MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Physics Department

Physics 8.286: The Early Universe — October 13, 2018
Prof. Alan Guth

Lecture Notes 5
INTRODUCTION TO NON-EUCLIDEAN SPACES

INTRODUCTION:

The history of non-Euclidean geometry is a fascinating subject, which is described very well in the introductory chapter of *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* by Steven Weinberg. Here I would like to summarize the important points. Although historical in its organization, this section describes some essential mathematics and should be read carefully.

Euclid showed in his *Elements* how geometry could be deduced from a few definitions, axioms, and postulates. One of Euclid's assumptions, however, seemed to generations of mathematicians to be somewhat less obvious than the others. This assumption, known as Euclid's fifth postulate, was stated by Euclid as follows:

> "If a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines if produced indefinitely meet on that side on which the angles are less than two right angles." *[This statement is interpreted to imply that the two straight lines will never meet if extended on the opposite side.]*

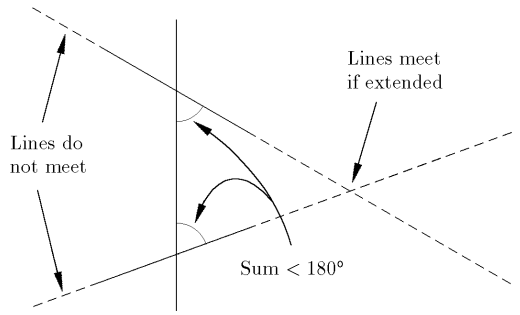

Figure 5.1: Euclid's fifth postulate.

Many mathematicians attempted to prove this postulate from the other assumptions, but all of these attempts ended in failure. It was discovered, however, that the fifth postulate could be replaced by any of a number of equivalent statements, such as:

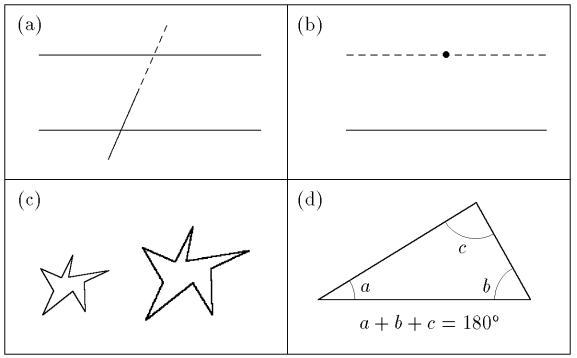

Figure 5.2: Statements equivalent to the fifth postulate.

(a) "If a straight line intersects one of two parallels (i.e, lines which do not intersect however far they are extended), it will intersect the other also."

(b) "There is one and only one line that passes through any given point and is parallel to a given line."

(c) "Given any figure there exists a figure, similar* to it, of any size."

(d) "There is a triangle in which the sum of the three angles is equal to two right angles (i.e., 180°)."

Given Euclid's other assumptions, each of the above statements is equivalent to the fifth postulate.

The attitude of mathematicians toward the fifth postulate underwent a marked change during the eighteenth century, when mathematicians began to consider the possibility of abandoning the fifth postulate. In 1733 the Jesuit Giovanni Geralamo Saccheri (1667–1733) published a study of what geometry would be like if the postulate were false. He, however, was apparently convinced that the fifth postulate must be true, and he pursued this work because he hoped to discover an inconsistency — he didn't.

Carl Friedrich Gauss (1777-1855) seems to have been the first to really take seriously the possibility that the fifth postulate could be false. He, János Bolyai (an Austrian army officer, 1802-1860), and Nikolai Ivanovich Lobachevsky (a Russian mathematician, 1793-1856) independently discovered and explored a geometry which in modern terms is described as a two-dimensional space of constant negative curvature. The space is infinite

* Two polygons are similar if their corresponding angles are equal, and their corresponding sides are proportional.

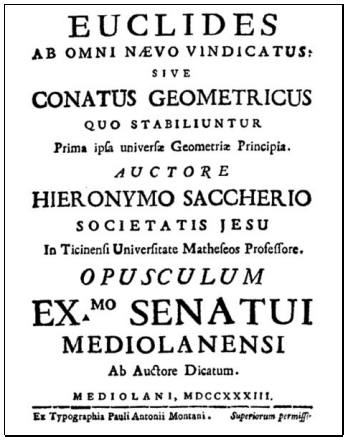
EUCLIDES
AB OMNI NÆVO VINDICATUS:
SIVE
CONATUS GEOMETRICUS
QUO STABILIUNTUR
Prima ipſa univerſæ Geometriæ Principia.
AUCTORE
HIERONYMO SACCHERIO
SOCIETATIS JESU
In Ticinenſi Univerſitate Matheſeos Profeſſore.
OPUSCULUM
EX.MO SENATUI
MEDIOLANENSI
Ab Auctore Dicatum.
MEDIOLANI, MDCCXXXIII.
Ex Typographia Pauli Antonii Montani. Superiorum permiſſu

Figure 5.3: The frontispiece of Giovanni Geralamo Saccheri's 1733 book titled *Euclides ab omni naevo vindicatus* (Euclid Freed of Every Flaw). Saccheri pursued the consequences of assuming that the fifth postulate was false, hoping to find a contradiction.

in extent, is homogeneous and isotropic, and satisfies all of Euclid's assumptions except for the fifth postulate. In this space every one of the statements of the fifth postulate and its equivalents listed above are false — through a given point there can be drawn ***infinitely*** many lines parallel to a given line; ***no*** figures of different size are similar; and the sum of the angles of any triangle is ***less*** than 180°.

The surface of a sphere, it should be pointed out, satisfies all the postulates of Euclid except for the fifth and the second, which states that "Any straight line segment can be extended indefinitely in a straight line." From a modern point of view the surface of a sphere provides a perfectly interesting example of a non-Euclidean geometry. Historically, however, this example was not taken very seriously, apparently because it seemed too simple. The great circles would be the objects that play the role of straight lines, but since any two great circles intersect, there could be no such thing as parallel lines.

Despite the work of Gauss, Bolyai, and Lobachevsky, it was still not clear that their non-Euclidean geometry was logically consistent. This problem was not solved until 1870, when Felix Klein (1849-1925) developed an "analytic" description of this geometry. In Klein's description, a "point" of the Gauss-Bolyai-Lobachevsky (G-B-L) geometry can be described by two real number coordinates (x,y), with the restriction

$$x^2 + y^2 < 1 \,. \tag{5.1}$$

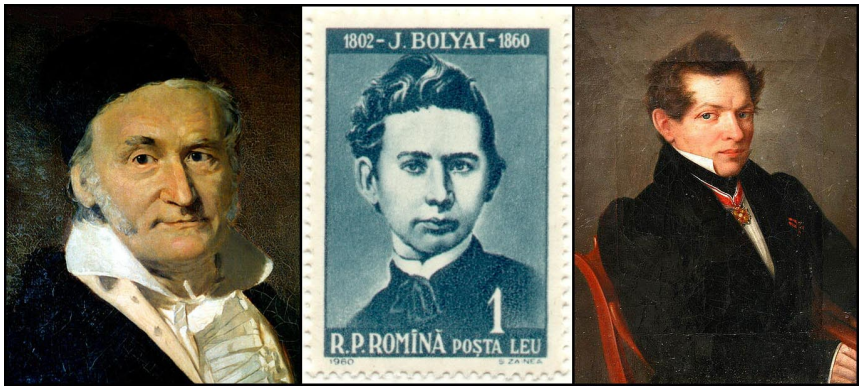

Figure 5.4: Carl Friedrich Gauss, János Bolyai, and Nikolai Ivanovich Lobachevsky independently developed the first example of a mathematical theory in which Euclid's fifth postulate is false, now known as the Gauss–Bolyai–Lobachevsky geometry. Gauss (1777–1855) was the son of poor working-class parents in Germany, but by the time he was 15 his mathematical talents were noticed by the Duke of Brunswick, who sent Gauss to the Collegium Carolinum and then the University of Göttingen. Gauss remained at Göttingen for the rest of his life, becoming Professor of Astronomy and director of the astronomical observatory in 1807. His students included Richard Dedekind, Bernhard Riemann, Peter Gustav Lejeune Dirichlet, Gustav Kirchhoff, August Ferdinand Möbius, and Friedrich Bessel. Bolyai (1802–1860) was the son of Farkas Bolyai, a teacher of mathematics, physics, and chemistry at the Calvinist College in Marosvásárhely, Hungary (now Tirgu-Mures, Romania). Although his father was well-educated, he was nonetheless not well paid, so János attended Marosvásárhely College and later studied military engineering at the Academy of Engineering at Vienna, because that is what they could afford. He then entered the army engineering corps, where he served for 11 years, during which time he carried out his now-famous investigation of non-Euclidean geometry. The work was published in 1831 as an appendix in a book written by his father. Bolyai resigned from the army in 1833 due mainly to health problems, and lived the rest of his life in relative poverty, dying at the age of 57 of pneumonia. The Romanian postage stamp shown here honored the 100th anniversary of Bolyai's death; the picture was apparently fabricated, as no authentic picture of Bolyai is known to exist. Lobachevsky (1792–1856) was the son of Polish parents living in Russia. His father was a clerk in a land-surveying office, who died when Lobachevsky was only seven. His mother relocated the family to Kazan, Russia, where Lobachevsky attended Kazan Gymnasium and later was given a scholarship to Kazan University, where one of his professors was Martin Bartels, who was a teacher and friend of Gauss. Lobachevsky remained at Kazan University for the rest of career, becoming rector of the university in 1827. His work on non-Euclidean geometry was published in the *Kazan Messenger* in 1829, but was rejected for publication by the St. Petersburg Academy of Sciences. Lobachevsky was asked to retire in 1846, and after that his health and financial situation deteriorated, he became blind, and his favorite eldest son died. Lobachevsky himself died before the importance of his work in mathematics was appreciated.

The distance $d(1,2)$ between two points (x_1,y_1) and (x_2,y_2) is then defined to be

$$\cosh\left[\frac{d(1,2)}{a}\right] = \frac{1 - x_1x_2 - y_1y_2}{\sqrt{1-x_1^2-y_1^2}\,\sqrt{1-x_2^2-y_2^2}}\ , \tag{5.2}$$

where a is a fundamental length which sets a scale for the geometry. Note that the space is infinite despite the coordinate restriction of Eq. (5.1), because the distance approaches infinity as either $x_1^2+y_1^2 \to 1$ or $x_2^2+y_2^2 \to 1$. Klein showed that with this definition of point and distance the model satisfies all of the assumptions of the G-B-L geometry. Thus, assuming the consistency of the real number system, the consistency of the G-B-L geometry was established. In addition, this work reinforced the important idea of analytic geometry which had been introduced by Descartes. It has since proven to be very useful to describe a geometry not by listing axioms, but instead by giving an explicit description in terms of a coordinate system and distance function.

Gauss went on to develop two very central ideas in non-Euclidean geometry. The first is the distinction between the "inner" and "outer" properties of a surface. The inner properties of a surface are those distance relationships that can be measured within the surface itself, such as in Eq. (5.2). The outer properties refer to the way in which a space might be embedded in a higher dimensional space. For example, the surface of a sphere is a two-dimensional space which we visualize by embedding in a three-dimensional space. Gauss emphasized that the distance relationships within the two-dimensional surface itself provide a complete mathematical system which can be studied independently of any assumptions about the embedding in the three-dimensional space. Gauss wrote in 1827 that it is the inner properties of the surface that are "most worthy of being diligently explored by geometers." Note that the G-B-L geometry cannot be fully embedded in a three-dimensional Euclidean space, although finite patches of it can be so embedded. To describe the whole space, it is necessary to describe it in terms of its inner properties.

Gauss's second central idea had to do with the form of the distance function $d(1,2)$. It turns out that if one allows this function to have any form, then the class of geometries is so unconstrained that nothing very interesting results. Gauss realized first that one need not specify $d(1,2)$ for arbitrary points 1 and 2. It is sufficient to consider only infinitesimal line segments. Such a line segment can be described as extending from the point (x,y) to $(x+\mathrm{d}x, y+\mathrm{d}y)$. The length of a finite segment of a curve is then defined by summing up (integrating) the lengths of the infinitesimal segments that make it up. The distance $d(1,2)$ between two arbitrary points can then be defined as the length of the shortest curve which joins the two points. The concept of a *line* is replaced by a *geodesic,* defined to be any curve that is the shortest path between its endpoints. More precisely, a geodesic is not necessarily the true minimum of the path length — it is only necessary that the path is ***stationary***, in the sense that the first derivative with respect to any variation of the path between the two endpoints must vanish. The path length might then be a minimum, a maximum, or a saddle point.

For the length of the infinitesimal line segment from (x,y) to $(x+\mathrm{d}x, y+\mathrm{d}y)$, Gauss realized that the interesting case is to restrict one's attention to functions for which the squared segment length $\mathrm{d}s^2$ is quadratic in $\mathrm{d}x$ and $\mathrm{d}y$ (i.e., functions for which each term contains two powers of $\mathrm{d}x$ and/or $\mathrm{d}y$). Such functions can be written as

$$\mathrm{d}s^2 = g_{xx}\mathrm{d}x^2 + g_{xy}\mathrm{d}x\,\mathrm{d}y + g_{yx}\mathrm{d}y\,\mathrm{d}x + g_{yy}\mathrm{d}y^2\ , \tag{5.3}$$

where g_{xx}, g_{xy}, g_{yx}, and g_{yy} are functions of position (x,y) and are together called the metric of the space. (Since g_{xy} and g_{yx} both multiply $\mathrm{d}x\,\mathrm{d}y$, only their sum is relevant. By convention one sets $g_{xy} = g_{yx}$.) Gauss showed that the assumption that $\mathrm{d}s^2$ is quadratic is equivalent to the assumption that in any infinitesimal region it is possible to choose a coordinate system (x',y') in which the distance relation is Euclidean: $\mathrm{d}s^2 = \mathrm{d}x'^2 + \mathrm{d}y'^2$. Today spaces with a metric of this form are generally called either metric spaces or Riemannian spaces.

In Euclidean space one can use any coordinate system one wants, although one usually prefers a Cartesian system in which the metric has the form:

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2\ . \tag{5.4}$$

Any two systems with metrics of this form are related to each other by a translation and/or a rotation. For some purposes, however, it is convenient to use polar coordinates r and θ, for which the metric is given by

$$\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\mathrm{d}\theta^2\ . \tag{5.5}$$

Thus, the mere fact that the metric does not have the Cartesian form of Eq. (5.4) does ***not*** imply that the underlying space is non-Euclidean — one might simply be using a non-Cartesian coordinate system. It is therefore useful to have some way of describing the inner curvature of a space in a way which is not confused by the choice of a coordinate system. Such a method was developed for two-dimensional spaces by Gauss, who showed that the underlying space is Euclidean if and only if a somewhat complicated expression involving derivatives of the metric is equal to zero. The extension to more than two dimensions was carried out by Georg Friedrich Bernhard Riemann (1826-1866). The details of the Gaussian curvature and the Riemann curvature tensor are beyond the level of this discussion.

GENERAL RELATIVITY:

As I have mentioned before, Einstein's theory of general relativity is nothing more nor less than a theory of gravity. When Einstein invented the special theory of relativity in 1905, he realized immediately that it was inconsistent with Newton's theory of gravity. The inconsistency has nothing in particular to do with the inverse square nature of the force law, and it cannot be remedied by simply modifying the way that the force depends on the distance. Rather, the inconsistency is due to the fact that Newton's law of gravity assumes that the force between two bodies depends instantaneously on the distance between them. That is, to determine the force due to body B acting on body A at time t, one must merely know the position of the two bodies at time t. However, as we discussed in Lecture Notes 1, special relativity implies that the synchronization of clocks depends on the velocity of the observer. Thus, two observers who are moving relative to each other will not agree on what it means to measure the positions of A and B at the same time, and so a physically meaningful quantity like a force cannot be determined by these two positions. If special relativity is correct, then Newton's law of gravity must be modified.

The idea of an action-at-a-distance theory is not completely ruled out by special relativity, but it is very difficult to formulate such a theory. The electromagnetic force of one charged particle acting on another can be expressed by an action-at-a-distance law, but it is rather complicated. (The force law is stated, for example, in *The Feynman Lectures on Physics*, Volume 1, by R.P. Feynman, R.B. Leighton, and M. Sands.) The force on charge A at time t does not depend on the position of charge B at time t, but instead depends on the position (and velocity, and acceleration!) of charge B at a retarded time t'. The time t' is determined by the rule that a light pulse (moving at speed c) can just barely travel from B to A in the time interval from t' to t, as illustrated in the following diagram:

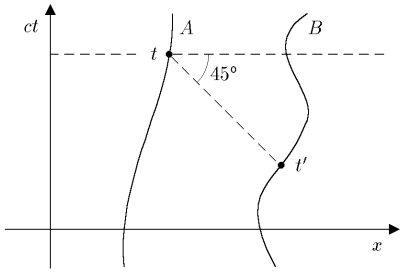

Figure 5.5: Definition of the retarded time t'. The electromagnetic force on particle A at time t, due to particle B, can be expressed in terms of the position, velocity, and acceleration of charge B at the retarded time t'.

Two different observers will agree when this relationship is met, since they agree on what it means for a trajectory to move at the speed of light. However, the two observers will measure different values for the positions, velocities, and accelerations, and it requires a very complicated force law such that both observers will conclude that the law is satisfied.

The simplest way to formulate electromagnetic theory is to avoid action-at-a-distance forces, but instead to use the concept of a field. The electric and magnetic fields are each defined at all points in space, and a charged particle interacts only with the fields at the location of the particle. The evolution of the fields is governed by Maxwell's equations. These equations allow information about the changing position of a particle to propagate in the form of waves which travel at the speed of light.

General relativity is also a theory of fields, similar in type to the Maxwell theory of electromagnetism. In the case of general relativity there is no known action-at-a-distance formalism. The "fields" which are involved in general relativity are of course not the electric and magnetic fields of the Maxwell theory. The fields of general relativity are in fact the metric functions defined earlier. Space and time must be considered together, and it is the metric functions on this "spacetime" which are the fields that general relativity uses to describe gravitation. We will see later that in this curved (i.e., non-Euclidean) spacetime, a freely falling particle is assumed to travel along a geodesic. The attractive effect of gravity then appears simply as a distortion of spacetime.

THE SURFACE OF A SPHERE:

As mentioned above, the surface of a sphere embedded in a three-dimensional Euclidean space is a perfectly good example of a non-Euclidean geometry. In order to develop some of the techniques of non-Euclidean geometry, we begin by studying this familiar system. Since the three-dimensional embedding space is Euclidean, we can use our knowledge of Euclidean geometry to learn about the non-Euclidean two-dimensional geometry of the surface of the sphere. Beware, however, that not all two-dimensional curved surfaces can be embedded in a three-dimensional Euclidean space.

The surface of the sphere can be described by using Cartesian coordinates (x, y, z)

in the three-dimensional space, in which case the surface is given by:

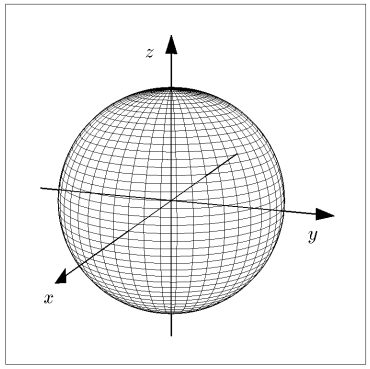

Figure 5.6: A sphere in Cartesian coordinates.

$$x^2 + y^2 + z^2 = R^2 \ , \tag{5.6}$$

where R is the radius of the sphere. We now want to take seriously the notion that the two-dimensional space of the surface defines a two-dimensional geometry with "inner" properties that are independent of the existence of the third dimension. We take the point of view that the third dimension has been introduced only as an aid in visualizing the two-dimensional surface. This third dimension can of course be useful, because in the three-dimensional picture the properties of homogeneity and isotropy are obvious. (Recall here that homogeneity and isotropy refer to properties of the *two-dimensional* space. Homogeneity means that all points on the surface of the sphere are equivalent. Isotropy means that if a two-dimensional creature living in the two-dimensional surface were to look in all directions within the two-dimensional surface, he would see the same thing in all directions.)

In order to describe the two-dimensional world without reference to the third dimension, it is useful to introduce a two-dimensional coordinate system. The most natural choice is to use the usual angular variables θ and ϕ, as shown in Fig. 5.7.

From the diagram we can see that x, y, and z can be expressed as

$$\begin{aligned} x &= R\sin\theta\cos\phi \\ y &= R\sin\theta\sin\phi \\ z &= R\cos\theta \ , \end{aligned} \tag{5.7}$$

where θ runs from 0 to π and ϕ runs from 0 to 2π.

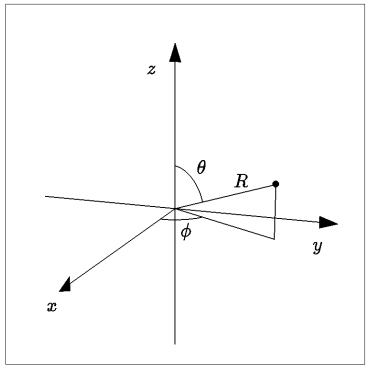

Figure 5.7: Spherical polar coordinates for the surfaces of a sphere.

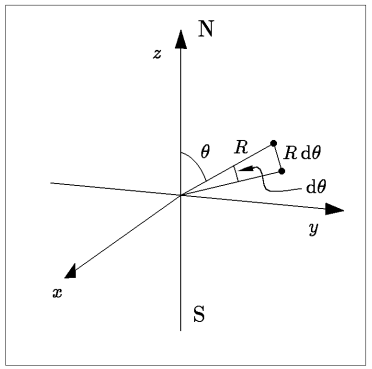

Figure 5.8: Variation of θ in spherical polar coordinates: $\mathrm{d}s = R\mathrm{d}\theta$.

To describe the inner properties of this two-dimensional space, we must write down an expression for the metric. That is, we need an expression for the distance $\mathrm{d}s$ between two points on the surface labelled by (θ,ϕ) and $(\theta+\mathrm{d}\theta,\phi+\mathrm{d}\phi)$. It is helpful to think about varying θ and ϕ one at a time. As θ is increased, the point moves a distance $R\,\mathrm{d}\theta$ toward the south (where I am using the positive z-axis to define a North pole), as can be seen in Fig. 5.8.

When ϕ is increased, the point moves toward the east, tracing out a circle at constant latitude. The radius of the circle is $R\sin\theta$, and so the distance moved is given by $R\sin\theta\,\mathrm{d}\phi$, as shown in the Fig. 5.9.

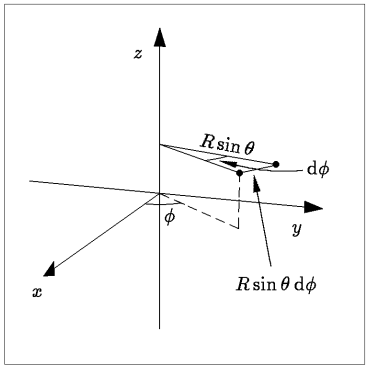

Figure 5.9: Variation of ϕ in spherical polar coordinates: $\mathrm{d}s = R\sin\theta\,\mathrm{d}\phi$.

Since these two displacements are in orthogonal directions, the total distance is given by the Pythagorean theorem:

$$\mathrm{d}s^2 = R^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right) \ . \tag{5.8}$$

Eq. (5.8) describes the metric of the two-dimensional space.

If one wishes to avoid the pictures, one can also derive Eq. (5.8) directly from Eqs. (5.7), by writing

$$\mathrm{d}x = \frac{\partial x}{\partial\theta}\mathrm{d}\theta + \frac{\partial x}{\partial\phi}\mathrm{d}\phi = R\cos\theta\cos\phi\,\mathrm{d}\theta - R\sin\theta\sin\phi\,\mathrm{d}\phi \ ,$$

$$\mathrm{d}y = \frac{\partial y}{\partial\theta}\mathrm{d}\theta + \frac{\partial y}{\partial\phi}\mathrm{d}\phi = R\cos\theta\sin\phi\,\mathrm{d}\theta + R\sin\theta\cos\phi\,\mathrm{d}\phi \ ,$$

and

$$\mathrm{d}z = \frac{\partial z}{\partial\theta}\mathrm{d}\theta + \frac{\partial z}{\partial\phi}\mathrm{d}\phi = -R\sin\theta\,\mathrm{d}\theta \ . \tag{5.9}$$

These expressions can then be substituted into

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \ , \tag{5.10}$$

and after some algebra involving repeated use of the identity $\sin^2\phi + \cos^2\phi = 1$, one again obtains Eq. (5.8).

A CLOSED THREE-DIMENSIONAL SPACE:

The goal here is to use the same techniques to describe a closed three-dimensional space. This space will be homogeneous and isotropic, and it will have a finite volume but no boundary. Since the space is homogeneous and isotropic, it is a candidate for the space in which we live.

To derive a metric for the three-dimensional space, one simply repeats the steps carried out above with one additional dimension. One begins therefore in a Euclidean space with four dimensions, and hence with four Cartesian coordinates which I will call (x,y,z,w). The surface of a sphere in this four-dimensional space is then described by the equation

$$x^2 + y^2 + z^2 + w^2 = R^2 \ . \tag{5.11}$$

Note that the surface of the sphere is a three-dimensional space, since it can be described by three coordinates.

To explicitly describe the surface by three coordinates, one can introduce one more angular variable in addition to θ and ϕ. We therefore introduce ψ, which will represent the angle between the point being described and the w-axis. Since ψ measures the angle from an axis, like θ it ranges from 0 to π. One can then look at the point projected into the x-y-z subspace and define the variables θ and ϕ as we did above. (By "project into the x-y-z subspace", I simply mean to ignore the w-coordinate.) Pictorially one would depict ψ as

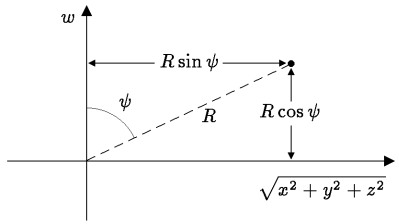

Figure 5.10: The new angular variable ψ, which measures the angle from the w-axis.

and in terms of equations it can be expressed as

$$\begin{aligned} x &= R\sin\psi\sin\theta\cos\phi \\ y &= R\sin\psi\sin\theta\sin\phi \\ z &= R\sin\psi\cos\theta \\ w &= R\cos\psi \ , \end{aligned} \tag{5.12}$$

where

$$0 \le \psi \le \pi \ , \quad 0 \le \theta \le \pi \ , \quad 0 \le \phi \le 2\pi \ , \tag{5.13}$$

and $\phi = 0$ is identified with $\phi = 2\pi$.

Since the coordinate system is to describe the surface, some point on the surface has to be chosen to be the origin of the coordinate system. For the two-dimensional spherical surface of the last section, we can consider the north pole to be the center, and then θ is the radial coordinate that measures the distance from the center. Here we are choosing the center of our coordinate system to be the positive w-axis, which we will also describe as the "north pole". The coordinates of the north pole in the four-dimensional embedding space are ($x=0$, $y=0$, $z=0$, $w=R$). In the polar coordinate system the north pole is described by $\psi = 0$, and the distance from the north pole is given by $R\psi$. Thus ψ plays the role of the radial coordinate in this system.

To derive the metric, one could proceed purely algebraically along the lines of Eq. (5.9) above, or one could use the geometric arguments which were used to motivate Eq. (5.8). For the geometric approach, one notes that a variation from ψ to $\psi + \mathrm{d}\psi$ results in a displacement by a distance $R\,\mathrm{d}\psi$. A variation in θ or ϕ results in a displacement contained entirely within the x-y-z three-space; $\mathrm{d}s^2$ is given by Eq. (5.8) times an overall factor of $\sin^2\psi$ due to the fact that the radius in the x-y-z space is given by $r\sin\psi$. Assuming that these two displacements are orthogonal to each other, the metric can be written as

$$\mathrm{d}s^2 = R^2\left[\mathrm{d}\psi^2 + \sin^2\psi\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right] \ . \tag{5.14}$$

To complete the justification of Eq. (5.14), we should verify that the infinitesimal displacement of the point when ψ is varied is orthogonal to the displacement caused by infinitesimal variation of θ or ϕ. To see this, let us use vector notation $\vec{r} \equiv (x, y, z, w)$ to describe the four-dimensional space. Then, as ψ is varied from ψ to $\psi + \mathrm{d}\psi$, the vector $\vec{r}$ varies from $\vec{r}$ to $\vec{r} + \mathrm{d}\vec{r}_\psi$, where we can see from Eq. (5.12) that

$$\begin{aligned}\mathrm{d}\vec{r}_\psi &= \left(\frac{\partial x}{\partial\psi}, \frac{\partial y}{\partial\psi}, \frac{\partial z}{\partial\psi}, \frac{\partial w}{\partial\psi}\right)\,\mathrm{d}\psi \\ &= R\cos\psi(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta, 0)\,\mathrm{d}\psi - R\sin\psi(0,0,0,1)\,\mathrm{d}\psi \ .\end{aligned} \tag{5.15}$$

Note that the components in the x-y-z subspace are proportional to $(x, y, z) = R\sin\psi(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, so within this subspace the vector points radially outward from the origin. Similarly, as θ is varied from θ to $\theta + \mathrm{d}\theta$, $\vec{r}$ varies from $\vec{r}$ to $\vec{r} + \mathrm{d}\vec{r}_\theta$, where

$$\begin{aligned}\mathrm{d}\vec{r}_\theta &= \left(\frac{\partial x}{\partial\theta}, \frac{\partial y}{\partial\theta}, \frac{\partial z}{\partial\theta}, \frac{\partial w}{\partial\theta}\right)\,\mathrm{d}\theta \\ &= R\sin\psi(\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta, 0) \ .\end{aligned} \tag{5.16}$$

This time there is no w-component, and we know that varying θ does not change $x^2 + y^2 + z^2$, and therefore the components within the x-y-z subspace make a tangential vector. Since a tangential vector is orthogonal to a radial vector, it follows that $\mathrm{d}\vec{r}_\psi \cdot \mathrm{d}\vec{r}_\theta = 0$, which is what we wanted to prove. The geometrical argument is easily verified by straightforward calculation:

$$\begin{aligned}\mathrm{d}\vec{r}_\psi \cdot \mathrm{d}\vec{r}_\theta = R^2 \sin\psi\cos\psi[&\sin\theta\cos\theta\cos^2\psi \\ &+ \sin\theta\cos\theta\sin^2\phi - \sin\theta\cos\theta + 0] = 0 \ .\end{aligned} \tag{5.17}$$

A similar argument guarantees that $\mathrm{d}\vec{r}_\psi$ is also orthogonal to $\mathrm{d}\vec{r}_\phi$, so the justification of Eq. (5.14) is complete.

Remember that the coordinate system that one uses to describe a curved space is totally arbitrary. Another choice that is frequently used to describe this space is to replace ψ by

$$u \equiv \sin\psi \ . \tag{5.18}$$

Note that u is double-valued: as ψ varies over its range from 0 to π, u varies from 0 to 1 and then decreases back to 0. The new metric can then be found by noting that

$$\mathrm{d}u = \cos\psi\,\mathrm{d}\psi = \sqrt{1-u^2}\,\mathrm{d}\psi \ , \tag{5.19a}$$

and so

$$\mathrm{d}\psi^2 = \frac{\mathrm{d}u^2}{1-u^2} \ , \tag{5.19b}$$

and then

$$\mathrm{d}s^2 = R^2\left\{\frac{\mathrm{d}u^2}{1-u^2} + u^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right\} \ . \tag{5.20}$$

In these coordinates it is particularly easy to see that in a small region about the origin, i.e., for $|u| \ll 1$, the u^2 in the denominator can be ignored, and the metric becomes the metric for Euclidean space in spherical polar coordinates. This is just an example of the general principle introduced by Gauss: as long as $\mathrm{d}s^2$ is expressed as a quadratic function of the coordinate differentials, then in any infinitesimal region it is possible to find coordinates for which the metric is Euclidean.

The geometry of this space will be pursued further in the next problem set.

IMPLICATIONS OF GENERAL RELATIVITY:

Eqs. (5.14) or (5.20) describe a curved three-dimensional space which is finite but without boundary. The length scale of this space is described by the parameter R, which can have any value. Since R corresponds to the radius of the sphere as embedded in the four-dimensional space, we will refer to R as the radius of curvature of the space.

Since general relativity describes gravity as a distortion of the spacetime metric, however, one might expect that the dynamics of general relativity would determine the curvature of the space, and hence determine the quantity R. The calculations are beyond these lectures, but the result is simple. General relativity requires that the geometry of the universe be non-Euclidean, except for the special case in which the parameter k defined in Lecture Notes 3 is zero. This is why the $k = 0$ model is called flat. When $k > 0$, which we have been calling a closed universe, general relativity requires that the geometry be a closed three-dimensional space, as described by the metric of Eqs. (5.14) or (5.20). Thus, if gravity is strong enough to cause the universe to recollapse, then it is also strong enough to curve the universe back on itself to create a universe that is finite but unbounded.*

Using Newtonian arguments, we have already calculated how the size of the model universe changes with time, proportional to the scale factor $a(t)$. The Friedmann equations that we obtained are identical to the predictions of general relativity, so the size of the universe will be proportional to the scale factor $a(t)$ that we already calculated. For the closed universe geometry, however, the size of the universe is proportional to the radius of curvature R, so consistency requires that R must be proportional to $a(t)$. Furthermore, we recall that the value of $a(t)$ depends on the size of the "notch." The radius of curvature R, however, is a physical length that must be measured in physical distance units, such as meters. Thus, dimensional consistency requires that $R(t)$ to be proportional to $a(t)/\sqrt{k}$, which also has the units of physical length. The constant of proportionality is fixed by the details of general relativity, but the answer is that the constant of proportionality is 1:

$$R^2(t) = \frac{a^2(t)}{k} \ . \tag{5.21}$$

Although the quantity $a^2(t)/k$ has been described in the context of a purely Newtonian calculation, the speed of light was inserted into the definition of k, which was given by Eq. (3.30) as

$$k = -\frac{2E}{c^2} \ , \quad \text{where} \ \ E = \frac{1}{2}\dot{a}^2 - \frac{4\pi}{3}\frac{G\rho_i}{a} \ .$$

* Warning: the simple correspondence between the closure of the universe in time and the closure of the universe in space holds for matter-dominated universes, and even for universes containing arbitrary mixes of matter and radiation. However, when we explore the consequences of a nonzero cosmological constant in Lecture Notes 7, we will find that the relation no longer holds. Universes which are spatially closed might nonetheless expand forever, and universes which are spatially open might nonetheless recollapse.

Thus Eq. (5.21) can be written as

$$R^2(t) = \frac{a^2(t)c^2}{2E} \ ,$$

which shows that curvature is explicitly a relativistic effect. In the nonrelativistic limit where c becomes infinitely large compared to all other velocities, $R(t)$ will approach infinity. Thus in the nonrelativistic limit the radius of curvature of the universe approaches infinity, so the space becomes closer and closer to Euclidean. (Note that the surface of a sphere of infinite radius is actually a plane.)

One can then rewrite the equations of evolution in terms of $R(t)$. Using

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3}G\rho - \frac{kc^2}{a^2} \tag{5.22}$$

from Eqs. (3.25) and (3.31), one has

$$\boxed{H^2 = \left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi}{3}G\rho - \frac{c^2}{R^2} \ .} \tag{5.23}$$

To express the value of $R(t)$ in terms of observables, one can replace ρ by $\Omega\rho_c$, where ρ_c is given by $3H^2/(8\pi G)$ as in Eq. (3.33). One then has

$$\boxed{R = \frac{cH^{-1}}{\sqrt{\Omega - 1}} \ ,} \tag{5.24}$$

which is the same as Eq. (4.32). Note that as Ω becomes closer to one (approaching from above), $R(t)$ becomes larger and larger, so the space becomes closer and closer to Euclidean. In addition, Eq. (5.24) shows explicitly that $R(t)$ is proportional to c, as we discussed in the previous paragraph. Thus, if the speed of light is taken to be infinitely larger than all other velocities, then again the space becomes Euclidean. Curvature is therefore a relativistic effect.

THE ROBERTSON-WALKER FORM OF THE METRIC:

When Eq. (5.21) is substituted into Eq. (5.20), the resulting metric is given by

$$\mathrm{d}s^2 = \frac{a^2(t)}{k}\left\{\frac{\mathrm{d}u^2}{1-u^2} + u^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right\} \ , \tag{5.25}$$

which is a little more complicated than necessary. It is convenient to replace the radial coordinate u (where $u \equiv \sin\psi$) with a new radial coordinate r defined by

$$r \equiv \frac{u}{\sqrt{k}} \equiv \frac{\sin\psi}{\sqrt{k}} \ . \tag{5.26}$$

Then $\mathrm{d}r = k^{-1/2}\mathrm{d}u$, and the metric can be rewritten as

$$\boxed{\mathrm{d}s^2 = a^2(t)\left\{\frac{\mathrm{d}r^2}{1-kr^2} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right\} \ .} \tag{5.27}$$

This is the standard form, called the Robertson-Walker metric. Since the coordinate r is proportional to u, and u is double-valued, so is r. That is, $r = 0$ at the center of the coordinate system, which is identified with the north pole of the sphere that describes the closed universe. As r grows the point described by (r, θ, ϕ) moves away from the north pole, and r reaches its maximum value of $1/\sqrt{k}$ when the point reaches the equator of the sphere. If one continues to move the point in the same direction, then r decreases back to zero as the point moves from the equator to the south pole, where r again is zero.

THE OPEN UNIVERSE:

We have seen that when $k > 0$ the universe is spatially closed (finite volume), and that it approaches an infinite volume Euclidean space as $k \to 0$ (i.e., in this limit the radius of the sphere approaches infinity). What happens if $k < 0$?

As you have probably learned from your experience in physics, in many cases the same equations will hold whether the variables that occur in those equations are positive or negative. Thus, we might expect that the formulas derived above would be valid for $k < 0$, and this is indeed the case. However, there is one complication which should be pointed out. Above we made the change of variables given by Eq. (5.26), involving the quantity $\sqrt{k}$. This quantity would be imaginary if k were negative, and thus it would not be possible for both u and r to be real. One can see from Eq. (5.25) that the metric in terms of u is pathological when k is negative, since $\mathrm{d}s^2$ is not positive definite. For $u < 1$ it is in fact negative definite, and for $u > 1$ the sign is indeterminate, since the angular pieces contribute negatively while the radial piece contributes positively. Thus, it seems clear that the u variable must be discarded when $k < 0$. On the other hand, the metric in the form of Eq. (5.27) remains perfectly well behaved for negative values of k. To minimize the possible confusion of dealing with negative quantities, we can define $\kappa = -k$, and rewrite the Robertson-Walker metric (5.27) for open universes as

$$\boxed{\mathrm{d}s^2 = a^2(t)\left\{\frac{\mathrm{d}r^2}{1+\kappa r^2} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right\} \ .} \tag{5.28}$$

(Open universe, $\kappa > 0$)

While it is reasonable to assume that Eq. (5.28) is correct, our derivation was certainly far from rigorous. I will not try to give a rigorous derivation, but I will try at least to sketch how a rigorous derivation could be constructed. If we wanted to be more rigorous, we would begin by summarizing the goal: to construct a metric describing a homogeneous and isotropic space. While the θ and ϕ angular coordinates are not very obviously isotropic, we are sufficiently familiar with this construction to be convinced that the angular dependence of the metric above is isotropic. Although the coordinate system makes the north pole ($\theta = 0$) look like a special direction, we know that the coordinates could be redefined to put the north pole of the coordinate system at any angle. The homogeneity of the Robertson-Walker metric is similar, but less familiar to us. For the closed Robertson-Walker metric we know that the space is homogeneous, because we derived the metric by starting with the manifestly homogeneous 3-dimensional sphere embedded in four Euclidean dimensions. But the Robertson-Walker coordinates make the origin ($r = 0$) look special, just as the angular coordinates make the north pole look special. As in the case of the angular coordinates, we know that the origin of the closed Robertson-Walker coordinate system is not really special, and that we could redefine our coordinate system so that the origin can be put at any location.

To show that the open Robertson-Walker metric in Eq. (5.28) is homogeneous, we would start by studying the homogeneity of the closed universe metric in detail, turning the verbal statements in the previous paragraph into an explicit set of coordinate transformations that show how to move the origin to an arbitrary point. The details become rather complicated, as indeed they would if we tried to explicitly show how to construct a coordinate transformation to move the north pole of the (θ, ϕ) angular coordinates. Nonetheless, once the equations are written, it would become clear that they are just a set of algebraic relations: if they hold for all positive k, they will necessarily hold for negative k as well. Thus the same algebra that shows the closed Robertson-Walker universe to be homogeneous also shows that the open metric is homogeneous.

We will not try to show it, but it can be shown that any three-dimensional homogeneous and isotropic space can be described by the Robertson-Walker metric, Eq. (5.27), where k can be positive, negative, or zero. Other coordinate systems are of course possible, but geometrically different spaces are not.

Note that the sign of k affects the question of whether the space is finite or infinite. For $k > 0$, Eq. (5.27) implies that something peculiar happens when $kr^2 = 1$, at which point the metric is singular. Since r is related to the original ψ coordinate by $r = \sin(\psi)/\sqrt{k}$, one sees that this value of the radius variable corresponds to $\psi = \pi/2$, and hence the equator of the original sphere embedded in four dimensions. There is nothing singular about the space, but the metric becomes singular because the coordinate r behaves peculiarly, reaching a maximum value. Beyond the equator, r must get smaller and then approach zero at the "south pole" ($x = 0, y = 0, z = 0, w = -R$). Thus, the

space is finite. However, if $k < 0$ then the metric is given by Eq. (5.28), which remains perfectly well-defined for all values of r, and thus the range of the r-coordinate is infinite. This does not by itself prove that the space is infinite, since the value of a coordinate is not directly measurable. However, one can calculate the physical distance from the origin to a point with radial coordinate r by integrating the metric of Eq. (5.28) along a radial path (with $\mathrm{d}\theta = \mathrm{d}\phi = 0$):

$$\ell_{\mathrm{phys}}(r) = a(t) \int_0^r \frac{\mathrm{d}r'}{\sqrt{1+\kappa r'^2}} = \frac{\sinh^{-1}\sqrt{\kappa}\, r}{\sqrt{\kappa}} \ , \tag{5.29}$$

where the integration can be carried out by substituting $r' = \sinh(\psi)/\sqrt{\kappa}$. Since the inverse sinh function can become arbitrarily large, the space is infinite.

The G-B-L geometry discussed in the introduction is simply the two-dimensional version of the space of an open universe at some arbitrary fixed time. The realization by Klein described in Eqs. (5.1) and (5.2) represents a somewhat peculiar choice of coordinate system.

THE GENERALIZATION FROM SPACE TO SPACETIME

Eq. (5.27) actually shows only a **spatial** metric, while I said earlier that general relativity describes the gravitational field in terms of a **spacetime** metric. To put the spacetime metric into context, we recall that in special relativity it is possible to define a Lorentz-invariant separation between two events. Specifically, if the coordinates of an event A are (x_A, y_A, z_A, t_A), and the coordinates of an event B are (x_B, y_B, z_B, t_B), then the Lorentz-invariant separation between A and B is defined by

$$s^2 \equiv (x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2 - c^2 (t_A - t_B)^2 \ . \tag{5.30}$$

By saying that this expression is Lorentz-invariant, we mean that it has the same value in all inertial references frames, even though the individual terms may very well have different values.

While the value of s^2 is the same in all inertial frames, the intuitive meaning of s^2 is easiest to see by considering its value in particular frames. If $s^2 > 0$, then the separation between the events is called *spacelike*. In that case it is always possible to find an inertial reference frame in which the two events are simultaneous, and in that frame s is equal to the spatial distance between the two events. Equivalently, we can say that it is always possible to find an inertial observer to whom the two events appear simultaneous. s is then equal to the distance between these events, as measured by a ruler at rest with respect to this observer. s can be called the *proper distance* between the events. If $s^2 < 0$ then the separation is called *timelike*, and in that case it is always possible to find an

inertial observer to whom it appears that the two events occur at the same position. If she defines

$$s^2 = -c^2\tau^2 \ , \tag{5.31}$$

then τ is the time separation between the events when measured on her clock. τ is often called the *proper time* between the two events. Note that if the two events happen to the same object, such as two flashes of the same strobe light, and the object is moving at constant velocity, then the proper time between the flashes is just the time as measured by a clock at rest with respect to the strobe light. If $\mathrm{d}s^2 = 0$, then the separation between the two events is called *lightlike*, and in that case a light pulse leaving the earlier event will arrive at the location of the latter event just as it occurs.

If you are not familiar with the Lorentz-invariant separation, you may want to look at the Appendix at the end of this set of Lecture Notes. There I start with the three basic effects of special relativity, as described in Lecture Notes 1, and show how to construct the Lorentz transformation. The Lorentz transformation is the set of equations that describe how to relate the coordinates of an event in two different inertial coordinate frames, where one frame is moving relative to the other. Using the Lorentz transformation, the Appendix goes on to show that the expression defined by Eq. (5.30) is truly Lorentz-invariant. (For purposes of this course, however, the Appendix can be considered outside the course requirements. It is okay for you to just accept the result that s^2 is Lorentz-invariant.)

The spacetime metric of general relativity is the curved-spacetime generalization of the Lorentz-invariant separation of special relativity. Following the ideas of Gauss discussed near the beginning of these lecture notes, we will restrict our attention to describing the separation between two infinitesimally separated spacetime points (x, y, z, t) and $(x + \mathrm{d}x, y + \mathrm{d}y, z + \mathrm{d}z, t + \mathrm{d}t)$. For special relativity the metric of Eq. (5.30) reduces in the infinitesimal case to

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 - c^2\,\mathrm{d}t^2 \ , \tag{5.32}$$

which is known as the Minkowski metric. Continuing with Gauss's approach, we insist — even when we describe arbitrary curved spacetimes — that $\mathrm{d}s^2$ be expressed as a quadratic expression in the coordinate differentials. This implies (although we will not show it) that for any spacetime point P it is always possible to choose a coordinate system (x', y', z', t') so that the metric reduces to the Minkowski metric in an infinitesimal region around that point. If the spacetime is curved the metric will not have the Minkowski form outside this infinitesimal region, however, so the metric will be called *locally Minkowskian* at the point P.

In curved spacetimes there is no coordinate system in which the metric has the Minkowski form everywhere. Thus, to infer the separation between two points one must know not only the values of the coordinates, but also the metric. The coordinates are

then not themselves direct measurements of distance, but instead are just an arbitrary way of labeling points. Since one needs to introduce a metric for any coordinate system, there is nothing that forces us to use any particular coordinate system or set of coordinate systems. This is different from special relativity, where the metric (5.32) is valid only for a special class of coordinate systems, called inertial coordinate systems, which are related to each other by a special class of transformations, called Lorentz transformations. If I were to replace the coordinate x by $x' \equiv \sinh x$, then the metric would no longer look like Eq. (5.32). The coordinate transformation $x' \equiv \sinh x$ is therefore not allowed in the standard formulation of special relativity (although one could use the formalism of general relativity for a special relativity problem if one chose to.) In general relativity, on the other hand, there is usually no coordinate system in which the metric is particularly simple, so the formalism is designed to allow any choice of coordinates, and hence any kind of coordinate transformation. In general relativity, therefore, $x' = \sinh x$ is a perfectly acceptable coordinate transformation. As long as the coordinates allow a unique way to label each point in spacetime, they are acceptable. If I change coordinate systems, I can always change the metric so that the value of ds^2 between any two points remains the same. For this reason ds^2 is said to be *coordinate-invariant.*

When we introduced the two-dimensional spatial metric in Eq. (5.3), we assumed that ds^2 represented the distance between the two points, where the meaning of "distance" was no different from what it would mean in Euclidean geometry — it is what one would measure with a ruler. Here we are trying to generalize this method, so we want to define ds^2 to have the same meaning it would have in special relativity. In special relativity we were able to define ds^2 in terms of the observations made by inertial observers, which means observers for whom the law of inertia is valid, which in turn means observers to whom no net force is applied. In general relativity, forces other than gravity are treated in essentially the same way as in special relativity, so there is no problem defining what it means for the net **nongravitational** force on an observer to vanish. But gravity is trickier. Consider, for example the homogeneously expanding universe that we discussed in Lecture Notes 2, 3, and 4. If I am moving with the expansion of the universe (i.e., if I am at rest with respect to the comoving coordinate system), then I can view myself as being at rest. If I look at the distant galaxies around me, however, they will appear to be slowing in their outward motion, and hence accelerating towards me, under the influence of gravity. But an observer on one of those galaxies would consider himself to be at rest, and I would appear to be accelerating. According to general relativity both points of view are equally valid, so the concept of gravitational acceleration becomes relative.

Another simple and famous example that illustrates the relative nature of gravitational forces is the elevator (thought) experiment. Suppose a man, holding a bag of groceries, is standing in an elevator. Now suppose that the elevator cables are cut, and the elevator free falls downward without friction or air resistance. The man will then accelerate downward with the same acceleration as the elevator, and he will feel no force between his feet and the elevator floor. If he lets go of the bag of groceries, the bag

would not move relative to him, but would appear to float in front of him. In the frame of the Earth, all the objects (the elevator, the man, and the groceries) are accelerating downward under the force of gravity. But in the frame of the elevator, everything appears weightless. (Everything is is weightless until the big crunch occurs in the building's basement — but remember, this in only a thought experiment. No living creatures were harmed in the writing of this paragraph.)

We are accustomed to thinking of the frame of the Earth as being the correct "physical" description, because the frame of the Earth is nearly inertial over a large region of space and time. In the context of general relativity, however, both frames are equally correct. Thus, the presence or absence of gravity is determined by which frame of reference we are using. This idea in fact is one of the foundational concepts of general relativity, known as the *equivalence principle.* The physics of the accelerating frame of the elevator, with no gravity, is equivalent to the physics in the rest frame of the Earth, with its gravitational field. The equivalence principle says that it is always possible, in a sufficiently small region, to find a frame of reference in which the force of gravity is absent.

The bottom line here is that if we are trying to generalize the notion of an inertial observer in special relativity, we cannot insist that the gravitational force on the observer vanishes, because this condition will appear to hold in some coordinate systems but not others. So, instead we insist only that the net **nongravitational** force on the observer vanish, and we say that such an observer is *free-falling.* Note that the man in the falling elevator is free-falling, while a man standing in an elevator that is at rest with respect to the Earth is not. In the latter case the floor is pushing upward on the man's feet, so the net nongravitational force is nonzero.

With the replacement of inertial observers by free-falling observers, the meaning of ds^2 in general relativity is the same as what we had in special relativity. If the value of ds^2 calculated between two events is positive, then there is always a free-falling observer to whom the events appear simultaneous. In this case, the proper distance ds between the events is the distance between them, as measured by a ruler at rest relative to this free-falling observer. If $ds^2 < 0$, then there is always a free-falling observer for whom the events appear to happen at the same location. One then defines

$$ds^2 \equiv -c^2\, d\tau^2 \ , \tag{5.33}$$

as in Eq. (5.31), where $d\tau$ is again called the proper time interval between the events. It is the time interval between the two events that would be measured by a clock carried by the free-falling observer mentioned above. If $ds^2 = 0$, then the two events can be connected by a light pulse, which leaves the first event and arrives at the second.*

* The concept of a free-falling observer is intimately linked to the concept of a locally Minkowskian coordinate system, so the meaning of ds^2 could also have been explained in terms of these coordinate systems. The free-falling observers are those that are at rest or moving at a constant velocity relative to a coordinate system that is locally Minkowskian at the location of the observer.

INCLUSION OF TIME IN THE ROBERTSON-WALKER METRIC

What happens when we add time to the Robertson-Walker metric of Eq. (5.27)? In general the answer can depend on how we choose to define our time variable, but we will hold with the choice called *cosmic time*, which we discussed in Lecture Notes 2 (in a section called "The Synchronization of Clocks"). We concluded there that it is possible to define a cosmic time variable t which can be measured locally. That is, each observer who is at rest with respect to the matter in her vicinity can measure t on her own wristwatch. The wristwatches throughout the universe can be synchronized, once and for all, by some choice of a cosmic event. For example, we can all agree to set our wristwatches to read 12 billion years when the temperature of the cosmic microwave background radiation reaches 3.0 K, or when the Hubble parameter reaches 85 km-sec^{-1}-Mpc^{-1}. Once the watches are synchronized, we argued that the homogeneity of the universe guarantees that they will stay synchronized: all watches will read the same time when the cosmic background radiation temperature reaches 2.0 K, or when the Hubble parameter reaches 75 km-sec^{-1}-Mpc^{-1}. In practice we usually define the synchronization of cosmic time so that $t = 0$ corresponds to our best estimate of when $a(t)$ was equal to zero, and the Hubble parameter and temperature were infinite. (More precisely, we choose $t = 0$ to correspond to the time when $a(t)$, as extrapolated in our mathematical model, was equal to zero. As discussed in *The Big Bang Singularity* section of Lecture Notes 4, there is no reason for us to have confidence in this extrapolation.)

I think it will be most straightforward for me to write the answer first, and then explain why it could not have been anything different. If the time variable t is taken to be cosmic time, and the metric is to be homogeneous and isotropic, then it can always be written as

$$\boxed{\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a^2(t)\left\{\frac{\mathrm{d}r^2}{1-kr^2} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right\}\ .} \tag{5.34}$$

So, why does this have to be the answer? Consider first the case in which the separation $\mathrm{d}t = 0$ (i.e., when the two events whose separation we are calculating have the same time coordinate). In that case Eq. (5.34) reduces to our previous expression, Eq. (5.27). Since we have already stated (albeit without proof) that Eq. (5.27) describes the most general possible three-dimensional space that is homogeneous and isotropic, the answer for the $\mathrm{d}t = 0$ case is settled. We could of course choose other coordinates that would make the spatial part of Eq. (5.34) look different, but Eq. (5.34) as written describes the most general possible geometry.

Now consider the interval defined by $\mathrm{d}t \neq 0$, but $\mathrm{d}r = \mathrm{d}\theta = \mathrm{d}\phi = 0$. This represents the motion of a comoving observer for an increment of coordinate time $\mathrm{d}t$. There are

no nongravitational forces acting on the comoving observer, so she is also a free-falling observer. This is a timelike separation, so we use the definition $\mathrm{d}s^2 = -c^2\,\mathrm{d}\tau^2$ from Eq. (5.33), and we deduce that $\mathrm{d}t = \mathrm{d}\tau$, where $\mathrm{d}\tau$ is the time measured on the comoving observer's wristwatch. But an interval of cosmic time is defined as the interval measured on the wristwatches of comoving observers, so the metric of Eq. (5.34) implies that t is precisely the time variable that we have called cosmic time. Note that if the coefficient of the $\mathrm{d}t^2$ term in the metric were anything other than $-c^2$, we would have found that the time coordinate interval $\mathrm{d}t$ is proportional to wristwatch time, but not equal to it.

We have now verified that the terms that are present in Eq. (5.34) must have the forms that they have. But what about the possibility of adding other terms. Since the metric is required to be a quadratic function of the coordinate differentials, the only possible new terms that could be added are terms proportional to $\mathrm{d}t\,\mathrm{d}r$, $\mathrm{d}t\,\mathrm{d}\theta$, or $\mathrm{d}t\,\mathrm{d}\phi$. (Recall that terms like $\mathrm{d}r\,\mathrm{d}\theta$ would contribute even when the time is fixed, $\mathrm{d}t = 0$, so such terms have already been ruled out by the statement that Eq. (5.27) is the most general possible homogeneous and isotropic space.) Let us consider first the possibility of adding a term $\mathrm{d}r\,\mathrm{d}t$ to the metric. The claim is that such a term would violate our assumption of isotropy, because it would create a distinction between the direction of increasing and decreasing r. To see this, consider two observers, Tweedledee and Tweedledum, who both start at $r = r_0$ at time $t = t_0$. Tweedledee is moving outward and Tweedledum is moving inward, both with coordinate speed $\mathrm{d}r/\mathrm{d}t = v$ (and with fixed values of θ and ϕ). At $t = t_0 + \mathrm{d}t$, Tweedledee will be located at $r = r_0 + v\,\mathrm{d}t$, while Tweedledum will be located at $r = r_0 - v\,\mathrm{d}t$. Thus the displacement vector of Tweedledee has $\mathrm{d}r > 0$, while that of Tweedledum has $\mathrm{d}r < 0$, and both have the same $\mathrm{d}t$. The hypothetical new term will therefore contribute to $\mathrm{d}s^2$ with opposite signs for the two cases, so the values of $\mathrm{d}s^2$ will be different for Tweedledee and Tweedledum. Since $\mathrm{d}s^2 = -c^2\,\mathrm{d}\tau^2$, and $\mathrm{d}\tau$ is the wristwatch time that each will measure, we conclude that each will have a different wristwatch time at the end of this interval. When they each compare with the comoving observers whose wristwatches read cosmic time, $t = t_0 + \mathrm{d}t$, the two will see different discrepancies. This means that there is a Tweedledee/Tweedledum asymmetry, but the only difference in the setup was their direction of travel. Thus, the addition of such a term would be a violation of isotropy. An identical argument can be made for $\mathrm{d}t\,\mathrm{d}\theta$ or $\mathrm{d}t\,\mathrm{d}\phi$ terms, so we conclude that Eq. (5.34) is necessarily the right answer.

EQUATIONS FOR A GEODESIC

As was stated earlier, in general relativity a freely falling particle is assumed to travel on a geodesic of the curved spacetime. Stated more precisely, the equations of motion in general relativity are derived from the assumption that the path length from the initial point to the final point should have a vanishing derivative with respect to any variation of the path that does not vary the endpoints. If the meaning of this statement is not clear to you at this point, then don't worry yet — it will hopefully become clear once we define some notation.

We will start by deriving the equation for a geodesic in a two-dimensional space with a positive-definite metric (i.e., with all lengths positive). The metric will be assumed to have the general form specified by Gauss, and given earlier as Eq. (5.3):

$$\mathrm{d}s^2 = g_{xx}\mathrm{d}x^2 + g_{xy}\mathrm{d}x\,\mathrm{d}y + g_{yx}\mathrm{d}y\,\mathrm{d}x + g_{yy}\mathrm{d}y^2 \ , \tag{5.3}$$

where g_{xx}, g_{xy}, g_{yx}, and g_{yy} are functions of position (x,y) and are together called the metric of the space. As explained earlier, we take $g_{yx} \equiv g_{xy}$.

The first step will be to simplify the notation, since Eq. (5.3) requires a lot of writing. To start, rename the coordinate x as x^1, and rename y as x^2. Then the two coordinates together can be described as x^i, where i is understood to take on the values 1 and 2. Eq. (5.3) can then be rewritten as

$$\mathrm{d}s^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} g_{ij}(x^k)\,\mathrm{d}x^i\,\mathrm{d}x^j \ , \tag{5.35}$$

where I write the metric as $g_{ij}(x^k)$ to indicate explicitly that it is a function of all of the coordinates x^k. One further simplification is known as the Einstein summation convention. This is no doubt Einstein's most important contribution to ecology, saving barrels of ink and tons of paper each year. The convention stipulates that whenever an index is repeated, it is automatically summed over the standard range (which in this case is from 1 to 2). Using this convention, Eq. (5.35) can be written compactly as

$$\mathrm{d}s^2 = g_{ij}(x^k)\,\mathrm{d}x^i\,\mathrm{d}x^j \ . \tag{5.36}$$

(In using this notation, it is important that the context makes it clear that the superscript i in x^i is to be interpreted as an index, and not a power. You might wonder why people tolerate this confusion, when it could be avoided by writing all indices as subscripts. The reason is that curved space geometers find it useful to use both superscripts and subscripts to denote indices. Quantities with upper indices (superscripts) are called contravariant, and quantities with lower indices (subscripts) are called covariant. These indices can always be arranged so that each summation over a repeated index involves one upper and one lower index, as has been done in Eq. (5.36). To understand fully the meaning of upper and lower indices, one must study how the equations of non-Euclidean geometry are transformed by a redefinition of the coordinate system. We will skip this topic, but I point out that the formalism is constructed so that the rules of transformation are indicated by whether the indices are upper or lower. Furthermore, the transformation rules guarantee that any sum over a repeated index, with one upper and one lower, is invariant under a change of coordinates.)

Now we can state the geodesic problem: given two points x^i_A and x^i_B, what equation determines the geodesic, or shortest path, between the two points? (In this case it will be the shortest path.)

An arbitrary path can be described by a function $x^i(\lambda)$, where λ is a parameter which we take to run between 0 and some final value λ_f. Thus, the statement that the path runs from x^i_A to x^i_B translates into the equations

$$x^i(0) = x^i_A, \qquad x^i(\lambda_f) = x^i_B \ . \tag{5.37}$$

Now focus attention on an infinitesimal segment of the curve, from λ to $\lambda + \mathrm{d}\lambda$. The change in the values of the two coordinates over this segment is given by

$$\mathrm{d}x^i = \frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\mathrm{d}\lambda \ . \tag{5.38}$$

Since $\mathrm{d}\lambda$ is infinitesimal, one need not consider terms in Eq. (5.38) that are higher order in $\mathrm{d}\lambda$. Combining this equation with Eq. (5.36), one has

$$\mathrm{d}s^2 = g_{ij}\big(x^k(\lambda)\big)\,\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\,\mathrm{d}\lambda^2 \ ,$$

and then

$$\mathrm{d}s = \sqrt{g_{ij}\big(x^k(\lambda)\big)\,\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}}\,\mathrm{d}\lambda \ . \tag{5.39}$$

The total length of the path is then

$$S[x^i(\lambda)] = \int_0^{\lambda_f} \sqrt{g_{ij}\big(x^k(\lambda)\big)\,\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}}\,\mathrm{d}\lambda \ . \tag{5.40}$$

The path length $S[x^i(\lambda)]$ is actually a function of the function $x^i(\lambda)$. A function of a function is usually called a ***functional***, and the argument of the functional is usually enclosed in square brackets.

Next we consider how the path length will vary if the path is changed infinitesimally. To formulate this precisely, we write the equation for a nearby path, with the same endpoints, as

$$\tilde{x}^i(\lambda) = x^i(\lambda) + \alpha w^i(\lambda) \ , \tag{5.41a}$$

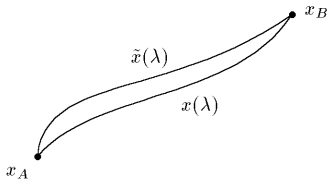

Figure 5.11: A path $x(\lambda)$ and a small variation of it, $\tilde{x}(\lambda)$.

where α is a number (which we will take to be small), and the path variation function $w^i(\lambda)$ is required to satisfy

$$w^i(0) = 0 \ , \qquad w^i(\lambda_f) = 0 \ , \tag{5.41b}$$

so that the new path $\tilde{x}^i(\lambda)$ has the same endpoints as original path $x^i(\lambda)$. The rule for a geodesic is that no matter how the path is varied, the original length is a minimum. This implies that if $w^i(\lambda)$ is held fixed, for any value that satisfies Eq. (5.41b), the path length of $\tilde{x}^i(\lambda)$ should have a minimum at $\alpha = 0$. Thus,

$$\left.\frac{\mathrm{d}\, S\left[\tilde{x}^i(\lambda)\right]}{\mathrm{d}\alpha}\right|_{\alpha=0} = 0 \qquad \text{for all } w^i(\lambda) \ . \tag{5.42}$$

The problem now is simply to calculate the derivative in Eq. (5.42). To simplify the notation, we define

$$A(\lambda,\alpha) = g_{ij}\left(\tilde{x}^k(\lambda)\right)\frac{\mathrm{d}\tilde{x}^i}{\mathrm{d}\lambda}\frac{\mathrm{d}\tilde{x}^j}{\mathrm{d}\lambda} \ , \tag{5.43}$$

so we can write

$$S\left[\tilde{x}^i(\lambda)\right] = \int_0^{\lambda_f}\sqrt{A(\lambda,\alpha)}\,\mathrm{d}\lambda \ . \tag{5.44}$$

Note that the derivative can be taken inside the integral that defines $S[\tilde{x}^i(\lambda)]$, since the limits of integration do not depend on α. Using the chain rule of differentiation, we find

$$\left.\frac{\mathrm{d}}{\mathrm{d}\alpha}g_{ij}\big(\tilde{x}^k(\lambda)\big)\right|_{\alpha=0} = \left.\frac{\partial g_{ij}}{\partial x^k}\right|_{x^k=x^k(\lambda)}\left.\frac{\partial\tilde{x}^k}{\partial\alpha}\right|_{\alpha=0} = \frac{\partial g_{ij}}{\partial x^k}\big(x^i(\lambda)\big)w^k \ , \tag{5.45}$$

where the Einstein summation convention applies to the sum over k. Differentiating Eq. (5.44), one then finds

$$\begin{aligned}\left.\frac{\mathrm{d}S\left[\tilde{x}^i(\lambda)\right]}{\mathrm{d}\alpha}\right|_{\alpha=0} = \frac{1}{2}\int_0^{\lambda_f}\frac{1}{\sqrt{A(\lambda,0)}}\Bigg\{&\frac{\partial g_{ij}}{\partial x^k}w^k\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda} + \\ &+g_{ij}\frac{\mathrm{d}w^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda} + g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}w^j}{\mathrm{d}\lambda}\Bigg\}\mathrm{d}\lambda \ ,\end{aligned} \tag{5.46}$$

where the metric g_{ij} is to be evaluated at $x^k(\lambda)$.

The expression can be further simplified by recognizing that the summed indices are "dummy" indices, in the sense that their names can be changed without changing the value of the expression. (When one does this, of course, it is essential that the name be changed in the same way for each occurrence of the index.) Suppose then that the third

term in curly brackets of the above equation is rewritten by substituting $i \to j$ and $j \to i$. It then becomes identical to the second term, except that the indices on g_{ij} are reversed. But g_{ij} is symmetric in the sense that $g_{ji} = g_{ij}$ (see the remarks following Eq. (5.3)), so the two terms are identical. Thus,

$$\left.\frac{\mathrm{d}S\left[\tilde{x}^i(\lambda)\right]}{\mathrm{d}\alpha}\right|_{\alpha=0} = \frac{1}{2}\int_0^{\lambda_f}\frac{1}{\sqrt{A(\lambda,0)}}\left\{\frac{\partial g_{ij}}{\partial x^k}w^k\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda} + 2g_{ij}\frac{\mathrm{d}w^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\right\}\mathrm{d}\lambda \ . \tag{5.47}$$

The next step is to simplify the dependence on $w^i(\lambda)$. The expression above depends explicitly on both the function $w^i(\lambda)$ and its derivative, but the dependence on the derivative can be removed by an integration by parts. Note that the term

$$\int_0^{\lambda_f}\left[\frac{1}{\sqrt{A}}g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\right]\frac{\mathrm{d}w^i}{\mathrm{d}\lambda}\,\mathrm{d}\lambda$$

can be integrated using

$$\int u\,\mathrm{d}v = -\int v\,\mathrm{d}u + [uv]_{\lambda=0}^{\lambda=\lambda_f} \ ,$$

where

$$u = \frac{1}{\sqrt{A}}g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda} \ , \qquad \mathrm{d}u = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{1}{\sqrt{A}}g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\right]\mathrm{d}\lambda$$

$$\mathrm{d}v = \frac{\mathrm{d}w^i}{\mathrm{d}\lambda}\,\mathrm{d}\lambda \ , \qquad v = w^i \ .$$

The surface term $[uv]_{\lambda=0}^{\lambda=\lambda_f}$ then vanishes, since $w^i(0) = w^i(\lambda_f) = 0$. So,

$$\int_0^{\lambda_f}\left[\frac{1}{\sqrt{A}}g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\right]\frac{\mathrm{d}w^i}{\mathrm{d}\lambda}\,\mathrm{d}\lambda = -\int_0^{\lambda_f}\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{1}{\sqrt{A}}g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\right]w^i\,\mathrm{d}\lambda \ . \tag{5.48}$$

Thus, Eq. (5.47) simplifies to

$$\left.\frac{\mathrm{d}S}{\mathrm{d}\alpha}\right|_{\alpha=0} = \frac{1}{2}\int_0^{\lambda_f}\left\{\frac{1}{\sqrt{A}}\frac{\partial g_{ij}}{\partial x^k}\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}w^k - 2\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{1}{\sqrt{A}}g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\right]w^i\right\}\mathrm{d}\lambda \ .$$

If one also renames the indices in the first term by $i \to j, j \to k, k \to i$, one can write

$$\left.\frac{\mathrm{d}S}{\mathrm{d}\alpha}\right|_{\alpha=0} = \int_0^{\lambda_f}\left\{\frac{1}{2\sqrt{A}}\frac{\partial g_{jk}}{\partial x^i}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\frac{\mathrm{d}x^k}{\mathrm{d}\lambda} - \frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{1}{\sqrt{A}}g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\right]\right\}w^i(\lambda)\,\mathrm{d}\lambda \ . \tag{5.49}$$

The next step is to set the quantity in curly brackets in the expression above equal to zero. To justify this, one must of course realize that the vanishing of an integral

does not in general require that the integrand is zero — that is, it is very easy to find nonzero functions that integrate to zero over some specified range. However, we need to require that the derivative above vanish not merely for some particular value of $w^i(\lambda)$, but rather that it vanish for ***all*** values of $w^i(\lambda)$ that are consistent with Eq. (5.41b). This stronger requirement implies that the integrand must vanish. Note that if the quantity in curly brackets did not vanish, one could choose $w^i(\lambda)$ to equal the quantity in curly brackets, so the integral in Eq. (5.49) becomes the integral of a perfect square. Since then the integrand is nonnegative, the integral can vanish only if the integrand is identically zero. (Technically, the integrand can still be nonzero on a set of measure zero, such as a discrete set of points, since the integral over such a set gives zero in any case. We will restrict ourselves, however, to continuous functions, and then such a quantity must vanish everywhere.) Thus,

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{1}{\sqrt{A}}g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\right] = \frac{1}{2\sqrt{A}}\frac{\partial g_{jk}}{\partial x^i}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\frac{\mathrm{d}x^k}{\mathrm{d}\lambda}\ .} \tag{5.50}$$

The above equation is actually quite complicated, since the quantity A defined by Eq. (5.43) is complicated. However, the equation also has more generality than we really need: as we derived it, it will be valid for any parameterization $x^i(\lambda)$ of the path. If we instead make a specific choice about how the path is to be parameterized, then the equation can be simplified. In particular, we can simplify the equation tremendously by choosing λ to be the path length, as measured along the curve. Recalling that

$$\mathrm{d}s = \sqrt{g_{ij}\big(x^k(\lambda)\big)\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}}\,\mathrm{d}\lambda = \sqrt{A}\,\mathrm{d}\lambda\ ,$$

one sees that $\mathrm{d}\lambda = \mathrm{d}s$ requires

$$A = 1 \qquad (\text{for } \lambda = \text{path length}). \tag{5.51}$$

Then the geodesic equation becomes

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}s}\left[g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}s}\right] = \frac{1}{2}\frac{\partial g_{jk}}{\partial x^i}\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\mathrm{d}x^k}{\mathrm{d}s}\ ,} \tag{5.52}$$

where I have replaced λ by s to indicate clearly that it is the physical path length.

Eq. (5.52) is in many cases the most convenient form of the geodesic equation, but it is nonetheless not the standard way that the geodesic equation is written in general

relativity books. Instead, the standard form is to write an explicit equation for $\mathrm{d}^2x^i/\mathrm{d}s^2$. One begins by expanding the left-hand side of Eq. (5.52), using the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}s}\left[g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}s}\right] = g_{ij}\frac{\mathrm{d}^2x^j}{\mathrm{d}s^2} + \partial_k g_{ij}\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\mathrm{d}x^k}{\mathrm{d}s}\ , \tag{5.53}$$

where I have used the standard abbreviation

$$\partial_k \equiv \frac{\partial}{\partial x^k}\ . \tag{5.54}$$

The geodesic equation then becomes

$$g_{ij}\frac{\mathrm{d}^2x^j}{\mathrm{d}s^2} = \frac{1}{2}\left(\partial_i g_{jk} - 2\partial_k g_{ij}\right)\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\mathrm{d}x^k}{\mathrm{d}s}\ . \tag{5.55}$$

Using the symmetry of the factor on the right, $-2\partial_k g_{ij}$ can be rewritten more symmetrically as $-\partial_k g_{ij} - \partial_j g_{ik}$. Eq. (5.55) can then be turned into an equation of the desired form by inverting the matrix g_{ij} that appears on the left-hand side. One defines g^{ij} as the matrix inverse of g_{ij}, which in index notation translates into the statement

$$g^{i\ell}g_{\ell j} = \delta^i_j\ , \tag{5.56}$$

where δ^i_j denotes the Kronecker δ-function (which is defined to be one if $i = j$, and zero otherwise). One can then change the free index in Eq. (5.55) to ℓ, and then multiply by $g^{i\ell}$. The result is written standardly in the form

$$\boxed{\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2} = -\Gamma^i_{jk}\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\mathrm{d}x^k}{\mathrm{d}s}\ ,} \tag{5.57}$$

where

$$\boxed{\Gamma^i_{jk} = \frac{1}{2}g^{i\ell}\left(\partial_j g_{\ell k} + \partial_k g_{\ell j} - \partial_\ell g_{jk}\right)\ .} \tag{5.58}$$

The quantity Γ^i_{jk} is called the affine connection.

THE SCHWARZSCHILD METRIC

General relativity includes a set of equations known as the Einstein field equations, which describe how a gravitational field is produced by matter. These equations are the analogue of the Maxwell equations of electromagnetism, which describe how an electromagnetic field is produced by charges and currents. The Einstein field equations are beyond the scope of this course, but it will nonetheless be useful to describe some features of the solutions to the field equations.

Of particular interest are the solutions for spherically symmetric objects, such as planets, stars, or black holes. In Newtonian mechanics, you will recall, the gravitational field outside a spherical distribution of matter has the peculiar property that it is independent of the details of the mass distribution. Outside of a spherical distribution, the field is uniquely determined if the total mass is known, independent of how this mass is distributed with radius. In general relativity, it turns out, the same feature is found — the metric is determined solely by the total mass enclosed. The metric for a spherically symmetric distribution of mass, in the region outside the mass, is given by the Schwarzschild metric,

$$\begin{aligned} \mathrm{d}s^2 = -c^2\mathrm{d}\tau^2 = -\left(1-\frac{2GM}{rc^2}\right)c^2\mathrm{d}t^2 + \left(1-\frac{2GM}{rc^2}\right)^{-1}\mathrm{d}r^2 \\ + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2 \ , \end{aligned} \tag{5.59}$$

where M is the total mass of the object, and θ and ϕ are the usual polar coordinates. Their range is given by $0 \le \theta \le \pi$, $0 \le \phi < 2\pi$, and $\phi = 2\pi$ is identified with $\phi = 0$.

Note that the metric becomes singular at $r = 2GM/c^2$, which is known as the Schwarzschild radius:

$$R_S = \frac{2GM}{c^2} \ . \tag{5.60}$$

A metric is said to be singular if any of the coefficients become infinite, or if any of the coefficients vanish; in this case both happen: the coefficient of the $\mathrm{d}t^2$ term vanishes at the Schwarzschild radius, and the coefficient of $\mathrm{d}r^2$ becomes infinite. The singularity at the Schwarzschild radius, however, does not indicate any true singularity in the structure of space. If a person or instrument fell through the Schwarzschild radius, nothing peculiar would be felt. In this case the singularity is caused only by the choice of the coordinate system, and other coordinate systems can be constructed for which there is no singularity. In this course, however, we will not have time to look at such coordinate systems. The Schwarzschild metric is also singular at $r = 0$; unlike the singularity at $r = R_S$, the singularity at $r = 0$ is a true physical singularity. Physically measurable quantities, such as the tidal forces associated with nonuniform gravitational fields, become infinite at $r = 0$.

Although the singularity at $r = R_S$ is only an artifact of the coordinate system, it can be shown nonetheless that $r = R_S$ represents the point of no return for an object falling into a black hole. If any object (even a photon) falls inside the Schwarzschild radius, then it will never be able to escape. Thus, an object that is contained within its Schwarzschild radius is called a black hole. The sphere at $r = R_S$ is called the "Schwarzschild horizon," meaning that it is impossible, from the outside, to see anything beyond $r = R_S$.

The distinction between a black hole and a star is simply the question of whether this Schwarzschild horizon exists. If the matter extends to radii beyond the value of R_S indicated by Eq. (5.60), then the Schwarzschild metric will not be valid at the Schwarzschild radius. In this case the horizon may or may not exist, depending on the distribution of matter inside the object. However, if the mass distribution is so compact that it is contained within the Schwarzschild radius, then the Schwarzschild metric will describe the space outside of the matter, and the Schwarzschild horizon will be guaranteed to exist.

Just for orientation, we can compute the Schwarzschild radius of the sun, which has a mass of 1.989×10^{30} kg. Thus,

$$\begin{aligned} R_{S,\odot} &= \frac{2 \times 6.673 \times 10^{-11}\ \text{m}^3\text{-kg}^{-1}\text{-s}^{-2} \times 1.989 \times 10^{30}\ \text{kg}}{(2.998 \times 10^8\ \text{m-s}^{-1})^2} \\ &= 2.95\ \text{km} \ . \end{aligned} \tag{5.61}$$

So if the sun were compressed to a size smaller than 2.95 km, it would become a black hole.

GEODESICS IN THE SCHWARZSCHILD METRIC

Our purpose in introducing the Schwarzschild metric is mainly to provide an example of the calculation of a geodesic in a realistic general relativity setting.

In this section we will calculate the geodesic, and hence the trajectory, for a particle that is released from rest at $r = r_0$ in the Schwarzschild metric of Eq. (5.59). Note that r is a radial coordinate, in the sense that it provides a measure of how far a spacetime point is from the center of symmetry. However, it would be misleading to call r the radius, since it does not literally measure the distance from the center. If r is varied by an amount $\mathrm{d}r$, the new point is separated from the first not by $\mathrm{d}r$, but instead by $\mathrm{d}r/\sqrt{1-2GM/rc^2}$. r is sometimes called the circumferential radius, since the term $r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$ in the metric implies that the circumference of a circle at a fixed value of r is equal to $2\pi r$, as in Euclidean geometry.

By spherical symmetry, we know that the particle will fall straight toward the center of the sphere, so the coordinates θ and ϕ will remain constant. Thus, the terms in the metric proportional to $\mathrm{d}\theta^2$ and $\mathrm{d}\phi^2$ will give no contribution as the particle moves along

the trajectory. Since the spherical symmetry also guarantees that the other terms in the metric are independent of θ and ϕ, these two angles can be completely ignored in solving the problem; the values of the two angles will remain constant at their initial values.

The trajectory of such a particle is timelike, and can be parameterized by the proper time as it would be measured on a clock that moves with the particle. The trajectory can be described by the functions $r(\tau)$ and $t(\tau)$, where the latter function gives the value of the coordinate t as a function of the proper time. The metric (5.59) gives the separation $d\tau^2$ between two neighboring points along the trajectory. Dividing Eq. (5.59) by $d\tau^2$, one finds the relation

$$c^2 = \left(1 - \frac{2GM}{rc^2}\right) c^2 \left(\frac{dt}{d\tau}\right)^2 - \left(1 - \frac{2GM}{rc^2}\right)^{-1} \left(\frac{dr}{d\tau}\right)^2 . \tag{5.62}$$

This allows one to determine $dt/d\tau$ in terms of $dr/d\tau$. To be more compact, we introduce the notation

$$h(r) \equiv 1 - \frac{R_S}{r} = 1 - \frac{2GM}{rc^2} , \tag{5.63}$$

so Eq. (5.62) can be rewritten as

$$c^2 \left(\frac{dt}{d\tau}\right)^2 = c^2 h^{-1}(r) + h^{-2}(r) \left(\frac{dr}{d\tau}\right)^2 . \tag{5.64}$$

To generalize the geodesic equation (5.52) to spacetime trajectories, there is nothing significant that needs to be changed. We are changing the number of dimensions and we are switching to a metric that is not positive definite, but neither of these changes affect the derivation of the geodesic equation in any way. Since the trajectories of particles are timelike, we parameterize the path not by s, which would be imaginary, but instead by τ. This does not change the form of the equation either, since the only place where the parameterization mattered was when we assumed that $A = 1$, in deriving Eq. (5.52) from Eq. (5.50). But the derivation depended only on the prescription that $A = constant$, and not on $A = 1$. In this case we will be using $A = -c^2$, but the geodesic equation will be unaffected. So, we can rewrite the geodesic equation as

$$\boxed{\frac{d}{d\tau}\left[g_{\mu\nu}\frac{dx^\nu}{d\tau}\right] = \frac{1}{2}\frac{\partial g_{\lambda\sigma}}{\partial x^\mu}\frac{dx^\lambda}{d\tau}\frac{dx^\sigma}{d\tau} ,} \tag{5.65}$$

where I followed a common convention of using Greek letters for spacetime indices. The letters μ, ν, λ, σ, etc. are summed from 0 to 3 when they are repeated, where $x^0 \equiv t$.

Note that of the 4 components of $dx^\mu/d\tau$, only two are nonzero: $dr/d\tau$ and $dt/d\tau$. Since Eq. (5.64) allows us to find $dt/d\tau$ in terms of $dr/d\tau$, it will be sufficient for us to look at only the geodesic equation for $dr/d\tau$. Writing Eq. (5.65) for $\mu = r$, one finds

$$\frac{d}{d\tau}\left[g_{rr}\frac{dr}{d\tau}\right] = \frac{1}{2}\partial_r g_{rr}\left(\frac{dr}{d\tau}\right)^2 + \frac{1}{2}\partial_r g_{tt}\left(\frac{dt}{d\tau}\right)^2 , \tag{5.66}$$

where

$$g_{rr} = h^{-1}(r) , \tag{5.67}$$

and

$$g_{tt} = -c^2 h(r) . \tag{5.68}$$

Using the fact that $\partial_r h(r) = -R_S/r^2$, Eq. (5.66) becomes

$$\begin{aligned} h^{-1}(r)\frac{d^2r}{d\tau^2} - h^{-2}(r)\frac{R_S}{r^2}\left(\frac{dr}{d\tau}\right)^2 = \\ -\frac{1}{2}h^{-2}(r)\frac{R_S}{r^2}\left(\frac{dr}{d\tau}\right)^2 - \frac{1}{2}c^2\frac{R_S}{r^2}\left(\frac{dt}{d\tau}\right)^2 . \end{aligned} \tag{5.69}$$

Now use Eq. (5.64) to eliminate $dt/d\tau$, and notice that the terms involving $dr/d\tau$ cancel against each other. The only remaining terms are proportional to $h^{-1}(r)$, so one can multiply by the inverse of this quantity to obtain

$$\frac{d^2r}{d\tau^2} = -\frac{c^2}{2}\frac{R_S}{r^2} = -\frac{GM}{r^2} . \tag{5.70}$$

This equation is identical in form to the corresponding equation in Newtonian mechanics, but the physics is far from identical. In the Newtonian case the time variable denotes a universal time that can be read on any clock, while in the general relativity case the time variable τ represents the proper time that would be measured by a clock that is moving with the falling particle. The time that would be measured on a stationary clock would be different.

Since Eq. (5.70) is a familiar differential equation, we can integrate it without difficulty. The first step is to obtain a conservation of energy equation, which can be done by multiplying the equation by $dr/d\tau$. The equation can then be written as

$$\frac{d}{d\tau}\left\{\frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 - \frac{GM}{r}\right\} = 0 , \tag{5.71}$$

which implies that the quantity in curly brackets is conserved. If the particle is released from rest at $r = r_0$, then the initial value of this conserved quantity is $-GM/r_0$, so Eq. (5.71) becomes

$$\frac{\mathrm{d}r}{\mathrm{d}\tau} = -\sqrt{2GM\left(\frac{1}{r} - \frac{1}{r_0}\right)} = -\sqrt{\frac{2GM(r_0 - r)}{rr_0}} \ . \tag{5.72}$$

This equation can be reduced to a definite integral by bringing all of the r-dependent factors to one side and integrating:

$$\tau = -\int_{r_0}^{r_f} \mathrm{d}r \sqrt{\frac{rr_0}{2GM(r_0 - r)}} \ . \tag{5.73}$$

This integral can be carried out, so finally we have an expression for the proper time $\tau(r_f)$ at which the particle is at the radius coordinate r_f:

$$\tau(r_f) = \sqrt{\frac{r_0}{2GM}} \left\{ r_0 \tan^{-1}\left(\sqrt{\frac{r_0 - r_f}{r_f}}\right) + \sqrt{r_f(r_0 - r_f)} \right\} \ . \tag{5.74}$$

So, from the point of view of a person riding on the falling particle, the Schwarzschild horizon will be reached in a finite length of time.

However, if we ask how the trajectory evolves as a function of coordinate time t, we will see a very different picture. The velocity with respect to coordinate time can be found by the chain rule:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{\mathrm{d}r}{\mathrm{d}\tau}\frac{\mathrm{d}\tau}{\mathrm{d}t} = \frac{\mathrm{d}r/\mathrm{d}\tau}{\mathrm{d}t/\mathrm{d}\tau} \ , \tag{5.75}$$

and then Eq. (5.64) can be used to eliminate $\mathrm{d}t/\mathrm{d}\tau$:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{\mathrm{d}r/\mathrm{d}\tau}{\sqrt{h^{-1}(r) + c^{-2}h^{-2}(r)\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2}} \ . \tag{5.76}$$

It is possible to find an exact solution for t as a function of r, which can be obtained by using Eq. (5.72) to eliminate $\mathrm{d}r/\mathrm{d}\tau$ from the above equation, and then expressing t as an integral over r, similar to Eq. (5.73). The result is very cumbersome, however, and not very illuminating. We are most interested, however, in how Eq. (5.76) behaves when r is near the horizon, and that behavior can be extracted rather easily. Near the horizon $h(r)$ approaches zero so $h^{-1}(r)$ blows up, with

$$h^{-1}(r) = \frac{r}{r - R_S} \approx \frac{R_S}{r - R_S} \ . \tag{5.77}$$

The argument of the square root in the denominator of Eq. (5.76) is then dominated by the second term, which with Eq. (5.77) gives

$$\frac{\mathrm{d}r}{\mathrm{d}t} \approx c\left(\frac{r - R_S}{R_S}\right) \ . \tag{5.78}$$

Rearranging and integrating to some final $r = r_f$, one finds

$$t(r_f) \approx -\frac{R_S}{c}\int^{r_f} \frac{\mathrm{d}r'}{r' - R_S} \approx -\frac{R_S}{c}\ln(r_f - R_S) \ . \tag{5.79}$$

Thus t diverges logarithmically as $r_f \to R_S$, so the object does not reach R_S for any finite value of t. Thus, even though a person falling into a black hole would pass the horizon in a finite amount of time, from the outside the person will never be seen to reach the horizon.

APPENDIX 5A: THE LORENTZ TRANSFORMATION AND THE LORENTZ-INVARIANT INTERVAL

THE LORENTZ TRANSFORMATION:

The kinematic results of special relativity which were discussed in Lecture Notes 1 — time dilation, Lorentz-Fitzgerald contraction, and the relativity of simultaneity — can all be neatly summarized in a set of equations called the Lorentz transformation. These equations relate the coordinates of an event as seen by one inertial observer to the coordinates of the same event as seen by another inertial observer in relative motion.

The Lorentz transformation can be easily derived from the principles that have already been established. Suppose that a space ship observer constructs a physical coordinate system by carrying with him an entire network of measuring rods oriented along his x- and y-axes, as in Fig. 5A.1. He also has a network of clocks. He determines the spatial coordinates of an event by observing where in this network of measuring rods it occurs, and he determines the time by reading it from a clock located at the site of the event. We will refer to these coordinates as x', y', and t', using the primes to distinguish them from our own coordinate system, which we will continue to call x, y, and t. (To simplify the discussion I am assuming that everything happens in the 2-dimensional plane spanned by the x- and y-axes. The z direction can be reinstated very easily, since its properties are the same as those of the y direction.)

Let us suppose that the moving coordinate system is oriented so that its x'-axis moves to the right along our x-axis, and the clocks are synchronized so that the clock at the origin of each system is set to zero at the time when the two origins cross each other.

Notice, that since there is no contraction of the measuring rods that are oriented perpendicular to the motion, the y-coordinate of an event has the same value in either frame. This leads to the first transformation equation,

$$y' = y \ . \tag{5A.1}$$

If there was a third spatial dimension in the problem, one would similarly conclude that $z' = z$.

Suppose now that an event A occurs in our coordinate system at a spacetime point (x,t), where we will set $y \equiv 0$ for simplicity. We now wish to calculate the coordinates as measured by the moving (primed) system. Since $y = y' = 0$, the event will occur on the measuring rod which constitutes the x'-axis of the moving system, so we can for now ignore the existence of the other measuring rods.

Fig. 5A.2 shows the trajectory of the origin of the primed coordinate system, which we will call O'. It starts at the origin of our system at $t = 0$, and then moves to the right at speed v. The diagram also shows that the moving measuring rod which connects

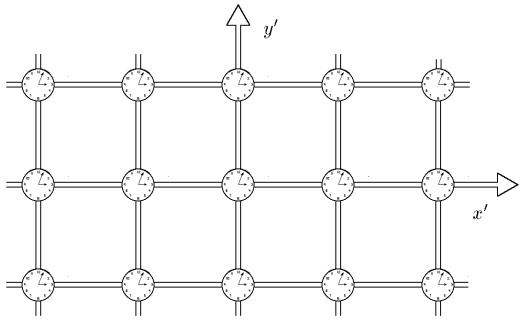

Figure 5A.1: A "physical" coordinate frame, made of clocks and measuring rods.

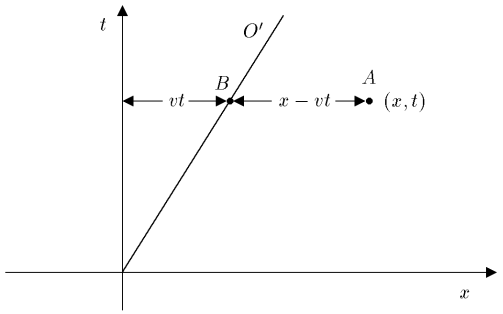

Figure 5A.2: Trajectory of O', the origin of the primed coordinate system. It starts at the origin of our system at $t = 0$, and moves to the right at speed v.

the event A to O' has length $x - vt$, when measured in our frame. However, since the measuring rod is contracted by a factor

$$\gamma \equiv \frac{1}{\sqrt{1-\beta^2}} \ , \tag{5A.2}$$

it follows that the length that one would read off from the rod itself must be $\gamma(x - vt)$. Thus,

$$x' = \gamma(x - vt) \ . \tag{5A.3}$$

To determine t', we must find the time on the moving clock which coincides with the event A. To do this, consider first the event B which occurs at the same time as event A in our frame, but which is located at the origin O' of the moving system. Since the clock at O' is synchronized with ours at $t = 0$ and then runs slowly by a factor of γ, we know that

$$t'(B) = t/\gamma \ . \tag{5A.4}$$

However, the clock at B is trailing the clock at A, and therefore the two clocks will not appear to us to be synchronized. Instead, we have learned that the trailing clock will read a time that is later than the leading clock by an amount $\beta\ell_o/c$, where ℓ_o is rest length of the rod that joins the two clocks. In this case $\ell_o = x' = \gamma(x - vt)$, so

$$\begin{aligned} t'(A) &= t'(B) - \beta\gamma(x - vt)/c \\ &= \left(1 - \beta^2\right)\gamma t - \beta\gamma\left(\frac{x}{c} - \beta t\right) \\ &= \gamma\left(t - \frac{vx}{c^2}\right) \ . \end{aligned} \tag{5A.5}$$

This completes the derivation of the Lorentz transformation equations, which can be summarized as follows:

$$\boxed{\begin{aligned} x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \\ t' &= \gamma\left(t - \frac{vx}{c^2}\right) \ . \end{aligned}} \tag{5A.6}$$

We have already verified that there is no distinction between the moving reference frame and ours, so that the moving observer observes the same distortion in our measuring devices that we observe in his. In the formalism of the Lorentz transformation, this fact is verified by inverting the transformation. That is, the above equations can be solved to express the unprimed variables in terms of the primed variables. When this exercise is carried out, it is found that the equations have exactly the same form, except that the sign of the relative velocity v is reversed.

THE LORENTZ-INVARIANT INTERVAL:

So far we have considered only pulses of light that move either parallel or perpendicular to the direction of motion. However, the Lorentz transformation allows us to easily verify that the measured speed of light is the same in ***all*** directions. To see this,

consider a spherical light pulse that emanates from the origin. In our system, the wave front moves at the speed of light and therefore satisfies the equation

$$x^2 + y^2 + z^2 = c^2t^2 \ . \tag{5A.7}$$

We need to verify that the same equation holds for the coordinates of the wave front in the primed reference frame. We therefore use the Lorentz transformations to calculate the quantity

$$x'^2 + y'^2 + z'^2 - c^2t'^2 \ .$$

When we carry out this somewhat complicated but straightforward calculation, we find the following remarkable relation:

$$x'^2 + y'^2 + z'^2 - c^2t'^2 = x^2 + y^2 + z^2 - c^2t^2 \ . \tag{5A.8}$$

This quantity,

$$x^2 + y^2 + z^2 - c^2t^2 \ ,$$

is therefore called the "Lorentz invariant interval" between the event (x, y, z, t) and the origin.

The origin is of course not really a special point, so one can just as well define the Lorentz invariant interval between any two events A and B:

$$\boxed{s^2 \equiv (x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2 - c^2(t_A - t_B)^2 \ .} \tag{5A.9}$$

Although I am calling the Lorentz invariant interval s^2, I obviously do not mean to imply that it is always positive — it can have either sign. I call it s^2 only because it has the units of cm^2. If s^2 is positive, then the two events are said to be spacelike separated. In that case, it can be shown that there exists a frame of reference in which the two events occur at the same time, and the value of s^2 represents the square of the distance between the events in that frame. If s^2 is negative, the two events are said to be timelike separated. In that case there exists a frame of reference in which the two events occur at the same position, and the value of s^2 represents $-c^2$ times the square of the time separation in that frame. Note also that whenever s^2 is negative one can imagine a clock that moves between the two events at a uniform speed — s^2 is then equal to $-c^2$ times the time interval as measured by the clock. This time interval is sometimes called the proper time between the two events.