EXPANSION OF THE UNIVERSE - STANDARD BIG BANG MODEL

Matts Roos
Department of Physics, University of Helsinki, Helsinki, Finland

Keywords: Relativity, gravitational lensing, black holes, cosmic inflation, Cosmic Microwave Background, Dark Matter, Dark Energy

Contents

Summary

After a brief introduction to the sixteenth and seventeenth century views of the Universe and the nineteenth century paradox of Olbers, we start the history of the cosmic expansion with Hubble's epochal discovery of the recession velocities of spiral galaxies. By then Einstein's theories of relativity were well known, but no suitable metric. Prior to introducing General Relativity we embark on a non-chronological derivation of the Robertson-Walker metric directly from Special Relativity and the Minkowski metric endowed with a Gaussian curvature. This permits the definition of all relativistic distance measures needed in observational astronomy. Only thereafter do we come to General Relativity, and describe some of its consequences: gravitational lensing, black holes, various tests, and the cornerstone of the standard Big Bang model, the Friedmann-Lemaître equations. Going backwards in time towards Big Bang we first have to trace the thermal history, and then understand the needs for a cosmic inflation and its predictions. The knowledge of the Big Bang model is based notably on observations of the Cosmic Microwave Background Radiation, large scale structures,

and the redshifts of distant supernovae. They tell us that gravitating matter is dominated by a dark and dissipationless component of unknown composition, and that the observable part of the Universe exhibits an accelerated expansion representing a fraction of the energy even larger than gravitating matter.

1. Historical Cosmology

The history of ideas on the structure and origin of the Universe shows that humankind has always put itself at the center of creation. As astronomical evidence has accumulated, these *anthropocentric* convictions have had to be abandoned one by one. From the natural idea that the solid Earth is at rest and the celestial objects all rotate around us, we have come to understand that we inhabit an average-sized planet orbiting an average-sized sun, that the Solar System is in the periphery of The Milky Way, a rotating galaxy of average size, flying at hundreds of kilometers per second towards an unknown goal in an immense Universe, containing billions of similar galaxies.

Cosmology aims to explain the origin and evolution of the entire contents of the Universe, the underlying physical processes, and thereby to obtain a deeper understanding of the laws of physics assumed to hold throughout the Universe. Unfortunately, we have only one universe to study, the one we live in, and we cannot make experiments with it, only observations. This puts serious limits on what we can learn about the origin. If there were other universes we would never know.

Although the history of cosmology is long and fascinating we shall neither trace it in detail, nor any further back than *Isaac Newton* (1642–1727). In the early days of cosmology when little was known about the Universe, the field was really just a branch of philosophy. At the time of Newton the heliocentric Universe of *Nicolaus Copernicus* (1473–1543), *Galileo Galilei* (1564–1642) and *Johannes Kepler* (1571–1630) had been accepted because no sensible description of the motion of the planets could be found if the Earth was at rest at the center of the Solar System. However, this anthropocentric view persisted, locating the Solar System at the center of the Universe. The Milky Way had been resolved into an accumulation of faint stars with the telescope of Galileo. Copernicus had formulated the *cosmological* or *Copernican principle,* according to which

- *The Universe is homogeneous and isotropic in three-dimensional space, has always been so, and will always remain so.*

Obviously, matter introduces lumpiness which violates homogeneity on the scale of stars and on the scale of the Milky Way, but on some larger scale isotropy and homogeneity is still taken to be a good approximation.

The first theory of gravitation appeared when Newton published his *Philosophiae Naturalis Principia Mathematica* in 1687, explaining the empirical laws of Kepler: that the planets moved in elliptical orbits with the Sun at one of the focal points. Newton considered the stars to be suns like ours, evenly distributed in a static, infinite Universe. The total number of stars could not be infinite because then their attraction would also be infinite, making the static Universe unstable. There were controversial opinions whether the number of stars was finite or infinite, and whether a finite universe was bounded and an infinite one unbounded. Later *Immanuel Kant* (1724–1804) claimed that the question of infinity was irrelevant because neither type of system embedded in infinite space could be stable and homogeneous. The right conclusion is that the Universe cannot be static, an idea which would have been too revolutionary at Newton's time. The infinity argument was, however, not properly understood until *Bernhard Riemann* (1826–1866) pointed out that the world could be *finite* yet *unbounded*, provided the geometry of the space had a positive curvature, however small.

The first description of the Milky Way as a rotating galaxy can be traced to *Thomas Wright* (1711–1786). Wright's galactic picture had a direct impact on Kant who suggested in 1755 that the diffuse nebulae observed by Galileo could be distant galaxies rather than nearby clouds of incandescent gas. This implied that the Universe could indeed be homogeneous on the scale of galactic distances. This view was also defended by *Johann Heinrich Lambert* (1728–1777) who came to the conclusion that the Solar System, along with the other stars in our Galaxy, orbited around the galactic center, thus departing from the heliocentric view. Kant and Lambert thought that matter is clustered on ever larger scales of hierarchy and that matter is endlessly being recycled. This leads to the question of the origin of time: what was the first cause of the rotation of the galaxy and when did it all start? This is the question modern cosmology attempts to answer by tracing the evolution of the Universe backwards in time.

Newton's first law states that *inertial systems* on which no forces act, are either at rest or in uniform motion. He considered that these properties implicitly referred to an absolute space that was unobservable, yet had a real existence. In 1883 *Ernst Mach* (1838–1916) rejected the concept of absolute space, precisely because it was unobservable: the laws of physics should be based only on concepts which could be related to observations. Since motion still had to be referred to some frame at rest, he proposed replacing absolute space by an idealized rigid frame of fixed stars. Although Mach clearly realized that all motion is relative, it was left to *Albert Einstein* (1879–1955) to take the full step of studying the laws of physics as seen by observers in inertial frames in relative motion with respect to each other. On the basis of Riemann's geometry, Einstein subsequently established the connection between the geometry of space and the distribution of matter.

In spite of the work of Kant and Lambert, the heliocentric picture of the Galaxy remained well into the 20th century. A decisive change came with the observations in 1915–1919 by *Harlow Shapley* (1895–1972) of the distribution of *globular clusters* hosting 10^5–10^7 stars. He found that perpendicular to the galactic plane they were uniformly distributed, but along the plane these clusters had a distribution which peaked in the direction of the Sagittarius. This defined the center of the Galaxy to be quite far from the Solar System: we are at a distance of about two-thirds of the galactic radius. Thus the anthropocentric world picture received yet another blow, and not the last one. Shapley still believed our Galaxy to be at the center of the astronomical Universe.

2. Olbers' Paradox

An early problem still discussed today is the paradox of *Wilhelm Olbers* (1758–1840):

why is the night sky dark if the Universe is infinite, static and uniformly filled with stars? They should fill up the total field of visibility so that the night sky would be as bright as the Sun, and we would find ourselves in the middle of a heat bath of the temperature of the surface of the Sun. Obviously, at least one assumption about the Universe must be wrong.

Olbers' own explanation was that invisible interstellar dust absorbed the starlight so as to make its intensity decrease exponentially with distance. But one can show that the amount of dust needed would be so great that the Sun would also be obscured. Moreover, radiation heats dust so that it becomes visible in the infrared.

A large number of different solutions to this paradox have been proposed, and indeed several effects can be invoked (see ref. Harrison). One possible explanation evokes expansion and special relativity. If the Universe expands, starlight redshifts, so that each arriving photon carries less energy than when it was emitted. At the same time, the volume of the Universe grows, and thus the energy density decreases. The observation of the low level of radiation in the intergalactic space has in fact been evoked as a proof of the expansion.

The dominant effect is, however, that stars radiate only for a finite time, they burn their fuel at well-understood rates. Each galaxy has existed only for a finite time, whether the age of the Universe is infinite or not. Also, the volume of the observable Universe is not infinite; it is in fact too small to contain sufficiently many visible stars. When the time perspective grows, an increasing number of stars become visible because their light has had time to reach us, but at the same time stars which have burned their fuel disappear.

3. Hubble's Law

In a static universe the galaxies should move about randomly, but early galaxy observations had shown that atomic spectral lines of known wavelengths λ exhibited a systematic redward shift to λ' by a factor $1+z=\lambda'/\lambda$ (an exception is the blueshifted *Andromeda nebula* M31), thus these galaxies were receding from us with velocity $v=cz$. In an expanding homogeneous Universe distant galaxies should appear to recede faster than nearby ones.

In the 1920s *Edwin P. Hubble* measured the recession velocities of 18 spiral galaxies with a reasonably well-known distance, and found that all the velocities increased linearly with distance, $v=H_0 r$ or

$$z=H_0\frac{r}{c}. \tag{1}$$

This is *Hubble's law*, and H_0 is called the *Hubble parameter* (present values are always subscripted 0). The message of Hubble's law is that the Universe is expanding and a static Universe is thus ruled out. Einstein had until then firmly believed in a static universe, but when he met Hubble in 1929 he was overwhelmed. This moment marks the beginning of modern cosmology, and sets the primary requirement on theory.

The expansion affects the wavelengths of radiation and the distances between galaxies, but it does not affect the size and internal distances of gravitationally bound systems such as the Solar system, the Milky Way or other galaxies. The expansion appears as if all astronomical objects were receding from us and we were at the center of the Universe. But the Cosmological Principle does not allow a center, and therefore every observer, regardless of position, will have the same impression. Thus the observed recession is really a general expansion.

Equation (1) shows that the Hubble parameter has the dimension of inverse time. Thus a characteristic timescale for the expansion of the Universe is the *Hubble time* $\tau_{\mathrm{H}} = H_0^{-1}$, and the size scale of the observable Universe is the *Hubble radius* $r_{\mathrm{H}} = \tau_{\mathrm{H}}\, c$. In Section 5 we shall discuss measurements of H_0. Using the dimensionless quantity $h = H_0/(100\,\mathrm{km\,s^{-1}Mpc^{-1}})$ which has the value $h \approx 0.72$, we can derive

$$\tau_{\mathrm{H}} \equiv H_0^{-1} = 9.78\,h^{-1} \times 10^9\,\mathrm{yr}, \quad r_{\mathrm{H}} \equiv \tau_{\mathrm{H}}\, c = 3000 h^{-1}\mathrm{Mpc}. \qquad (2)$$

Radiation traveling with the speed of light c reaches r_{H} in time τ_{H}. Note that Hubble's law is non-relativistic, objects beyond r_{H} would be expected to attain recession velocities exceeding c, which is an absolute limit in the theory of special relativity.

The size of the expanding Universe is unknown and immeasurable, but it is convenient to express distances at different epochs in terms of a *cosmic scale factor*: at time t the scale was $a(t)$ when the present value is $a(t_0) \equiv a_0 \equiv 1$. The rate of change of the scale factor can then be identified with the Hubble parameter, $H(t) = \dot{a}(t)/a(t)$ (to first-order time differences).

4. Special Relativity and Metrics

In Einstein's theory of special relativity one studies how signals are exchanged between inertial frames in motion with constant velocity with respect to each other. Einstein postulated that

- The results of measurements in different frames must be identical, and
- Light travels at a constant velocity in vacuum, c, in all frames.

Consider two linear axes x and x' in one-dimensional space, x' being at rest and x moving with constant velocity v in the positive x' direction. Time increments are measured in the two coordinate systems as $\mathrm{d}t$ and $\mathrm{d}t'$ using two identical clocks. Neither the spatial increments $\mathrm{d}x$ and $\mathrm{d}x'$ nor the time increments are invariants – they do not obey the first postulate. Let us replace $\mathrm{d}t$ and $\mathrm{d}t'$ with the *temporal distances* $c\,\mathrm{d}t$ and $c\,\mathrm{d}t'$ and look for a linear transformation between the primed and unprimed frames under which the two-dimensional *space-time distance* element $\mathrm{d}\mathbf{s}$ between two *space-time events*,

$$\mathrm{d}s^2 = c^2\,\mathrm{d}\tau^2 \equiv c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 = c^2\,\mathrm{d}t'^2 - \mathrm{d}x'^2 \equiv c^2\,\mathrm{d}\tau'^2, \tag{3}$$

is invariant. The quantity $\mathrm{d}\tau$ is called the *proper time* and $\mathrm{d}s$ the *line element*.

Invoking the second postulate it is easy to show that the transformation must be of the form

$$\mathrm{d}x' = \gamma\,(\mathrm{d}x - v\,\mathrm{d}t), \quad c\,\mathrm{d}t' = \gamma\,(c\,\mathrm{d}t - v\,\mathrm{d}x/c), \tag{4}$$

where

$$\gamma = (1-(v/c)^2)^{-1/2}. \tag{5}$$

Equation (4) defines the *Lorentz transformation*, after *Hendrik Antoon Lorentz* (1853–1928). Scalar products (such as $\mathrm{d}\tau^2$ and $\mathrm{d}x^2$) in this two-dimensional (ct, x)-space-time are invariants under Lorentz transformations. For example, a particle with mass m moving with velocity three-vector $\mathbf{v}$ and three-momentum $\mathbf{p} = m\mathbf{v}$ is described in four-dimensional space-time by the four-vector $P = (E/c, \mathbf{p})$. The scalar product P^2 is an invariant related to the mass, $P^2 = (E/c)^2 - p^2 = (\gamma mc)^2$. For a particle at rest, this gives Einstein's famous formula

$$E = mc^2. \tag{6}$$

It follows that time intervals measured in the two frames are related by $\mathrm{d}t = \gamma\,\mathrm{d}t'$. This *time dilation effect* is only noticeably when v approaches c. It has been confirmed in particle accelerators and by muons produced in cosmic ray collisions in the upper atmosphere. These unstable particles have well-known lifetimes in the laboratory, but when they strike Earth with relativistic velocities, they appear to have a longer lifetime by the factor γ.

The Lorentz transformations (4) can immediately be generalized to three spatial coordinates x, y, z, so that the *metric* (3) is replaced by the four-dimensional metric of *Hermann Minkowski* (1864–1909),

$$\mathrm{d}s^2 = c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \equiv c^2\,\mathrm{d}t^2 - \mathrm{d}l^2. \tag{7}$$

The trajectory of a body moving in space-time is called its *world line*. A body at a fixed location in space follows a world line parallel to the time axis in the direction of increasing time. A moving body follows a world line making a slope with respect to the time axis. Since the speed of a body or a signal traveling from one event to another cannot exceed the speed of light, there is a maximum slope to such world lines. All world lines for which $ct < 0$ and arriving at $t = 0$ form our *past light cone*, thus they enclose the present observable universe. All world lines for which $ct > 0$ and starting from where we are now can influence events inside our *future light cone*. Two separate events in space-time can be causally connected provided their spatial separation $\mathrm{d}\mathbf{l}$ and their temporal separation $\mathrm{d}t$ (in any frame) obey $|\mathrm{d}\mathbf{l}/\mathrm{d}t| \leq c$. Their world line is then inside the light cone. In Figure (1) we draw this four-dimensional cone in t, x, y-space (suppressing the z direction).

Special relativity thus revised our concept of space-time and made it four-dimensional. Riemann and others realized that Euclidean geometry was just a particular choice suited to flat space, but not necessarily correct in the space we inhabit. Consider the path in three-space followed by a free body obeying Newton's first law of motion. This path represents the shortest distance between any two points along it, called a *geodesic* of the space. In flat Euclidean space the geodesics are straight lines. But measurements of distances depend on the geometric properties of space, as has been known to navigators ever since Earth was understood to be spherical. A spherical surface is characterized by its radius of curvature which causes the geodesics to be great circles.

Suppose an observer wants to make a map of points in the expanding Universe. It is then no longer convenient to use the coordinates x, y, z in Equations (3) and (7) nor the spherical coordinates R, θ, ϕ, because the cosmic expansion would quickly outdate the map. Instead it is convenient to factor out the expansion $a(t)$ and replace the radial distance R by $a(t)\sigma$, where σ is a dimensionless stationary *comoving* coordinate.

If the four-dimensional space happens to be curved just like the surface of Earth, a *Gaussian curvature k* may be included in the Minkowski metric. The parameter k can take on the values $+1$, 0, -1, corresponding to a three-sphere, a flat three-space, and a three-hyperboloid, respectively. The metric of four-dimensional space-time can then be written in the form derived independently by *Howard Robertson* and *Arthur Walker* in 1934:

$$\begin{aligned} \mathrm{d}s^2 &= c^2\,\mathrm{d}t^2 - \mathrm{d}l^2 = \\ &c^2\,\mathrm{d}t^2 - a(t)^2\left(\frac{\mathrm{d}\sigma^2}{1-k\sigma^2} + \sigma^2\mathrm{d}\theta^2 + \sigma^2\sin^2\theta\,\mathrm{d}\phi^2\right). \end{aligned} \tag{8}$$

This metric (RW) can describe an expanding, spatially homogeneous and isotropic universe in accord with the cosmological principle.

Of course there was a motivation for introducing curvature: General Relativity, to which we shall come in Section 6.

5. Distance Measures

The *comoving distance* from us to a galaxy at comoving coordinates $(\sigma, 0, 0)$ is not an observable because the galaxy can only be observed by the light it emitted at an earlier time, $t < t_0$. In a space-time described by the RW metric the light signal propagates along a geodesic, $\mathrm{d}s^2 = 0$. Introducing an alternative comoving coordinate χ defined by $\mathrm{d}\chi = \mathrm{d}\sigma/\sqrt{1-k\sigma^2}$ in Equation (8) (with $\mathrm{d}\theta^2 = \mathrm{d}\phi^2 = 0$), the geodesic equation is

$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a(t)^2\mathrm{d}\chi^2 = 0$. From this,

$$\chi = c\int_t^{t_0} \frac{\mathrm{d}t}{a(t)}. \qquad (9)$$

The present *proper distance* to the galaxy is then $d_\mathrm{P} = a_0\chi$. In flat space $d_\mathrm{P} = \sigma$, but in curved spaces the function is more complicated (cf. Roos 2003). For practical astronomical measurements at small redshifts one uses an approximate expression for d_P in flat space ($k = 0$),

$$d_\mathrm{P}(z) \approx \frac{c}{H_0}\left(z - \frac{1}{2}(1+q_0)z^2\right), \qquad (10)$$

where q_0 is the present *deceleration parameter* $q \equiv -a\ddot{a}/\dot{a}^2 = -\ddot{a}/aH^2$. The first term on the right of Equation (10) gives Hubble's linear law (1), and thus the second term measures deviations from linearity to lowest order. The parameter value $q_0 = -1$ obviously corresponds to no deviation.

The largest comoving spatial distance from which a signal could have reached us is called the *particle horizon*, denoted σ_ph or alternatively χ_ph. This delimits the part of the Universe that has come into causal contact since time $t = 0$. If the lower integration limit in Eq. (9) is equal to the time when the Universe became transparent to light (the last scattering time), the particle horizon delimits the visible Universe.

In an analogous way, the comoving distance σ_eh or χ_eh to the *event horizon* is defined as the spatially most distant event at time t_0 from which a world line can ever reach our world line. By 'ever' we mean a finite future time, t_max. The particle horizon σ_ph at time t_0 lies on our past light cone, but with time our particle horizon will broaden so that the light cone at t_0 will move inside the light cone at $t > t_0$. The event horizon at this moment can only be specified given the time distance to the ultimate future, t_max. Only at t_max will our past light cone encompass the present event horizon. Thus the event horizon is our ultimate particle horizon.

The distances to relatively nearby stars can be measured by the *trigonometric parallax* up to about 30 pc away. This is the difference in angular position of a star as seen from Earth when at opposite points in its circumsolar orbit. The *parallax distance* is defined as $d_\mathrm{par} = d_\mathrm{P}/\sqrt{1-k\sigma^2}$.

Consider an astronomical object radiating photons isotropically with power or absolute luminosity L. At the *luminosity distance* d_L from the object we observe only the fraction $B = L/4\pi d_\mathrm{L}^2$, its surface brightness, given by the Euclidean inverse-square distance law. If the Universe does not expand and the object is stationary at proper distance d_{P}, a telescope with area A will receive a fraction $A/4\pi d_{\mathrm{P}}^2$ of the photons. But in a universe characterized by an expansion $a(t)$, the object is not stationary, so the energy of photons emitted at time t_{e} is redshifted by the factor $(1+z)=a^{-1}(t_{\mathrm{e}})$. Moreover, the arrival rate of the photons suffers time dilation by another factor $(1+z)$, often called the *energy effect*. The end result is that $d_{\mathrm{L}} = d_{\mathrm{P}}(1+z)$.

Astronomers usually replace L and B by two empirically defined quantities, *absolute magnitude* M of a luminous object and *apparent magnitude* m. The replacement rule is

$$m - M = -5 + 5\log d_{\mathrm{L}}, \tag{11}$$

where d_{L} is expressed in parsecs (pc) and the logarithm is to base 10.

Most stars in the Galaxy for which we know L from a kinematic distance determination exhibit a relationship between surface temperature T and L, the *Hertzsprung–Russell* relation. These *main-sequence stars* sit on a fairly well-defined curve in the $T-L$ plot, and temperature is related to color. From this relation one can derive distances to farther main-sequence stars: from their color one obtains the luminosity which subsequently determines d_{L}. By this method one gets a second rung in a ladder of estimates which covers distances within our Galaxy.

Yet another measure of distance is the *angular size distance* d_{A}. In Euclidean space an object of size D which is at distance d_{A} will subtend an angle θ such that $\theta \approx D/d_{\mathrm{A}}$ for small angles. In General Relativity we can still use this approximation to define d_{A}. From the RW metric (8) the diameter of a source of light at comoving distance σ is $D = a\sigma\theta$, so $d_{\mathrm{A}} = D/\theta = a\sigma = \sigma/(1+z)$.

As the next step on the distance ladder one chooses calibrators which are stars or astronomical systems with specific uniform properties, so called *standard candles*. The *RR Lyrae* stars all have similar absolute luminosities, and they are bright enough to be seen out to about 300 kpc. A very important class of standard candles are the *Cepheid* stars, whose absolute luminosity oscillates with a constant period $\log P \propto 1.3 \log L$. Globular clusters are gravitationally bound systems of 10^5–10^6 stars forming a spherical population orbiting the center of our Galaxy. They can also be seen in many other galaxies, and they are visible out to 100 Mpc. Various statistical properties of well-measured clusters, such as the frequency of stars of a given luminosity, the mean luminosity, and the maximum luminosity are presumably shared by similar clusters at all distances, so that clusters become standard candles. Similar statistical indicators can be used to calibrate clusters of galaxies; in particular the brightest galaxy in a cluster is a standard candle useful out to 1 Gpc.

A notable contribution to our knowledge of H_0 comes from the observation of Type Ia *supernova* explosions. The released energy is always nearly the same, in particular the peak brightness of Type Ia supernovae can serve as remarkably precise standard candles out to 1 Gpc. Additional information is provided by the color, the spectrum, and an empirical correlation observed between the timescale of the supernova light curve and the peak luminosity.

The existence of different methods of calibration covering similar distances is a great help in achieving higher precision. The expansion can be verified by measuring the surface brightness of standard candles at varying redshifts, the *Tolman test*. In an expanding universe, the intensity of the photon signal at the detector is further reduced by a factor $(1+z)^2$ due to an optical aberration which makes the surface area of the source appear increased. Such tests have been done and they do confirm the expansion.

The *Tully–Fisher* relation is a very important tool at distances which overlap those calibrated by Cepheids, globular clusters, galaxy clusters and several other methods. This empirical relation expresses correlations between intrinsic properties of whole spiral galaxies. It is observed that their absolute luminosity and their circular rotation velocity v_{c} are related by $L \propto v_{\mathrm{c}}^4$. (For more details on this Section, see the books by Peacock and Roos.)

The differential light propagation delay between two or more gravitationally lensed images of a background object such as a quasar establishes an absolute physical distance scale in the lens system. This is named the *Refsdal Method*, and it is the only direct way of measuring cosmological distances and the global expansion rate H_0 in a single step for each system, thus avoiding the propagation of errors along the distance ladder. There are about 10 such cases measured by now.

6. General Relativity

Although *Newton's second law*, $\mathbf{F} = m\mathbf{a}$, is invariant under special relativity in any inertial frame, it is not invariant in accelerated frames because it explicitly involves acceleration, $\mathbf{a}$. Einstein required that also observers in accelerated frames should be able to agree on the value of acceleration. Space-time derivatives in a curved RW metric are also not invariants because they imply transporting quantities along some curve and that makes them coordinate dependent. Thus the next necessary step is to search for invariant redefinitions of derivatives and accelerations, and to formulate the laws of physics in terms of them. Such a formulation is called *generally covariant*. Moreover, for a body of *gravitating mass* m_{G} at a distance r from another mass M, the force F specified by *Newton's law of gravitation*,

$$F = -GMm_{\mathrm{G}}/r^2, \tag{12}$$

where G is *Newton's constant*, is in serious conflict with special relativity in three ways. Firstly, there is no obvious way of rewriting the law in terms of invariants, since it only contains scalars. Secondly, it has no explicit time dependence, so gravitational effects propagate instantaneously to every location in the Universe. Thirdly, m_{G} is totally independent of the *inert mass* m appearing in Newton's second law, yet for unknown reasons both masses appear to be equal to a precision of 10^{-13} or better. Clearly a theory is needed to establish a formal link between them.

Einstein considered how Newton's laws would be understood by a passenger in a spacecraft, and realized that the passenger would not be able to distinguish between gravitational pull and local acceleration – this is called the *Weak Equivalence Principle* (WEP). This principle is already embodied in the *Galilean equivalence principle* in mechanics between motion in a uniform gravitational field and a uniformly accelerated frame of reference. What Einstein did was to generalize this to all of physics, in particular phenomena involving light. The more general formulation is the important *strong equivalence principle* (SEP): *to an observer in free fall in a gravitational field the results of all local experiments are completely independent of the magnitude of the field*.

In a suitably small spacecraft, curved space-time can always be locally approximated by flat Minkowski space-time. On a larger scale a nonuniform gravitational field can be replaced by a patchwork of locally flat frames which describe the curved space. Trajectories of bodies as well as rays of light follow geodesics, thus in a curved space-time also light paths are curved. Following SEP, this implies that photons in a gravitational field may appear to have mass.

In the gravitational field of Earth, two test bodies with a space-like separation clearly do not fall along parallels, but along different radii, so that their separation decreases with time. This phenomenon is called the *tidal effect*, or the tidal force, since the test bodies move as if an attractive exchange force acted upon them. A sphere of freely falling particles will be focused into an ellipsoid with the same volume, because the particles in the front of the sphere will fall faster than those in the rear, while at the same time the lateral cross-section of the sphere will shrink due to the tidal effect. This effect is responsible for the gravitational breakup of very nearby massive stars.

Since gravitating matter is distributed inhomogeneously (except on the largest scales) causing inhomogeneous gravitational fields, Einstein realized that the space we live in had to be curved, and the curvature had to be related to the distribution of matter. He then proceeded to search for a law of gravitation that was a generally covariant relation between mass density, as implied by the SEP, and curvature. The simplest form for such a relation is *Einstein's Equation*

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}. \tag{13}$$

The *Einstein tensor* $G_{\mu\nu}$ contains only terms which are either quadratic in the first space-time derivatives of the metric tensor $g_{\mu\nu}$, or linear in the second derivatives. (Higher-order derivatives are difficult to include without making the theory unstable.) The *stress–energy tensor* $T_{\mu\nu}$ contains the various components of energy densities, pressures and shears of matter and radiation.

The Einstein tensor vanishes for flat space-time and in the absence of matter and pressure, as it should. Thus the problems encountered by Newtonian mechanics have been resolved. The recession velocities of distant galaxies do not exceed the speed of light, and effects of gravitational potentials are not felt instantly. The discontinuity of homogeneity and isotropy at the boundary of the Newtonian universe also disappeared because four-space is unbounded, and because space-time in general relativity (GR) is generated by matter and pressure. Thus space-time itself ceases to exist where matter does not exist, so there cannot be any boundary between an 'inside' homogeneous universe and an 'outside' space-time void.

Einstein published his General Theory of Relativity in 1917, but the only solution he found to the highly nonlinear differential equations (13) was static. This was in good agreement with the then known Universe which comprised only the 'fixed' stars in our Galaxy, and some nebulae of poorly known distance and of controversial nature.

7. Tests of General Relativity

The classical testing ground of theories of gravitation, Einstein's among them, is celestial mechanics within the Solar System. The earliest phenomenon requiring general relativity for its explanation was noted in 1859, 20 years before Einstein's birth. The French astronomer *Urban Le Verrier* (1811–1877) found that the planet Mercury's elongated elliptical orbit precessed slowly around the Sun. As the innermost planet it feels the solar gravitation very strongly, but the orbit is also perturbed by the other planets. The total effect is that the *perihelion* of the orbit advances $574''$ (seconds of arc) per century. This is calculable using Newtonian mechanics and Newtonian gravity, but the result is $43''$ too little.

With the advent of general relativity the calculations could be remade. This time the discrepant $43''$ were successfully explained by the new theory, which thereby gained credibility. This counts as the first one of three 'classical' tests of GR.

The second classical test was the predicted deflection of a ray of light passing near the Sun. We shall come back to that test in Section 8 on gravitational lensing. The third classical test was the gravitational shift of atomic spectra: the frequency of emitted radiation makes atoms into clocks which run slower in a strong gravitational field. This was first observed in a cloud of plasma ejected by the Sun to an elevation of about 72000 km above the photosphere and an effect only slightly larger than that predicted by GR was found. Similar measurements have been made of radiation from the surface of more compact stars such as Sirius' companion, the white dwarf Sirius B.

A fourth test is based on the prediction that an electromagnetic wave suffers a time delay when traversing an increased gravitational potential. It was carried out in 1971 with the radio telescopes at the Haystack and Arecibo observatories by emitting radar signals towards Mercury, Mars and, notably, Venus, through the gravitational potential of the Sun. The round-trip time delay of the reflected signal was compared with theoretical calculations. Further refinement was achieved later by posing the Viking Lander on the Martian surface and having it participate in the experiment by receiving and retransmitting a radio signal from Earth. This experiment found the ratio of the delay observed to the delay predicted by GR to be 1.000 ± 0.002.

The most important tests of GR have been carried out on the radio observations of pulsars that are members of binary pairs, notably the PSR $1913+16$, a pair of rapidly rotating, strongly magnetized neutron stars discovered in 1974 by *R.A. HulseindexHulse, Hulse* and *J.H. Taylor*, awarded the Nobel prize in 1993. If the magnetic dipole axis does not coincide with the axis of rotation (just as is the case with Earth), the star would radiate copious amounts of energy along the magnetic dipole axis. These beams at radio frequencies precess around the axis of rotation like the searchlights of a beacon. As the beam sweeps past our line of sight, it is observable as a pulse with the period of the rotation of the star. Pulsars are the most stable clocks known in the Universe, the variation is about 10^{-14} on timescales of 6–12 months. The reason for this stability is the intense self-gravity of a neutron star that makes it almost undeformable until the very last few orbits when the binary pair coalesces into one star.

This system does not behave exactly as expected in Newtonian mechanics; hence the deviations provide several independent confirmations of GR. The largest relativistic effect is the apsidal motion of the orbit which is analogous to the advance of the perihelion of Mercury. A second effect is the counterpart of the relativistic clock correction for an Earth clock. The travel time of light signals from the pulsar through the gravitational potential of its companion provides a further effect.

The slowdown of the binary pulsar is indirect evidence that this system loses its energy by radiating gravitational waves. Such waves travel through space-time with the speed of light, traversing matter unhindered and unaltered, and producing ripples of curvature, oscillatory stretching and squeezing of the web of space-time analogously to the tidal effect of the Moon on Earth. Any matter they pass through will feel this effect. Thus a detector for gravitational waves is similar to a detector for the Moon’s tidal effect, but the waves act on an exceedingly weak scale.

In GR, the inertial and centrifugal forces felt on Earth are due to our accelerations and rotations with respect to the local inertial frames which, in turn, are determined, influenced and *dragged* by the distribution and flow of mass densities in the Universe. A spinning mass will ‘drag’ inertial frames and gyroscopes along with it. This also influences the flow of time around a spinning body, so that synchronization of clocks around a closed path near it is not possible. This effect is predicted by GR to be quite small.

Note that the expansion of the Universe and Hubble’s linear law (1) are not tests of GR. Objects observed at wavelengths ranging from radio to gamma rays are close to isotropically distributed over the sky. Either we are close to a center of spherical symmetry—an anthropocentric view—or the Universe is close to homogeneous. In the latter case, and if the distribution of objects is expanding so as to preserve homogeneity and isotropy, the expansion velocities satisfy Hubble’s law.

-
-
-

Bibliography

Harrison E. (1987). *Darkness at night*. Harvard University Press, Cambridge, MA. [A review of all historical explanations of Olbers' paradox.]

Peacock J.A. (1999). *Cosmological physics*. Cambridge University Press, Cambridge. [A comprehensive introduction to modern cosmology.]

Roos M. (2003). *Introduction to Cosmology*, Third Edition. John Wiley & Sons, Ltd, England. [A textbook of cosmology at an introductory level.

Biographical Sketch

Matts Roos was born in Helsinki, Finland, in 1931. He studied technical physics at the Technical University of Helsinki to become Master of Engineering (1956), Licentiate of Technology (1960), and Doctor of Technology (1967). He studied undergraduate physics and atomic physics at Union College, Schenectady, NY 1950-51, entered the Finnish army Signal corps 1955, and concluded as Ensign (1956). Subsequently he worked as Nuclear reactor engineer in Stockholm (1957-59), as Research assistant in the Institution of Theoretical Physics, University of Stockholm (1959-62), as Fellow at NORDITA (1962-64), as Visiting scientist at the Niels Bohr Institute (1964-65), as Fellow (1965-67) and Staff member at CERN (1957-71), as Associate professor of nuclear physics, University of Helsinki (1970-77), as Personal extraordinary professor of particle physics, University of Helsinki 1977-96, as Director of the High Energy Physics Laboratory, University of Helsinki (1992-96) until retirement in 1996.

A founding member of the Particle Data Group (1963-2004), he has published over 150 research papers, co-authored two books (one translated twice), and authored three editions of the book Introduction to Cosmology, John Wiley & Sons, Ltd, England, 1993, 1997, 2003. His current research interest is cosmology, previous interests particle physics (meson spectroscopy, weak interactions, neutrino physics) and quantum mechanics.

Emeritus Prof. Roos has been awarded the Finnish State Order of Merit SVR R1 in 1995. Member of the European Physical Society (1971–2004), the Finnish Physical Society (1971–), the Finnish National Committee of IUPAP (1980–83), Societas Scientiarum Fennica (1990–), and the International Astronomical Union (1994–). Through his membership in the Finnish Painters' Union (1997–) and the Artists' Association of Finland he is also a member of the International Association of Art IAA.