec}^2}{1-V_{\rm pec}^2}\right]^{1/2} . \end{aligned} \tag{7.27}$$

This last expression uses eq. (7.19) to relate the proper time of the observer, $\delta\tau_{\rm obs}$, and of the emitter, $\delta\tau_{\rm em}$, to the corresponding coordinate time differences, δt and δT.

Eq. (7.27) is the main result. For negligible peculiar motions it reduces to a simple expression for the redshift due to the Hubble flow

$$1+z = \frac{a_0}{a(T)} , \tag{7.28}$$

which is a *red*shift – *i.e.* $z > 0$ – if the universe expands – *i.e.* $a_0 > a(T)$. This expression gives a good method for measuring the universe's scale factor, $a(t)$, since it shows that it is simply related to the redshift of the light received from distant galaxies.

For non-relativistic peculiar velocities this generalizes to the approximate formula

$$1+z \approx \frac{a_0}{a(T)}\Big[1 + (V_{\rm pec} - v_{\rm pec})\Big] . \tag{7.29}$$

Notice that (as expected) relative peculiar motion also generates a redshift – $z > 0$ – if $V_{\rm pec} > v_{\rm pec}$ – that is, if the emitting galaxy is receding from the observing one.

In principle, the dependence of z on peculiar velocity complicates the inference of the universal scale factor from measurements of redshift, since in principle it requires knowledge of the peculiar velocity of the distant emitting galaxy. In practice, however, this complication is only important for relatively nearby galaxies, for which the redshift due to the peculiar velocities are not dominated by that due to the universal expansion.

7.2 Distance vs redshift

In LFRW cosmology the expansion of the universe is characterized by the time dependence of the scale factor, $a(t)$, which we shall see is in most circumstances a

monotonic function of t. In principle, predictions for $a(t)$ can be tested by measuring the proper distances, $D(L, -T)$, to distant celestial objects and comparing this with the look-back time, T, to these objects. Measurements of $D(L, -T)$ vs T allow the inference of $a(t)$ because of the connection between L and T — *i.e.* the relation $L(T)$ given implicitly by eq. (7.23) — which expresses the fact that all of our observations about the distant universe lie along our past light cone, because they rely on our detecting photons which have come to us from the far reaches of space.

In practice, however, it is much easier to directly measure a than it is to measure T because of the direct relationship between a and redshift. So inferences about the geometry of spacetime instead are founded on measuring the dependence of distance on redshift, z, for distant objects, rather than on look-back time, T. z and T carry the same information provided $a(t)$ is a monotonic function of time, and so it is more convenient to use z itself as an operational measure of the universe's age and size.

The remainder of this section derives expressions for the dependence of various measures of distance on redshift, given a universal expansion history, $a(t)$.

Proper Distance

Consider, then, a galaxy which at event $(-T, L, \theta_0, \phi_0)$ sends light to us which we receive at the origin at $t = 0$. Writing $a_0 = a(0)$, the present-day proper distance to this galaxy is given by

$$D(T) = D(L(T), -T) = a_0\, L = \int_0^T \left(\frac{a_0}{a(t)}\right) \mathrm{d}t\,. \tag{7.30}$$

This may be changed into an expression in terms of redshift by changing integration variable from t to z using the relations

$$1 + z = \frac{a_0}{a(t)} \qquad \text{and so} \qquad \mathrm{d}z = -\left(\frac{a_0\,\dot{a}}{a^2}\right) \mathrm{d}t = -(1+z)\,H\,\mathrm{d}t\,, \tag{7.31}$$

where as before $H = \dot{a}/a$. This leads to the desired result

$$D(z) = \int_0^z \frac{dz'}{H(z')}\,. \tag{7.32}$$

Unfortunately, proper distance is also not particularly convenient since it is not easily obtained from observations. There are two other notions of distance which are more practical, whose dependence on z is now derived.

Luminosity Distance

One way of inferring how far away a distant object is becomes possible if the object's intrinsic rate of energy release per unit time — *i.e.* luminosity, $\mathcal{L}$ — is known. If $\mathcal{L}$ is

known then it may be compared with the observed energy flux, f, which is received at Earth from the object, with the distance to the object obtained by assuming that the flux is related to $\mathcal{L}$ only by the geometrical solid angle which the Earth subtends at the source. For instance in Euclidean space the flux received by a source of luminosity $\mathcal{L}$ situated a distance D away is given by

$$f = \frac{\mathcal{L}}{4\pi D^2}\,, \tag{7.33}$$

provided the source sends its energy equally in all directions and that there is no absorption or scattering of the light while it is *en route* from the source. The *luminosity distance*, D_L, to the object may then be defined in terms of $\mathcal{L}$ and f by $D_L = [\mathcal{L}/(4\pi f)]^{1/2}$. This is the distance measure which is used, for example, in recent measurements of the universal expansion using distant Type I supernovae.

Suppose, then, that the source emits a packet of light having energy, $\delta E_{\rm em}$, in a time, $\delta t_{\rm em}$, and so has luminosity $\mathcal{L} = \delta E_{\rm em}/\delta t_{\rm em}$. In an LFRW universe the relation between $\mathcal{L}$ and the flux, f, we observe depends differently on distance than in flat spacetime, in the following ways.

- Because the wavelength of the light is stretched by the universal expansion, and the energy of a light wave is inversely proportional to its wavelength ($E = h\nu = hc/\lambda$) this packet of energy arrives to us having a red-shifted energy $\delta E_{\rm obs} = \delta E_{\rm em}/(1+z)$.

- Because of the expansion of space the wavelength of the light stretches as space expands while it is en route. As a result the spatial extent of the packet also stretches by a factor $1+z$ during its passage between the source and us. The means that on its arrival the time taken for the packet to deliver its energy is $\delta t_{\rm obs} = \delta t_{\rm em}(1+z)$.

- The total energy from the source is sent in all directions, and so (using the LFRW metric) it is spread over a sphere having surface area $\mathcal{A} = 4\pi r^2(L)a^2$ at a proper distance $D = La$ from the source, where $r(L)$ is given by eq. (7.3).

The flux observed at Earth is therefore given by

$$\begin{aligned} f &= \frac{1}{4\pi r^2(L)\,a_0^2}\left(\frac{\delta E_{\rm obs}}{\delta t_{\rm obs}}\right) \\ &= \frac{1}{4\pi r^2(L)\,a_0^2}\left(\frac{\delta E_{\rm em}/(1+z)}{\delta t_{\rm em}(1+z)}\right) \\ &= \left(\frac{\mathcal{L}}{4\pi r^2(L)\,a_0^2}\right)\frac{1}{(1+z)^2}\,, \end{aligned} \tag{7.34}$$

and so the luminosity distance becomes

$$D_L(z) \equiv \left[\frac{\mathcal{L}}{4\pi f}\right]^{1/2} = a_0\, r(L(z))\,(1+z)\,. \tag{7.35}$$

Notice that the present-day proper distance to the same galaxy would be $D = L\,a_0$. Since in the special case of a spatially-flat universe, $\kappa = 0$, we have $r(\ell) = \ell$, in this case D_L is related to this proper distance by

$$D_L(z) = D(z)\,(1+z) \qquad (\text{if } \kappa = 0)\,. \tag{7.36}$$

Angular-Diameter Distance

A second measure of distance becomes possible if an object of known proper length is observed at a distance, since the angle which the object subtends as seen from Earth is geometrically related to its distance from us. In Euclidean geometry an object of length $\mathrm{d}s$ placed a distance $D \gg \mathrm{d}s$ from us subtends an angle

$$\mathrm{d}\theta = \frac{\mathrm{d}s}{D}\ (\text{radians})\,, \tag{7.37}$$

which motivates defining the angular-diameter distance by $D_A = \mathrm{d}s/\mathrm{d}\theta$ in terms of the (assumed) known length $\mathrm{d}s$ and measured angle $\mathrm{d}\theta$. This notion of distance comes up in the study of the temperature fluctuations of the cosmic microwave background radiation (about which more will be said later).

The connection between $\mathrm{d}s$ and $\mathrm{d}\theta$ differs in the LFRW geometry in the following ways.

- At any given time, within an LFRW geometry the proper length of an object which subtends an angle $\mathrm{d}\theta$ when placed a proper distance $D = a\,\ell$ away is given by $\mathrm{d}s = a\,r(\ell)\,\mathrm{d}\theta$, with $r(\ell)$ given by eq. (7.3).

- When an object is observed from a great distance it is the proper distance at the time its light was emitted which appears in the previous argument. Due to the overall expansion of space this corresponds to a proper distance at present which is a factor $a_0/a(-T) = 1+z$ larger.

With these two effects in mind, the angle subtended by an object having proper length $\mathrm{d}s$ when observed from a present-day proper distance $D = a_0 L$ away is given by

$$\mathrm{d}\theta = \frac{\mathrm{d}s}{a(-T)\,r(L)} = \frac{\mathrm{d}s}{a_0\,r(L)/(1+z)}\,, \tag{7.38}$$

and so the angular-diameter distance of such an object is

$$D_A(z) \equiv \frac{\mathrm{d}s}{\mathrm{d}\theta} = \frac{a_0\,r(L(z))}{1+z} = \frac{D_L(z)}{(1+z)^2}\,, \tag{7.39}$$

where the last equality uses eq. (7.35).

Notice that in the special case of a spatially-flat universe ($\kappa = 0$), we have $r(\ell) = \ell$ and so the angular-diameter distance to an object situated a proper distance $D = a_0 L$ away is

$$D_A(z) = \frac{D(z)}{1+z} \qquad (\text{if } \kappa = 0)\,. \tag{7.40}$$

This is equivalent to the object's proper distance as measured at the time of the light's emission rather than its present proper distance.

Exercise 31: Measurements of the total number, N, of distant objects as a function of their redshift, z, provide another way to measure $a(t)$. Show that if the objects in question have a density $n(t)$, then

$$\begin{aligned} \mathrm{d}N &= 4\pi n(t) a^3(t) r^2(\ell(t))\,\mathrm{d}\ell \\ &= 4\pi n(t) a^2(t) r^2(\ell(t))\,\mathrm{d}t \\ &= 4\pi n[t(z)] a_0^2 r^2[\ell(t(z))]\,\frac{\mathrm{d}z}{(1+z)^3 H(z)}\,. \end{aligned} \tag{7.41}$$

The Recent Universe

For later purposes it is useful to evaluate the above distance-redshift expressions for various choices for the time-dependence of the universal expansion, $a(t)$. For simplicity (and because this appears to be a good description of the present-day universe) in the case of D_L and D_A we provide formulae for the special case $\kappa = 0$.

A great many cosmological observations are restricted to the comparatively nearby universe, for which the observed red-shifts are small. For such small red-shifts it is useful to evaluate the distance-redshift expressions by expanding about the present epoch, for which $z = 0$. Consider, therefore, a scale factor of the form

$$a(t) = a_0 + \dot{a}_0\,(t - t_0) + \frac{1}{2}\,\ddot{a}_0\,(t - t_0)^2 + \cdots\,, \tag{7.42}$$

where $t = t_0$ denotes the present epoch. In what follows it is convenient to measure the time difference in units of H_0^{-1}, where $H_0 = \dot{a}_0/a_0$ by defining $\zeta = -H_0\,(t - t_0)$, in which case the above expansion is expected to furnish a good approximation for $|\zeta| \lesssim 1$. (Notice that as defined $\zeta \geq 0$ when applied to $a(t)$ in the past universe, for which $t \leq t_0$.)

In terms of this expansion the redshift of light becomes

$$1 + z = \frac{a_0}{a(t)} = 1 + \zeta + \left(1 + \frac{q_0}{2}\right)\,\zeta^2 + \cdots\,, \tag{7.43}$$

where $q_0 \equiv -a_0\ddot{a}_0/\dot{a}_0^2 = -\ddot{a}/(a_0\,H_0^2)$, with the sign chosen so that $q_0 > 0$ for a decelerating universe (for which $\ddot{a}_0 < 0$).

The distance-redshift relations are governed by $H(z)$, which is given by

$$\begin{aligned} H &= H_0\left[1 + \left(\frac{\ddot{a}_0}{\dot{a}_0} - \frac{\dot{a}_0}{a_0}\right)(t - t_0) + \cdots\right] \\ &= H_0\Big[1 + (1 + q_0)\,z + \cdots\Big]\,. \end{aligned} \tag{7.44}$$

Using this in eq. (7.32) leads to the following expression for $D(z)$ near $z = 0$

$$D(z) = H_0^{-1}\left[z - \frac{1}{2}\,(1 + q_0)\,z^2 + \cdots\right]\,, \tag{7.45}$$

which for $\kappa = 0$ also imply the following small-z expansions for the luminosity and angular-diameter distances

$$\begin{aligned} D_L(z) &= H_0^{-1}\left[z + \frac{1}{2}\,(1 - q_0)\,z^2 + \cdots\right] \\ D_A(z) &= H_0^{-1}\left[z - \frac{1}{2}\,(3 + q_0)\,z^2 + \cdots\right]\,. \end{aligned} \tag{7.46}$$

For small z the leading distance-redshift dependence is therefore predicted to be linear — $D(z) \simeq H_0^{-1}z$ — for all of the distance definitions given above, a result which expresses Hubble's Law in the form observers really test it (such as in fig. 25). It is the measurement of this slope, such as by the Hubble Key Project [7], that lead to the current best value $H_0 = 72 \pm 8$ km/sec/Mpc.

Clearly a precise determination of distance vs redshift for objects out to larger redshifts permits the extraction of the deceleration parameter (q_0) in addition to both the present-day Hubble constant (H_0). This has proven quite difficult to do reliably, but has recently been accomplished (see fig. 27) using the luminosity-distance vs redshift relation measured for Type IA supernovae, which are bright enough to be seen at enormous distances but for which the intrinsic luminosity is known. It is these measurements that discovered that the universal expansion is *accelerating* – that is, $q_0 < 0$ so $\ddot{a}_0 > 0$.

Power-Law Expansion

Another situation of considerable practical interest is the case where the expansion varies as a power of t, as in

$$1 + z = \frac{a_0}{a(t)} = \left(\frac{t_0}{t}\right)^{\alpha}\,, \tag{7.47}$$

for some choices for the parameters a_0, t_0 and α. In later sections we shall find this law is produced (if $\kappa = 0$) with $\alpha = \frac{1}{2}$ for a universe full of radiation, and with $\alpha = \frac{2}{3}$

for a universe consisting dominantly of non-relativistic matter (like atoms or stars). For such a universe the Hubble and deceleration parameters become

$$H(t) = \frac{\dot{a}}{a} = \frac{\alpha}{t} = H_0 \left(\frac{t_0}{t}\right) = H_0\,(1+z)^{1/\alpha} \qquad \text{and} \qquad q(t) = -\frac{a\,\ddot{a}}{\dot{a}^2} = \frac{1-\alpha}{\alpha}\,. \tag{7.48}$$

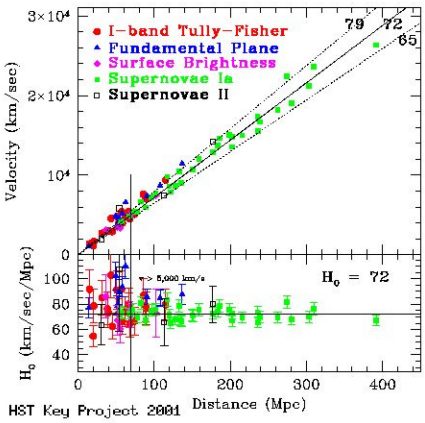


Figure 26: Measurements of the present-day Hubble scale, H_0, as obtained from distance-redshift measurements by the Hubble Key Project.

Notice that this kind of power law implies that a vanishes for $t = 0$ provided only that $\alpha > 0$ (and so in particular does so for the cases $\alpha = \frac{1}{2}$ and $\frac{2}{3}$ mentioned above). This is the Big Bang which underlies much of modern cosmology. In terms of $q = q_0$ and the present value of the Hubble parameter, H_0, this occurs a time

$$t_0 = \alpha H_0^{-1} = \frac{H_0^{-1}}{q_0 + 1} \tag{7.49}$$

in the past.

Using the above expressions for q and $H(z)$ in eq. (7.32) gives the following expression for the proper distance

$$D(z) = H_0^{-1} \int_0^z \frac{\mathrm{d}z'}{(1+z')^{1/\alpha}} = \frac{H_0^{-1}}{q}\left[1 - \frac{1}{(1+z)^q}\right], \tag{7.50}$$

which with $D_L(z) = D(z)\,(1+z)$ and $D_A(z) = D(z)/(1+z)$ give the luminosity and angular-diameter distances when $\kappa = 0$.

Radiation-Dominated Universe (if $\kappa = 0$):

As mentioned above, the special case where the universe is dominated by radiation with $\kappa = 0$ turns out to correspond to a power-law expansion with $\alpha = \frac{1}{2}$, and so we have $H(z) = H_0(1+z)^2$ and $q(z) = q_0 = 1$. This leads to the following proper distance

$$D(z) = H_0^{-1}\left(\frac{z}{1+z}\right) = \begin{cases} H_0^{-1}\left[z - z^2 + \cdots\right] & \text{if} \quad z \ll 1 \\ H_0^{-1}\left[1 - \frac{1}{z} + \cdots\right] & \text{if} \quad z \gg 1 \end{cases}. \tag{7.51}$$

Since $\kappa = 0$ the luminosity and angular-diameter distances become

$$D_L(z) = H_0^{-1}\,z\,, \qquad D_A(z) = H_0^{-1}\left[\frac{z}{(1+z)^2}\right] = \begin{cases} H_0^{-1}\left[z - 2\,z^2 + \cdots\right] & \text{if} \quad z \ll 1 \\ \frac{H_0^{-1}}{z}\left[1 - \frac{2}{z} + \cdots\right] & \text{if} \quad z \gg 1 \end{cases}. \tag{7.52}$$

Matter-Dominated Universe (if $\kappa = 0$):

The special case where $\kappa = 0$ and the universe is dominated by non-relativistic matter corresponds to power-law expansion with $\alpha = \frac{2}{3}$, and so $H(z) = H_0(1+z)^{3/2}$ and $q(z) = q_0 = \frac{1}{2}$. This leads to the proper distance

$$D(z) = 2\,H_0^{-1}\left[\frac{(1+z)^{1/2}-1}{(1+z)^{1/2}}\right] = \begin{cases} H_0^{-1}\left[z - \frac{3}{4}\,z^2 + \cdots\right] & \text{if} \quad z \ll 1 \\ 2H_0^{-1}\left[1 - \left(\frac{1}{z}\right)^{1/2} + \cdots\right] & \text{if} \quad z \gg 1 \end{cases} . \tag{7.53}$$

Because $\kappa = 0$ the luminosity and angular-diameter distances are

$$D_L(z) = 2H_0^{-1}\left[(1+z) - \sqrt{1+z}\right] = \begin{cases} 2H_0^{-1}\left[z + \frac{1}{4}\,z^2 + \cdots\right] & \text{if} \quad z \ll 1 \\ 2H_0^{-1}z\left[1 - \left(\frac{1}{z}\right)^{1/2} + \cdots\right] & \text{if} \quad z \gg 1 \end{cases} , \tag{7.54}$$

and

$$D_A(z) = 2\,H_0^{-1}\left[\frac{(1+z)^{1/2}-1}{(1+z)^{3/2}}\right] = \begin{cases} H_0^{-1}\left[z - \frac{7}{4}\,z^2 + \cdots\right] & \text{if} \quad z \ll 1 \\ \frac{2H_0^{-1}}{z}\left[1 - \left(\frac{1}{z}\right)^{1/2} + \cdots\right] & \text{if} \quad z \gg 1 \end{cases} . \tag{7.55}$$

Notice that for both matter- and radiation-dominated universes the present-day proper distance approaches a limiting value of order H_0^{-1} when $z \to \infty$. This implies that we do not learn about arbitrarily large distances when we look into the past at objects having larger and larger redshifts. A related observation is the fact that the angular-diameter distance is not a monotonic function of z, since it grows like z for small z but vanishes asymptotically for large z, proportional to $1/z$. Since (when $\kappa = 0$) angular-diameter distance is the proper distance to the source measured at the time the light is emitted rather than observed, this vanishing of D_A for large z shows that our observations are limited to a vanishingly small local region in the very distant past. This limitation to our view is called our local *particle horizon*. It arises because for these geometries the universe becomes vanishingly small at a finite time in our past and the universal expansion can be fast enough to permit objects to be sufficiently distant that light cannot reach us from them given the limited age of the universe.

Exponential Expansion

The next special case of interest corresponds to exponential expansion

$$1 + z = \frac{a_0}{a(t)} = \exp[-H_0\,(t - t_0)]\,, \tag{7.56}$$

which may be regarded as the limiting case of a power law for which $\alpha \to \infty$. We shall find this kind of expansion can be produced when the universal energy density is dominated by the energy of the vacuum.

In this case the Hubble and deceleration parameters are time-independent, with

$$H(t) = \frac{\dot{a}}{a} = H_0 \qquad \text{and} \qquad q(t) = q_0 = -1\,, \tag{7.57}$$

and the redshift-dependence of the proper distance is $D(z) = H_0^{-1}\int_0^z \mathrm{d}z' = H_0^{-1} z$. The luminosity and angular-diameter distances (when $\kappa = 0$) then become.

$$D_L(z) = H_0^{-1} z(1+z) \qquad \text{and} \qquad D_A(z) = H_0^{-1}\left(\frac{z}{1+z}\right)\,. \tag{7.58}$$

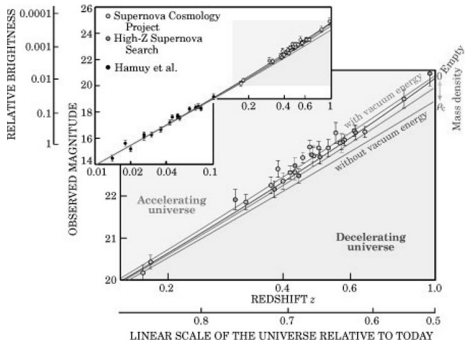


Figure 27: Measurements of the luminosity-distance/redshift relation to higher redshifts, with evidence for $q_0 < 0$, by the Supernova Cosmology Project and the High-z Supernova Search.

Notice that, unlike for the previous examples, the expansion in this case is *accelerated*, with $\ddot{a} > 0$ and so $q_0 < 0$. This kind of expansion is particularly interesting because of recent tests of Hubble's Law out to comparatively large redshifts, which indicate q_0 really is negative (see fig. 27). We shall see later that this kind of expansion can also be generated by plausible kinds of matter, and in particular would arise if the vacuum itself were to have a nonzero energy density.

Unlike the case of matter- and radiation-domination considered earlier, in this case the present-day proper distance grows without bound but the proper-distance at emission approaches a fixed limit, $D_A \to H_0^{-1}$, as $z \to \infty$. This distance represents an apparent *horizon* beyond which we are unable to penetrate with observations, and differs from the particle horizon considered above because it is not tied to there only having been a finite proper time since the universe had zero size. For the exponentially-expanding universe only a finite proper distance in the past is accessible to us even though t can run back to $-\infty$. The existence of this horizon can be traced to the enormous speed of the exponential expansion, with which light waves travelling at finite speed cannot keep up.

7.3 Dynamics of an expanding Universe

The previous sections described the kinematics of how various distance-redshift relationships depend on the universal expansion history, $a(t)$. The present section instead addresses the question of how this expansion history depends on the energy content of the matter which lives inside the universe. This connection has its roots in the Einstein field equations

$$R_{\mu\nu} - \frac{1}{2} R\, g_{\mu\nu} = 8\pi G\, T_{\mu\nu}\,, \tag{7.59}$$

which relate the curvature of spacetime to its energy-momentum content — *i.e.* "matter tells space how to curve".

Homogeneous and Isotropic Stress Energy

The conditions of homogeneity and isotropy strongly restrict the distribution of matter and energy within the universe, in the same way that they restrict the metric to take the Friedmann-Robertson-Walker form, given by eq. (7.2). For the stress-energy tensor, $T_{\mu\nu}$, the analogous conditions have the following form.

- Isotropy permits the energy density, $\rho = T^{tt}$, to be an arbitrary function of time, t, and radial position, ℓ, but homogeneity forbids any dependence on the position ℓ. The most general energy density can therefore only be time dependent: $T^{tt} = \rho(t)$.

- Isotropy permits a net energy flux, $s^i = T^{ti}$ with $i = 1, 2, 3$, so long as it points purely in the radial direction.[14] In LFRW coordinates this implies $T^{t\theta} = T^{t\phi} = 0$ while $T^{t\ell}$ can be a nonzero function of t and ℓ. Homogeneity, however, requires $T^{t\ell} = 0$ because having a nonzero energy flux would necessarily allow one to distinguish between the directions from which and to which the energy is flowing. The same conclusions equally apply to the momentum density: $\pi^i = T^{it} = 0$.

- Isotropy permits the 3-dimensional stress tensor, T^{ij}, to be nonzero provided it is built from the metric tensor itself, or from the radial direction vector, x^i. That is, isotropy allows $T^{ij} = p\, g^{ij} + q\, x^i x^j$, where p and q can be functions of both t and ℓ. However homogeneity precludes p from depending on ℓ, and does not permit a nonzero q at all, since the radial vector picks out a preferred place as its origin. It follows that the stress tensor must have the diagonal form $T^j_i = g_{ik} T^{kj} = p(t)\, \delta^j_i$.

We are led to the conclusion that homogeneity and isotropy only permit a stress-energy of the form

$$T^{tt} = \rho(t)\,, \quad T^{ti} = T^{it} = 0 \qquad \text{and} \qquad T^{ij} = p(t)\, g^{ij}\,, \tag{7.60}$$

which is characterized by two functions of time: $\rho(t)$ and $p(t)$. As is clear from the definition of $T^{\mu\nu}$, ρ represents the (average) energy density as seen by co-moving observers who are situated at fixed values of (ℓ, θ, ϕ). As we saw in earlier sections, the interpretation of T^{ij} as a momentum flux together with stress-energy conservation

[14] This can be removed by changing the radial coordinate, but we do not do so in order not to lose the simple connection between proper distance and coordinate distance, $D = a(t)\Delta\ell$.

implies that the net rate of change in momentum of a volume V — *i.e.* the net force acting on V — is given by the flux of momentum current through the boundary, ∂V:

$$F^i \equiv \frac{\mathrm{d}P^i}{\mathrm{d}t} = \int_V \frac{\partial \pi^i}{\partial t}\,\mathrm{d}^3V = -\int_{\partial V} T^{ij} n_j\,\mathrm{d}^2S = -\int_{\partial V} p\,n^i\,\mathrm{d}^2S\,, \tag{7.61}$$

which shows that p represents the total (average) *pressure* of the matter whose stress energy is under consideration.

Our goal now is to see how Einstein's equations relate these quantities to $a(t)$.

Einstein's Equations

In order to determine how $a(t)$ is connected to $\rho(t)$ and $p(t)$ we require the Ricci tensor for the LFRW metric, eq. (7.2). It is convenient to write the metric in terms of the time coordinate, t, and the space coordinates, $x^i = \{r, \theta, \phi\}$, as

$$g_{tt} = -1\,, \quad g_{ti} = 0 \quad \text{and} \quad g_{ij} = a^2(t)\,\hat{g}_{ij}\,, \tag{7.62}$$

where $\hat{g}_{ij}\,\mathrm{d}x^i\mathrm{d}x^j = \mathrm{d}r^2/(1-\kappa r^2/r_0^2) + r^2\,(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$ denotes the spatial metric with the scale factor, $a(t)$, removed. In terms of these, the only nonzero Christoffel symbols are

$$\Gamma^t_{ij} = a\dot{a}\,\hat{g}_{ij}\,, \quad \Gamma^i_{tj} = \Gamma^i_{jt} = \frac{\dot{a}}{a}\,\delta^i_j \quad \text{and} \quad \Gamma^i_{jk} = \hat{\Gamma}^i_{jk}\,, \tag{7.63}$$

where $\hat{\Gamma}^i_{jk}$ denotes the Christoffel symbols built from the spatial metric, $\hat{g}_{ij}$. The components of the Ricci tensor are similarly given by

$$R_{tt} = -\frac{3\,\ddot{a}}{a}\,, \quad R_{ti} = 0 \quad \text{and} \quad R_{ij} = \hat{R}_{ij} + \left(a\ddot{a} + 2\,\dot{a}^2\right)\hat{g}_{ij}\,, \tag{7.64}$$

where the Ricci tensor for the spatial metric is

$$\hat{R}_{ij} = \frac{2\kappa}{r_0^2}\,\hat{g}_{ij}\,. \tag{7.65}$$

In the same basis the components of the stress energy are

$$T_{tt} = \rho\,, \quad T_{ti} = 0 \quad \text{and} \quad T_{ij} = p\,g_{ij} = p\,a^2\,\hat{g}_{ij}\,, \tag{7.66}$$

and so specializing the Einstein field equations, eq. (4.4), to homogeneous and isotropic geometries leads to the following two independent differential equations which relate $a(t)$ to $\rho(t)$ and $p(t)$:

$$\begin{aligned} 3\left(\frac{\ddot{a}}{a}\right) &= -4\pi G\,(\rho + 3p) \\ \frac{\ddot{a}}{a} + 2\left(\frac{\dot{a}}{a}\right)^2 + \frac{2\kappa}{a^2 r_0^2} &= 4\pi G\,(\rho - p)\,. \end{aligned} \tag{7.67}$$

In particular, a particularly useful combination of these may be chosen for which $\ddot{a}$ is eliminated, and is called the *Friedmann equation*,

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{\kappa}{a^2 r_0^2} = \frac{8\pi G}{3}\,\rho\,. \tag{7.68}$$

Rather than directly using an equation involving second derivatives as our second equation it is more convenient to instead use the equation describing the *Conservation of Stress-Energy* in curved space, eq. (4.6):

$$\nabla_\mu T^{\mu\nu} = \partial_\mu T^{\mu\nu} + \Gamma^\mu_{\mu\alpha} T^{\alpha\nu} + \Gamma^\nu_{\mu\alpha} T^{\mu\alpha} = 0\,. \tag{7.69}$$

Once specialized to the stress energy and connection given above, eqs. (7.63) and (7.66), the $\nu = i$ components of this equation vanish for any ρ or p (because of the assumed homogeneity and isotropy). But the $\nu = t$ component of this equation carries some content:

$$\begin{aligned} 0 &= \frac{\partial T^{tt}}{\partial t} + \Gamma^i_{it} T^{tt} + \Gamma^t_{ij} T^{ij} \\ &= \dot{\rho} + 3\left(\frac{\dot{a}}{a}\right)(\rho + p)\,. \end{aligned} \tag{7.70}$$

The physical meaning of this last equation as energy conservation is more easily seen if it is rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\rho\, a^3\Big) + p\,\frac{\mathrm{d}}{\mathrm{d}t}(a^3) = 0\,, \tag{7.71}$$

since in this form it relates the rate of change of the total energy, ρa^3, to the work done by the pressure as the universe expands. For matter in thermal equilibrium, a comparison of this last equation with the 1st Law of Thermodynamics shows that the expansion of the universe is adiabatic, inasmuch as the total entropy of the matter in the universe does not change in a homogeneous and isotropic expansion.

Cosmic Acceleration and Matter

In what follows we use the easier-to-use first-order Friedmann and energy-conservation equations, eqs. (7.68) and (7.70), rather than the original second-order equations, eq. (7.67), that directly arise from the Einstein equations.

To see that these are equivalent it is instructive to rederive the second-order equations, eqs. (7.67), from eqs. (7.68) and (7.70). To this end differentiate eq. (7.68) and use eq. (7.70) to eliminate $\dot{\rho}$. This gives (if $\dot{a} \neq 0$) the first of eqs. (7.67):

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\,(\rho + 3p)\,. \tag{7.72}$$

Notice that this last equation implies that $\ddot{a} < 0$ for most forms of matter, since for these ρ and p are typically positive. This corresponds physically to the statement

that gravity is always attractive, and so the mutual attraction of the galaxies in the universe always acts to slow down the universal expansion. As we shall see there can be exceptions to this general rule, for which $\rho + 3p < 0$, and so whose presence could cause the universal expansion to accelerate rather than decelerate.

Another application of eq. (7.72) is to use it to see what may be learned about the present-day values of ρ and p from measurements of the present-day expansion rate, H_0, and deceleration parameter, q_0. To this end notice that the Friedmann equation evaluated at the present epoch implies

$$H_0^2 + \frac{\kappa}{(a_0 r_0)^2} = \frac{8\pi G}{3}\rho_0 \quad \text{or} \quad 1 + \frac{\kappa}{(a_0 H_0 r_0)^2} = \frac{\rho_0}{\rho_c} \equiv \Omega_0 \,, \tag{7.73}$$

where the *critical density* is defined by $\rho_c \equiv 3H_0^2/(8\pi G)$ and the last equality defines Ω_0 to be the energy density in units of this critical density, $\Omega_0 = \rho_0/\rho_c$. Given the current measurement $H_0 = 70 \pm 10$ km/sec/Mpc, the critical density's numerical value becomes $\rho_c = 5200 \pm 1000$ MeV m^{-3} $= (9 \pm 2) \times 10^{-30}$ g cm^{-3}.

ρ_c is defined in the way it is because if $\rho_0 = \rho_c$ then $\kappa = 0$. Similarly if $\kappa = +1$ then we must have $\rho_0 > \rho_c$ and if $\kappa = -1$ then $\rho_0 < \rho_c$. Evaluating the acceleration equation, eq. (7.72), at the present epoch similarly gives

$$q_0 = -\frac{\ddot{a}_0}{a_0\, H_0^2} = \frac{4\pi G}{3H_0^2}(\rho_0 + 3p_0) = \frac{\rho_0 + 3p_0}{2\rho_c} = \frac{\Omega_0}{2}(1 + 3w_0)\,, \tag{7.74}$$

where we define $w_0 = p_0/\rho_0$. Clearly a measurement of H_0 and q_0 allows the inference of both ρ_0 and p_0, and knowledge of ρ_0 also allows the determination of κ, since $\kappa = +1$ if and only if $\Omega_0 > 1$ and $q_0 > \frac{1}{2}$ while $\kappa = -1$ requires both $\Omega_0 < 1$ and $q_0 < \frac{1}{2}$. In particular, distance-redshift measurements that indicate $q_0 < 0$ also imply $w_0 < -\frac{1}{3}$ (given that $\rho_0 \simeq \rho_c > 0$).

Equations of State

Mathematically speaking, finding the evolution of the universe as a function of time requires the integration of eqs. (7.68) and (7.70), but in themselves these two equations are inadequate to determine the evolution of the three unknown functions, $a(t)$, $\rho(t)$ and $p(t)$. Another condition is required in order to make the problem well-posed.

The missing condition is furnished by the equation of state for the matter in question, which for the present purposes may be regarded as being an expression for the pressure as a function of energy density, $p = p(\rho)$. As we shall see this expression is typically characteristic of the microscopic constituents of the matter whose stress energy is of interest. Such an equation of state naturally arises for matter which is in local thermodynamic equilibrium, since this often allows both p and ρ to be expressed in terms of a single quantity like the local temperature, T. But it may also

arise for matter which was only in equilibrium in the past, even if it is no longer in equilibrium at present.

Most of the equations of state of interest in cosmology have the general form

$$p = w\,\rho\,, \tag{7.75}$$

where w is a t-independent constant. Given an equation of state of this form it is possible to integrate eqs. (7.68) and (7.70) to determine how a, ρ and p vary with time, as we now see.

The first step is to determine how p and ρ depend on a, since this is dictated by energy conservation. Using eq. (7.75) to eliminate p allows eq. (7.70) to be written

$$\frac{\dot{\rho}}{\rho} + 3(1+w)\,\frac{\dot{a}}{a} = 0\,, \tag{7.76}$$

which may be integrated to obtain

$$\rho = \rho_0 \left(\frac{a_0}{a}\right)^{\sigma} \qquad \text{with} \quad \sigma = 3(1+w)\,. \tag{7.77}$$

The pressure satisfies an identical dependence on a by virtue of the equation of state: $p = w\rho$.

If eq. (7.77) is now used to eliminate ρ from eq. (7.68), the following differential equation for $a(t)$ is obtained

$$\dot{a}^2 = \frac{8\pi G \rho_0 a_0^2}{3}\left(\frac{a_0}{a}\right)^{\sigma-2} - \frac{\kappa}{r_0^2}\,, \tag{7.78}$$

In the special case that $\kappa = 0$ this equation is easily integrated to give

$$a(t) = a_0 \left(\frac{t}{t_0}\right)^{\alpha} \qquad \text{with} \quad \alpha = \frac{2}{\sigma} = \frac{2}{3(1+w)}\,. \tag{7.79}$$

We now apply the above expressions to a few examples of the equations of state which are known to be relevant to cosmology.

Empty Space

The simplest cosmology possible is obtained in the absence of matter, in which case $\rho = p = 0$. In this case we have $\dot{a}^2 = -\kappa$, from which we see that $\kappa \neq +1$. Two distinct solutions are possible, depending on whether $\kappa = 0$ or $\kappa = -1$.

If $\kappa = 0$ we have $\dot{a} = 0$ and so we may choose $a = 1$ for all t. In this case the LFRW metric simply reduces to the flat metric of Minkowski space, written in polar coordinates.

If $\kappa = -1$ then we have $\dot{a} = \pm 1$ and so $a = \pm(t - t_0) + a_0$. This negatively-curved geometry is known as the *Milne Universe*, but so far as we know it does not play any role in Big Bang cosmology.

Radiation

A gas of relativistic particles, like photons or neutrinos (or other particles for sufficiently high temperatures), when in thermal equilibrium has an energy density and pressure given by

$$\rho = a_B\, T^4 \qquad \text{and} \qquad p = \frac{1}{3}\, a_B\, T^4\,, \tag{7.80}$$

where $a_B = \pi^2/15 = 0.6580$ is the Stefan-Boltzmann constant (in units where $k_B = c = \hbar = 1$) and T is the temperature. These two expressions ensure that ρ and p satisfy the relation

$$p = \frac{1}{3}\,\rho \qquad \text{and so} \qquad w = \frac{1}{3}\,. \tag{7.81}$$

Since $w = \frac{1}{3}$ we see that $\sigma = 3(1+w) = 4$ and so $\rho \propto a^{-4}$. This has a simple physical interpretation for a gas of noninteracting photons, since for these the total number of photons is fixed (and so $n_\gamma \propto a^{-3}$), but each photon energy also redshifts like $1/a$ as the universe expands, leading to $\rho_\gamma \propto a^{-4}$.

Since $\sigma = 4$ we have $\alpha = 2/\sigma = 1/2$, and so if $\kappa = 0$ then $a(t) \propto t^{1/2}$. Explicit expressions are given in previous sections for the proper, luminosity and angular-diameter distance as functions of redshift for this type of expansion.

Non-relativistic Matter

An ideal gas of non-relativistic particles in thermal equilibrium has a pressure and energy density given by[15]

$$p = n\,T \qquad \text{and} \qquad \rho = n\,m + \frac{n\,T}{\gamma - 1}\,, \tag{7.82}$$

where n is the number of particles per unit volume, m is the particle's rest mass and $\gamma = c_p/c_v$ is its ratio of specific heats, with $\gamma = \frac{5}{3}$ for a gas of monatomic atoms.

For non-relativistic particles the total number of particles is usually also conserved, which implies that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big[n\,a^3\Big] = 0\,. \tag{7.83}$$

Since $m \gg T$ (or else the atoms would be relativistic) the equation of state for this gas may be taken to be

$$p/\rho \approx 0 \qquad \text{and so} \qquad w \approx 0\,. \tag{7.84}$$

Notice that although this equation of state is derived for a thermal gas, it applies much more generally, such as for the cosmic fluid of galaxies, or for other forms of non-relativistic matter that are not in thermal equilibrium. This because for all such systems the pressure is suppressed relative to the energy density by factors of v/c.

If $w = 0$ then energy conservation implies $\sigma = 3(1 + w) = 3$ and so ρa^3 is a constant. This is appropriate for non-relativistic matter for which the energy density is dominated by the particle rest-masses, $\rho \approx n\,m$, because in this case energy conservation is equivalent to conservation of particle number, which we've seen is equivalent to $n \propto a^{-3}$ (since this leaves the total number of particles, $N \sim n\,a^3$, fixed).

Given that $\sigma = 3$ we have $\alpha = 2/\sigma = \frac{2}{3}$ and so if $\kappa = 0$ then the universal scale factor expands like $a \propto t^{2/3}$. Explicit expressions for the proper, luminosity and angular-diameter distances for this type of expansion are all given in earlier sections.

Nonrelativistic Solutions for General κ:

When $\sigma = 3$ it is also possible to solve eq. (7.68) analytically even when $\kappa \neq 0$. We pause here to display these solutions in some detail because most of the history of the universe from $z \sim 10^4$ down to $z \sim 1$ appears to have been governed by a universe whose energy density was dominated by non-relativistic matter.

As was described in earlier sections, we may expect the solutions for general κ to be described by two integration constants, which we may take to be Ω_0 and H_0, or equivalently to be $q_0 = \Omega_0/2$ and H_0. The value of κ is related to these parameters

[15] Units are again used for which Boltzmann's constant is unity: $k_B = 1$.

because $\Omega_0 = 2q_0 = 1$ if and only if $\kappa = 0$, and $\kappa = +1$ if $\Omega_0 > 1$ and $\kappa = -1$ if $\Omega < 1$.

For $\kappa = +1$ (and so $\rho_0 > \rho_c$) the solution for $a(t)$ is most compactly given in parametric form, as the formula for a cycloid:

$$\begin{aligned} \frac{a(\zeta)}{a_0} &= \frac{q_0}{2q_0 - 1}\Big(1 - \cos\zeta\Big) = \frac{1}{2}\left(\frac{\Omega_0}{\Omega_0 - 1}\right)\Big(1 - \cos\zeta\Big) \\ H_0\, t(\zeta) &= \frac{q_0}{(2q_0 - 1)^{3/2}}\Big(\zeta - \sin\zeta\Big) = \frac{\Omega_0}{2(\Omega_0 - 1)^{3/2}}\Big(\zeta - \sin\zeta\Big)\,. \end{aligned} \tag{7.85}$$

Here the initial conditions which parameterize this solution are given in terms of the physically measurable parameters, $q_0 = \Omega_0/2$ and H_0.

As ζ increases from 0 to 2π, t increases monotonically from an initial value of 0 to $t_{\rm end} = \pi\Omega_0 H_0^{-1}/(\Omega_0 - 1)^{3/2}$, but a/a_0 rises from 0 at $t = 0$ to a maximum value, $\Omega_0/(\Omega_0 - 1)$ when $t = t_{\rm max} = t_{\rm end}/2$. After this point a/a_0 decreases monotonically until it again vanishes at $t = t_{\rm end}$. This describes a universe which begins in a Big Bang at $t = 0$, stops expanding at $t = t_{\rm max}$ and then finally recollapses and ends in a Big Crunch at $t = t_{\rm end}$.

For $\kappa = -1$ (and so $\Omega_0 < 1$ and $q_0 < \frac{1}{2}$) the solution for $a(t)$ is given by a very similar expression

$$\begin{aligned} \frac{a(\zeta)}{a_0} &= \frac{q_0}{1 - 2q_0}\Big(\cosh\zeta - 1\Big) = \frac{1}{2}\left(\frac{\Omega_0}{1 - \Omega_0}\right)\Big(\cosh\zeta - 1\Big) \\ H_0\, t(\zeta) &= \frac{q_0}{(1 - 2q_0)^{3/2}}\Big(\sinh\zeta - \zeta\Big) = \frac{\Omega_0}{2(1 - \Omega_0)^{3/2}}\Big(\sinh\zeta - \zeta\Big)\,. \end{aligned} \tag{7.86}$$

This time both t and a increase monotonically with ζ, whose range runs from 0 to infinity. In this case the universe begins in a Big Bang at $t = 0$ and then continues expanding (and cooling) forever, leading to a Big Chill in the remote future.

The Vacuum

If the vacuum is Lorentz invariant, as the success of special relativity seems to indicate, then its stress energy must satisfy $T_{\mu\nu} = \rho\, g_{\mu\nu}$. This implies the vacuum pressure must satisfy the only possible Lorentz-invariant equation of state:

$$p = -\rho \qquad \text{and so} \qquad w = -1\,. \tag{7.87}$$

Clearly either p or ρ must be negative with this equation of state, and unlike for other equations of state there is no reason of principle for choosing either sign for ρ *a priori*.

Because $w = -1$ when the vacuum energy is dominant, we see that $\sigma = 3(1 + w) = 0$ and so energy conservation implies that ρ is a constant, independent of a

or t. This kind of constant energy density is often called, for historical reasons, the *cosmological constant*.

In this situation $\alpha = 2/\sigma \to \infty$, which shows that the power-law solutions, $a \propto t^\alpha$, are not appropriate. Returning directly to the Friedmann equation, eq. (7.68), shows that if $\kappa = 0$ then $\dot{a} \propto \pm a$ and so the solutions are given by exponentials: $a \propto \exp[\pm H_0(t - t_0)]$. Explicit expressions for the proper, luminosity and angular-diameter distances as functions of z are given for this expansion in earlier sections.

Notice also that in this case $\rho + 3p = -2\rho$, which is negative if ρ is positive. As such this furnishes an explicit example of an equation of state for which the universal acceleration, $\ddot{a}/a = -\frac{4}{3}\pi G(\rho + 3p) = +\frac{8}{3}\pi G\rho$, can be positive if $\rho > 0$.

If all lengths are expanding, how can one tell?

We round out this section by taking a breather to address a basic conceptual question concerning the expanding universe. Since it is spacetime itself that is expanding, this question asks, how it is possible to measure the expansion of the universe if all of one's rulers are also expanding?

In a nutshell, the key to this puzzle is that time, t, is not expanding, and so energies that are defined relative to this time do not change as the universal length scale expands. For example, we have seen above that the rest masses of nonrelativistic particles do not change as the universe expands, and this is related to why the energy density of such particles fall with the universal expansion proportional to $1/a^3$. Because energies and masses do not change, neither do the sizes of bound states like atoms, and so small everyday objects do *not* grow along with the universal expansion.

To see this in more detail, imagine solving the Schrödinger equation for the ground state of the Hydrogen atom in a universe that is expanding, but doing so at a rate that is much smaller than any atomic frequencies.[16] In terms of rectangular co-moving coordinates, $\mathbf{x}$, physical (proper) distances, $\mathbf{y}$, are measured by including the (slowly varying) time-dependent scale factor, $\mathbf{y} = a\mathbf{x}$, where $a = a(t)$. In terms of these the Schrödinger equation is

$$-\frac{\hbar^2}{2m_e}\nabla_y^2\psi - \frac{\alpha}{|\mathbf{y}|}\,\psi = E\psi\,, \tag{7.88}$$

where $\alpha = e^2/4\pi$ is the electromagnetic fine-structure constant, m_e is the electron mass, $r^2 = \mathbf{y}^2 = a^2\mathbf{x}^2$ and so $\nabla_y^2 = a^{-2}\nabla_x^2$ relates the Laplace operators for the coordinates $\mathbf{y}$ and $\mathbf{x}$ respectively.

[16] This is an *extremely* good approximation, since the present-day Hubble scale, H_0, is roughly 10^{-34} times smaller than the frequency associated with the 13.6 eV binding energy of the Hydrogen atom.

Following the usual steps leads to ground-state wave functions of the form $\psi \propto \exp(-r/a_0)$, with the Bohr radius given by $a_0 = 1/(\alpha m_e)$ and the energy $E_0 = -\frac{1}{2}\,\alpha^2 m_e$. This shows that the atom's physical size, a_0, measured using the nominally expanding physical coordinates, $\mathbf{y}$, is fixed by the time-independent constants m_e and α. This is in contrast with the time-dependent separation between galaxies in the LFRW metric, which are situated at fixed values of $\mathbf{x}$ (because these are geodesics), and so separate as a gets larger.

But how do we see that it is the scale H that is the relevant comparison when deciding which bound systems do not expand with the universal expansion? And what about bound states where it is gravity itself that is doing the binding? Do the Schwarzschild radii of stars increase as the universe expands? These questions can be explicitly answered using an exact solution to Einstein's equations that describes a gravitating object (like a black hole) sitting within an expanding LFRW cosmology. The solution in question is called the *McVittie* solution [8], and for spatially flat cosmologies ($\kappa = 0$) has the form

$$\mathrm{d}s^2 = -\left(\frac{1-\mu}{1+\mu}\right)^2 \mathrm{d}t^2 + (1+\mu)^4\, a^2\left[\mathrm{d}\varrho^2 + \varrho^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right], \tag{7.89}$$

where ϱ is the radial coordinate, the dimensionless quantity μ is defined by

$$\mu(\varrho, t) = \frac{GM}{2a(t)\,\varrho}\,, \tag{7.90}$$

and the scale factor $a(t)$ is obtained by solving the Friedmann equation, as for the LFRW metric (with $\kappa = 0$):

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3}\,. \tag{7.91}$$

Here $\rho(t)$ is the homogeneous isotropic energy density that governs the time-dependence of the cosmological environment.

The limiting LFRW and Schwarzschild behaviours are easier to see if we change coordinates, $\varrho \to r$, where r is defined so that the area of the spheres at fixed r and t are $A = 4\pi r^2$. The desired coordinate change therefore is

$$r = (1+\mu)^2\, a\,\varrho\,. \tag{7.92}$$

The metric in these new coordinates then becomes

$$\mathrm{d}s^2 = -\left(1 - \frac{r_s}{r} - H^2 r^2\right)\mathrm{d}t^2 + \frac{\mathrm{d}r^2}{1 - r_s/r} - \frac{2Hr}{\sqrt{1 - r_s/r}}\,\mathrm{d}r\,\mathrm{d}t + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right), \tag{7.93}$$

where, as usual, $r_s = 2GM$ (which is independent of time). To see that this geometry approaches an LFRW metric at large distances, use $r_s/r \ll (Hr)^2$ to neglect the r_s/r terms in eq. (7.93). Then adopt the co-moving radius, ℓ, that is used in the standard form of the LFRW metric, defined (for $\kappa = 0$) by $r = a(t)\,\ell$. Using $\mathrm{d}r = a\,\mathrm{d}\ell + rH\,\mathrm{d}t$, we see that $-(1 - H^2r^2)\mathrm{d}t^2 + \mathrm{d}r^2 - 2Hr\,\mathrm{d}r\mathrm{d}t = -\mathrm{d}t^2 + a^2\mathrm{d}\ell^2$, and so

$$\mathrm{d}s^2 \simeq -\mathrm{d}t^2 + a^2\Big(\mathrm{d}\ell^2 + \ell^2\mathrm{d}\theta^2 + \ell^2\sin^2\theta\,\mathrm{d}\phi^2\Big) \quad \text{if } r^3 \gg r_s/H^2\,. \tag{7.94}$$

This is clearly the LFRW metric (with $\kappa = 0$), to which the McVittie solution asymptotes in the limit $(Hr)^2 \gg r_s/r$.

To identify that the metric, eq. (7.93), approaches the Schwarzschild metric in the opposite limit, $Hr \ll r_s/r$, it is worth defining the new time coordinate τ by

$$\mathrm{d}\tau = \mathrm{d}t + \frac{Hr\mathrm{d}r}{\sqrt{1 - r_s/r}\,(1 - r_s/r - H^2r^2)}\,, \tag{7.95}$$

since this allows the metric to be written in the diagonal form

$$\mathrm{d}s^2 = -\left(1 - \frac{r_s}{r} - H^2r^2\right)\mathrm{d}\tau^2 + \frac{\mathrm{d}r^2}{1 - r_s/r - H^2r^2} + r^2\Big(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\Big)\,. \tag{7.96}$$

Clearly this reduces to the Schwarzschild solution when $(Hr)^2 \ll r_s/r$, which is true for any distances that are small compared with the megaparsec scales of relevance to cosmology.

For the present purposes, the important thing is that the physical constants characterizing the size of the bound object ($a_0^{-1} = \alpha m_e$ for the atomic case, or $r_s = 2\,GM$ for gravitationally bound systems), are time-independent when expressed using the distance measure, r. But the distance between galaxies, that move along the geodesics corresponding to fixed values of ℓ, grow with time proportional to $a(t)$ in these same coordinates. The overall expansion of the universe can therefore be measured by using the sizes of the bound states as the rulers.

7.4 Present-day energy content

In general the universe contains more than one kind of matter, with some relativistic particles (like photons) mixed with non-relativistic particles (like atoms) plus possibly other more exotic forms, each of which satisfies its own equation of state and interacts fairly weakly with the others. This section summarizes what is known about the universe's contents now, and what may be said about the expansion of the universe in the presence of a mixture of matter of this sort.

Indeed, there is evidence that the universe now contains at least 4 independent types of matter. This section summarizes what is known about the abundance of various types of matter in our present best understanding of the universe.

Radiation

The universe is awash with radiation, with the following components.

The Cosmic Microwave Background Radiation:

The sky is full of photons, called the Cosmic Microwave Background (CMB), whose measured spectrum (see fig. 28) indicates that they are distributed in a thermal distribution whose temperature is $T_\gamma = 2.725$ K. These photons were first directly detected using a microwave horn on the Earth's surface, and their thermal properties have subsequently been precisely measured using balloon- and satellite-borne instruments.

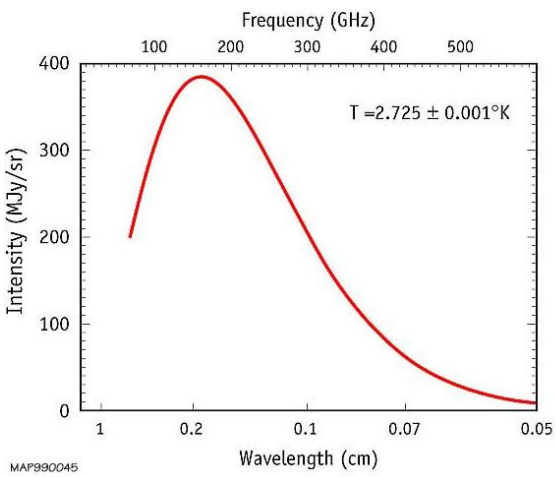


Figure 28: A plot of the measured spectrum of the cosmic microwave background radiation, as measured by the FIRAS instrument aboard the COBE satellite.

As we saw earlier, the number density and energy density of thermal photons are determined by the temperature, with $n_\gamma \propto T^3$ while $\rho_\gamma \propto T^4$. The number density corresponding to $T_0 = 2.725$ K turns out to be

$$n_{\gamma 0} = 4.11 \times 10^8 \text{ m}^{-3}\,, \qquad (7.97)$$

which is very high, much higher than the number density of ordinary atoms. The energy density carried by these photons similarly turns out to be

$$\rho_{\gamma 0} = 0.261 \text{ MeV m}^{-3}$$
$$\text{or} \quad \Omega_{\gamma 0} = 5.0 \times 10^{-5}\,, \qquad (7.98)$$

where as before $\Omega = \rho/\rho_c$ measures the density relative to the critical density, $\rho_c = 5200 \pm 1000$ MeV $^{-3} \simeq 9 \times 10^{-28}$ g/cm^3.

Starlight:

The CMB photons turn out to be somewhat more abundant and carry more energy than is the integrated number of photons emitted by stars since stars first formed, and so represent the dominant contribution of photons to the universal energy density. For instance, a very rough estimate of the density in starlight is obtained by multiplying the present-day luminosity density of galaxies,[17] $nL \simeq 2 \times 10^8\ L_\odot$ Mpc^{-3} by the approximate age of the universe, $H_0^{-1} \simeq 14$ Gy, which gives $\rho_\star \simeq 7 \times 10^{-3}$ MeV m^{-3}, or $\Omega_\star \simeq 1 \times 10^{-6}$.

[17] $L_\odot$ here denotes the luminosity of the Sun.

Relic Neutrinos:

Neutrinos are elementary particles whose mass is small enough to also make them relativistic during most of the universe's history, meaning they also count as radiation when tallying the universe's total energy density. There are three species of neutrino, but because they are electrically neutral they interact very weakly with matter: they can penetrate the entire earth without interacting once. Their existence is known because they take part in radioactive decays, such as in the conversion of a neutron into a proton,

$$n \leftrightarrow p + e^- + \overline{\nu}_e \,, \tag{7.99}$$

in beta decay.

It is believed on theoretical grounds (more about these grounds in subsequent sections) that there is also an almost equally large population of cosmic relic neutrinos filling the universe, although these neutrinos have never been detected. They are expected to have been relativistic throughout most of the universe's history, although they may have perhaps become non-relativistic very recently. They are also expected to be thermally distributed, as are the photons. The neutrinos are expected to have a slightly lower temperature, $T_{\nu 0} = 1.9$ K, than the photons, and because neutrinos are fermions they have a slightly different energy-density/temperature relation than do photons (which are bosons).

These properties make their contribution to the present-day cosmological energy budget not negligible, being predicted to be

$$\rho_{\nu 0} = 0.18 \text{ MeV m}^{-3} \qquad \text{or} \qquad \Omega_{\nu 0} = 3.4 \times 10^{-5} \,. \tag{7.100}$$

If the neutrinos are relativistic, the total radiation density becomes $\rho_{R0} = \rho_{\gamma 0} + \rho_{\nu 0}$, which is of order

$$\rho_{R0} = 0.44 \text{ MeV m}^{-3} \qquad \text{or} \qquad \Omega_{R0} = 8.4 \times 10^{-5} \,. \tag{7.101}$$

Nonrelativistic Matter

There are two qualitatively different kinds of matter present in the universe that we know are not moving at relativistic speeds.

Baryons

The main constituents of the matter we see around us on Earth are atoms, which are themselves made up of protons, neutrons and electrons, and these are predominantly non-relativistic at the present epoch. Furthermore the abundance of electrons is very likely to precisely equal that of protons, since these carry opposite electrical charge, and a precise equality of abundance is required to ensure that the universe carries

no net charge. (The penalty for not having charges locally balance is huge electric forces that ensure that charges move until the local charge density vanishes.)

The mass of the proton and neutron is 940 MeV, which is about 1840 times more massive than the electron, and so the energy density in ordinary non-relativistic particles is likely to be well approximated by the total energy in protons and neutrons. This is also called the total energy in *baryons*, since protons and neutrons carry an approximately conserved charge called baryon number.

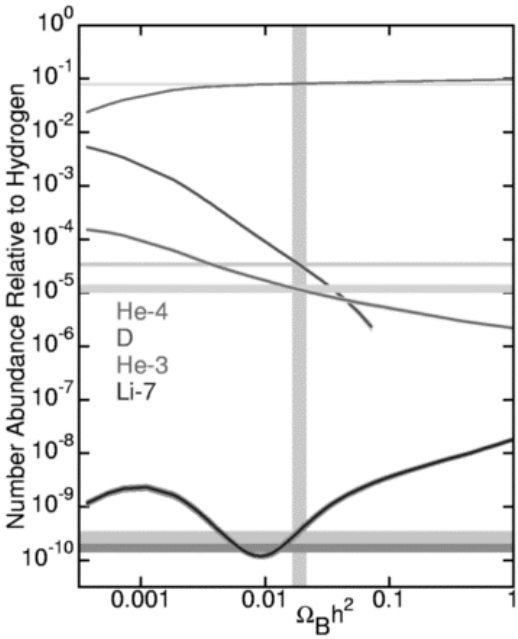


Figure 29: The predictions for light-nuclei abundance as a function of baryon density, with the vertical strip indicating the baryon abundance that gives agreement with observations for all of the light elements. (Courtesy of Ned Wright's cosmology page.)

For reasons to become clear in later sections, it is possible to determine the total number of baryons in the universe (regardless of whether or not they are presently visible) from the success of the predictions of the abundances of light elements due to primordial nucleosynthesis during the very early universe (see fig. 29). This indicates that there is about one baryon for every 10^{10} photons, leading to the following contribution to the total energy density in baryons (*i.e.* ordinary protons, neutrons and electrons)

$$\rho_{B0} = 210 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{B0} = 0.04\,. \tag{7.102}$$

For comparison, the amount of *luminous* matter is considerably smaller than this. Using the previously-quoted luminosity density for galaxies, $nL = 2 \times 10^8$ $L_\odot$ Mpc^{-3}, together with a typical mass-to-luminosity ratio of $M/L = 4M_\odot/L_\odot$, gives an energy density in luminous baryons which is roughly 10% of the total amount in baryons

$$\rho_{L0} = 20 \text{ MeV m}^{-3} \qquad \text{or} \qquad \Omega_{L0} = 0.004\,. \tag{7.103}$$

It should be emphasized that although there is more energy in baryons than in CMB photons, the *number density* of baryons is much smaller. That is

$$n_{B0} = \frac{210 \text{ MeV m}^{-3}}{940 \text{ MeV}} = 0.22 \text{ m}^{-3} = 5 \times 10^{-10}\, n_{\gamma 0}\,, \tag{7.104}$$

and this plays an important role in the physics of the early universe.

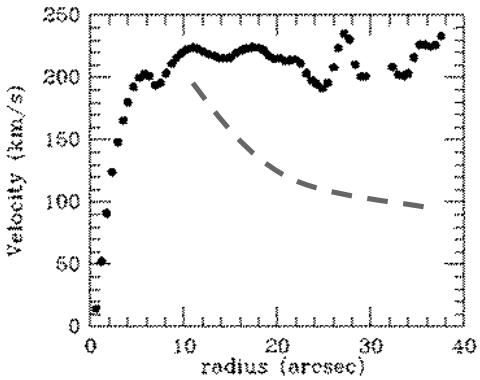


Figure 30: A measurement of a galactic rotation speed vs distance from the galactic center. The dashed line indicates what would be expected if the visible matter were the only matter present.

There several lines of evidence that point to the existence of another form of non-relativistic matter besides baryons, called *Dark Matter*, which appear to carry more energy density than the baryons.

Some of this evidence comes from different independent measures of the total amount of gravitating mass in galaxies. This can be inferred by measuring the rotation rates of galaxies as a function of distance from the galactic center, since this gives speed as a function of radius, $v(r)$, for objects orbiting the galactic center (see fig. (30)). For circular orbits about a point mass Newton's Laws would imply $a = v^2/r = F/m \propto 1/r^2$, and so $v \propto 1/\sqrt{r}$, and a similar fall-off is expected for gas and stars within galaxies (indicated by the dashed line in fig. 30) if only the matter that is visible were present. The disagreement between predictions and observations — which is the rule for large luminous galaxies — indicates that there is 10 – 100 times as much gravitating mass present than would be inferred by counting the luminous matter.

A similar result holds for the total mass in galaxy clusters, as estimated in three independent ways:

- The mass in a galaxy cluster can be inferred by measuring the motions of its constituent galaxies, and comparing this to Newton's Laws (much as was done for stars and gas orbiting in galaxies).

- Alternatively, it can be inferred from the temperature of the hot intergalactic gas that is seen when the galaxy cluster is viewed in x-ray wavelengths (see fig. 31).[18] This temperature gives the average speed of the hydrogen ions present, and the cluster mass must be large enough to have kept this gas bound to the cluster to prevent its dispersal.

- Finally, the mass of a cluster can be inferred by measuring the amount of gravitational lensing that it produces in the images of more distant galaxies, such as revealed by micro-lensing surveys.

[18]Typically, there are more baryons in the intergalactic gas than in the galaxies themselves.

Two further lines of evidence also point towards the existence of Dark Matter, based on the picture that the *large-scale structure* of galaxies and clusters of galaxies first arose as gravity amplified initially small primordial density fluctuations that were already present in the early universe. They start from the realization that these primordial fluctuations are revealed to us by detailed measurements of the temperature of the Cosmic Microwave Background (CMB) as a function of direction, seen from Earth (see fig. 35). Since the CMB represents light that last scattered from matter as the universe cooled through the temperature when electrons and protons were first combining into hydrogen nuclei, these temperature fluctuations represent density fluctuations in the primordial hydrogen gas. Since it is these same fluctuations that are later amplified by gravity to form the galaxies, the properties of the CMB can be related to those of the observed distribution of galaxies we see in the later universe.

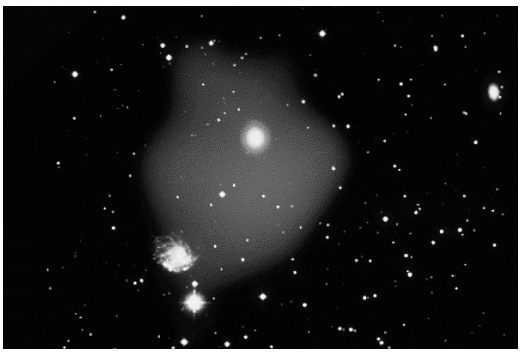


Figure 31: A visible-light photograph of a cluster of galaxies overlaid by an x-ray picture indicating the presence of hot intergalactic gas. The orbital speeds of the galaxies and the gas molecules both indicate the presence of Dark Matter.

Since it turns out that gravity can only amplify density fluctuations if the universe is dominated by nonrelativistic matter, the first piece of evidence asks how long it would take to produce the observed galaxies from the initially small (10^{-5}) amplitude of temperature fluctuations seen in the CMB. It turns out that there has been insufficient time if baryons were the only nonrelativistic matter in the universe, but galaxies would have had time to form if there were sufficiently much Dark Matter.

Similarly, since galaxies form by amplifying fluctuations seen in the CMB, the correlations of the CMB should be mirrored by correlations amongst the positions of the subsequent galaxies. These correlations have been seen and are known as *baryon acoustic oscillations*. The properties of these oscillations agree with predictions only given the right amount of Dark Matter.

All of these estimates appear to be consistent with one another, and indicate a Dark Matter density that is of order

$$\rho_{DM0} = 1350 \text{ MeV m}^{-3} \qquad \text{or} \qquad \Omega_{DM0} = 0.26\,. \tag{7.105}$$

Furthermore it turns out that whatever this gravitating matter is, it must be non-relativistic since it otherwise would not take part in the gravitational collapse that makes galaxies and their clusters in the first place. This indicates that it should

have the same equation of state, $p \approx 0$, as have the baryons, meaning that the total energy density in non-relativistic matter is the sum of the baryonic and Dark Matter abundances: $\Omega_{M0} = \Omega_{B0} + \Omega_{DM0}$. Combining the above estimates gives a total that is of order

$$\rho_{M0} = 1600 \text{ MeV m}^{-3} \qquad \text{or} \qquad \Omega_{m0} = 0.30\,. \tag{7.106}$$

Dark Energy

Finally, there are two lines of evidence which point to a second form of unknown matter in the universe, which does not share the same equation of state of either relativistic or nonrelativistic matter. As mentioned above, one line is based on the recent measurements of the deceleration parameter, q_0, that were made by detecting the expected deviation from Hubble law for very distant supernovae (see fig. 27). This shows that the universal expansion is *accelerating*, rather than decelerating, and so requires the universe must now be dominated by a form of matter for which $\rho + 3p < 0$.

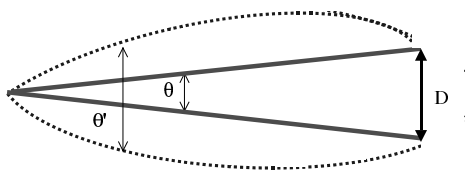


Figure 32: A sketch of the relation between the measured angle on the sky, θ, of a known length, D, seen from across a flat ($\kappa = 0$) universe. The dotted lines indicate how the angle would change (for fixed D) if the intervening geometry of space were positively curved ($\kappa = 1$).

The second line of argument is based on the evidence in favor of the universe being spatially flat: $\kappa = 0$ and so $\Omega_0 = 1$. This evidence comes from measurements of the angular distance between hot and cold fluctuations in the temperature of the CMB photon distributions, as has been measured by satellite experiments (see fig. 35). Since these fluctuations are due to sound waves in the primordial hydrogen gas, their physical size can be computed in terms of the known speed of sound in Hydrogen: it is as if someone has held up a ruler of known length for us at the other end of the universe. Furthermore, we also know the distance to this fluctuation from measurements of H_0. In Euclidean geometry knowledge of these two distances would not be independent of the angle since the geometry of an isosceles triangle is over-determined by a measurement of its length, breadth and angular width. Such a triangle is similarly over-determined in a curved ($\kappa = \pm 1$) geometry, but with a different angle predicted for a given length triangle (as is shown in fig. 32). Consequently, the geometry of space can be inferred by comparing the physical distances with the measured angular separation, leading to the conclusion that $\kappa = 0$ to within the errors.

But the Friedmann equation tells us that $\kappa = 0$ implies $\Omega_0 = 1$ and so $\rho_0 = \rho_c$. And this requires the existence of something besides Dark Matter, since the evidence for Dark Matter indicates that its abundance is too small to give $\Omega_0 = 1$. These two lines of evidence are consistent with one another (within sizeable errors) and point to a *Dark Energy* density which is of order

$$\rho_{DE0} = 3600 \text{ MeV m}^{-3} \qquad \text{or} \qquad \Omega_{DE0} = 0.70\,. \tag{7.107}$$

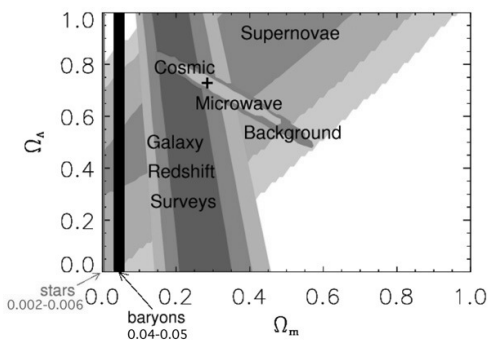


Figure 33: A plot of the amount of Dark Energy and Dark Matter as indicated by supernova measurements, properties of the CMB and direct measurements for Dark Matter. The fact that the regions overlap indicates that all evidence is consistent.

The equation of state for the Dark Energy is not known, apart from the remark that the observations indicate both that at present $\rho_{DE0} \sim 0.7\,\rho_c > 0$ and $w \lesssim -0.8$. If w is constant, it is likely on theoretical grounds that $w = -1$ and the Dark Energy is simply the Lorentz-invariant vacuum energy density. Although it is not yet known whether the vacuum need be Lorentz invariant to the precision required to draw cosmological conclusions of sufficient accuracy, in what follows it will be assumed that the Dark Energy equation of state is $w = -1$.

What emerges is a universe consisting of 70% Dark Energy, 26% Dark Matter and 4% baryons, with many different lines of evidence converging to paint the same picture. It is the very consistency of these many lines of evidence — what has become known as *concordance cosmology* — that helps give confidence that the overall framework is healthy even though it involves the existence of two completely new kinds of unknown matter.

7.5 Earlier epochs

Given the present-day cosmic ingredients described in the previous section, this section uses the equations of state for each type of ingredient to extrapolate the relative abundances into the past in order to estimate what can be said about the cosmic environment during earlier epochs. The main assumption for this extrapolation is that the various components of the cosmic fluid are weakly coupled to one another, and so cannot transfer energy directly to one another.

Under these circumstances the equation of energy conservation, eq. (7.70), applies separately to each component of the fluid. The relative energy densities then

change as these components respond differently to the expansion of the universe, as follows.

- **Radiation:** For photons, starlight and relic neutrinos of sufficiently small mass we have $w = \frac{1}{3}$ and so $\rho(a)/\rho_0 = (a_0/a)^4$;
- **Non-relativistic Matter:** For both ordinary matter (baryons and electrons) and for the Dark Matter we have $w = 0$ and so $\rho(a)/\rho_0 = (a_0/a)^3$;
- **Vacuum Energy:** Assuming the Dark Energy has the equation of state $w = -1$ we have $\rho(a) = \rho_0$ for all a.

This implies the total energy density and pressure have the form

$$\begin{aligned} \rho(a) &= \rho_{DE0} + \rho_{M0}\left(\frac{a_0}{a}\right)^3 + \rho_{R0}\left(\frac{a_0}{a}\right)^4 \\ p(a) &= -\rho_{DE0} + \frac{1}{3}\rho_{R0}\left(\frac{a_0}{a}\right)^4 . \end{aligned} \tag{7.108}$$

As the universe is run backwards to smaller sizes it is clear that these results imply that the Dark Energy becomes less and less important, while relativistic matter becomes more and more important (see fig. 34). Although the Dark Energy now dominates, non-relativistic matter is the next most abundant contribution, and when extrapolated backwards would have satisfied $\rho_M(a) > \rho_{DE}(a)$ relatively recently, at a redshift

$$1 + z = \frac{a_0}{a} > \left(\frac{\Omega_{DE0}}{\Omega_{M0}}\right)^{1/3} = \left(\frac{0.7}{0.3}\right)^{1/3} = 1.3\,. \tag{7.109}$$

The energy density in baryons alone becomes larger than the Dark Energy density at a slightly earlier epoch

$$1 + z > \left(\frac{\Omega_{DE0}}{\Omega_{B0}}\right)^{1/3} = \left(\frac{0.7}{0.04}\right)^{1/3} = 2.6\,. \tag{7.110}$$

For times earlier than this the dominant component of the energy density is due to non-relativistic matter, and this remains true back until the epoch when the energy density in radiation became comparable with that in non-relativistic matter. Since $\rho_R \propto a^{-4}$ and $\rho_M \propto a^{-3}$ radiation-matter equality occurs when

$$1 + z > \frac{\Omega_{M0}}{\Omega_{R0}} = \frac{0.3}{8.4 \times 10^{-5}} = 3600\,. \tag{7.111}$$

This crossover would have occurred much later in the absence of Dark Matter, since the radiation energy density equals the energy density in baryons when

$$1 + z > \frac{\Omega_{B0}}{\Omega_{R0}} = \frac{0.04}{8.4 \times 10^{-5}} = 480\,. \tag{7.112}$$

Knowing how ρ depends on a immediately gives, with the Friedmann equation, H as a function of a, and so also an explicit form for the proper, luminosity and angular-diameter distances. For example, eq. (7.108) implies

$$H(a) = H_0 \left[\Omega_{DE0} + \Omega_{\kappa 0} \left(\frac{a_0}{a} \right)^2 + \Omega_{M0} \left(\frac{a_0}{a} \right)^3 + \Omega_{R0} \left(\frac{a_0}{a} \right)^4 \right]^{1/2}, \tag{7.113}$$

where we define

$$\Omega_{\kappa 0} \equiv - \frac{\kappa}{(H_0 r_0 a_0)^2} \,. \tag{7.114}$$

Using $1 + z = a_0/a$ to eliminate a in favour of z then allows the present-day proper distance in such a universe to be written

$$D(z) = H_0^{-1} \int_0^z dz' \left[\Omega_{DE0} + \Omega_{\kappa 0}(1+z')^2 + \Omega_{M0}(1+z')^3 + \Omega_{R0}(1+z')^4 \right]^{-1/2}, \tag{7.115}$$

with D_L and D_A being related to this by powers of $(1 + z)$ if $\kappa = 0$. It is clear from this expression how measurements of $D_L(z)$ or $D_A(z)$ for a range of z's can allow an inference of the relative present-day density abundances, Ω_{i0}, for $i = DE, M, R$ and κ.

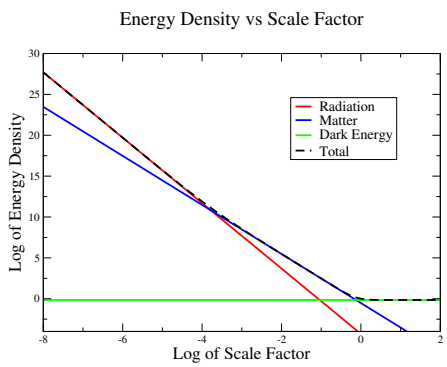


Figure 34: A plot of the energy density, ρ, vs universal scale factor, a, for radiation, matter and dark energy.

Given the dependence, eq. (7.113) of H on a, it is possible to integrate to obtain the t-dependence of a. Although in general this dependence must be obtained numerically, many of its features may be understood on simple analytic grounds based on the recognition that for most epochs there is only a single component of the cosmic fluid which is dominating the total energy density. We expect that for redshifts larger than several thousand $a(t)$ should be well approximated by the expansion in a universe which is filled purely by radiation. Once a/a_0 rises to above 1/3600 there should be a brief transition to the time dependence which describes the universal expansion in a universe dominated by non-relativistic matter. This should apply right up to the very recent past, when a/a_0 is around 0.8, after which there is a transition to vacuum-energy domination, during which the universal expansion accelerates to become exponential with t. In all likelihood we are at present still living in the transition period from matter to vacuum-energy domination.

Although the detailed relationship of a on t in principle depends on the value taken by κ, in practice the contribution of κ is only important in the very recent past. This is because the best information available at present indicates that $\Omega_0 = \Omega_{DE0} + \Omega_{m0} + \Omega_{r0} = 1$, which is consistent with $\kappa = 0$. But even if $\kappa \neq 0$, since the curvature term in eq. (7.68) varies like a^{-2}, it falls more slowly than does either the contribution of matter ($\rho_m \propto a^{-3}$) or radiation ($\rho_r \propto a^{-4}$). So given that the curvature term is at best only comparable to the other energy densities at present, it becomes more and more negligible the further one looks into the universe's past.

As a result it is a very good approximation to use $\kappa = 0$ in the expression for $a(t)$ during the matter-dominated and the earlier radiation-dominated epoch, in which case it has the very simple form $a(t) = a_0(t/t_0)^\alpha$, with $\alpha = \frac{1}{2}$ during radiation domination and $\alpha = \frac{2}{3}$ during matter domination. It may not be valid to neglect κ for the more recent periods of matter domination, and so in this case the more detailed expressions given in the previous section should instead be used. For the present-day epoch it is best to include both $\kappa \neq 0$ and $\rho_{DE} \neq 0$, although the best evidence remains consistent (within largish errors) with $\kappa = 0$.

When $\kappa = 0$ it is also possible to give simple analytic expressions for the time dependence of a in the two transition regions: between radiation- and matter-domination; and between matter- and dark-energy domination. Neglecting radiation during the matter/dark-energy transition gives a Friedmann equation of the form

$$\left(\frac{\dot{a}}{a}\right)^2 = H_{de}^2 \left[1 + \left(\frac{a_{eq}}{a}\right)^3\right] , \tag{7.116}$$

where a_{eq} is the value of the scale factor when the energy densities of the matter and dark energy are equal to one another, and $H_{de}^2 = 8\pi G\rho_{de}/3$ is the (constant) Hubble scale during the pure dark-energy epoch. Integrating this equation (assuming $\dot{a} > 0$), with the boundary condition that $a = 0$ when $t = 0$ then gives the solution

$$a(t) = a_0 \sinh^{2/3}\left(\frac{3H_{de}t}{2}\right) , \tag{7.117}$$

where a_0 is a constant. Notice that when $H_{de}t \gg 1$ this approaches the exponential solution, $a/a_0 \propto \exp(H_{de}t)$ of the dark-energy epoch, while for $H_{de}t \ll 1$ it instead implies $a/a_0 \propto t^{2/3}$, as is appropriate for the matter-dominated epoch.

More generally, the transition from an epoch for which

$$\left(\frac{\dot{a}}{a}\right)^2 = H_{de}^2 \left[1 + \left(\frac{a_{eq}}{a}\right)^p\right] , \tag{7.118}$$

is given by the solution

$$a(t) = a_0 \sinh^{2/p}\left(\frac{pH_{de}t}{2}\right) . \tag{7.119}$$

The transition from radiation to matter domination may be handled in a similar way. It is convenient to write the Friedmann equation during this transition as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{H_{eq}^2}{2}\left[\left(\frac{a_{eq}}{a}\right)^3 + \left(\frac{a_{eq}}{a}\right)^4\right], \tag{7.120}$$

where the constants a_{eq} and H_{eq} are the scale factor and Hubble scale at the instant where radiation and matter have equal energy densities. This may be integrated directly (with $\dot{a} > 0$ and the initial condition $a = 0$ when $t = 0$) to give

$$\left(\frac{a}{a_{eq}} + 1\right)^{1/2}\left(\frac{a}{a_{eq}} - 2\right) = \frac{3H_{eq}t}{2\sqrt{2}} - 2\,. \tag{7.121}$$

Again this has the correct limits: $a \propto t^{2/3}$ when $a \gg a_{eq}$ and $a \propto t^{1/2}$ when $a \ll a_{eq}$.

Exercise 32: Derive eqs. (7.119) and (7.121) by respectively integrating eqs. (7.118) and (7.120).

7.6 Hot Big Bang cosmology

The equations of state for radiation and non-relativistic matter used in the previous discussion are based on those which arise for radiation and atoms which are in thermal equilibrium, and for the case of CMB photons the photons can be seen explicitly to have a thermal distribution. This all points to matter being hot and dense at some point in the universe's past. As we shall see there is also other evidence that the matter in the universe was once as hot as 10^{10} K or more, at which time nuclei were once synthesized from a hot soup of protons, neutrons and electrons.

The Big Bang theory of cosmology starts with the idea that the universe was once small and hot enough that it contained just a soup of elementary particles, in order to see if this leads to a later universe that we recognize in cosmological observations. This picture turns out to describe well many of the features we see around us, which are otherwise harder to understand. This section starts the discussion of the Big Bang theory by exploring the properties of a thermal bath of particles in an expanding universe, in order to understand the conditions under which equilibrium might be expected to hold, and to see what happens as such a bath cools as the universe expands.

The Known Particle Content

The starting point of any such description is a summary of the various types of elementary particles which are known, and their properties. These are well-known from experimental and theoretical study over more than 40 years.

As mentioned earlier, the highest temperature there is direct observational evidence the universe has attained in the past is $T \sim 10^{10}$ K, which corresponds to thermal energies of order 1 MeV. The elementary particles which might be expected to be found within a soup having this temperature are the following.

- **Photons (γ):** are bosons that have two spin (or polarization) states, and have no electric charge or mass. They can be singly emitted and absorbed by any electrically-charged particles.

- **Electrons and Positrons ($e^{\pm}$):** are fermions that each have two spin states and have charge $\pm e$, where e denotes the proton charge.[19] Their masses are the same size as one another, and equal numerically to $m_e = 0.511$ MeV. Because the positron, e^+, is the antiparticle for the electron, e^-, (and vice versa), these particles can completely annihilate into photons through the reaction

$$e^+ + e^- \leftrightarrow 2\gamma\,. \tag{7.122}$$

- **Protons (p):** are fermions that have two spin states, charge $+e$ and a mass $m_p = 938$ MeV. Unlike all of the other particles described here (except the neutron, which is next), the proton can take part in the *strong interactions*, which are what hold nuclei together. For example, this permits reactions like

$$p + n \leftrightarrow D + \gamma\,, \tag{7.123}$$

  in which a proton and neutron combine to produce a *deuterium* nucleus, which is a heavy isotope of Hydrogen that consists of a bound state of one proton and one neutron. The photon which appears in this expression simply carries off any excess energy which is released by the reaction.

- **Neutrons (n):** are fermions having two spin states, no electric charge and a mass $m_n = 940$ MeV. Like protons, neutrons participate in the strong interactions. Isolated neutrons are unstable, and left to themselves decay through the *weak interactions* into a proton, an electron and an electron-antineutrino (see below).

$$n \to p + e^- + \overline{\nu}_e\,. \tag{7.124}$$

- **Neutrinos and Anti-neutrinos (ν_e, $\overline{\nu}_e$, ν_μ, $\overline{\nu}_\mu$, ν_τ, $\overline{\nu}_\tau$):** are fermions which are electrically neutral, and have been found to have nonzero masses whose precise values are not known, but which are known to be smaller than 1 eV.

[19] Superscripts '$\pm$' or context should allow the use of e both as the symbol of the electron and to denote its charge.

Although each has two spin states, it is not yet known whether or not the neutrino and antineutrino are distinct particles (like for electrons) or not (as for photons).

- **Gravitons (G):** are bosons which are not electrically charged and are massless. Gravitons are the quanta which carry the energy packets in a gravitational wave, in the same way that photons do for electromagnetic waves. Gravitons only interact with other particles with gravitational strength, which is very weak compared to the strength of the other interactions. As a result they turn out never to have been in thermal equilibrium for any of the temperatures to which we have observational access in cosmology.

The next sections ask how the temperature of a bath of particles would evolve on thermodynamic grounds as the universe expands.

Cooling Rate

We have found (for several choices for the equation of state) how the energy density in different forms of matter varies with a as the universe expands, and we have seen how to find from this how a varies with time, t. We now ask how thermodynamics relates the temperature to a (and so also t), in order to quantify the rate with which a hot bath cools due to the universal expansion. Since most of the universe's history was dominated by radiation (whose energy density was more important in the past than it is now), we do so here for relativistic particles.

The energy density and pressure appropriate to a gas of relativistic particles (like photons) when in thermal equilibrium at temperature T_R are given by

$$\rho_R = a_B\, T_R^4 \qquad \text{and} \qquad p_R = \frac{1}{3}\, a_B\, T_R^4\,, \tag{7.125}$$

where a_B is $g/2$ times the Stefan-Boltzmann constant, where g counts the number of internal (spin) states of the particles of interest (and so $g = 2$ for a gas of photons).

The evolution of T_R as the universe expands is simply determined by these expressions together with energy conservation, which for relativistic particles we have seen implies $\rho_R\, a^4$ does not change as a increases. It is clear that because $\rho_R \propto T_R^4$ and $\rho_R \propto a^{-4}$, consistency implies the product $a\, T_R$ is constant, and so

$$T_R = T_{R0} \left(\frac{a_0}{a}\right) = T_{R0}(1+z)\,. \tag{7.126}$$

Notice that this assumes only that $\rho_R \propto T_R^4 \propto a^{-4}$, and so (unlike the expression for a vs t) it does *not* assume that the total energy density is radiation-dominated. One way to see why this is so is to recognize that eq. (7.126) is equivalent to the

statement that the expansion is *adiabatic*, since the entropy per unit volume of a relativistic gas is $s_R \propto T_R^3$, and so the total entropy in this gas is

$$S_R \propto s_R\, a^3 \propto (T_R\, a)^3 = \text{constant}\,. \tag{7.127}$$

A Thermal History of the Universe

An important consequence of the falling of the temperature as the universe expands is that it makes interactions amongst the various particles run more slowly. This happens because the lower temperature means there is less energy available per collision on average, but also because it means that there are fewer particles about (per unit volume) with which to interact. Eventually for all interactions there comes a point where reactions run slowly enough that they are so rare as to be nonexistent. When this happens the very equilibrium of the particles involved breaks down, and they are said to *freeze out*. That is, they coast along without interacting down until the present day.

What is spectacular about the study of cosmology now is the ability to test cosmological ideas with observations, and these tests largely rely on detecting those particles which have fallen out of equilibrium to persist to the present day as residual relics of the early universe. This section provides a brief history of the early universe with a focus on describing the various types of relics which arise. Our starting point is the epoch when the universe has a temperature of about 10 MeV, at which point it consists of a hot soup of non-relativistic protons and neutrons, in equilibrium with a population of relativistic electrons, positrons, photons and three species of neutrino.

At MeV temperatures we have approximately equal numbers of protons and neutrons. Since all of the other particles satisfy $m \ll T$ at these temperatures, equipartition of energy in a thermal environment ensures that there are roughly equal numbers of electrons, positrons, photons and each species of neutrino. Furthermore, agreement with observations requires the relativistic particles to be considerably more numerous, with $\eta_B = n_B/n_\gamma = (n_n + n_p)/n_\gamma \sim 10^{-10}$. There must also be a slight excess of electrons over positrons so that $n_e - \overline{n}_e = n_p$ in order to ensure the electrical neutrality of the cosmic environment. This enormous excess of relativistic particles over non-relativistic ones ensures that the entropy of the equilibrium bath which they all share is dominated by the relativistic particles, and so the temperature of the bath falls like $T \propto a^{-1}$, as discussed above. The excess of relativistic matter over non-relativistic matter also ensures that the energy density is radiation-dominated, and so $\rho_{\rm tot} \propto T^4 \propto a^{-4}$.

We now list a number of landmarks in the thermal history of the universe, which make an important impact on the relics we see today that are left over from this earlier and hotter time.

1. **Neutrino Freeze-out:** Once the temperatures fall below a few MeV, the weak interactions are not sufficiently strong to keep the three types of neutrino species in thermal equilibrium. After this point these neutrinos continue to run around the universe without scattering, and are still present during the present epoch as a *Cosmic Neutrino Background*. Since the neutrinos are relativistic, however, their number density remains in its equilibrium form with the temperature simply red-shifting, $T_\nu \propto a^{-1}$, as the universe expands. Since this is precisely the same time-dependence as for the thermal bath containing the rest of the particles, T_ν continues to track the temperature of the thermal bath as the universe expands. Although these neutrinos are in principle all around us, they have so far escaped detection due to their extremely small interaction cross sections.

2. **Electron-Positron Annihilation:** Once the temperature falls below twice the electron mass, $2m_e = 1.2$ MeV, the abundance of electrons and positrons begins to decline relative to photons due to the reaction $e^+e^- \to \gamma\gamma$ beginning to predominate over the inverse process of pair creation. This ends up removing essentially all of the positrons, leaving the same number of residual electrons as there are protons. This has an important consequence for the later universe, because this process of annihilation dumps a considerable amount of energy which reheats the equilibrium bath of photons, neutrons and charged particles relative to the neutrino temperature, which continues to redshift without experiencing any heating (because it is no longer in equilibrium).

3. **Formation of Nuclei:** The thermal evolution at temperatures lower than 1 MeV is richer than would be believed from previous sections due to the possibility which arises of forming bound states. In particular, nuclear interactions can bind a neutron and proton into deuterium, with a binding energy of 2.22 MeV, and so once temperatures reach this energy range light nuclei begin to form and so change the chemical composition of the cosmic fluid. The residual abundance of these nuclei predicted by this process agrees well with the observed primordial abundances, which provides strong evidence for the validity of the Big Bang picture of cosmology, and gives important information about the total abundance, n_B, of baryons (protons and neutrons). A constraint on the total number of baryons is possible because the nuclear reaction rates are proportional to the density of reactants, with more baryons leading to faster reactions. But the total number of nuclei formed depends on how long it takes for temperatures to cool to the point that nuclear reactions also stop happening, and this is controlled in part by the size of the reaction rates (and so also by

the baryon density). The result is usually normalized to the density of photons since the result is then time-independent, leading to $\eta_B := n_B/n_\gamma \simeq 10^{-10}$.

4. **Formation of Atoms:** Electromagnetic interactions furnish another important set of bound states which complicate the picture of the universe at lower temperatures. In particular, electrons can bind with nuclei to form neutral atoms once the temperature falls below the relevant binding energies, $E \sim 10$ eV. In practice atoms don't actually form until the temperature is somewhat cooler than this, $T \simeq 1$ eV, because the large number ($\sim 10^{10}$) of photons for each electron and proton, makes the reactions where photons dissociate bound atoms initially more common than those where atoms are formed. At this point the equilibrium conditions for charged particles and photons changes dramatically, since once atoms form the cosmic fluid becomes electrically neutral, and so largely transparent to photons. The cosmic microwave background (CMB) consists of those photons which last scattered from matter at this point, and have survived unscathed to be observed during the present epoch. The observation of these photons gives a direct measure of the temperature of the heat bath from which the photons eventually decoupled, a map of which is given in fig. 35.

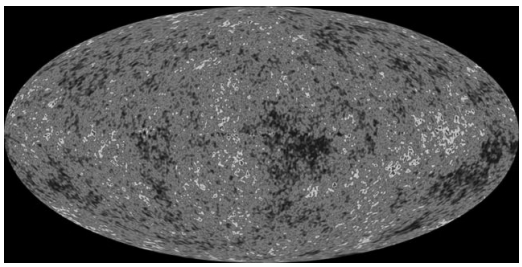


Figure 35: The temperature of the cosmic microwave background radiation as a function of direction, as measured by the WMAP collaboration. The difference between the hottest and coolest points in this map are of order 10 μK.

In all, the Hot Big Bang provides an outstandingly successful description what we see around ourselves in cosmology, but only if we start with just the right initial conditions sometime before nucleosynthesis. These initial conditions require the early universe to be very homogeneous and isotropic, since this is what is observed to be true for the cosmic microwave background. Indeed, small temperature (and so also density) fluctuations in the primordial hydrogen environment are directly observed in precision measurements of the cosmic microwave background temperature as a function of direction in the sky (as seen in fig. 35). Since these fluctuations are only about 10 μK in size, compared with the CMB's average temperature of 2.725 K they show that density perturbations were at most as big as 1 part in 10^5 when atoms were first forming in the early universe.

But because primordial fluctuations have been seen, the initial universe cannot be perfectly homogeneous. This is also a good thing, because the amplitude of these small density fluctuations is ultimately amplified by gravitational collapse to form the galaxies and stars we find ourselves surrounded by. An important piece of evidence for Dark Matter is that there has not been sufficient time for this amplification to take place if the only non-relativistic particles around are baryons.

It turns out that these initial conditions are not natural, in that they do not automatically arise unless they are put by hand into the initial conditions. Furthermore, because time evolution moves the universe away from homogeneity and isotropy, the universe at still-earlier times must be smooth to a much higher accuracy than at present. It is hoped that these initial conditions may be the relics of a still-earlier epoch of the universe about which physicists have long speculated, called the *inflationary epoch*. The speculations center around the observation that the special initial conditions of the Big Bang would emerge very naturally if the universe were to have undergone a period of near exponential expansion (much like the Dark Energy dominated epoch we now appear to be entering, but with much higher energies and densities) at much earlier times.

Here is a selection of textbooks on General Relativity, and cosmology.

1. C.M. Will, *Theory and Experiment in Gravitational Physics (Revised Edition)*, Cambridge University Press, 1993.

2. S. Carroll, *An Introduction to General Relativity Spacetime and Geometry*, Addison Wesley 2004. [Modern and well written]

3. S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*, Wiley 1972. [The timeless classic – very physical]

4. C. Misner, K. Thorne and J. Wheeler, *Gravitation*, Freeman and Company 1970. [Encyclopedic, with many layers of insight]

5. R. Wald, *General Relativity*, University of Chicago 1984. [More mathematical, with an emphasis on modern differential geometry]

6. P.J.E. Peebles, *Principles of Physical Cosmology*, Princeton University Press (1993).

7. B. Ryden, *Introduction to Cosmology*, Pearson Education 2003. [A good undergraduate introduction to modern cosmology]

8. S. Dodelson, *Modern Cosmology*, Academic Press 2003. [A good, but more advanced, introduction to modern cosmology.]

9. A. Linde, *Particle Physics and Inflationary Cosmology*, Harwood Academic Publishers (1990).

10. E. W. Kolb and M. S. Turner, *The Early Universe*, Addison-Wesley (1990).

11. A. R. Liddle and D. H. Lyth, *Cosmological Inflation and Large-Scale Structure*, Cambridge University Press (2000).

12. S. Weinberg, *Cosmology*, Oxford University Press (2008).

13. S. Chandrashekhar, *The Mathematical Theory of Black Holes*, Oxford University Press 1992.

14. S.L. Shapiro and S.A. Teukolsky, *Black Holes, White Dwarfs and Neutron Stars: The physics of compact objects*, Wiley 1983.

References

[1] S. Baessler *et.al.*, Physical Review Letters **83** (1999) 3585;
E. Adelberger, Classical and Quantum Gravity **18** (2001) 2397.

[2] I. I. Shapiro, Fourth Test of General Relativity, Physical Review Letters **13** (1964) 789-791.

[3] R. D. Reasenberg, et al., Viking Relativity Experiment: Verification of Signal Retardation by Solar Gravity, Astrophysical Journal **234**, (1979) L219-L221.

[4] B. Bertotti, L. Iess and P. Tortora, A Test of General Relativity Using Radio Links with the Cassini Spacecraft, Nature **425**, (2003) 374-376 (2003);

John D. Anderson, Eunice L. Lau, and Giacomo Giampieri, "Measurement of the PPN Parameter with Radio Signals from the Cassini Spacecraft at X- and Ka-Bands," in the proceedings of the 22nd Texas Symposium on Relativistic Astrophysics, Stanford, 2004.

[5] Reflections on Relativity, http://www.mathpages.com/rr/rrtoc.htm.

[6] S. Gillessen *et.al.*, arXiv:0810.4674 (astro-ph).

[7] W.L. Freedman *et.al.*, Ap. J. **553** (2001) 47–72, e-print (arXiv:astro-ph/0012376).

[8] G.C. McVittie, Mon. Not. Roy. Aston. Soc. **93** (1933) 325;

B.C. Nolan, Phys. Rev. **D58** (1998) 064006 [gr-qc/9805041].