ASTR 3740: Relativity and Cosmology

Spring 2001

MWF, 2:00–2:50 PM, Duane G131

Instructor: Dr. Ka Chun Yu

Office: Duane C-327, Phone: (303) 492-6857

Office Hours: MW 3:00–4:00 PM or by appointment

Email: kachun@casa.colorado.edu

Course Page: http://casa.colorado.edu/~kachun/3740/

This is a upper division introduction to Special and General Relativity, with applications to theoretical and observational cosmology. This course is an APS minor elective, and is intended for science majors. We will delve into the reasons why relativity is important in studying cosmology, work through applications of SR and GR, and then jump from there to theoretical and observational cosmology. Because this is an astrophysics course, there will be strong emphasis on observational confirmations of Einstein's theories, astrophysical applications of relativity including black holes, and finally the evidence for a Big Bang cosmology. We will follow this with discussion of the evolution of the universe, including synthesis of the elements, and the formation of structure. We will conclude (if time allows) with advanced topics on the inflationary period of the early universe and analyzing primordial fluctuations in the cosmic microwave background.

Although a year each of calculus and freshman physics are the only required prerequisites for this course, be warned that we will be moving quickly through a wide range of quantitative material, and hence you are expected to have a *firm and thorough* understanding of the prerequisite classwork. It is also helpful to have taken or have an understanding commensurate with having taken a sequence of the 1000 level astronomy courses. (Although not required, some level of familiarity with thermodynamics, quantum mechanics, electromagnetism, and topics in mathematical physics would be useful.) We will *not* be covering GR with full-blown tensor calculus. Students interested in this more rigorous approach should take one of the graduate-level GR courses. If this course sounds a bit too mathematical for you, you might be better off taking ASTR 2010, Modern Cosmology, taught by Prof. Nick Gnedin at the same time and down the hall.

There is no required textbook for this course. Instead I will be lecturing out of a set of notes that will be available online at the course webpage (http://casa.colorado.edu/~kachun/3740/). A number of titles are suggested for optional reading, and are available for short-term loan from the Lester Math-Physics Library, or can be purchased from the CU bookstore or other booksellers. These are

Spacetime Physics, 2nd edition, by Edwin Taylor & John R. Wheeler, 1992, W. H. Freeman & Co., $45.30 (paperback)

Principles of Cosmology and Gravitation, by M. V. Berry, 1989, Adam Hilger, $25.00 (paperback)

The Big Bang, 3rd edition, by Joseph Silk, 2000, W. H. Freeman & Co., $19.95 (paperback)

Grading

Weekly homework assigments will be given out, where you will have a week to turn in the assignment for full credit. **Assignments turned in past the 5:00 PM deadline on the due date will have points deducted.** (My box can be found amongst the mailboxes across from the CASA

office in Duane C-333.) **Although you are free to work together, the work you turn in *must* be your own.** If I detect copying between homeworks, I will penalize all parties involved.

In addition to the homeworks, we will also have an in-class midterm and final. These will be closed book tests. **The final is Wednesday, May 9, 7:30 am to 10:00 am.**

The last major component of the grade will be a 12–15 page term paper (including equations, figures, references, etc.) on a topic in relativity and/or cosmology. For this paper, I want you to look up one or more papers appearing in peer-reviewed journals that are related to the topic you wish to discuss. Although you may use secondary sources of information (such as textbooks, books written for the general public, articles in *Astronomy* or *Sky & Telescope*, websites, etc.) to help write your report, your main goal is to report on a scientific result appearing in a scientific paper. I will give out a list of suggested topics, as well as ways to research and look up scientific papers later in the semester. **This project will be due on the last day of classes, May 4.** Because of the technical nature of this project, I want you to turn into me bibliographic information for the paper (title, authors, journal, volume number, etc.) and its abstract, preferably by March 16, but no later than the last day of classes before Spring Break (March 23). *It is highly recommended that you consult with me in person or via email before making a final decision on what to write about.*

The final breakdown for the grades will be roughly:

Homeworks	25%
Midterm	25%
Term Paper	25%
Final	25%

For borderline grades, class participation will be used to nudge numbers up or down. The final total class grade will be based on a curve.

Schedule

Here is a rough breakdown of the topics that will be covered during the course of this semester:

1. Early Ideas of Our Universe
2. Special Relativity
    - Length Contraction
    - Time Dilation
    - Velocity Transformations
    - Relativistic Doppler Effect
    - Gravitational Redshift
    - Spacetime
3. General Relativity
    - Geodesics and Spatial Curvature
    - The Schwarzschild Solution
    - Motion of Particles and Light in the Schwarzschild Metric
    - Effective Potentials
    - Effective Potentials in the Schwarzschild Metric
4. Black Holes

- Gravitational Collapse
- Evidence for the Existence of Black Holes
- Massive Black Holes in Galaxies

5. Theoretical Cosmology
    - Cosmological Principle
    - Comoving Coordinates
    - Friedmann-Robertson-Walker Metric
    - Horizons
    - Deceleration Parameter q_0
    - Friedmann Equations
6. Observational Cosmology
    - Nucleosynthesis in the Big Bang
    - Cosmic Problems
    - Dark Matter
7. Formation of Structure in the Universe
    - Jeans Mass
    - Spectrum of Perturbations; Linear/Non-Linear Perturbations
    - Primordial Spectrum of Perturbations
    - Structure Formation: The Virial Theorem
    - Cooling of Baryonic Gas
    - Galaxy Formation
    - Correlation Function
8. Inflation
9. Analyzing the Cosmic Microwave Background

Chapter 1

Early Ideas of Our Universe

1.1 The Ancients

The Babylonians were some of the earliest astronomers. They invented a sexagesimal (base 60) numbering system that is reflected in our modern day usage of seconds, minutes, and hours. Babylonian astronomers kept careful logs of the motions of the Moon and the planets in the sky in order to predict the future using astrology. They also believed in a cosmology where the Earth was at the center of the universe, bound below by water. The seven heavenly bodies that moved in the sky represented dieties, with each one moving in a progressively further sphere from the Earth. (In order, they were the Moon, Mercury, Venus, the Sun, Mars, Jupiter, and Saturn.) The fixed stars lay beyond Saturn, and beyond that was more water binding the outer edge of the known universe.

The Rig Vedas were Hindu texts that date back to 1000 BC. Part of them discussed the cyclical nature of the universe. The universe underwent a cycle of rebirth followed by fiery destruction, as the result of the dance of Shiva. The length of each cycle is a "day of Brahma" which lasts 4.32 billion years (which coincidentally is roughly the age of our Earth and only a factor a few off from the actual age of the universe). The cosmology has the Earth resting on groups of elephants, which stand on a giant turtle, who in turn is supported by the divine cobra Shesha-nāga.

The Ancient Greeks: Although early Greek thought on the heavens mirrored that of the Babylonians, with a reliance on gods and myths, by the 7th century BC, a new class of thinkers, relying in part on observations of the world around them, began to use logic and reason to arrive at theories of the natural world and of cosmology. These ancient Greek philosophers had a variety of ideas about the nature of the universe.

- **Thales** of Miletus (634–546 BC) believed the Earth was a flat disk surrounded by water.

- **Anaxagoras** (ca. 500–ca. 428 BC) believed the world was cylindrically shaped, where we lived on the flat-topped surface. This world cylinder floats freely in space on nothingness, with the fixed stars in a spherical shell that rotated about the cylinder.

By Ka Chun Yu.

The Moon shone as a result of reflected light from the Sun, and lunar eclipses were the result of the Earth's shadow falling on the Moon.

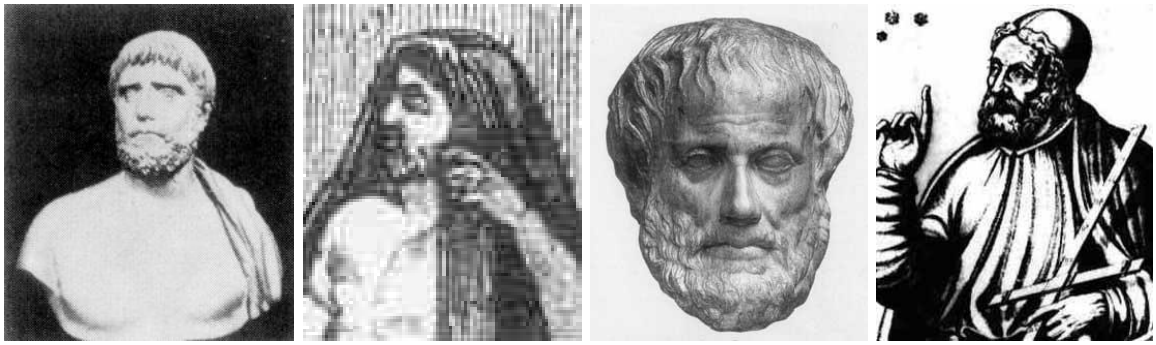


Figure 1.1: Left to right: Thales, Anaxagoras, Aristotle, and Claudius Ptolemy.

- **Eudoxus** of Cnidus (ca. 400–ca. 347 BC) also had a geocentric model for the Earth, but added in separate concentric spheres for each of the planets, the Sun, and the Moon, to move in, with again the fixed stars located on an outermost shell. Each of the shells for the seven heavenly bodies moved at different rates to account for their apparent motions in the sky. To keep the model consistent with observations of the planets' motions, Eudoxus' followers added more circles to the mix—for instance, seven were needed for Mars. The complexity of this system soon made this model unpopular.

- **Aristotle** (384–322 BC) refined the Eudoxus model, by adding more spheres to make the model match the motions of the planets, especially that of the retrograde motions seen in the outermost planets. Aristotle believed that "nature abhors a vacuum," so he believed in a universe that was filled with crystalline spheres moving about the Earth. Aristotle also believed that the universe was eternal and unchanging. Outside of the fixed sphere of stars was "nothingness."

- **Aristarchus** (ca. 310–ca. 230 BC) made a first crude determination of the relative distance between the Moon and the Sun. His conclusion was that the Sun was 20× further, and the only reason they appeared to be of the same size was that the Sun was also 20× larger in diameter. Aristarchus then wondered, if the Sun was so much larger, would it make sense for it to move around in the universe? Would it make more sense for the Earth to move around it?

- **Claudius Ptolemy** (ca. 100–ca. 170 AD) writing in *Syntaxis* (aka *Almagest*; ∼ 140 AD) took the basic ideas of Eudoxus' and Aristotle's cosmology, but had the planets move in circular *epicycles*, the centers of which then moved around the Earth on the *deferent*, an even bigger orbit. Ptolemy's ideas gave the most accurate explanations for the motion of the planets (as best as their positions were known at the time). (Ptolemy's and Aristotle's ideas about the universe and its laws of motion remained the dominant idea in Western thought until the 15th century AD!)

1.2 European Thought Before the 20th Century

Nicolaus Copernicus (1473–1543) made a radical break from Ptolemaic thought by proposing that the Earth was *not* at the center of the universe. In his *De Revolutionibus Orbium Celestium*, he believed a Sun-centered universe to be more elegant:

> In no other way do we perceive the clear harmonious linkage between the motions of the planets and the sizes of their orbs.

However to preserve a model that accurately reflected the actual motions of the planets, he still had to use additional smaller circles, known as an *epicyclet*, that orbited an offset circle.

Figure 1.2: Left to right: Nicolaus Copernicus, Giordano Bruno, and Tycho Brahe.

Thomas Digges (1546–1595), a leading English admirer of Copernicus, published *A Perfect Description of the Celestial Orbes*, which re-stated Copernicus' heliocentric theory. However Digges went further by claiming that the universe is infinitely large, and filled uniformly with stars. This is one of the first pre-modern statements of the *cosmological principle*.

Giordano Bruno (1548–1600) goes even further: not only are there an infinite number of stars in the sky, but they are also suns with their own solar systems, and orbited by planets filled with life. These and other heretical ideas (e.g., that all these other life-forms, planets, and stars also had their own souls) resulted in him being imprisoned, tortured, and finally burned at the stake by the Church.

Tycho Brahe (1546–1601) made and recorded very careful naked eye observations of the planets, which revealed flaws in their positions as tabulated in the Ptolemaic system. He played with a variety of both geocentric and heliocentric models.

Johannes Kepler (1571–1630) finally was able to topple the Ptolemaic system by proposing that planets orbited the Sun in ellipses, and not circles. He proposed his three laws of planetary motion. In 1610, Kepler also first pointed out that an infinite universe with an infinite number of stars would be extremely bright and hot. This issue was taken up again by Edmund Halley in 1720 and Olbers in 1823. Olbers suggested that the universe was filled with dust that obscured light from the most distant stars. Only 20 years later, John

Herschel showed that this explanation would not work. The problem of *Olber's paradox* would not be resolved until the 20th century.

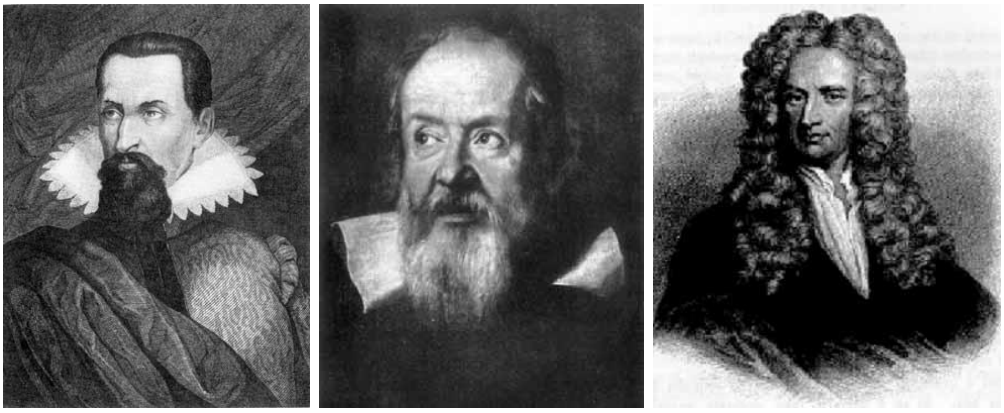


Figure 1.3: Left to right: Johannes Kepler, Galileo Galilei, and Sir Isaac Newton.

Galileo Galilei (1564–1642) found observational evidence for heliocentric motion, including the phases of Venus and the moons of Jupiter. He not only supported a heliocentric view of the universe in his book *Dialogue on the Two Great World Systems*, but his work on motion also attacked Aristotelian thought.

Sir Isaac Newton (1642–1727) discovered the mathematical laws of motion and gravitation that today bear his name. His *Philosophiae Naturalis Principia Mathematica*—or simply, the *Principia*—was the first book on theoretical physics, and provided a framework for interpreting planetary motion. He was thus the first to show that the laws of motion which applied in laboratory situations, could also apply to the heavenly bodies.

Newton also wrote about his own view of a cosmology with a static universe in 1691: he claimed that the universe was infinite but contained a finite number of stars. Self gravity would cause such a system to be unstable, so Newton believed (incorrectly) that the finite stars would be distributed infinitely far so that the gravitational attraction of stars exterior to a certain radius would keep the stars interior to that radius from collapsing.

The English astronomer **Thomas Wright** (1711–1786) published *An Original Theory or New Hypothesis of the Universe* (1750), in which he proposed that the Milky Way was a grouping of stars arranged in a thick disk, with the Sun near the center. The stars moved in orbits similar to the planets around our Sun.

Immanuel Kant (1724–1804), the German philosopher, inspired by Wright, proposed that the Milky Way was just one of many "island universes" in an infinite space. In his *General Natural History and Theory of Heaven* (1755), he writes of the nebulous objects that had been observed by others (including Galileo!), and reflects on what the true scale of the universe must be:

> Because this kind of nebulous stars must undoubtedly be as far away from us as the other fixed stars, not only would their size be astonishing (for in this respect they would have to exceed by a factor of many thousands the largest star), but the strangest point of all would be that with this extraordinary size, made up

of self-illuminating bodies and suns, these stars should display the dimmest and weakest light.

Figure 1.4: Immanuel Kant (left) and Sir William Herschel (right).

Sir William Herschel (1738–1822) and his son John used a telescope, based on a design by Newton, to map the nearby stars well enough to conclude that the Milky Way was a disk-shaped distribution of stars, and that the Sun was near the center of this disk. He mapped some 250 diffuse nebulae, but thought they were really gas clouds inside our own Milky Way. Others however took Kant's view that the nebulae were really distant galaxies. The German mathematician **Johann Heinrich Lambert** (1728–1777) adopted this idea, plus he discarded heliocentrism, believing the Sun to orbit the Milky Way like all of the other stars.

1.3 Early This Century

The argument over the location of the Sun inside the Milky Way, and the nature of the nebulae remained unresolved until early this century.

Harlow Shapley (1885–1972), an American astronomer, observed globular clusters and the RR Lyrae variable stars in them. From their directions and distances, he was able to show that they placed in a spherical distribution not centered on the Sun, but at a point nearly 5000 light years away. (We know today that Shapley over-estimated his distance by a factor of two.) The Copernican revolution was almost complete: not only was the Earth not at the center of the universe, but the Sun was far from the center of the Milky Way as well.

The American astronomer **Vesto Slipher** (1875–1969), working at Lowell Observatory, used spectroscopy to study the Doppler shift of spectral lines in the "spiral nebulae," thus establishing the rotation of these objects (1912–1920). Most of the galaxies (as they are known today) in his sample, except for M31, the Andromeda Galaxy, were found to be moving away from the Milky Way.

Albert Einstein (1879–1955) publishes his General Theory of Relativity in 1916, which explains how matter causes space and time to be warped. The resulting force of gravity

Figure 1.5: Harlow Shapley (left) and Herbert Curtis (right).

can now be thought as the motion of objects moving in a warped space-time. He realized that General Relativity could be used to explain the structure of the entire universe. He assumed that the universe obeyed the *cosmological principle*: it was infinite in size with the same average density of matter everywhere, with spacetime in the universe warped by the presence of matter within it. However he found that his equations predicted a universe to be either expanding or contracting, which appeared to contradict his sensibilities. Einstein as a result added a term into his equations, the *cosmological constant* to keep his model universe static.

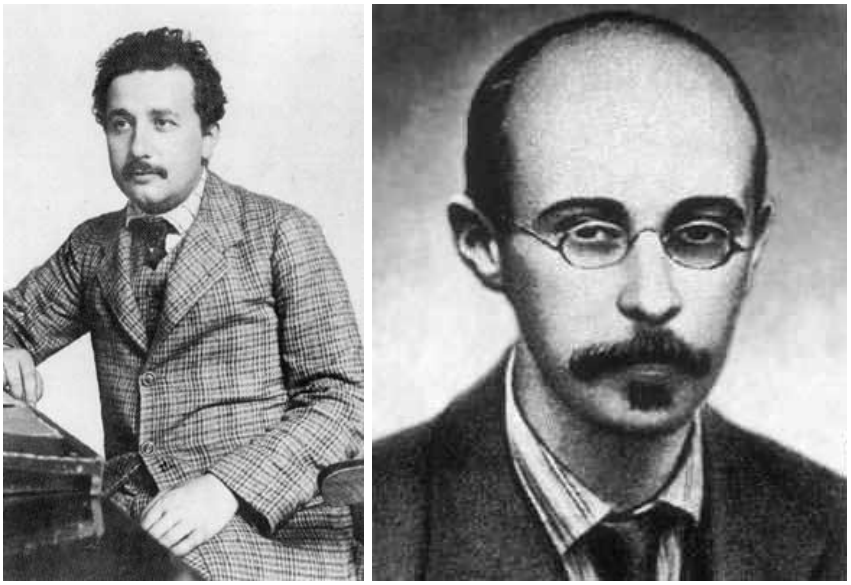


Figure 1.6: Albert Einstein (left) and Aleksandr Friedmann (right).

Dutch astronomer **Willem de Sitter** (1872–1934) used Einstein's General Relativity equations with a low (or zero) matter density but without the cosmological constant to arrive at an expanding universe (1916–1917). His view was that the cosmological constant:

> ... detracts from the symmetry and elegance of Einstein's original theory, one of whose chief attractions was that it explained so much without introducing any new hypothesis or empirical constant.

Russian mathematician **Aleksandr Friedmann** (1888–1925) finds a solution to Einstein's equation with no cosmological constant (1920), but with any density of matter. Depending on the matter density, his model universes either expanded forever or expanded and collapsed in a manner that was periodic with time. His work was dismissed by Einstein and generally ignored by other physicists.

In 1920, **Harlow Shapley** and **Herbert Curtis** held a debate on the "Scale of the Universe," or really about the nature of the "spiral" nebulae. Shapley argued that these were gas clouds inside our own Milky Way and that the universe consisted just of our Milky Way. Curtis on the other hand argued that they were other galaxies just like the Milky Way, but much further away. Although the debate laid open the positions of the two sides, nothing was immediately resolved. (That same year, Johannes Kapteyn was arguing that the Sun was in the center of a small Milky Way, based on star counts.) It was only in the following decade that as Edwin Hubble and other astronomers found novae and Cepheid variable stars in nearby galaxies, that Curtis' view was slowly adopted. (When a letter from Hubble describing the period-luminosity relation for Cepheids in M31 arrived at Shapley's office, Shapley held out the letter and said, "Here is the letter that destroyed my universe!")

Edwin Hubble (1889–1953) worked at Mt. Wilson Observatory, California in 1923–1925, to systematically survey spiral galaxies, following up on Slipher's work. In 1929 he published his observations showing that the galaxies around us appeared to be expanding, and this expansion followed "Hubble's Law:" $v = H_0 D$, which related the radial velocity of the galaxy with its distance. His *Hubble constant* $H_0 = 500$ km s^{-1}Mpc-1, nearly 10 times the current value. In 1927, the Belgian astronomer **Georges Lemaître** (1894–1966) independently arrived at Friedmann's solutions to Einstein's equations, and realized they must correctly describe the universe, given Hubble's recent discoveries. Lemaître was the first person to realize that if the universe has been expanding, it must have had a beginning, which he called the "Primitive Atom." This is the precursor to what is today known as the "Big Bang."

Figure 1.7: Edwin Hubble.

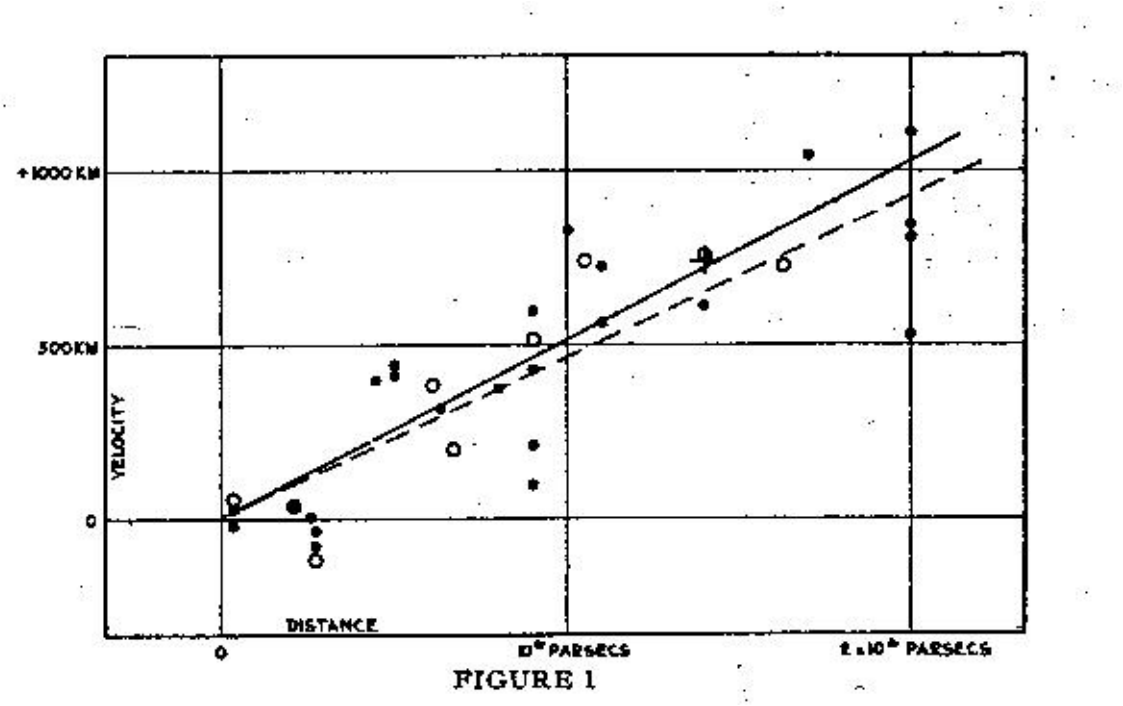


Figure 1.8: The figure from Edwin Hubble's original paper showing a linear relationship between the distance and the redshift of galaxies. From the March 15, 1929 issue of the *Proceedings of the National Academy of Sciences*, 15, 3.

By 1932, **Einstein** had come around to excepting the idea of an expanding universe. When he went to Mt. Wilson to meet Hubble, he said the invention of the cosmological constant was the "the biggest blunder of my life." That same year, he and **de Sitter** published a joint paper on their Einstein-de Sitter universe, an expanding universe without a cosmological constant.

Chapter 2

Overview of Modern Cosmology and Relativity

Cosmology requires a theory of gravity. Why? Because gravity is the dominant force in the universe, even though it is the weakest of the four fundamental forces (the strong nuclear force, the weak nuclear force, electromagnetism, and gravity):

1. The strong and weak nuclear forces fall off exponentially with distance.
2. Electrostatic and gravitational forces fall off as $1/r^2$. The Coulomb force is vastly more powerful, e.g., for two electrons:

$$\frac{F_{\rm Coul}}{F_{\rm grav}} = \frac{e^2/r^2}{Gm_e^2/r^2} = \frac{e^2}{Gm_e^2} = 4.2 \times 10^{42}$$

However, precisely because Coulomb forces are so strong, matter is neutral in bulk. *Gravity, on the other hand, dominates on large scales.*

2.1 Newtonian Gravity and Mechanics

Newton's Law of Gravity: $F = \dfrac{Gm_1m_2}{r_{12}^2}$

Newton's Laws of Mechanics:

1. Free particles move with $\boldsymbol{v}$ = constant ("Law of inertia").

2. $\boldsymbol{F} = m\boldsymbol{a}$

3. Reaction forces are equal and opposite:

$$\boldsymbol{F}_{21} = \boldsymbol{F}_{12}$$

Note that (1) is really a special case of (2).

By Phil Maloney.

Velocities and accelerations must be specified with respect to some reference frame, e.g., a rigid Cartesian frame. (This assumes Euclidean geometry—as everybody did before 1915!)

Newton's 1st Law singles out one class of reference frames as special—inertial frames. Only in inertial frames do Newton's Laws apply. A reference frame in which there are gravitational forces is *not* an inertial frame. Classically, the frame of the "fixed" stars was believed to represent an inertial frame.

2.2 Transformations Between Frames

Consider two Cartesian frames, S and S', with coordinates (x, y, z, t) and (x', y', z', t'), respectively. And assume S' moves in the x-direction of S with velocity v; the axes remain parallel at all times, and the origins coincide at time $t = t' = 0$.

Let some event happen at (x, y, z, t) relative to S and (x', y', z', t') relative to S'. The classical (common-sense) relation between the coordinates in the two frames is the *Galilean transform*:

$$x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t \tag{2.1}$$

where the spatial origins of the two frames are separated in the x-direction by a distance vt.

Differentiating Eq. 2.1 gives the classical velocity transformation law:

$$u'_1 = u_1 - v, \quad u'_2 = u_2, \quad u'_3 = u_3 \tag{2.2}$$

where $u_1 = dx/dt$, $u_2 = dy/dt$, $u_3 = dz/dt$, etc.

If S is an inertial frame, then so is S': linear equations of motion of S (of free particles) are transformed into similar linear equations of motion in S'. Furthermore, the acceleration is the same in both frames, as can be seen by differentiating Eq. 2.2.

Conversely, any inertial frame must move uniformly with respect to any other inertial frame, e.g., the frame of the "fixed stars." Newton's Laws apply in any inertial frame (in Newtonian mechanics), e.g., a moving ship.

However, something is fishy here. Note that acceleration is absolute—it is the same in all inertial frames. This raises the question—absolute with respect to what? The answer is—with respect to any inertial frame. But what singles out inertial frames as the standard of non-acceleration?

To answer this question, Newton postulated *absolute space*—this is supposed to act on every particle to resist changes in its velocity—that is, absolute space is the source of inertia. Newton identified it with the center of mass of the solar system. Later it was identified with the frame of the fixed stars. This isn't very satisfactory—there appears to be nothing to single out absolute space from the class of inertial frames.

2.3 Maxwell and the Ether

In Maxwell's equations of electromagnetism, a constant c with units of speed arises. The equations predict that electromagnetic waves propagate with speed c in vacuum. This constant is easily measured in laboratory experiments.

c coincided precisely with the known value for the speed of light in a vacuum. $\Longrightarrow$ Light is electromagnetic waves.

Maxwell postulated an "ether" to support electromagnetic waves. This ether was identified with absolute space. However, the Michelson-Morley experiment failed to detect any sign of the ether.

2.4 Einstein and Special Relativity

Einstein's solution to this puzzle is embodied in the *Equivalence Principle:* all inertial frames are completely equivalent. Combining this with the observed constancy of the speed of light in all frames leads to *Special Relativity*.

In Special Relativity, the Galilean transformation between reference frames is no longer correct (except in the limit of $v \ll c$). Instead, the relations between coordinates are given by the *Lorentz transformations* (more on these below). The Lorentz transformations lead to many apparently bizarre predictions—time dilation, length contraction, etc., which have been experimentally verified.

There is still sometihing missing, however: it is not possible to "patch" gravity onto Special Relativity. We can't put an inertial frame *around* a gravitating mass. Why are inertial frames singled out as special?

An additional clue was provided by the anomalous precession of the perihelion of Mercury ($43''$ century^{-1}), which was discrepant with Newtonian gravitation.

2.5 Mach's Principle

In developing General Relativity, Einstein was heavily influenced by the physicist-philosopher Ernst Mach. In particular, Mach denied the existence of absolute space, and proposed that inertia was the result of the mass of the rest of the universe acting on a particular body.

2.6 Inertial and Gravitational Mass

Newton's 2nd Law can be regarded as the definition of *inertial mass*:

$$\boldsymbol{F} = m_I \boldsymbol{a}$$

while Newton's Law of Gravity defines *gravitational mass*:

$$\boldsymbol{F}_{\rm grav} = \frac{G m_1 m_2}{r^2}$$

or

$$\boldsymbol{F}_{\rm grav} = \frac{G M m_G}{r^2}$$

where $F = GM/r^2$ is the force on a mass m_G due to some other mass M. Note that this means the acceleration of an object in a gravitational field is independent of its mass, *if* $m_I = m_G$. Experimentally, inertial and gravitational masses are identical to very high precision.

Einstein raised this equivalence to a postulate, which is the foundation of General Relativity. All local, freely falling, non-rotating laboratories (frames of reference) are completely equivalent as far as the laws of physics are concerned. ("Local" means small compared to gradients in the gravitational field.)

How does $m_I = m_G$ enter into this? Consider a laboratory which is freely falling towards the Earth, in a gravitational field $\boldsymbol{g}$. Suppose there is some mass (inertial mass m_I) in the lab, which is being acted on by total force $\boldsymbol{f}$, while $\boldsymbol{f}_G$ is the gravitational force acting on it; m_G is the gravitational mass. Then

$$\begin{aligned} \boldsymbol{f} &= m_I \boldsymbol{a} \\ \boldsymbol{f}_G &= m_G \boldsymbol{g} \end{aligned} \tag{2.3}$$

$\boldsymbol{g}$ is the acceleration of the lab; thus the acceleration of the mass relative to the lab is $\boldsymbol{a} - \boldsymbol{g}$, and so the force acting on it (relative to the lab) is

$$\boldsymbol{f}_{\rm rel} = (\boldsymbol{a} - \boldsymbol{g}) m_I .$$

The non-gravitational force acting on it is:

$$\boldsymbol{f}_{\rm NG} = \boldsymbol{f} - \boldsymbol{f}_G = m_I \boldsymbol{a} - m_G \boldsymbol{g} .$$

These are identical if $m_I = m_G$. Thus gravity has been transformed away, and we can construct *local* inertial frames anywhere, even near massive objects. (Note that this also means we can create gravity by acceleration.)

The Principle of Equivalence leads to two immediate predictions:

1. **Light bends in gravitational fields.** Consider a person standing in an elevator pointing a flashlight horizontally so that its beam points towards one of the side walls. Now give the elevator some constant acceleration upwards. The beam will appear to the person inside the elevator to curve downward. Since the Principle of Equivalence says that we cannot tell the difference between gravity and accelerated motion, such a beam of light should be bent in a gravitational field as well. Equivalently, space is curved in the presence of gravity.
2. **Light climbing out of a gravitational field suffers a red-shift; conversely, light falling down a gravitational field is blue-shifted.** Assume the lab is dropped just as light enters the top of the lab; then $\boldsymbol{v}_{\rm obs} = -\boldsymbol{g}t$:

   $$\nu_B = \nu_e (1 + v_{\rm obs}/c) = \nu_e (1 - gt/c)$$

   A observes the light ray just as he passes B. A observes no Doppler shift, so B must.

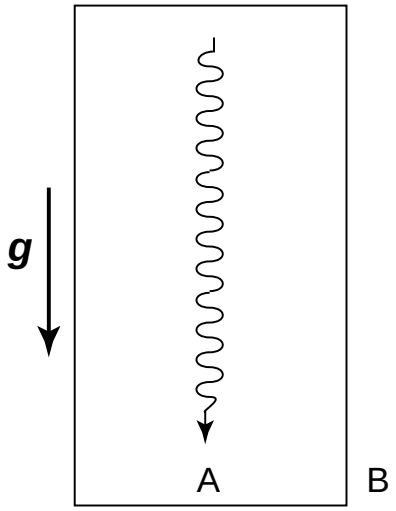


Figure 2.1: Light falling down a gravitational well.

Chapter 3

Special Relativity

To investigate the Lorentz transformations, consider two frames, S and S', in standard configuration:

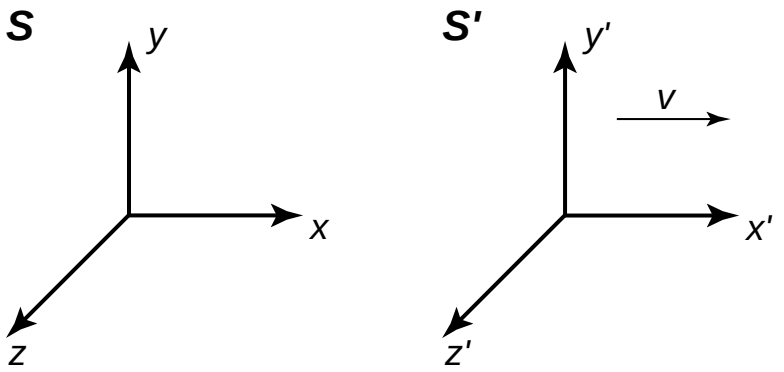


Figure 3.1: Coordinate frames S and S'.

We fix the axes parallel at all times; we also set the clocks in S and S' such that the origins coincide at $t = t' = 0$.

The transformation must be linear in coordinates. Trivially, $y = y'$, and $z = z'$. By linearity, and since $x = vt$ must correspond to $x' = 0$, x' must be of the form

$$x' = \gamma(x - vt); \qquad \gamma \text{ is a (possibly } v\text{-dependent) constant.} \tag{3.1}$$

Similarly, t' must be of the form

$$t' = mt - nx. \tag{3.2}$$

Now, suppose a light pulse is emitted at time $t = 0$, from the origin; this also occurs at time $t' = 0$ at the origin of the S' frame. Let r and r' denote the coordinates perpendicular to the direction of motion (i.e., $r = (z^2 + y^2)^{1/2}$). Since the speed of light c is a constant,

By Phil Maloney.

the distance travelled will be the same in both frames: at time t, the light pulse will have reached the surface of a sphere of radius ct in frame S, and of ct' in frame S'.

Hence

$$x^2 + r^2 = c^2\, t^2,$$

and similarly

$$x'^2 + r'^2 = c^2\, t^2,$$

or

$$\begin{aligned} c^2\, t^2 - x^2 &= r^2 \\ c^2\, t'^2 - x'^2 &= r'^2. \end{aligned}$$

Since perpendicular coordinates (y, z) are unaffected by motion in the x-direction, r and r' must be equal at all times. Thus,

$$c^2\, t^2 - x^2 = c^2\, t'^2 - x'^2. \tag{3.3}$$

If we substitute Eqs. 3.1 and 3.2 into 3.3, we get

$$(c^2m^2 - v^2\gamma^2)t^2 + 2(v\gamma^2 - c^2mn)tx - (\gamma^2 - c^2n^2)x^2 = c^2t^2 - x^2.$$

This must hold for *all* values of x and t. This can only be true if the coefficients on both sides are equal:

$$\begin{aligned} c^2m^2 - v^2\gamma^2 &= c^2 && \text{(a)} \\ v\gamma^2 - c^2mn &= 0 && \text{(b)} \\ \gamma^2 - c^2n^2 &= 1 && \text{(c).} \end{aligned}$$

This gives us 3 algebraic equations for the three unknowns γ, m, and n.

$$\begin{aligned} \gamma^2 &= 1 + c^2n^2 \\ c^2m^2 - v^2(1 + c^2n^2) &= c^2 \\ m^2 - \frac{v^2}{c^2}(1 + c^2n^2) &= 1 \\ m^2 &= 1 + \frac{v^2}{c^2}(1 + c^2n^2) = 1 + \frac{v^2}{c^2} + v^2n^2 \\ m^2 - v^2n^2 &= 1 + \frac{v^2}{c^2} \\ v(1 + c^2n^2) - c^2mn &= 0 \\ c^2nm - c^2n^2v &= v \\ c^2n(m - vn) &= v \\ n &= \frac{v}{c^2}(m - vn)^{-1} \\ \Longrightarrow (m - vn)(m + vn) &= 1 + \frac{v^2}{c^2}. \end{aligned}$$

Now,

$$\begin{aligned}
(m+vn) = m - vn + 2vn &= m - vn + 2v \cdot \frac{v}{c^2}(m-vn)^{-1} \\
&= m - vn + 2\frac{v^2}{c^2}(m-vn)^{-1} \\
\implies (m-vn)\left(m - vn + 2\frac{v^2}{c^2}(m-vn)^{-1}\right) &= 1 + \frac{v^2}{c^2} \\
(m-vn)^2 + 2\frac{v^2}{c^2} &= 1 + \frac{v^2}{c^2} \\
(m-vn)^2 = 1 - \frac{v^2}{c^2}.
\end{aligned}$$

Since we want to approach the Galilean transform as $v/c \to 0$, we must take the positive root:

$$\begin{aligned}
m - vn &= (1 - v^2/c^2)^{1/2} \\
n &= \frac{v}{c^2}(1 - v^2/c^2)^{-1/2}.
\end{aligned}$$

From (c):

$$\begin{aligned}
\gamma^2 = 1 + c^2 n^2 &= 1 + c^2 \frac{v^2}{c^4}(1 - v^2/c^2)^{-1} \\
&= (1 - v^2/c^2)^{-1} \\
\gamma &= [1 - v^2/c^2]^{-1/2}.
\end{aligned}$$

From (a):

$$\begin{aligned}
m^2 - \frac{v^2}{c^2}\gamma^2 &= 1 \\
m^2 = 1 + \frac{v^2}{c^2}\gamma^2 &= (1 - v^2/c^2)^{-1} = \gamma^2.
\end{aligned}$$

Thus the Lorentz transformations for x' and t' are:

$$x' = \frac{x - vt}{(1 - v^2/c^2)^{1/2}} \tag{3.4}$$

$$\begin{aligned}
t' &= \frac{t}{(1 - v^2/c^2)^{1/2}} - \frac{vx/c^2}{(1 - v^2/c^2)^{1/2}} \\
&= \frac{t - vx/c^2}{(1 - v^2/c^2)^{1/2}}.
\end{aligned} \tag{3.5}$$

The notation $\gamma = (1 - v^2/c^2)^{-1/2}$ for the *Lorentz factor* is standard, hence:

$$x' = \gamma(x - vt), \qquad y' = y, \qquad z' = z \tag{3.6}$$

$$t' = \gamma(t - vx/c^2). \tag{3.7}$$

Oddly enough, the Lorentz transformations were known before the advent of Special Relativity! They were known to be the transformations which formally left Maxwell's equations invariant, but their physical significance was not recognized. Thus Maxwell's equations were regarded as non-relativistic.

Special Relativity eliminates absolute time; instead we have a relativity of simultaneity.

The Lorentz transformations have a number of radical and counter-intuitive implications which will be discussed below.

3.1 Length Contraction

Consider a rod of length L_0 at rest in the S' frame; the rod is oriented along the x' axis. What is its length in the S frame?

Let $\Delta x = x_2 - x_1$, $\Delta y = y_2 - y_1$, etc.; denote the coordinate differences of two events in the S frame, and similarly in the S' frame. If we substitute these coordinates successively into Eqs. 3.6 and 3.7 and subtract, we get

$$\Delta x' = \gamma(\Delta x - v\Delta t), \qquad \Delta y' = \Delta y, \qquad \Delta z' = \Delta z \tag{3.8}$$

$$\Delta t' = \gamma(\Delta t - v\Delta x/c^2). \tag{3.9}$$

Let $\Delta x' = L_0$. To determine its length in the S frame, we must observe the ends at the *same time* in the S frame. This means $\Delta t = 0$ from Eq. 3.8, and so

$$\Delta x = L(S) = L_0/\gamma.$$

Since v/c is always < 1, γ is always > 1.

The rod is shortened in the direction of its motion by

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

Note that dimensions perpendicular to the direction of motion are unaffected. The rod has its greatest length in the frame in which it is at rest ($v = 0$). This is known as its *rest frame*.

3.2 Time Dilation

The analog of Eq. 3.9 giving Δt in terms of $\Delta t'$ and $\Delta x'$ is:

$$\Delta t = \gamma(\Delta t' + v\,\Delta x'/c^2). \tag{3.10}$$

Consider a clock which is fixed in the S' frame. Two events in the S' frame are separated by $\Delta t' = t'_2 - t'_1$. What time interval does an observer in the S frame see for these events?

Since the clock is fixed in S', $\Delta x' = 0$, and so

$$\Delta t = \gamma \Delta t' \equiv \gamma \Delta t_0,$$

where Δt_0 is the rest-frame time interval.

Clocks moving with velocity v with respect to an inertial frame S run slow by a factor $1/\gamma = (1 - v^2/c^2)^{1/2}$ relative to stationary clocks in S.

There is an analogous time dilation in a gravitational field, which leads to a *gravitational redshift*, which we will discuss shortly.

3.3 Velocity Transformations

In the Galilean transform, velocity addition is simple—this, however, is no longer the case in Lorentz transforms. Consider again two frames S and S' in standard configuration. Suppose a particle in S has velocity $\boldsymbol{u} = (u_x, u_y, u_z)$. What is its velocity $\boldsymbol{u'}$ in S'?

Assume the particle moves uniformly. then we can write its velocity in the two frames as:

$$\boldsymbol{u} = (u_x, u_y, u_z) = \left(\frac{\Delta x}{\Delta t}, \frac{\Delta y}{\Delta t}, \frac{\Delta z}{\Delta t}\right) \tag{3.11}$$

$$\boldsymbol{u'} = (u'_x, u'_y, u'_z) = \left(\frac{\Delta x'}{\Delta t'}, \frac{\Delta y'}{\Delta t'}, \frac{\Delta z'}{\Delta t'}\right). \tag{3.12}$$

Substituting from Eqs. 3.8 and 3.10 into Eq. 3.12:

$$\begin{aligned} u'_x &= \frac{\gamma(\Delta x - v\Delta t)}{\gamma(\Delta t - v\Delta x/c^2)} \\ &= \frac{\Delta x/\Delta t - v}{1 - v\Delta x/\Delta t/c^2} = \frac{u_x - v}{1 - u_x v/c^2} \qquad (3.13) \\ u'_y &= \frac{\Delta y}{\gamma(\Delta t - v\Delta x/c^2)} = \frac{\Delta y/\Delta t}{\gamma(1 - v\Delta x/\Delta t/c^2)} \\ &= \frac{u_y}{\gamma(1 - u_x v/c^2)} \qquad (3.14) \\ u'_z &= \frac{u_z}{\gamma(1 - u_x v/c^2)}. \qquad (3.15) \end{aligned}$$

3.4 Relativistic Doppler Effect

Consider first the *classical Doppler effect*. Suppose we have a light source emitting radiation with rest-frame wavelength λ_0. Consider an observer S, relative to whose frame the source is in motion with radial (towards the observer) velocity u_r.

Let the time between two successive pulses (i.e., wavecrests) in the source's rest frame be $\Delta t'$. The distance these two pulses have to travel to reach S differs by $u_r \Delta t'$. Since the pulses travel with speed c, they arrive at S with a time difference

$$\begin{aligned} \Delta t &= \Delta t' + u_r \, \Delta t'/c \\ \frac{\Delta t}{\Delta t'} &= 1 + \frac{u_r}{c}. \end{aligned}$$

Since the frequency is just $\nu_0 = 1/\Delta t'$, $\nu = 1/\Delta t$:

$$\lambda = c\Delta t \qquad\qquad \lambda_0 = c\Delta t'$$

$$\Longrightarrow \frac{\lambda}{\lambda_0} = 1 + \frac{u_r}{c}. \tag{3.16}$$

Now consider the *relativistic Doppler effect.* Since S' is in motion with respect to S, the time interval between pulses according to S is $\gamma \Delta t'$, due to time dilation. Thus

$$\begin{aligned} \Delta t &= \gamma \, \Delta t' + u_r \gamma \, \Delta t'/c \\ \Longrightarrow \frac{\lambda}{\lambda_0} &= \gamma \left(1 + \frac{u_r}{c}\right) = \frac{1 + u_r/c}{(1 - v^2/c^2)^{1/2}}. \end{aligned} \tag{3.17}$$

If the velocity is purely radial, $u_r = v$, so

$$\frac{\lambda}{\lambda_0} = \frac{1 + v/c}{(1 - v^2/c^2)^{1/2}} = \left(\frac{1 + v/c}{1 - v/c}\right)^{1/2}. \tag{3.18}$$

However there is a Doppler shift even if $u_r = 0$! If $u_r = 0$ (e.g., if S' is in a circular orbit about S), then

$$\frac{\lambda}{\lambda_0} = \frac{1}{(1 - v^2/c^2)^{1/2}}. \tag{3.19}$$

This is the *transverse Doppler effect*; it is a purely relativistic effect due to time dilation in the moving source.

3.5 Relativistic Mass

In Newtonian mechanics,

$$\boldsymbol{F} = m\boldsymbol{a} = m\frac{d\boldsymbol{v}}{dt} = \frac{d}{dt}\boldsymbol{p}, \tag{3.20}$$

where $\boldsymbol{p} = m\boldsymbol{v}$ is the linear momentum, and we have *assumed* that m is constant.

In relativistic mechanics, things are more complicated, as we will now see. We will *assume* that Newton's 2nd Law, in the form $\boldsymbol{F} = d\boldsymbol{p}/dt$, still holds, and also that mass is conserved, and see where this gets us.

Consider a perfectly inelastic collision (i.e., the particles stick together), from the point of view of our usual two frames S and S'. Let one of the particles be at rest in frame S and the other have velocity u, before they collide. After the collision, the particles stick together and move with velocity U.

We are free to pick our inertial frame any way we want, so for simplicity, pick S' to be the *center of mass* frame.

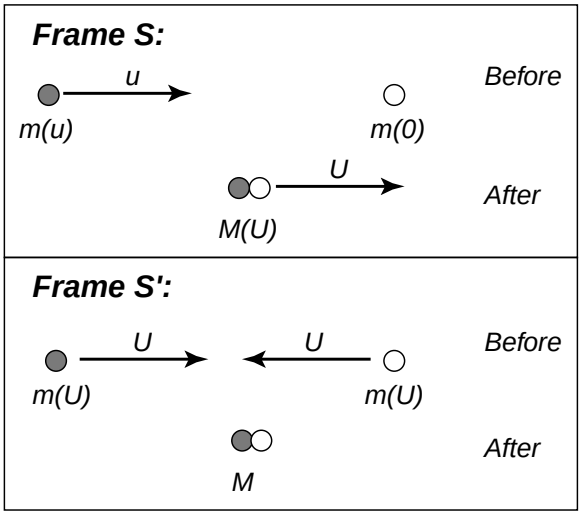


Figure 3.2: Collisions in a center of mass frame.

In the center of mass frame, a particle of mass $M(0)$ is at rest after the collision; the two particles collide with equal and opposite velocities. Remember that S' **must move at velocity** U **with respect to** S**.** From the conservation of mass in frame S:

$$m(u) + m(0) = M(U). \tag{3.21}$$

From conservation of momentum:

$$\begin{aligned} m(u)\,u - M(U)U &= 0 \\ m(u)\,u - [m(u) + m(0)]\,U &= 0, \end{aligned}$$

or

$$m(u) = m(0)\left(\frac{U}{u-U}\right). \tag{3.22}$$

The left hand particle has a velocity U relative to S'; S' in turn has a velocity U relative to S. Adding these two velocities must give the particle velocity u in S.

Recall the velocity transformation law (Eq. 3.13):

$$u'_x = \frac{u_x - v}{1 - u_x v/c^2} \tag{3.23}$$

for frame S' moving with velocity v. Here we want u_x in terms of u'_x. Recall frames are symmetric: to an observer in S', frame S is moving with velocity $-v$. Therefore, replace v with $-v$ and swap primes:

$$u_x = \frac{u'_x + v}{1 + u_x\, v/c^2}. \tag{3.24}$$

Here $u'_x = v$, and $v = U$, while $u_x = u$, so

$$u = \frac{2U}{1 + U^2/c^2};$$

solving for U in terms of u, we get a quadratic equation:

$$U^2 - \left(\frac{2c^2}{u}\right) U + c^2 = 0$$

which has roots

$$\begin{aligned} U &= \frac{c^2}{u} \pm \left[\left(\frac{c^2}{u}\right)^2 - c^2\right]^{1/2} \\ &= \frac{c^2}{u}\left[1 \pm \left(1 - \frac{u^2}{c^2}\right)^{1/2}\right]. \end{aligned} \tag{3.25}$$

In the limit $u \to 0$, this must produce a finite result, so we have to take the negative sign.

Substituting this into Eq. 3.22:

$$\begin{aligned}
m(u) &= m(0)\left[\frac{\frac{c^2}{u}\left[1-\left(1-\frac{u^2}{c^2}\right)^{1/2}\right]}{u-\frac{c^2}{u}\left[1-\left(1-\frac{u^2}{c^2}\right)^{1/2}\right]}\right] \\
&= m(0)\left[\frac{\frac{c^2}{u}\left(1-1/\gamma\right)}{u-\frac{c^2}{u}\left(1-1/\gamma\right)}\right] \\
&= m(0)\left[\frac{\frac{c^2}{u}\left(\gamma-1\right)}{\gamma u-\frac{c^2}{u}\left(\gamma-1\right)}\right] \\
&= m(0)\left[\frac{\gamma u}{\frac{c^2}{u}(\gamma-1)}-1\right]^{-1} \\
&= m(0)\left[\frac{\gamma\left(\frac{u^2}{c^2}\right)}{\gamma-1}-1\right]^{-1} \\
&= m(0)\left[\frac{\gamma\left(\frac{u^2}{c^2}\right)-(\gamma-1)}{\gamma-1}\right]^{-1} \\
&= m(0)\left[\frac{\gamma\left(\frac{u^2}{c^2}\right)-\gamma+1}{\gamma-1}\right]^{-1} \\
&= m(0)\left[\frac{\gamma\left(1-\gamma^{-2}\right)-\gamma+1}{\gamma-1}\right]^{-1} \\
&= m(0)\left[\frac{\gamma-\gamma^{-1}-\gamma+1}{\gamma-1}\right]^{-1} \\
&= m(0)\left[\frac{1-\gamma^{-1}}{\gamma-1}\right]^{-1} \\
&= \gamma\, m(0). \qquad (3.26)
\end{aligned}$$

Thus, **mass is not independent of velocity.** Here, $m(0)$ is the "rest mass," or

$$m(0) = \gamma^{-1}\, m = \left(1-\frac{u^2}{c^2}\right)^{1/2} m. \qquad (3.27)$$

Eq. 3.27 implies that **photons have zero rest mass.** This is why they can move at c; for any particle with non-zero rest mass, $m(u) \to \infty$ as $u \to c$.

Now assuming that u/c is small, let's expand Eq. 3.26:

$$\begin{aligned} m(u) = \gamma\, m_0 &= m_0 \left(1 - \frac{u^2}{c^2}\right)^{-1/2} \\ &= m_0 + \frac{1}{2} m_0 \frac{u^2}{c^2} + \cdots \end{aligned} \tag{3.28}$$

Multiplying both sides by c^2, we get:

$$mc^2 = m_0 c^2 + \frac{1}{2} m_0 u^2 + \cdots \text{ (higher order terms)}. \tag{3.29}$$

Note that the right-hand side just looks like a constant plus kinetic energy. thus the relativistic mass contains within it the expression for classical kinetic energy. In fact, conservation of relativistic mass just leads to conservation of energy in the Newtonian limit.

For example, suppose we have two particles with rest mass $m_{0,1}$ and $m_{0,2}$ which collide; their initial velocities are $v_{i,1}$ and $v_{i,2}$ and their final velocities are $v_{f,1}$ and $v_{f,2}$. Conservation of relativistic mass requires:

$$m_{0,1}\, \gamma(v_{i,1}) + m_{0,2}\, \gamma(v_{i,2}) = m_{0,1}\, \gamma(v_{f,1}) + m_{0,2}\, \gamma(v_{f,2}). \tag{3.30}$$

In the Newtonian limit ($v/c \ll 1$ for all v), we can expand the γs in Eq. 3.30:

$$\begin{aligned} m_{0,1} \left(1 + \frac{1}{2}\frac{v_{i,1}^2}{c^2}\right) &+ m_{0,2} \left(1 + \frac{1}{2}\frac{v_{i,2}^2}{c^2}\right) \\ &= m_{0,1} \left(1 + \frac{1}{2}\frac{v_{f,1}^2}{c^2}\right) + m_{0,2} \left(1 + \frac{1}{2}\frac{v_{f,2}^2}{c^2}\right). \end{aligned}$$

Multiplying by c^2 and subtracting the constant terms from both sides, we get

$$\frac{1}{2} m_{0,1}\, v_{i,1}^2 + \frac{1}{2} m_{0,2}\, v_{i,2}^2 = \frac{1}{2} m_{0,1}\, v_{f,1}^2 + \frac{1}{2} m_{0,2}\, v_{f,2}^2 \tag{3.31}$$

which is just the usual conservation of energy equation.

Eq. 3.29 suggests that we regard $E = mc^2$ as the total energy of a particle; this consists of the kinetic energy plus the *rest-mass energy* $m_0 c^2$. The latter is a huge quantity; one gram of rest mass is equivalent to 9×10^{20} erg $\approx$ 20 kilotons. Let's *define* the kinetic energy of a particle to be:

$$K = mc^2 - m_0 c^2 = m_0 c^2 (\gamma - 1). \tag{3.32}$$

For $u/c \ll 1$, this reduces to the usual $K = \frac{1}{2} m_0 u^2$. Similarly, the relativistic momentum is:

$$\boldsymbol{p} = m\boldsymbol{u} = \gamma m_0 \boldsymbol{u}.$$

Classically, energy and momentum are related by $E = \frac{1}{2}mv^2 = p^2/2m$. What is the relativistic relation?

Squaring the expression for relativistic momentum,

$$\begin{aligned}
p^2 &= m_0^2\, u^2\, \gamma^2 = \frac{m_0^2\, u^2}{1 - u^2/c^2} \\
\frac{p^2}{c^2} &= \frac{m_0^2\, u^2/c^2}{1 - u^2/c^2} \\
\left(1 - \frac{u^2}{c^2}\right)\frac{p^2}{c^2} &= \frac{m_0^2 u^2}{c^2} \\
\left(m_0^2 + \frac{p^2}{c^2}\right)\frac{u^2}{c^2} &= \frac{p^2}{c^2} \\
\frac{u^2}{c^2} &= \frac{p^2}{m_0^2\, c^2 + p^2} \\
\gamma^2 &= (1 - u^2/c^2)^{-1} \\
&= \left[\frac{m_0^2\, c^2 + p^2 - p^2}{m_0^2\, c^2 + p^2}\right]^{-1} \\
&= 1 + p^2/m_0^2\, c^2 \\
\Longrightarrow E^2 &= m_0^2\, c^4 \gamma^2 \\
&= m_0^2\, c^4 + p^2\, c^2 \\
&= (m_0\, c^2)^2 + (p\, c)^2. \qquad (3.33)
\end{aligned}$$

Note again that we can have particles with zero rest mass (e.g., the photon, and maybe the neutrino(s)) with non-zero energy and momentum; these obey:

$$E = p\, c. \qquad (3.34)$$

In order to have non-zero momentum, the relativistic momentum

$$\boldsymbol{p} = \gamma\, m_0\, \boldsymbol{u} = \frac{m_0 \boldsymbol{u}}{(1 - u^2/c^2)^{1/2}}$$

must go to a finite value as $m_0 \to 0$; this requires that $u \to c$ as $m_0 \to 0$. Hence all massless particles must travel with speed c.

The relativistic mass of the photon is non-zero:

$$\begin{aligned}
E = m\, c^2 &= p\, c \\
\Longrightarrow m &= p/c. \qquad (3.35)
\end{aligned}$$

And since $E = h\nu$,

$$p = h\nu/c \qquad \Longrightarrow \qquad m = h\nu/c^2. \qquad (3.36)$$

Since the relativistic, inertial mass of the photon is non-zero, photons are acted on by gravity (as we have seen already from the Equivalence Principle).

3.6 Gravitational Redshift

Suppose a photon of frequency ν_e is emitted at the surface of a body of mass M and radius R. the photon escapes to infinity. What is the frequency as observed at infinity?

In order to escape from the gravitational field, the photon must do work. The work done per unit mass is just

$$\int_R^\infty \frac{GM}{r^2}\,dr = \frac{GM}{R},$$

where we have just the potential difference. And so, since the photon's inertial mass is $m = h\nu_0/c^2$, the energy loss is just

$$\Delta E = \frac{GM}{R}\frac{h\nu_0}{c^2}.$$

Denote the frequency observed at infinity as ν_0. then

$$h\nu_e - h\nu_0 = \frac{GM}{R}\frac{h\nu_0}{c^2}, \tag{3.37}$$

so

$$\frac{\nu_e - \nu_0}{\nu_0} = \frac{GM}{Rc^2}.$$

It is conventional to define the *redshift* by

$$\begin{aligned} z &= \frac{\lambda_0 - \lambda_e}{\lambda_e} \\ &= \frac{c/\nu_0 - c/\nu_e}{c/\nu_e} = \frac{\nu_e}{c}\left[\frac{c}{\nu_0} - \frac{c}{\nu_e}\right] \\ &= \frac{\nu_e}{\nu_0} - 1 = \frac{\nu_e - \nu_0}{\nu_0}. \end{aligned}$$

Thus the *gravitational redshift* is

$$z = \frac{GM}{Rc^2}. \tag{3.38}$$

Completely equivalently, since we can regard emitting atoms as clocks, we can regard this as *gravitational time dilation*: clocks run slower in a gravitational field.

$$\frac{\Delta t(r)}{\Delta t(\infty)} = 1 - \frac{GM}{rc^2}. \tag{3.39}$$

3.7 Spacetime

As we have seen,

Relativity $\Longrightarrow$ Lorentz Transforms $\Longrightarrow$
Elimination of Absolute Time and Absolute Space

Spatial and time coordinates are "mixed" for different observers. It no longer makes sense to talk about space and time separately, as in classical physics. Instead, we have a single, 4-D entity called *spacetime*.

The fundamental quantity in spacetime is not position, or time, but an *event*. An event is specified by four quantities, e.g., x, y, z, t.

Consider some event O, which we take as the origin of our coordinate system. Fire off a light pulse at O. What will we see with increasing time? The wavefront should expand outward at speed c; hence at time t, it has reached a distance ct from O. We can't draw in four dimensions, so let's drop one of the spatial dimensions. What does a spacetime diagram look like?

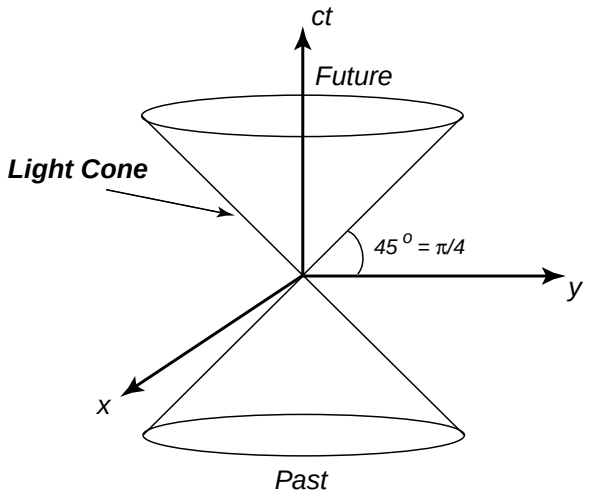


Figure 3.3: A spacetime diagram.

With one spatial dimension suppressed, the wavefront of light pulse looks like a cone—this is called the *light cone*. With ct for the vertical axis (hence letting all the coordinate axes have the same units or dimensions), the light cone makes an angle of 45° with the spatial and ct axes.

Similarly, we can consider some time $-t$ *before* event O. Only photons at a distance ct from O can reach O between times $-t$ and O. As $-t$ gets closer to O, the size of the light wavefront which can reach O shrinks.

Thus, the wavefront collapses to zero at O, then re-expands (symmetrically).

With one spatial dimension suppressed, we thus have a *past light cone* and a *future light cone*. Since nothing can travel faster than liight, the light cones divide spacetime (as seen by an observer at O) into accessible and inaccessible regions; not all directions are equivalent in the spacetime diagram.

Spacetime is not isotropic.

We can quantify this: If we send out a light pulse from the origin of a coordinate system at $t = 0$ (assuming Euclidean space or Cartesian coordinates), its radial distance from the origin is ct:

$$c^2\,t^2 - x^2 - y^2 - z^2 = 0.$$

If we consider two events which are connected by a light ray, then

$$c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2 = 0.$$

(This just says the distance traveled by the light ray, $c\Delta t$, is equal to the spatial distance between the two events, $[\Delta x^2 + \Delta y^2 + \Delta z^2]^{1/2}$.) Now recall the difference form of our standard Lorentz transformation:

$$\begin{aligned} \Delta x' &= \gamma(\Delta x - v\Delta t), \qquad \Delta y' = \Delta y, \qquad \Delta z' = \Delta z \\ \Delta t' &= \gamma(\Delta t - v\,\Delta x/c^2) \\ \gamma &= (1 - v^2/c^2)^{-1/2}. \end{aligned}$$

Using these expressions, we can show that:

$$\begin{aligned} c^2\Delta t'^2 - \Delta x'^2 - \Delta y'^2 - \Delta z'^2 &= c^2\gamma^2(\Delta t^2 - 2v\,\Delta x\,\Delta t/c^2 + v^2\Delta x^2/c^4) \\ &\quad -\gamma^2(\Delta x^2 - 2v\,\Delta x\,\Delta t + v^2\Delta t^2) - \Delta y^2 - \Delta z^2 \\ &= c^2\gamma^2\Delta t^2 - 2\gamma^2 v\,\Delta x\,\Delta t + \gamma^2 v^2\Delta x^2/c^2 \\ &\quad -\gamma^2\Delta x^2 + 2\gamma^2 v\,\Delta x\,\Delta t - \gamma^2 v^2\Delta t^2 - \Delta y^2 - \Delta z^2 \\ &= \gamma^2 c^2\Delta t^2 - \gamma^2\Delta x^2 - \gamma^2 v^2\Delta t^2 + \gamma^2 v^2\Delta x^2/c^2 - \Delta y^2 - \Delta z^2 \\ &= \gamma^2(c^2\Delta t^2 - \Delta x^2 - v^2\Delta t^2 + v^2\Delta x^2/c^2) - \Delta y^2 - \Delta z^2 \\ &= \frac{c^2}{c^2 - v^2}(c^2\Delta t^2 - \Delta x^2 - v^2\Delta t^2 + v^2\Delta x^2/c^2) - \Delta y^2 - \Delta z^2 \\ &= \frac{c^2}{c^2 - v^2}\left(\Delta t^2(c^2 - v^2) - \frac{\Delta x^2}{c^2}(c^2 - v^2)\right) - \Delta y^2 - \Delta z^2 \\ &= c^2\,\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2. \end{aligned}$$

Thus the interval $\Delta s^2 \equiv c^2\,\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$ between the two events is unchanged by a Lorentz transformation; it is *Lorentz invariant*. (Note that Δs^2 is a scalar quantity.) This is analogous to the spatial separation between two points in Euclidean space, $\Delta r^2 = \Delta x^2 + \Delta y^2 + \Delta z^2$, staying unchanged after a change of coordinates.

There is an important distinction, however: Δr^2 is *always* positive. This is *not* true of the interval $\Delta s^2 = c^2\,\Delta t^2 - \Delta r^2$, which may be positive, negative, or zero.

We have already seen that for two events separated by a light ray,

$$\Delta s^2 = 0.$$

For obvious reasons, this is called a *light-like* separation.

If $\Delta s^2 > 0$, that means that

$$c^2\,\Delta t^2 > \Delta r^2,$$

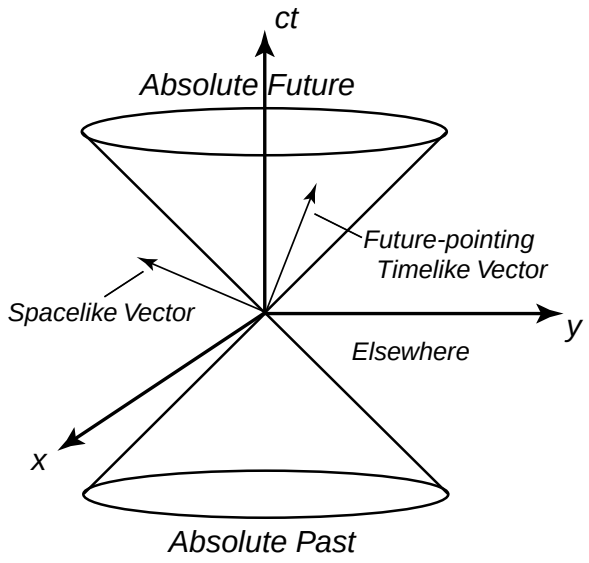


Figure 3.4: Past and future in a spacetime diagram.

or

$$\frac{\Delta r^2}{\Delta t^2} < c^2$$

in any inertial frame (since Δs^2 is Lorentz invariant). This means that it is possible for an observer moving with uniform velocity $v < c$ to travel from one event to the other; in this observer's rest frame $\Delta r = 0$, and the time separation between events is just $\Delta t = \Delta s/c$.

Thus when $\Delta s^2 > 0$, Δs is equal to c times the time difference Δt between the events, as seen by the inertial observer for whom the events take place at the same point. Thus events for which $\Delta s^2 > 0$ can lie on the world line of a material particle.

$\Delta s^2 > 0$ **is called a *time-like* separation.**

If $\Delta s^2 < 0$, then

$$\frac{\Delta r^2}{\Delta t^2} > c^2$$

which again is true in any inertial frame. It is impossible for two events with $\Delta s^2 < 0$ to be connected by a light ray, or to lie on the world line of a material particle, as this would require superluminal travel.

There is still a physical meaning in this case, however:

$$\Delta s^2 = c^2\,\Delta t^2 - \Delta r^2,$$

which implies that $|\Delta s^2|$ is the spatial separation between the events in an inertial frame in which the events are simultaneous; such a frame always exists, as can be seen from the Lorentz transformation.

$\Delta s^2 < 0$ is called a *space-like* separation.

This is known as *Minkowski spacetime*, or *flat* spacetime, since the geometry is Euclidean.

Since Δs^2 is invariant, light cones in one inertial frame are mapped into light cones in any other inertial frame. *All inertial observers agree on the past and future of an event.*

3.8 Spacetime Continued

Suppose we have some particle in motion (for convenience, along the x-axis of our coordinate system). if we plot its position vs. time, we construct a *spacetime* diagram shown in Fig. 3.5.

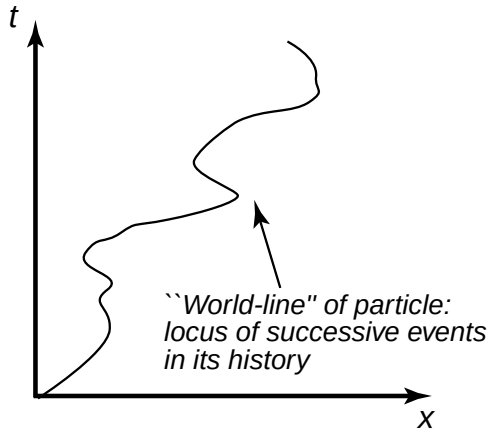


Figure 3.5: Spacetime diagram for a particle.

In the Lorentz transformations, space and time get "blended" together, analogous to a rotation of coordinate axes. How are the spacetime diagrams of S and S' related?

Let the vertical axis units be ct; then a light ray has slope $\pi/4 = 45°$. As usual, we synchronize clocks at $t = t' = 0$. What are the ct' and x' axes in this diagram?

From the Lorentz transformations,

$$x' = \gamma(x - vt) \tag{3.40}$$

$$ct' = c\gamma(t - vx/c^2). \tag{3.41}$$

The ct' axis is the line $x' = 0$; from Eq. 3.40, this means

$$\begin{aligned} t &= \frac{x}{v} \\ ct &= \left(\frac{c}{v}\right).x \end{aligned}$$

Thus the ct' axis is the straight line $ct = (c/v)x$ with slope $c/v > 1$. Similarly the x' axis is the line $ct' = 0$; from Eq. 3.41 we obtain

$$ct = \left(\frac{v}{c}\right) x$$

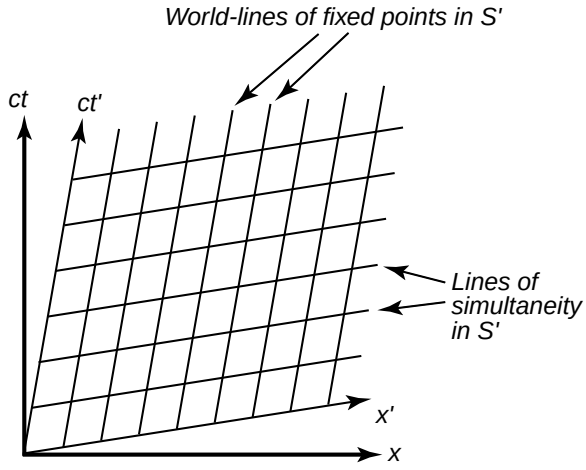


Figure 3.6: Spacetime diagram for frames S and S'.

so the x' axis is the line $ct = (v/c)x$ with slope $v/c < 1$.

Since the ct and x axes are orthogonal and the slopes of the ct' and x' axes are reciprocals of one another, the angles between the x' and x axes and the ct' and ct axes are equal.

Chapter 4

General Relativity

4.1 General Relativity and Curved Space Time

Last time we talked about the spacetime interval

$$\Delta s^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$$

and showed that this is Lorentz-invariant, where:

$$\begin{array}{ll} \Delta s^2 = 0 & \text{Light-like separation;} \\ \Delta s^2 > 0 & \text{Time-like separation;} \\ \Delta s^2 < 0 & \text{Space-like separation.} \end{array}$$

We had assumed Euclidean geometry (i.e., Cartesian coordinates). This is true for *flat* spacetimes (also known as Minkowski space); this is the standard geometry for *Special Relativity*.

However in the presence of matter, spacetime does not have Euclidean geometry. We have seen hints of this via Einstein's Equivalence Principle where we found that in a reference frame that in a gravitational field, light rays tend to bend. Thus in general spacetime will not be flat, and coordinates will not be Euclidean.

Directly related to the spacetime interval Δs^2 is the *proper time interval* $\Delta\tau^2$:

$$\Delta\tau^2 = \frac{\Delta s^2}{c^2} = \Delta t^2 - \frac{\Delta x^2 + \Delta y^2 + \Delta z^2}{c^2}. \tag{4.1}$$

This gets its name from the fact that this is the time measured by a clock moving with a particle, for in the particle's instantaneous rest frame, $\Delta x = \Delta y = \Delta z = 0$.

Equivalently, suppose that we have an observer whose velocity at time t is $\boldsymbol{v}$, where

$$\boldsymbol{v} = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right).$$

By Phil Maloney.

Relative to a Cartesian set of coordinates, t is the coordinate time. Then

$$\begin{aligned}\Delta\tau^2 &= \Delta t^2\left(1-\frac{1}{c^2}\left(\frac{\Delta x^2}{\Delta t^2}+\frac{\Delta y^2}{\Delta t^2}+\frac{\Delta z^2}{\Delta t^2}\right)\right)\\ &= \Delta t^2\left(1-\frac{1}{c^2}\left(v_x^2+v_y^2+v_z^2\right)\right)\\ &= \Delta t^2\left(1-\frac{v^2}{c^2}\right).\end{aligned}$$

And we recover the usual time dilation expression.

As with Δs^2, the separations are:

$$\begin{aligned}\Delta\tau^2 = 0 &\quad \text{Light-like separation;}\\ \Delta\tau^2 > 0 &\quad \text{Time-like separation;}\\ \Delta\tau^2 < 0 &\quad \text{Space-like separation.}\end{aligned}$$

We need four coordinates to designate the position of a particle in spacetime. Denote these as (x^0, x^1, x^2, x^3). The convention is to take x^0 as the time coordinate, so x^1, x^2, and x^3 are the spatial coordinates. In Euclidean geometry, these are x, y, and z, but this will not in general be the case.

The Principle of General Covariance: The laws of physics must take the same form no matter what coordinates we use to describe events.

This is not true for example, Newton's Laws, or the equations of Special Relativity, as these hold only in inertial frames—i.e., the coordinates cannot rotate or accelerate. Einstein however produced a completely covariant set of equations for General Relativity, and they are very complicated.

We have already seen that relativity tosses out absolute time. We have also seen, however, that the spacetime interval Δs^2, or equivalently the proper time interval $\Delta\tau^2$, is a Lorentz-invariant quantity. We will therefore use the proper time τ (the time read by a clock traveling with a material body along its world-line) as the time coordinate. τ is *always* a good coordinate, even in non-inertial frames, as τ increases monotonically along a body's world-line.

Suppose the particle is in a gravitational field. By the Equivalence Principle, we can choose a local inertial frame which is freely falling, in which Special Relativity applies; we can then use local Cartesian coordinates, and the proper time interval is given by Eq. 4.1. We will not get into the full equations of General Relativity since that would involve tensor calculus. We do however need to have general expressions for the separation $\Delta\tau^2$ of two events in spacetime.

To do this, consider regular 3-D space, where the spatial separation between a pair of points is just:

$$\Delta r^2 = \Delta x^2 + \Delta y^2 + \Delta z^2. \tag{4.2}$$

This distance in space is independent of our choice of coordinates.

Suppose we introduce a new set of general coordinates, x^1, x^2, and x^3, and write the original Cartesian coordinates x, y, and z, in terms of these new coordinates:

$$x = x(x^1, x^2, x^3,), \; y = y(x^1, x^2, x^3), \; z = z(x^1, x^2, x^3).$$

By simple calculus, the separation Δx, for example, is then given by

$$\Delta x = \frac{\partial x}{\partial x^1}\Delta x^1 + \frac{\partial x}{\partial x^2}\Delta x^2 + \frac{\partial x}{\partial x^3}\Delta x^3. \tag{4.3}$$

With similar expressions for Δy and Δz in terms of Δx^1, Δx^2, and Δx^3, if we then substitute Eq. 4.3 and its Δy and Δz analogs into Eq. 4.2, then we get:

$$\begin{aligned}
\Delta r^2 &= \left[\left(\frac{\partial x}{\partial x^1}\right)^2 + \left(\frac{\partial y}{\partial x^1}\right)^2 + \left(\frac{\partial z}{\partial x^1}\right)^2\right](\Delta x^1)^2 \\
&\quad +2\left[\frac{\partial x}{\partial x^1}\frac{\partial x}{\partial x^2} + \frac{\partial y}{\partial x^1}\frac{\partial y}{\partial x^2} + \frac{\partial z}{\partial x^1}\frac{\partial z}{\partial x^2}\right]\Delta x^1\Delta x^2 + \cdots \\
&= \sum_{\mu=1}^{3}\sum_{\nu=1}^{3} g_{\mu\nu}(x^1, x^2, x^3)\Delta x^\mu \Delta x^\nu,
\end{aligned} \tag{4.4}$$

where

$$g_{\mu\nu} = \left(\frac{\partial x}{\partial x^\mu}\frac{\partial x}{\partial x^\nu} + \frac{\partial y}{\partial x^\mu}\frac{\partial y}{\partial x^\nu} + \frac{\partial z}{\partial x^\mu}\frac{\partial z}{\partial x^\nu}\right). \tag{4.5}$$

We can make this more compact by using Einstein's summation convention; we automatically sum over any index which appears twice. This is true for both μ and ν in Eq. 4.4, so

$$\Delta r^2 = g_{\mu\nu}(\boldsymbol{r})\Delta x^\mu \Delta x^\nu, \tag{4.6}$$

where $\boldsymbol{r} = (x^1, x^2, x^3)$.

Eq. 4.6 tells us how to find the spatial separation of two points given the coordinate differences between them.

Note again that Δr^2 is invariant, but the coordinate differences are not, as they depend on the coordinate system we choose (which can be arbitrary).

In 3-D space, there are nine functions in $g_{\mu\nu}$:

$$g_{\mu\nu} = \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{pmatrix}. \tag{4.7}$$

But only 6 of these terms are independent, since (as is obvious from Eq. 4.5), $g_{\mu\nu} = g_{\nu\mu}$. $g_{\mu\nu}$ is the *metric tensor*.

Tensors are quantities which transform between coordinate systems in a particular way. A tensor of *rank* 0 is just a scalar, i.e., a single function of position which is the same in all coordinate systems. In an n-dimensional space, a tensor of rank one is an n-dimensional vector, i.e., a set of n functions. For example, $\Delta\boldsymbol{r} = (\Delta x^1, \Delta x^2, \Delta x^3)$ is a rank 1 tensor.

A tensor of rank 2 is a set of n^2 functions, e.g., the metric tensor. The simplest form of the metric tensor is found if we use Cartesian coordinates (x, y, z). Then from Eq. 4.5,

$$\begin{aligned} g_{\mu\nu} &= 0 \qquad \mu \neq \nu \\ g_{\mu\nu} &= 1 \qquad \mu = \nu, \end{aligned}$$

and the metric tensor takes the diagonal form:

$$g_{\mu\nu} = g^0_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.8}$$

We can make a completely analogous argument for $\Delta\tau^2$ in spacetime; the proper time interval in an arbitrary reference frame is given by

$$\Delta\tau^2 = g_{ij}\Delta x^i \Delta x^j = \sum_{i=1}^{4}\sum_{j=1}^{4} g_{i,j}\Delta x^i \Delta x^j. \tag{4.9}$$

The spacetime metric tensor g_{ij} has sixteen components; like $g_{\mu\nu}$, not all of these are independent: $n_{\text{indep}} = (16-4)/2 + 4 = 10$. As for the spatial metric tensor $g_{\mu\nu}$, there is a simplest possible form for g_{ij}; this occurs in a freely-falling reference frame. In this case, from Eq. 4.1, using x^0 as the time coordinate,

$$g_{ij} = g^0_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1/c^2 & 0 & 0 \\ 0 & 0 & -1/c^2 & 0 \\ 0 & 0 & 0 & -1/c^2 \end{pmatrix}. \tag{4.10}$$

In general however, for arbitrary (e.g., accelerating, rotating) reference frames, the components of g_{ij} depend on the event coordinates x^i; this in general is why General Relativity is so complicated.

There are two important differences between $g_{\mu\nu}$ and g_{ij}:

1. Elements of $g_{\mu\nu}$ are always positive; thus $\Delta\boldsymbol{r}$ can never be zero for $\Delta x, \Delta y, \Delta z \neq 0$.

   In regular 3-D space, there are no such distinctions as time-like and space-like. Metrics such as the spacetime g_{ij} for which $\Delta\tau^2$ can be zero are called indefinite; in contrast, $g_{\mu\nu}$ is definite.

2. The spatial metric $g_{\mu\nu}$ can always be put in diagonal form $g^0_{\mu\nu}$ by a suitable choice of coordinates.

   In contrast, g_{ij} can be reduced to g^0_{ij} only locally, via the Equivalence Principle. This is just a restatement of the fact that it is not possible to surround a gravitating mass with a single inertial reference frame. In general, g_{ij} is much more complicated than g^0_{ij} in presence of gravitational fields: spacetime is *curved*.

4.2 Geodesics and Spatial Curvature

Consider once again normal space, rather than spacetime. We will define a *geodesic* to be the shortest distance between two points. In a plane, this is obviously a straight line; on a sphere, it is a great circle. Geodesics are *intrinsic* properties of a surface; that is, they can be determined entirely by measurements made within that surface (e.g., by 2-D beings on the surface of a 2-D sphere, without making references to the fact that the sphere is embedded in 3-D space). Since they are intrinsic, they remain unaltered even if the surface is bent.

Consider a plane, a sphere, and a saddle; on each surface draw a *geodesic circle* of radius r.

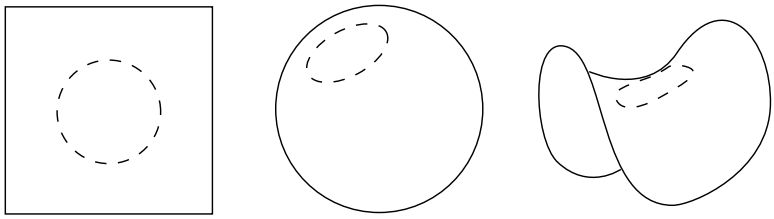


Figure 4.1: A geodesic circle on a plane, a sphere, and a saddle.

(By geodesic circle, we mean the locus of points connected by geodesics of length r to a common center.) If we cut these circles out of each surface and tried to flatten them into a plane, we would get Fig. 4.2.

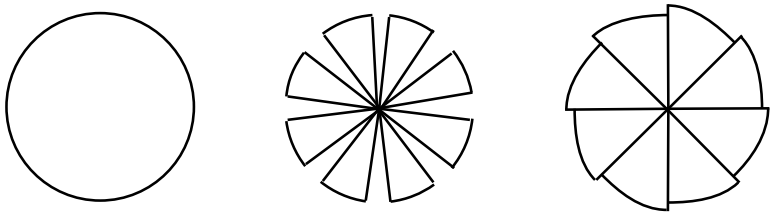


Figure 4.2: Flattened geodesic circles from a plane, a sphere, and a saddle.

Clearly the circle on the sphere has too little surface area relative to the planar circle, while the saddle has too much.

In plane (Euclidean) geometry, the circumference of a circle is $C = 2\pi r$, while the area is $A = \pi r^2$. Clearly on a sphere, C and A are smaller than the Euclidean values, while on the saddle C and A are larger than the Euclidean values.

We can quantify these differences as follows: consider two geodesics on a sphere (i.e., great circles), which pass through the pole. Let the radius of the sphere be a; the angle subtended at the pole by the two geodesics is θ, and $\theta \ll 1$. At a distance r along the geodesics from the pole P, let their perpendicular separation be η.

From simple geometry, perpendicular distance x from surface (at r) to midline of the

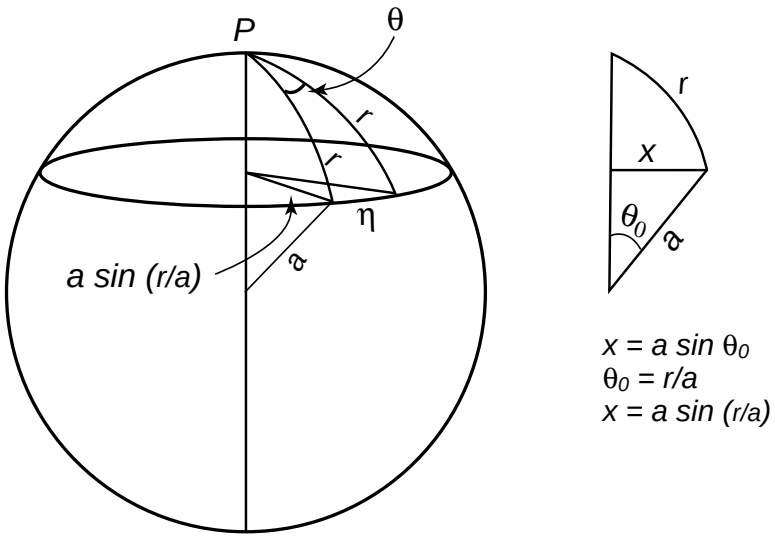


Figure 4.3: Geodesics on a sphere.

sphere is $x = a \sin r/a$, while η is just given by $\eta = \theta x$, or

$$\eta = \theta a \sin\left(\frac{r}{a}\right).$$

Expand the sine function into a Taylor series:

$$\eta = \theta a \left(\frac{r}{a} - \frac{1}{3!}\frac{r^3}{a^3} + \cdots\right) = \theta\left(r - \frac{r^3}{6a^2} + \cdots\right).$$

The circumference of the circle is obtained by letting $\theta \to 2\pi$:

$$C = 2\pi\left(r - \frac{r^3}{6a^2} + \cdots\right).$$

Define the curvature of a sphere of radius a to be $K = 1/a^2$; then to 3rd order in r,

$$C = 2\pi\left(r - \frac{r^3}{6}K\right) = 2\pi r\left(1 - \frac{r^2}{6}K\right). \tag{4.11}$$

To the same order r, our expression for η becomes:

$$\eta = \theta\left(r - \frac{r^3}{6}K\right). \tag{4.12}$$

It can be proved (although we won't) that Eq. 4.12 is valid to $\mathcal{O}(r^3)$ for *any* surface, (provided that it is sufficiently differentiable, i.e., non-pathological). Eq. 4.11 therefore provides us with a general definition for the curvature K:

$$K = \frac{3}{\pi}\lim_{r\to 0}\frac{2\pi r - C}{r^3}, \tag{4.13}$$

and so the curvature can be determined by local measurements.

So what is the curvature? It is a measure of the *spread* of the geodesic in any direction from a point.

1. For a plane, $K = 0$, and $C = 2\pi r$, or *parallel lines remain parallel*; equivalently the area A is given by $A = \pi(r^2 - \frac{r^4}{12}K)$ $(= \int C\, dr)$.
2. For a sphere, $K > 0$; $C < 2\pi r$, $A < \pi r^2$.
3. For a saddle, $K < 0$; $C > 2\pi r$, $A > \pi r^2$.

We now want to relate the curvature to the metric tensor, $g_{\mu\nu}$, where in two dimensions μ, ν range from 1 to 2. $g_{\mu\nu}$ gives the spatial separation between two points on the surface in terms of their coordinate differences Δx^μ:

$$\Delta r^2 = g_{\mu\nu}\Delta x^\mu \Delta x^\nu.$$

In Cartesian coordinates, with $x^1 = x$, $x^2 = y$, $\Delta r^2 = (\Delta x^1)^2 + (\Delta x^2)^2 = \Delta x^2 + \Delta y^2$, and the metric tensor is

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

In this case, the geometry is everywhere flat, as is obvious from the form of Δr^2 and the fact that the non-zero components of $g_{\mu\nu}$ are *constants*, independent of x^1 and x^2.

Suppose however that we use plane polar coordinates instead, with radius R and polar angle θ. In this case, $x^1 = R$ and $x^2 = \theta$, and

$$\begin{aligned} \Delta r^2 &= \Delta R^2 + R^2\Delta\theta^2 \\ &= (\Delta x^1)^2 + (x^1)^2(\Delta x^2)^2. \end{aligned} \tag{4.14}$$

In this case, the metric tensor corresponding to Eq. 4.14 is

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & (x^1)^2 \end{pmatrix}. \tag{4.15}$$

Thus the metric tensor is now position-dependent. However we are still dealing with the same flat surface.

If we were just given the metric tensor (Eq. 4.15), how could we tell if the surface was flat or not? We search for a coordinate transformation which puts $g_{\mu\nu}$ back into Cartesian form. In the case of polar coordinates, this is trivial; we define:

$$\begin{aligned} x^{1\prime} &= x^1 \cos x^2 \\ x^{2\prime} &= x^1 \sin x^2. \end{aligned}$$

This is just $x = R\cos\theta$, $y = R\sin\theta$, and $g_{\mu\nu}$ is back in the form $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. This is also easy for a cylinder of radius a. cylindrical coordinates are just R, θ, and z, where

$$\Delta r^2 = \Delta z^2 + a^2\Delta\theta^2, \tag{4.16}$$

since the surface of the cylinder is just defined by $R = a$. If we define,

$$\begin{aligned} x^1 &= z \\ x^2 &= a\,\theta, \end{aligned}$$

the distance formula becomes

$$\Delta r^2 = (\Delta x^1)^2 + (\Delta x^2)^2.$$

As before, a cylindrical surface is also flat. (A cylinder is just a plane which has been rolled up.) If we tried to do this for the surface of a sphere, we would not be able to find any such transformation.

The formal solution to this problem was obtained by Gauss, who showed how to obtain the circumference of a circle of radius r in arbitrary coordinates in terms of the components of $g_{\mu\nu}$ and their derivatives. If the components of $g_{\mu\nu}$ are constants, it is always possible to find a trivial change of coordinates which diagonalizes the metric tensor: if $g_{\mu\nu}$'s components are constants, *the surface must be flat.*

Curvature arises from variations in the components of $g_{\mu\nu}$, so it is not surprising that Gauss' curvature formula involves the derivatives of $g_{\mu\nu}$.

Any metric tensor in which the off-diagram components are zero (even if the diagonal components are not constants) is called *orthogonal*: there is no "mixing" of coordinates, i.e., the coordinate axes are orthogonal. For 2 dimensions and orthogonal metrics (which is always the case for $g_{\mu\nu}$), Gauss' curvature forumula is given by Eq. 4.3.5 of M. V. Berry's *Principles of Cosmology and Gravitation* and is derived in Appendix B of that same volume:

$$\begin{aligned} K = \frac{1}{2g_{11}g_{22}} \Bigg\{ &-\frac{\partial^2 g_{11}}{\partial (x^2)^2} - \frac{\partial^2 g_{22}}{\partial (x^1)^2} + \frac{1}{2g_{11}} \left[\frac{\partial g_{11}}{\partial x^1} \frac{\partial g_{22}}{\partial x^1} + \left(\frac{\partial g_{11}}{\partial x^2} \right)^2 \right] + \\ &\frac{1}{2g_{22}} \left[\frac{\partial g_{11}}{\partial x^2} \frac{\partial g_{22}}{\partial x^2} + \left(\frac{\partial g_{22}}{\partial x^1} \right)^2 \right] \Bigg\} \end{aligned}$$

Gauss also proved that K is invariant; it has the same value no matter what coordinate system we choose for evaluating it.

For more than two dimensions, things are more complicated. It is not possible to describe the curvature by a single function K. Physically this is because for a 2-D surface there is only a single plane which passes through a point; for 3 or more dimensions, there are an infinite number. In general, curvature is described by the curvature tensor R_{ijkl} of rank 4; i.e., there are n^4 components in an n-dimensional space. Not all are independent; in 2-D, only one is—this is K.

4.3 Curved Space in 3-D

Let's consider the simplest possible curved space in 3-dimensions: the 3-D isotropic space of constant curvature. (If the curvature is the same in every direction about a point, the point is an isotropic point. If *every* point in a space is an isotropic point, then the curvature must be the same everywhere and the space has constant curvature.)

We use the coordinates R, θ, and ϕ. The radial coordinate R defines a "hyper-spherical" surface whose area *by definition* is $4\pi R^2$. (For 2-D surfaces, the analog to R is the coordinate x we discussed earlier. The circumference is $C = 2\pi x$ in that case, as usual.)

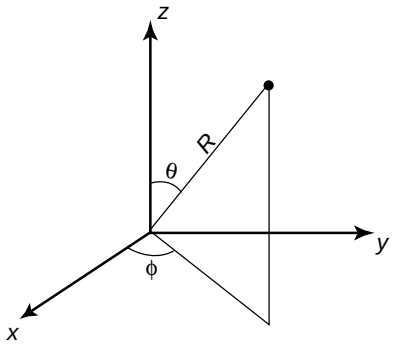


Figure 4.4: The spherical coordinate system.

The hyper-spherical surface R = constant forms a 2-D sub-space, on which positions are given by θ and ϕ. The metric of this 2-D sub-space is just the usual expression for the surface of a sphere:

$$\Delta r^2 = R^2\,\Delta\theta^2 + R^2\,\sin^2\theta\,\Delta\phi^2 \qquad (R = \text{constant}).$$

The 3-D space is then made up of a succession of these "R-spheres." *However* because of curvature, R is *not* the proper radius of each sphere (just as x was not the radius of the circle we considered in the 2-D case).

Therefore write the 3-D metric in this space as

$$\Delta r^2 = f(R)\,\Delta R^2 + R^2\,\Delta\theta^2 + R^2\,\sin^2\theta\,\Delta\phi^2, \tag{4.17}$$

where the function $f(R)$ allows for the fact that the proper distance between points (R, θ, ϕ) and $(R + \Delta R, \theta, \phi)$ is *not* ΔR. We want to find the function $f(R)$.

As always, we want to do this in the simplest possible way. Since the curvature is the same everywehere, all geodesics—which in this case are surfaces, not lines as in 2-D—have the same curvature. So we can pick any geodesic surface to find $f(R)$.

For simplicity, pick the "equatorial" surface $\theta = \pi/2$ (i.e., the x-y plane in the figure on on this page). In this case $\Delta\theta = 0$, and the metric on this surface is

$$\Delta r^2 = f(R)\,\Delta R^2 + R^2\,\Delta\phi^2. \tag{4.18}$$

The coordinates on this surface are thus $x^1 = R$, $x^2 = \phi$, and the metric tensor is

$$g_{\mu\nu} = \begin{pmatrix} f(R) & 0 \\ 0 & R^2 \end{pmatrix} = \begin{pmatrix} f(x^1) & 0 \\ 0 & (x^1)^2 \end{pmatrix}.$$

Using Eq. 4.3.5 of Berry for the curvature gives

$$K = \frac{df(x^1)/dx^1}{2f^2(x^1)\,x^1} = \frac{df(R)/dR}{2f^2(R)\,R}. \tag{4.19}$$

This is a simple differential equation for $f(R)$, since K is constant:

$$\begin{aligned} \frac{df(R)}{dR} &= 2Kf^2(R)R \\ \frac{df(R)}{f^2(R)} &= 2KR\,dR. \end{aligned}$$

Since the derivative of $-1/f(R)$ is just $df(R)/f^2(R)$, the left-hand side is just

$$\begin{aligned} d\left(-1/f(R)\right) &= 2KR\,dR \\ \Longrightarrow \frac{d}{dR}\left(-1/f(R)\right) &= 2KR. \end{aligned}$$

Integrating gives:

$$\begin{aligned} \frac{1}{f(R)} &= -KR^2 + C; \qquad C \text{ is a constant.} \\ f(R) &= \frac{1}{C - KR^2}. \end{aligned} \tag{4.20}$$

To determine the value of the constant C, we require that $f \to 1$ as $K \to 0$, i.e., in normal flat space R *is* the proper radial distance coordinate. This requires that $C = 1$, so

$$f(R) = \frac{1}{1 - KR^2}. \tag{4.21}$$

And the metric in 3-D in this space is Eq. 4.17

$$\Delta r^2 = \frac{\Delta R^2}{1 - KR^2} + R^2\ \Delta\theta^2 + R^2\ \sin^2\theta \Delta\phi^2. \tag{4.22}$$

Recall that, by definition, the area of an R-sphere is $4\pi r^2$, while from Eq. 4.22, we can now determine its *proper radius* $a(R)$:

$$a(R) = \int_0^{a(R)} dr = \int_0^R \frac{dR}{(1 - KR^2)^{1/2}} = \frac{1}{K^{1/2}} \arcsin\left(R\,K^{1/2}\right). \tag{4.23}$$

We finally arrive at:

$$R = \frac{1}{K^{1/2}} \sin\left(aK^{1/2}\right). \tag{4.24}$$

Since $A = 4\pi R^2$, we can now write the relation between area and *proper* radius for these "hyper-spheres:"

$$A = \frac{4\pi}{K} \sin^2\left(aK^{1/2}\right). \tag{4.25}$$

For small x, $\sin x \approx x$, and so for small spheres (where "small" means $a \ll K^{-1/2}$),

$$A \simeq 4\pi a^2,$$

i.e., the usual Euclidean form. With increasing a, A departs from the Euclidean value in a way which depends on the value and sign of K.

If $K < 0$, then the argument of sine squared in Eq. 4.25 is imaginary. For imaginary z, $\sin z$ is just $\sinh |z|$, so for $K < 0$, Eq. 4.25 is

$$A = \frac{4\pi}{|K|} \sinh^2 \left(a|K|^{1/2} \right), \tag{4.26}$$

where $|K| = -K$.

Since $\sinh x = \frac{e^x - e^{-x}}{2}$, for large x, $\sinh x \approx \frac{1}{2} e^x$. Thus for large a, area increases with a faster than in Euclidean space (where $A \propto r^2$), and becomes infinite as $a \to \infty$.

If $K > 0$, then Eq. 4.25 shows that A increases with a more slowly than in Euclidean space (for $x = 0$ to $\pi/2$, $x/\sin x > 1$), and reaches a maximum at:

$$a_{\max} = \frac{\pi}{2K^{1/2}}, \tag{4.27}$$

at which

$$A_{\max} = \frac{4\pi}{K}.$$

With increasing radius beyond $a_{\max}$, A *decreases*, and reaches zero at $a = \pi/K^{1/2}$. The behavior of A with increasing a is periodic.

$\therefore$ Space with positive K is *closed*; the periodic behavior of A corresponds to successive circumnavigations of the surface.

To understand what this means, recall our 2-D sphere of radius a. The sphere's surface has constant curvature $K = 1/a^2$. If we start from the pole, we find the circumference of circles *increases* with proper radius r until we reach the equator:

$$C = 2\pi a \sin \left(\frac{r}{a} \right).$$

This has a maximum value when $r/a = \pi/2$ or $r = \pi a/2$, which is when we have reached the equator.

With further increase in r, C decreases, and goes to zero at $r = \pi a$ (the opposite pole).

4.4 Geodesics in Spacetime

What path does a body subject to no non-gravitational forces follow, i.e., what is its world-line $x(\tau)$?

In a local inertial frame, this is easy—just use Special Relativity. The body moves uniformly in a straight line, so the equations of motion are

$$\frac{d^2 t}{d\tau^2} = 0 \qquad \frac{d^2 \boldsymbol{r}}{d\tau^2} = 0.$$

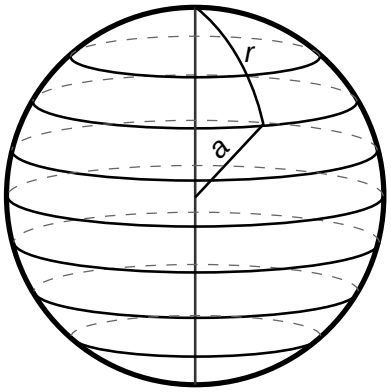


Figure 4.5: Circles with increasing circumferences on a sphere.

The first equation just says that the frame is not accelerating; the second says that the body in the inertial frame is unaccelerated. In general of course, we can't describe everything with inertial frames in presence of matter.

As we have already noted, in general, spacetime is curved by the presence of matter. The General Relativistic answer to this question (the generalization of the Special Relativistic inertial frame result) is:

Bodies subjected to no non-gravitational forces follow *time-like geodesics* in spacetime.
Light rays follow light-like (or *null*) geodesics. (The name comes from the fact that Δs^2 or $\Delta\tau^2 = 0$ for light rays.)

The tricky part is determining the spacetime metric tensor g_{ij} from the mass distribution; this requires solving the *field equations* of General Relativity. Remarkably, the first exact solution of Einstein's field equations was found by Karl Schwarzschild in 1916 in the trenches of World War I.

In a sense, gravity has disappeared as a force: the shape of spacetime is determined by the presence of matter, and a *free particle* is now redefined to mean a particle which is affected by no non-gravitational forces. Just as the fictitious forces associated with rotation can be locally transformed away by switching to an inertial frame, the "fictitious" force of gravity can be transformed away by switching to a freely falling frame. Recall that we can always construct local inertial frames even in the presence of gravitational fields; these are local freely-falling frames. In these local inertial frames, particles unaffected by non-gravitational forces will obey the law of inertia, i.e., will move at constant velocity in straight lines.

Locally, geodesics will look like straight lines in spacetime.

Globally however, the geodesics in curved spacetime will be curved lines (just as a geodesic on the surface of a sphere looks, locally, like a straight line in a plane).

Spatial geodesics are the shortest distances between any pair of points; equivalently, a geodesic is the *straightest possible path* between two points. The latter is still true of spacetime geodesics, but the former is not. This is connected to the fact that, while the Euclidean metric Δr^2 is *positive definite* ($\Delta r^2 > 0$ for Δx, Δy, $\Delta z \neq 0$), the metric of spacetime (in both Special and General Relativity) is *indefinite* (Δs^2 can be positive, negative, or zero).

In fact, it turns out that the spacetime separation along a time-like geodesic between two points (events) is *greatest*. This is easy to see in Special Relativity: take one event O to be the origin of a Cartesian coordinate system, and let P have the coordinates $(t_P, 0, 0, 0)$. (We can *always* set up our coordiantes like this in Special Relativity.) Obviously, OP is the straightest world line connecting O and P, and the spacetime separation is just $\tau_{OP} = t_{OP}$.

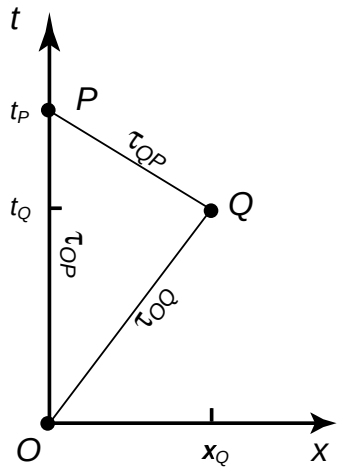


Figure 4.6: A spacetime diagram showing three events O, P, and Q.

Now consider some other possible path OQP, where we Q has the coordinates $(t_Q, x_Q, 0, 0)$. We must have $\tau_{OQP} = \tau_{OQ} + \tau_{QP}$. (This must be a time-like path.) Using the expression for $\Delta\tau^2$ from Special Relativity:

$$\Delta\tau^2 = \Delta t^2 - \Delta r^2/c^2 = \Delta t^2 - \Delta x^2/c^2. \tag{4.28}$$

In this case, we find that $\tau_{OQ}^2 = t_Q^2 - x_Q^2/c^2$, and similarly,

$$\tau_{QP}^2 = (t_P - t_Q)^2 - x_Q^2/c^2.$$

Now,

$$\tau_{OP} = t_Q + (t_P - t_Q),$$

and so

$$\begin{array}{rcccl} \tau_{OQP} & = & \left(t_Q^2 - x_Q^2/c^2\right)^{1/2} & + & \left[(t_P - t_Q)^2 - x_Q^2/c^2\right]^{1/2} < \tau_{OP}. \\ & & < t_Q & & < (t_P - t_Q) \end{array}$$

It turns out that this is also true in General Relativity.

Whether geodesics are the longest or shortest separations of two points (or events) depends on the metric; in either case the geodesics are *extremal* paths. For light rays, $\Delta s^2 = 0$ in Special Relativity; in General Relativity, the paths followed by light rays are *null geodesics.* Bodies subject to non-gravitational forces will follow time-like paths which are *not* time-like geodesics.

4.5 The Schwarzschild Solution

We will now consider the simplest possible General Relativity problem, namely, a single body of mass M and radius r, in otherwise empty spacetime. (The body is not rotating.) The exact solution, from Einstein's field equations, is

$$\Delta\tau^2 = \left(1 - \frac{r_s}{R}\right)\Delta t^2 - \frac{1}{c^2}\left[\frac{\Delta R^2}{\left(1 - \frac{r_s}{R}\right)} + R^2\,\Delta\theta^2 + R^2\,\sin^2\theta\,\Delta\phi^2\right], \tag{4.29}$$

where $r_s = 2GM/c^2$ is the *Schwarzschild radius.* We won't derive this, but let's try to understand it and motivate it.

Consider good old Newtonian gravitation for a moment. If we have a body of mass M and radius r, the potential at the surface is just

$$\phi = -\frac{GM}{r},$$

and the escape velocity is

$$v_{\rm esc} = \left(\frac{2GM}{r}\right)^{1/2}. \tag{4.30}$$

Keeping the mass fixed while decreasing the radius causes $v_{\rm esc}$ to increase, and it is equal to c when the radius is

$$r_s = \left.\frac{2GM}{c^2}\right|_{1{\rm M}_\odot} = 3 \times 10^5 \text{ cm},$$

i.e., the Schwarzschild radius. This calculation is actually a swindle, since the potential will not be Newtonian. However, the result equals the correct General Relativistic result, as we will see when we talk about black holes.

Clearly, if R is close to r_s, gravity (and hence curvature) will be very strong. In general astrophysical situations, this is not the case; e.g., the surface of the Sun is at $\mathrm{R}_\odot/r_s = 7 \times 10^{10}\,\mathrm{cm}/3 \times 10^5\,\mathrm{cm} = 2.3 \times 10^5$, so General Relativistic effects are small (i.e., the curvature is small), and so we might expect the correction to be $(1 - r_s/R)$.

There is another, more subtle way to understand this correction. The mass M here is at rest and non-rotating. This implies that the metric must be *static*, that is, the coefficients of all the metric coordinates must be independent of time. Consider two clocks sitting at distances r_1 and r_2, respectively, from a body of mass M, and suppose there is an observer O_2 at the position of Clock 2.

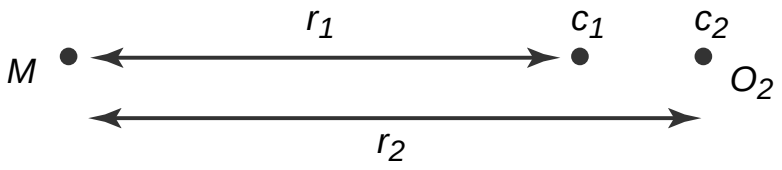


Figure 4.7: Mass M and two clocks, with an observer O_2 at clock c_2.

If O_2 "looks at" (i.e., receives a signal from) Clock 1, O_2 will not see it reading the same time as C_2. this is expected, since it takes a finite time for the signal from C_1 to reach C_2.

However, we have already seen that in a gravitational field there is a gravitational redshift, which is completely equivalent to gravitational time dilation: clocks run slow in a gravitational field, or (Eq. 3.39)

$$\frac{\Delta t(r)}{\Delta t(\infty)} = 1 - \frac{GM}{rc^2}. \tag{4.31}$$

Thus O_2 does not see C_1 running at the same rate as C_2, and after a given time interval Δt_2 on O_2's clock, C_1 has only elapsed a time

$$\Delta t_1 = \Delta t_2 \left[1 - \left(\frac{GM}{c^2 r_1} - \frac{GM}{c^2 r_2}\right)\right]. \tag{4.32}$$

For our purposes, this is a terrible property for a time coordinate, as O_2 will see C_1 lagging farther and farther behind C_2. This means that the time difference between C_1 and C_2 will depend on when O_2 looks (i.e., when the clocks were synchronized), which means the coefficient of Δt^2 in the metric (Eq. 4.29) would have to be a function of time.

To eliminate this problem, we speed up all the clocks on each R-sphere by the factor $(1 - GM/c^2R)^{-1}$. Then any observer will see all clocks running at the same rate as the observer's clock, which will be the same rate as the clocks at infinity.

Since we have sped the clocks up, this means that an indicated time difference Δt will correspond to a smaller proper time difference $\Delta\tau$:

$$\Delta\tau = \left(1 - \frac{GM}{c^2 R}\right) \Delta t = \left(1 - \frac{r_s}{2R}\right) \Delta t, \tag{4.33}$$

which leads to:

$$\Delta\tau^2 = \left(1 - \frac{2GM}{c^2 R}\right) \Delta t^2 = \left(1 - \frac{r_s}{R}\right) \Delta t^2, \tag{4.34}$$

where we have neglected 2nd-order terms in r_s/R. Again, although our derivation is approximate, this is the exact General Relativistic result.

4.6 Motion of Particles and Light in the Schwarzschild Metric

Let's briefly review motion in central force potentials.

Let the position of a particle be $\boldsymbol{r} = r\,\hat{\boldsymbol{r}}$, where $\hat{\boldsymbol{r}}$ is just the radial unit vector. By assumption, the force per unit mass is

$$\boldsymbol{F} = F(r)\,\hat{\boldsymbol{r}}, \tag{4.35}$$

which depends only on radius r. The equation of motion is then just

$$\frac{d^2\boldsymbol{r}}{dt^2} = F(r)\,\hat{\boldsymbol{r}}. \tag{4.36}$$

Recall that the cross-product of any vector with itself is zero. We can then write:

$$\begin{aligned} \frac{d}{dt}\left(\boldsymbol{r}\times\frac{d\boldsymbol{r}}{dt}\right) &= \frac{d\boldsymbol{r}}{dt}\times\frac{d\boldsymbol{r}}{dt} + \boldsymbol{r}\times\frac{d^2\boldsymbol{r}}{dt^2} \\ &= F(r)\,(\boldsymbol{r}\times\hat{\boldsymbol{r}}) = 0. \end{aligned} \tag{4.37}$$

Letting dots denote derivatives with respect to time, Eq. 4.37 says that the vector $\boldsymbol{r}\times\dot{\boldsymbol{r}}$ is some constant vector:

$$\boldsymbol{r}\times\frac{d\boldsymbol{r}}{dt} = \boldsymbol{L}, \tag{4.38}$$

where in fact $\boldsymbol{L}$ is simply the angular momentum per unit mass. Since $\boldsymbol{L}$, which is perpendicular to the instantaneous position and velocity, is constant, this says that the motion is confined to a plane.

If we use plane polar coordinates, where $r = 0$ is the attracting source and θ is the angle in the orbital plane, we can write the equation of motion as

$$\ddot{r} - r\,\dot{\theta}^2 = F(r) \tag{4.39}$$

$$2\dot{r}\,\dot{\theta} + r\,\ddot{\theta} = 0 \qquad \text{(Eq. 4.37 again)}. \tag{4.40}$$

Eq. 4.40 can be trivially integrated by multiplying by r to get:

$$r^2\,\dot{\theta} = \text{constant} = L, \tag{4.41}$$

which is just conservation of angular momentum again.

Now, Eq. 4.41 implies that

$$\frac{d}{dt} = \frac{L}{r^2}\,\frac{d}{d\theta}. \tag{4.42}$$

So using this to switch from d/dt to $d/d\theta$ in Eq. 4.39, we get

$$\frac{L^2}{r^2}\,\frac{d}{d\theta}\left(\frac{1}{r^2}\,\frac{dr}{d\theta}\right) - \frac{L^2}{r^3} = F(r). \tag{4.43}$$

This equation can be greatly simplified by introducing the new variable $u \equiv 1/r$,

$$\frac{d^2u}{d\theta^2} + u = -\frac{F(1/u)}{L^2\,u^2}. \tag{4.44}$$

To get a better idea of what this means, assume that $F(r)$ is derivable from a potential $\Phi(r)$:

$$F(r) = -\frac{d\Phi}{dr} = u^2 \frac{d\Phi}{du}. \tag{4.45}$$

If we multiply Eq. 4.44 by $du/d\theta$, we can integrate it once to obtain

$$\left(\frac{du}{d\theta}\right)^2 + \frac{2\Phi}{L^2} + u^2 = \text{constant} \equiv \frac{2E}{L^2}. \tag{4.46}$$

The reason for picking this form for the constant is apparent when we multiply through by $L^2/2$ and, noting from Eq. 4.41 that $L = r^2 d\theta/dt$, find

$$\begin{aligned} \left(\frac{r^2}{L}\right)^2 \left(\frac{d(1/r)}{dt}\right)^2 + \frac{2\Phi}{L^2} + \frac{1}{r^2} &= \frac{2E}{L^2} \\ E &= \frac{r^4}{2}\left[-\frac{1}{r^2}\frac{dr}{dt}\right]^2 + \Phi + \frac{L^2}{2r^2} \\ &= \frac{1}{2}\left(\frac{dr}{dt}\right)^2 + \frac{1}{2}\left(r\frac{d\theta}{dt}\right)^2 + \Phi. \end{aligned} \tag{4.47}$$

The first term on the right-hand side is just the radial kinetic energy per unit mass; the second term is the tangential kinetic energy per unit mass. So clearly, the constant E is just the energy per unit mass.

The solutions to Eq. 4.44 are of two types: *unbound* orbits, in which $r \to \infty$ ($u \to 0$), and *bound* orbits, for which r oscillates back and forth between finite limits.

For bound orbits, $du/d\theta = 0$ at the turning points, so from Eq. 4.46, we get

$$u^2 + \frac{2}{L^2}\left[\Phi\left(\frac{1}{u}\right) - E\right] = 0, \tag{4.48}$$

which has two roots, the *pericenter* (closest approach) and the *apocenter* (greatest distance).

In general, orbits in spherically symmetric potentials are not *closed*, that is, the radial period T_r (the time to go from pericenter to apocenter and back) is not equal to the azimuthal period T_θ (the time for the body to travel 2π radians). One important exception to this is the Keplerian potential $\Phi = -GM/r$, in which case, Eq. 4.44 becomes (using $F(r) = -GM/r^2 = -GMu^2$):

$$\frac{d^2u}{d\theta^2} + u = \frac{GM}{L^2}. \tag{4.49}$$

The solution of this equation is just:

$$u = \frac{GM}{L^2}\left[1 + e\cos(\theta - \theta_0)\right], \tag{4.50}$$

where θ_0 is a constant of integration and the eccentricity is:

$$e^2 \equiv 1 + \frac{EL^4}{G^2M^2}. \tag{4.51}$$

For convenience, set $\theta_0 = 0$:

$$u = \frac{GM}{L^2}\left[1 + e\cos\theta\right]. \tag{4.52}$$

We expect in General Relativity that the orbit will *not* be closed. Why? Because of $E = mc^2$: the gravitational energy associated with the central mass is itself a contributor to the mass of the system; since the gravitational energy density is non-zero, the effective mass seen by an orbiting body will vary as a function of r, and therefore $F \neq \text{constant}/r^2$.

The analog of Eq. 4.49, derived from the Schwarzschild metric, is

$$\frac{d^2u}{d\theta^2} + u = \frac{GM}{L^2} + 3\frac{GM}{c^2}u^2. \tag{4.53}$$

Clearly the effect of General Relativity is in the form of the correction term $3GMu^2/c^2$; also unsurprisingly this correction is:

- Proportional to r_s/r.
- A function of radius.

For the solar system, as we have already noted, r_s/r is very small, and so the correction term is very small. We know that in the absence of the General Relativistic correction term, the solution is given by Eq. 4.52:

$$u = \frac{GM}{L^2}(1 + e\cos\theta).$$

Treat the General Relativistic term as a small perturbation. Substituting Eq. 4.52 into the right-hand side of Eq. 4.53 gives

$$\begin{aligned}\frac{d^2u}{d\theta^2} + u &= \frac{GM}{L^2} + \frac{3\,GM}{c^2}\cdot\left(\frac{GM}{L^2}\right)^2(1 + 2e\cos\theta + e^2\cos^2\theta) \\ &= \frac{GM}{L^2} + \frac{3(GM)^3}{c^2L^4}(1 + 2e\cos\theta + e^2\cos^2\theta).\end{aligned} \tag{4.54}$$

The right-hand side does not involve u, so this is a linear equation. The general solution is given by solving the left-hand side for the individual terms on the right-hand side. Let $A = 3(GM)^3/c^2L^4$. We already know the answer for the right-hand side should be GM/L^2, so we just need to solve the equations:

$$\frac{d^2u}{d\theta^2} + u = \begin{cases} A & \text{(a)} \\ 2A\cos\theta & \text{(b)} \\ Ae^2\cos^2\theta & \text{(c)} \end{cases}. \tag{4.55}$$

Eq. 4.55a has the solution $u_a = A$, Eq. 4.55b has the solution $u_b = Ae\,\theta\sin\theta$, and Eq. 4.55c has the solution $u_c = Ae^2\left(\frac{1}{2} - \frac{1}{6}\cos 2\theta\right)$.

The solution to Eq. 4.54 is thus given by Eq. 4.52 plus u_a, u_b, and u_c. The solutions u_a and u_c are uninteresting: they are just tiny constants and tiny oscillations. However

$u_b \propto \theta$ so that its contribution steadily builds up. Including this solution, the approximate solution to Eq. 4.53 is then

$$u \approx \frac{GM}{L^2}\left(1 + e\cos\theta + \frac{3(GM)^2}{c^2L^2}e\theta\sin\theta\right). \tag{4.56}$$

Define a new angle $\theta' = 3(GM/cL)^2\theta$; here θ' is very small. For small angles, $\cos\theta \approx 1$, $\sin\theta \approx \theta$, and so Eq. 4.56 is approximately

$$u \approx \frac{GM}{L^2}\left(1 + e\cos\theta\cos\theta' + e\sin\theta\sin\theta'\right). \tag{4.57}$$

The difference-angle formula for cosine from simple trigonometry is:

$$cos(\theta - \theta') = \cos\theta\cos\theta' + \sin\theta\sin\theta', \tag{4.58}$$

and so Eq. 4.57 can be written as:

$$u \approx \frac{GM}{L^2}\left[1 + e\cos\left(1 - \frac{3(GM)^2}{c^2L^2}\right)\theta\right]. \tag{4.59}$$

Eq. 4.59 shows that u (and therefore r) is a periodic function of θ with period

$$2\pi\left(1 - \frac{3\,G^2M^2}{c^2L^2}\right)^{-1} > 2\pi. \tag{4.60}$$

Thus if we take $\theta = 0$ to be, say, the pericenter, r does not return to the pericenter value until the body has travelled *more* than 2π in azimuth.

Effectively the location of perihelion *precesses*, by an amount

$$\begin{aligned}\Delta\theta = 2\pi\left(1 - \frac{3G^2M^2}{c^2L^2}\right)^{-1} - 2\pi &\simeq 2\pi\left(1 + \frac{3\,G^2M^2}{c^2L^2}\right) - 2\pi \\ &= \frac{6\pi G^2M^2}{c^2L^2}\end{aligned} \tag{4.61}$$

per orbit.

For a Keplerian potential, L is related to the semi-major axis by

$$a = \frac{L^2}{GM(1-e^2)}, \tag{4.62}$$

so we can write Eq. 4.61 as

$$\Delta\theta = \frac{6\pi GM/c^2}{a(1-e^2)} = \frac{3\pi r_s}{a(1-e^2)}. \tag{4.63}$$

Table 4.1: Anomalous perihelion advance per century

Planet	Predicted	Observed
Mercury	$43.03''$	$43.11 \pm 0.45''$
Venus	$8.6''$	$8.4 \pm 4.8''$
Earth	$3.8''$	$5.0 \pm 1.2''$

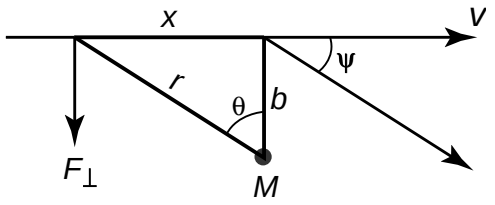


Figure 4.8: Deflection of light by a mass M.

4.7 Deflection of Light by Gravitating Bodies

In the Newtonian analysis, let us consider a light ray traveling past a body of mass M. Assume that the deflection will be very small (as we will see to be the case). In the absence of gravity, the smallest separation between the photon and the body (the "impact parameter") would be b. However because of the effect of gravity, the trajectory of the photon will be altered. To estimate the magnitude of this deflection, assume that the trajectory remains straight. What is the perpendicular force acting on the photon during the encounter?

Suppose we have a material particle instead of a photon, traveling with velocity $\boldsymbol{v}$. Take $t = 0$ to be the moment of closest approach. Clearly,

$$\begin{aligned} F_\perp &= \frac{GM}{b^2 + x^2} \cos\theta \\ &= \frac{GMb}{(b^2 + x^2)^{3/2}} \\ &= \frac{GM}{b^2} \left[1 + \left(\frac{vt}{b} \right)^2 \right]^{-3/2}, \end{aligned} \tag{4.64}$$

since $x = vt$. And because $m\dot{\boldsymbol{v}}_\perp = \boldsymbol{p}_\perp$, integration with respect to time gives:

$$|\delta \boldsymbol{v}_\perp| = \frac{GM}{bv} \int_{-\infty}^{\infty} (1 + s^2)^{-3/2}\, ds = \frac{2GM}{bv}, \tag{4.65}$$

since

$$\int (1 + s^2)^{-3/2}\, ds = \left. \frac{s}{(1 + s^2)^{1/2}} \right|_{-\infty}^{\infty} = 1 - (-1) = 2.$$

Thus $|\delta \boldsymbol{v}_\perp|$ is equal to the force at closest approach, $F_\perp = GM/b^2$, times an effective encounter time, $2b/v$. Now, the angle of deflection is given by $\psi \approx \Delta v/v = |\delta v_\perp|/v$, so

$$\psi \approx \frac{2\,GM}{bv^2}. \tag{4.66}$$

Now, consider the photon again. The situation is exactly the same, with v fixed at c. (Note that this also means that $|\delta \boldsymbol{v}_\perp|$ is entirely a deflection.) Hence in this case:

$$\psi_N = \frac{2\,GM}{bc^2}. \tag{4.67}$$

What happens in General Relativity? The analogue of Eq. 4.41 is

$$r^2 \frac{d\theta}{d\tau} = \text{constant}. \tag{4.68}$$

However, for photons (or other massless particles) which travel along null geodesics, $d\tau =$

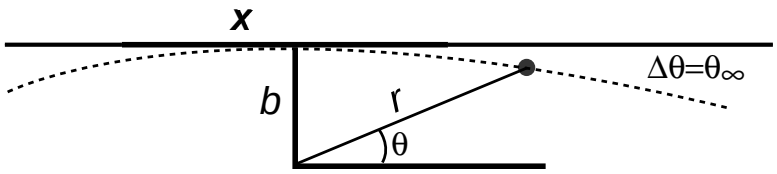


Figure 4.9: Deflection of light in General Relativity.

$ds = 0$. This implies that the right-hand side of Eq. 4.68 is infinite. Thus the analogue of Eq. 4.53, for particles moving on null geodesics, is

$$\frac{d^2u}{d\theta^2} + u = 3\frac{GM}{c^2}u^2. \tag{4.69}$$

Again, the term on the right-hand side is very small. If the right-hand side were zero, the solution to this equation would just be

$$u_0 = c\,\sin\theta, \tag{4.70}$$

where c is some constant. To determine c, note that when $\theta = \pi/2$, $1/u = r$ just equals b, the impact parameter. So

$$u_0 = \frac{\sin\theta}{b}. \tag{4.71}$$

If we substitute this into the right-hand side of Eq. 4.69 (just as we did earlier in deriving the precession of perihelion), we get

$$\begin{aligned} \frac{d^2u}{d\theta^2} + u &= \frac{3\,GM}{c^2b^2}\,\sin^2\theta \\ &= \frac{3\,GM}{c^2b^2}\,(1-\cos^2\theta). \end{aligned} \tag{4.72}$$

A particular solution of this equation is

$$u_1 = \frac{3\,GM}{2c^2b^2}\left(1 + \frac{1}{3}\cos 2\theta\right). \tag{4.73}$$

And so

$$u \simeq u_0 + u_1 = \frac{\sin\theta}{b} + \frac{3\,GM}{2c^2b^2}\left(1 + \frac{1}{3}\cos 2\theta\right). \tag{4.74}$$

We again expect that the deflection angle will be small. In the limit of large r, θ is a very small angle, so $\sin\theta \approx \theta$ and $\cos 2\theta \approx 1$, so as $r \to \infty$ ($u \to 0$), we get

$$\Delta\theta = \theta_\infty = -\frac{2GM}{c^2 b}. \tag{4.75}$$

By symmetry, the total deflection is twice this ($r \to -\infty$ gives the same result), so

$$|2\Delta\theta| = \psi_{\rm GR} = \frac{4GM}{bc^2} = 2\psi_N = 1.75'' \quad \text{for } M = 1\ \mathrm{M}_\odot, \quad b = \mathrm{R}_\odot. \tag{4.76}$$

4.8 Effective Potentials

Orbits are often easiest to understand if we use the effective potential, rather than simply the gravitational potential. What is the effective potential? Consider the Newtonian case again. The equation of motion is:

$$\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2 = -\frac{GM}{r^2}. \tag{4.77}$$

We also know from conservation of angular momentum that

$$\begin{aligned} r^2\,\frac{d\theta}{dt} &= L = \text{constant} \\ \Longrightarrow r\left(\frac{d\theta}{dt}\right)^2 &= \frac{L^2}{r^3}. \end{aligned}$$

And we can write the equation of motion as

$$\begin{aligned} \frac{d^2r}{dt^2} &= -\frac{GM}{r^2} + \frac{L^2}{r^3} \\ &= -\frac{d}{dr}\left(-\frac{GM}{r} + \frac{L^2}{2r^2}\right) \\ &= -\frac{d}{dr}\left(\Phi_g + \frac{L^2}{2r^2}\right) \\ &= -\frac{d\Phi_{\rm eff}}{dr}, \end{aligned} \tag{4.78}$$

where the *effective potential* $\Phi_{\rm eff} = \Phi_g + L^2/2r^2 = -GM/r + L^2/2r^2$.

The angular momentum term in $\Phi_{\rm eff}$ represents what is frequently termed the "centrifugal barrier." Conservation of angular momentum limits how closely a particle can approach a center of attraction.

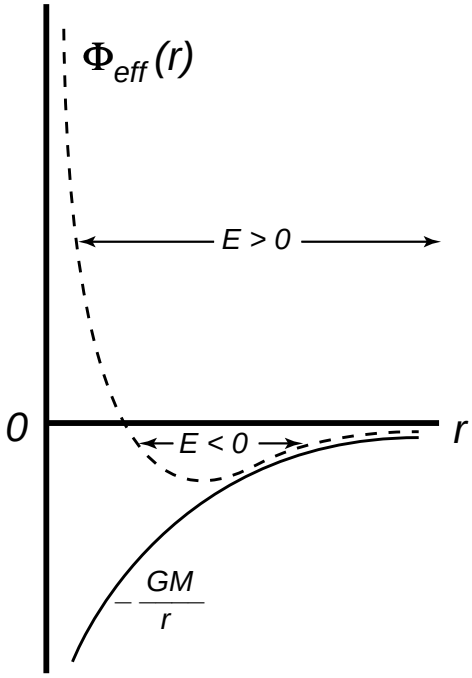


Figure 4.10: Bound and unbound particles in an effective potential $\Phi_{\rm eff}$ (dashed line) that is the sum of the Newtonian potential (solid line) and a "centrifugal" term (not shown).

Particles with $E < 0$ are *bound*, and move back and forth between pericenter and apocenter in an effective potential.

Particles with $E > 0$ are *unbound*; they come in from infinity along parabolic or hyperbolic orbits, and are reflected off the centrifugal barrier at $E = \Phi_{\rm eff}(r)$.

What happens in General Relativity? Noting that

$$\frac{d}{d\theta} = \frac{r^2}{L}\frac{d}{d\tau}$$

where the derivative is with respect to proper time, Eq. 4.53 can be rewritten as

$$\frac{d^2r}{d\tau^2} = -\frac{GM}{r^2} - \frac{3GML^2}{c^2r^4} + \frac{L^2}{r^3}, \tag{4.79}$$

and so the effective potential in the Schwarzschild metric is

$$\begin{aligned} \Phi_{\rm eff} &= -\frac{GM}{r} + \frac{L^2}{2r^2} - \frac{GML^2}{c^2r^3} \\ &= -\frac{GM}{r} + \frac{L^2}{2r^2}\left(1 - \frac{r_s}{r}\right). \end{aligned} \tag{4.80}$$

Aside: What is the meaning of the minimum in $\Phi_{\rm eff}$ in the Newtonian case? The minimum occurs where

$$\frac{d\Phi_{\rm eff}}{dr} = 0 = \frac{d\Phi_g}{dr} - \frac{L^2}{r^3}. \tag{4.81}$$

This is satisfied at the radius r_g where

$$\left(\frac{d\Phi_g}{dr}\right)_{r_g} = \frac{L^2}{r_g^3} = r_g\dot{\theta}^2. \tag{4.82}$$

This is simply a circular orbit with angular speed $\dot{\theta}$. Thus the minimum occurs at the radius at which a circular orbit has angular momentum L, and the value of $\Phi_{\rm eff}$ at this minimum is just the energy of this circular orbit.

Obviously in General Relativity, things are more complicated. In the Newtonian case, the $1/r^2$ term increases faster than the $1/r$ term, so for *any* non-zero value of L, there is a centrifugal barrier. Only $L = 0$ particles can reach the central body (for a point source).

In General Relativity, the additional term in r_s/r (of opposite sign) means that there is in general a maximum as well as a minimum in the potential. The behavior of orbiting bodies is quantitatively different in two respects:

1. Particles with non-zero angular momentum fall in.
2. Particles with large enough energy can always reach the center, no matter what their angular momentum L is.

4.9 Effective Potentials in the Schwarzschild Metric

Recall Eq. 4.80:

$$\Phi_{\rm eff} = -\frac{GM}{r} + \frac{L^2}{2r^2}\left(1 - \frac{r_s}{r}\right).$$

The presence of the term $-L^2 r_s/2r^3$ makes a crucial difference from the Newtonian case. The $L^2/2r^2$ term represents the "centrifugal barrier;" for any non-zero value of the angular momentum L, this will prevent a particle from reaching the origin. This is now multipled by $(1 - r_s/r)$, which effectively decreases the term; it goes to zero at $r = r_s$.

Most normal astrophysical objects have physical radii $r_B \gg r_s$, so the correction term in Eq. 4.80 is not particularly important. It is somewhat more important for neutron stars, which have $M \sim 1$ M$_\odot$, $R \sim 10$ km, so $r_B/r_s \sim 3$.

The correction term in Eq. 4.80 really becomes important for objects whose mass lies entirely within their Schwarzschild radii: as we will see shortly, these are *black holes*, and anything which crosses the Schwarzschild radius can never escape.

For a black hole then, what happens to particles?

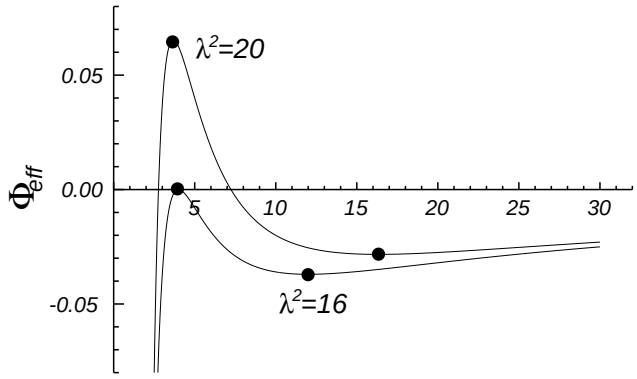


Figure 4.11: Effective potentials $\Phi_{\rm eff}$ for black holes with $\lambda^2 = 16$ and $\lambda^2 = 20$. The extrema for $\lambda^2 = 16$ are at $\rho = 4, 12$; the extrema for $\lambda^2 = 20$ are at $\rho = 3.67544, 16.3246$.

To see what the effective potentials look like, it is convenient to define new variables. Let:

$$\rho = \frac{r}{GM/c^2} = \frac{2r}{r_s} \tag{4.83}$$

$$\lambda = \frac{L}{GM/c} = \frac{2cL}{r_s}. \tag{4.84}$$

Then the effective potential is

$$\Phi_{\rm eff} = -\frac{c^2}{\rho} + \frac{c^2\lambda^2}{2\rho^2}\left(1 - \frac{2}{\rho}\right). \tag{4.85}$$

Note that this implies the depth of the potential at $\rho = 2$ ($r = r_s$) is just $\frac{1}{2}c^2$, i.e., 1/2 of the rest mass energy per unit mass. So define the new potential

$$\phi_{\rm eff} = \frac{\Phi_{\rm eff}}{c^2} = -\frac{1}{\rho} + \frac{\lambda^2}{2\rho^2}\left(1 - \frac{2}{\rho}\right). \tag{4.86}$$

The potential $\phi_{\rm eff}$ has maxima or minima at

$$\rho = \frac{\lambda^2}{2}\left[1 \pm (1 - 12/\lambda^2)^{1/2}\right]. \tag{4.87}$$

The behavior of particles depends on their energy and their angular momentum. Of particular interest are the cases $\lambda^2 = 12$ and $\lambda^2 = 16$. As you will see in Problem Set #5, bodies with $\lambda^2 < 12$ see effective potentials with no maxima or minima, and must reach $\rho = 2$, where they are lost.

At $\lambda^2 = 16$, the maxima and minima of the potential are at 4 and 12, respectively, and the effective potential has the values

$$\begin{aligned} \phi(\rho_{\rm min}) &= -3.704 \times 10^{-2} \\ \phi(\rho_{\rm max}) &= 0. \end{aligned}$$

This is the dividing line between bound ($E < 0$) and unbound ($E > 0$) particles. There are thus three different regimes of behavior:

1. $\boldsymbol{\lambda^2 < 12}$: There is no pericenter; all bodies reach $\rho = 2$ and will fall into the hole (more below).
2. $\boldsymbol{12 \leq \lambda^2 < 16}$: There are bound orbits which oscillate between pericenter and apocenter. Any particle which comes in from $\rho = \infty$ ($r = \infty$, $E > 0$) will necessarily be pulled into $\rho = 2$ ($r = r_s$) and fall into the hole.
3. $\boldsymbol{\lambda^2 > 16}$: There are "unbound" orbits with $E > 0$, as in the Newtonian case; however, any particle with $E > \phi_{\rm eff}(\rho_{\rm max})$, where

$$\phi_{\rm eff}(\rho_{\rm max}) = -\frac{2}{\lambda^2 \mu}\left[1 - \frac{1}{\mu}\left(1 - \frac{4}{\lambda^2 \mu}\right)\right], \tag{4.88}$$

and

$$\mu \equiv 1 - (1 - 12/\lambda^2)^{1/2}, \tag{4.89}$$

is necessarily pulled into the black hole; these particles have enough energy to overcome the centrifugal barrier.

Chapter 5

Black Holes

5.1 Gravitational Collapse and Black Holes

Stars like the Sun will eventually finish as white dwarf stars, supported by degenerate electron pressure. The maximum possible mass of a white dwarf, known as the *Chandrasekhar limit*, is $M_{\rm WD} \approx 1.5\ {\rm M}_\odot$. More massive stars can produce neutron star remnants (formed in supernova explosions) in which the pressures are so high that the protons and electrons are forced to form neutrons. Pulsars are undoubtedly neutron stars. The maximum possible mass of a neutron star is somewhat more uncertain (due to uncertainties inthe equation of state for nuclear matter at the relevant pressures), but it is certainly no more than $M_{\rm NS} \approx 3$–$5\ {\rm M}_\odot$.

For more massive stars, there appears to be no way to avoid unstoppable gravitational collapse once their nuclear fuel is exhausted. The Schwarzschild metric is the *vacuum* solution to the field equations, that is, it describes spacetime in the empty space outside the surface of a body of mass M and radius r_B. If the radius of a body $r_B < r_s$, then the Schwarzschild metric still describes the spacetime outside the surface. As we will see later, once an object collapses within r_s, it must inevitably collapse to a spacetime singularity—a black hole.

Since all the mass is then ar $r = 0$, the Scharzschild metric describes spacetime everywhere except $r = 0$. However, as we have already seen, Schwarzschild coordinates (i.e., the Schwarzschld metric) are singular at $r = r_s$; the coefficient of $dt^2 \to 0$, while the coefficient of $dr^2 \to \infty$.

For a long time, this was misunderstood, and it was thought that $r = r_s$ was a singular point. This is *not* true, and it is in fact just a coordinate singularity, as opposed to the genuine spacetime singularity of a black hole. For a simple example of a coordinate singularity, consider a 2-D flat (Euclidean) plane. In this case, the spatial separation (the line element) between two points is just

$$dr^2 = dx^2 + dy^2.$$

Now introducing a new coordinate

$$w = \frac{1}{3}x^3.$$

By Phil Maloney.

This is perfectly acceptable, as this gives a one-to-one mapping between w and x. In terms of w, the metric (the line element) is

$$dr^2 = (3w)^{-4/3}\, dw^2 + dy^2$$

This obviously has a coordinate singularity at $w = 0$, but it is not a physical singularity; we can get rid of it simply by transforming back to the original (x, y) coordinates.

The Schwarzschild metric holds for $0 < r < r_s$, $r_s < r \leq \infty$, just not right on the surface $r = r_s$. However, strange things happen when we cross $r = r_s$, and the meaning of the coordinates change. The Schwarzschild metric is again:

$$d\tau^2 = \left(1 - \frac{r_s}{r}\right) dt^2 - \frac{1}{c^2}\left[\frac{dr^2}{\left(1 - \frac{r_s}{r}\right)} + r^2\, d\theta^2 + r^2\, \sin^2\theta\, d\phi^2\right]. \tag{5.1}$$

For $r > r_s$, the coefficient of dt^2 is positive, and the coefficient of dr^2 is negative. For $r < r_s$, these signs flip: dt^2 has a negative coefficient, and dr^2 has a positive coefficient.

What does this mean?

Recall that, for a material particle, $d\tau^2 > 0$, while for a photon, $d\tau^2 = 0$. From Eq. 5.1, for $r_s/r > 1$, no particle or photon can have $r =$ constant. (Note the coefficients of $d\theta^2$ and $d\phi^2$ are unaltered.) For $r =$ constant, $d\tau^2$ would be negative (for any value of $r < r_s$). However, since this is the square of the proper time interval, this cannot be.

There can be no equilibrium inside $r = r_s$.

In a sense, r is now a "time" coordinate, as it is not stationary for any particle. Furthermore, since r *cannot* be constant, the coefficients in Eq. 5.1 are now functions of time: the Schwarzschild metric is no longer time-independent inside r_s. This is *not* due to a bad choice of coordinates, but due to the intrinsically non-stationary nature of spacetime inside r_s.

However, the coordinate singularity in the Schwarzschild metric, and the change in behavior of the coordinates across $r = r_s$, means that Schwarzschild coordinates are not very convenient for discussing spacetime in the vicinity of black holes. A variety of alternative coordinate systems have been constructed. A particularly useful set are the *Eddington-Finkelstein* coordinates.

The idea behind Eddington-Finkelstein coordinates is really very simple: change to a new time coordinate in which photons falling in purely radially (following *ingoing radial null geodesics*) travel in straight lines. This is obtained by changing the time variable from coordinate time t to the new time variable

$$\bar{t} = t + r_s \ln(r - r_s). \tag{5.2}$$

Differentiating this gives:

$$d\bar{t} = dt + \frac{r_s}{r - r_s} dr. \tag{5.3}$$

And substitution into the Schwarzschild metric (5.1) gives the new form

$$d\tau^2 = \left(1 - \frac{r_s}{r}\right) d\bar{t}^2 - 2\frac{r_s}{cr} d\bar{t}\, dr - \frac{1}{c^2}\left[\left(1 + \frac{r_s}{r}\right) dr^2 + r^2\, d\theta^2 + r^2\, \sin^2\theta\, d\phi^2\right]. \tag{5.4}$$

This metric is *regular* (no singularities) over the whole range $0 < r < \infty$.

What happens as we approach a black hole?

In Minkowski spacetime, the geometry is flat (Euclidean): light cones always make 45° angles with the ct axis. Suppose we suspend an object at some height above the surface in a gravitational field, and then release it. The trajectory is curved, as the object accelerates downward in the gravitational field until it hits the surface.

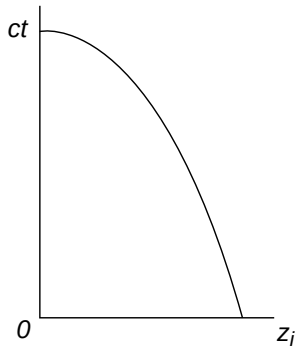


Figure 5.1: The trajectory of an object falling in a gravitational field.

Suppose the object is emitting photons at various points along its trajectory as it falls. These photons are also acted on by gravity, with the result that the light cones will no longer make 45° angles with the ct axis. Because the gravitational potential is deeper at the surface than at the initial height, the paths of photons emitted at $z = 0$ will be slightly different than photons emitted at z_i.

In curved space, light cones will change their shape and orientation.

Close to the Schwarzschild radius, where the curvature of spacetime becomes severe, this effect becomes severe.

It is easier to understand the behavior of photons by considering an "equatorial" spatial slice at some time. Consider the expanding light sphers around some arbitrary points in spacetime. In Minkowski space, the expanding spherical wavefronts are centered on their points of origin. In the Schwarzschild metric, this is not true: at large distance from the black hole, the picture is very similar to the Minkowski picture—the Schwarzschild metric is asymptotically flat.

As we move closer to the center, however, the photons—and therefore the wavefronts—are attracted towards the black hole, so that their points of origin are no longer at the origin. This effect becomes more pronounced as r decreases, until we reach $r = r_s$. At this point, only radially outgoing photons "stay put;" all the rest are dragged inward.

Within $r = r_s$, *all* photons are dragged towards the singularity. This is the *event horizon*. Since no material particle can have $v \geq c$, it follows that any particle which crosses the Schwarzschild radius must continue in to $r = 0$.

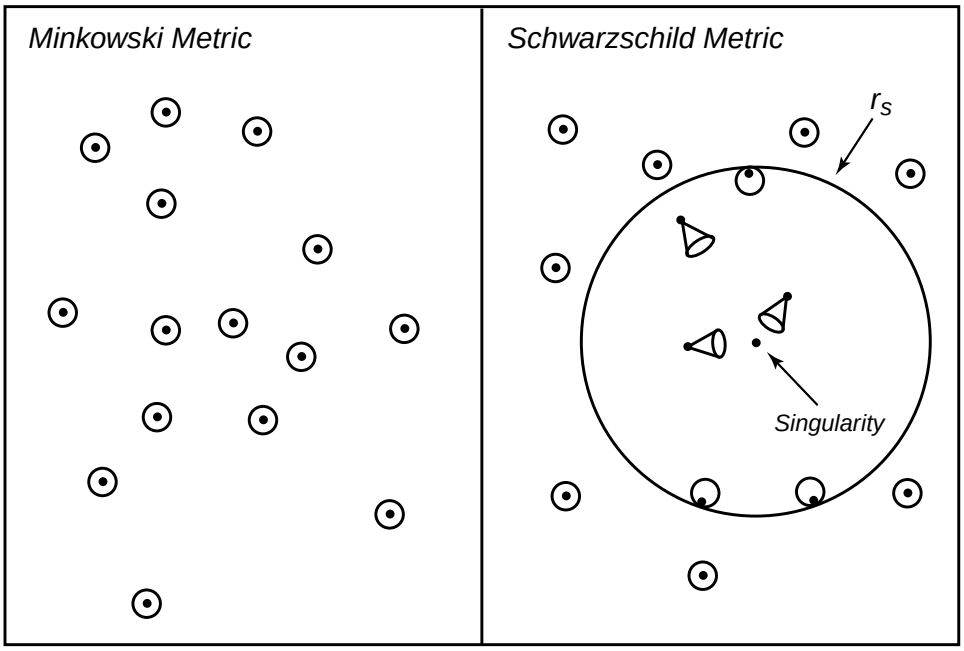


Figure 5.2: Light wavefronts in the Minkowski and Schwarzschild metrics.

Birkhoff's Theorem: A spherically symmetric vacuum solution is necessarily static.
Translation: If a spherically symmetric source is confined to some radius $r \leq a$, then spacetime outside the source ($r > a$) is given by the Schwarzschild solution. It doesn't matter if the body is collapsing, or pulsating, as long as it does it in a spherically symmetric fashion.

Combined with the fact that even light cones point inwards inside r_s, this implies that any object which collapses within its own Schwarzschild radius must collapse into a singularity.

What will such a collapse look like in a spacetime diagram?

What will the external observer see of the collapse?

Let the observer be at radius r_0. If a light signal is emitted at event r_e, t_e from the surface, and travels radially outward to reach the observer at event r_0, t_0, then (with $d\theta$, $d\phi = 0$), r and t are related along this outgoing radial null geodesic by:

$$0 = dt^2 \left(1 - \frac{r_s}{r}\right) - \frac{1}{c^2} \frac{dr^2}{1 - r_s/r}. \tag{5.5}$$

And so therefore

$$dt = \frac{1}{c} \frac{dr}{1 - r_s/r}.$$

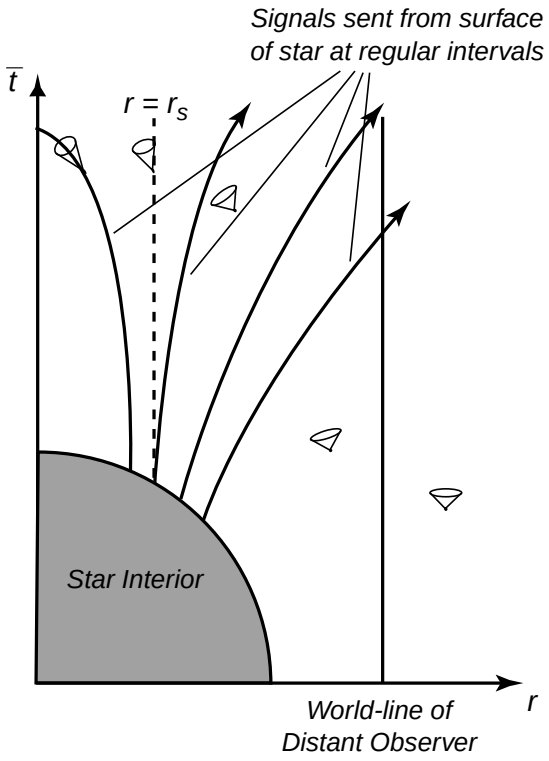


Figure 5.3: Spacetime diagram of a collapsing star.

This results in:

$$t_0 - t_e = \frac{1}{c}\int_{r_e}^{r_0} \frac{dr}{1 - r_s/r} = \frac{r_0 - r_e}{c} + \frac{r_s}{c}\ln\left(\frac{r_0 - r_s}{r_e - r_s}\right). \tag{5.6}$$

This light travel time becomes infinite when $r_e = r_s$, so to an external observer the collapse "freezes" at $r = r_s$.

The external observer does not perpetually see the surface as it was when it reached $r = r_s$, however, due to the gravitational redshift, since

$$\begin{aligned} \Delta\tau_e \equiv \frac{1}{\nu_e} &= \Delta t_e\sqrt{1 - r_s/r_e} \\ \Delta\tau_0 \equiv \frac{1}{\nu_0} &= \Delta t_0\sqrt{1 - r_s/r_0} = \Delta t_e\sqrt{1 - r_s/r_0}. \end{aligned}$$

Since $\Delta t_e = \Delta t_0$, we arrive at

$$z = \left[\frac{r_e(r_0 - r_s)}{r_0(r_e - r_s)}\right]^{1/2} - 1, \tag{5.7}$$

which goes to infinity as $r_e \to r_s$. Hence the surface very rapidly redshifts out of observability.

5.2 Falling Into a Black Hole

We saw previously that to an external observer, the collapse of a body (such as a star) to a black hole appears to "freeze" as the surface reaches the event horizon ar $r = r_s$, although this surface then very rapidly redshifts to invisibility. Suppose we now consider an existing black hole, and two observers: one stationary at some large radius r_0 from the black hole ($r_0 \gg r_s$), and the other falling in freely on a purely radial trajectory.

The infalling particle will follow a radial *time-like* geodesic. From the Schwarzschild metric (with $d\theta = d\phi = 0$),

$$d\tau^2 = \left(1 - \frac{r_s}{r}\right) dt^2 - \frac{1}{c^2}\left(1 - \frac{r_s}{r}\right)^{-1} dr^2. \tag{5.8}$$

This gives the equation of motion of the observer as

$$1 = \left(1 - \frac{r_s}{r}\right) \dot{t}^2 - \frac{1}{c^2}\left(1 - \frac{r_s}{r}\right)^{-1} \dot{r}^2, \tag{5.9}$$

where dots denote derivatives with respect to proper time τ.

We also need to relate $d\tau$ and dt for *motion along the geodesic*. This turns out to be

$$\left(1 - \frac{r_s}{r}\right) \dot{t} = K, \tag{5.10}$$

where K is a constant which depends on the initial conditions. If we assume the second (infalling) observer is at rest initially at ∞, then $K = 1$, as is apparent from Eq. 5.8 with $dr = 0$ and $r_s/r \to 0$. This just says that proper time goes to coordinate time for a stationary particle at infinity, as we would expect. With $K = 1$, then Eqs. 5.9 and 5.10 combine to give

$$\begin{aligned} 1 &= \left(1 - \frac{r_s}{r}\right)^{-1} - \frac{1}{c^2}\left(1 - \frac{r_s}{r}\right)^{-1} \dot{r}^2 \\ \left(1 - \frac{r_s}{r}\right) &= 1 - \frac{\dot{r}^2}{c^2} \\ \dot{r}^2 &= \left(\frac{dr}{d\tau}\right)^2 = c^2 \frac{r_s}{r} \Longrightarrow \frac{dr}{d\tau} = \pm c \left(\frac{r_s}{r}\right)^{1/2}. \end{aligned} \tag{5.11}$$

Take the negative root (since the observer is falling *in*) and integrate:

$$\begin{aligned} r^{1/2}\, dr &= -c\, r_s^{1/2}\, d\tau \\ \frac{2}{3} r^{3/2} &= -c\, r_s^{1/2} \tau + A. \end{aligned}$$

Assume that at $\tau = \tau_0$, the observer is at $r = r_0$; then

$$\tau - \tau_0 = \frac{2}{3c\, r_s^{1/2}} (r_0^{3/2} - r^{3/2}). \tag{5.12}$$

Eq. 5.12 shows that the infalling observer reaches $r = 0$ in a finite proper time, although the stationary observer will never see the infalling observer cross the event horizon.

5.3 Evidence for the Existence of Black Holes

The blackness of black holes makes trying to detect them directly impossible. This leaves two possible means of detecting black holes (aside from gravitational waves):

1. **Gravitational Effects** i.e., measurement of the black hole mass.
2. **Radiation** from material near the hole which is falling in.

What exactly does (2) mean?

Suppose we have a body of mass M and radius R, such as a star, and we drop a particle onto it from infinity. When it strikes the surface, it must give up all of its gravitational potential energy in the form of radiation, kinetic energy of fragments, etc. This gravitational potential energy is just:

$$U = -\frac{GM}{R} \tag{5.13}$$

per unit mass. Write this as

$$U = -\frac{1}{2}c^2\frac{2GM}{c^2R} = -\frac{1}{2}c^2\frac{r_s}{R}. \tag{5.14}$$

Now, $\frac{1}{2}c^2$ is just half of the rest-mass energy (per unit mass) of the particle, and this implies that the depth of the potential at the Schwarzschild radius is $\frac{1}{2}c^2$; although we have approximated the potential as Newtonian in Eq. 5.13, this is the correct General Relativistic answer (cf. Eq. 4.85 for the effective potential).

Eq. 5.14 indicates that if the radius $R \gg r_s$, then only a very small fraction of the rest-mass energy gets turned into kinetic energy or radiation: the binding energy $-U \ll c^2$. As R approaches r_s, however, this fraction gets larger, approaching $\frac{1}{2}$ at $R = r_s$.

Suppose we have spherical infall onto the body, with total mass accretion rate $\dot{M}$. The *accretion luminosity* generated as this accreted matter reaches the surface is then

$$L_{\rm acc} = \dot{M} \cdot \frac{GM}{R}. \tag{5.15}$$

In terms of the available rest-mass energy $\dot{M}c^2$, we can define an efficiency of radiating emission,

$$E = \frac{L_{\rm acc}}{\dot{M}c^2} = \frac{GM}{Rc^2} = \frac{1}{2}\frac{r_s}{R}. \tag{5.16}$$

Clearly, as noted above, the efficiency will be very low if $R \gg r_s$. For most astrophysical objects, such as stars, it is true that $R \gg r_s$. However, for neutron stars, where $R \sim$ a few r_s, it can be quite substantial.

There is another important implication of this, in terms of the temperature of the emitting material. Suppose that the accretion luminosity is radiated as thermal emission, i.e., the surface radiates as a blackbody of radius R. Then

$$L_{\rm acc} = 4\pi\, R^2\, \sigma\, T_b^4, \tag{5.17}$$

where σ is the Stefan-Boltzmann constant. Then we have

$$T_b = \left[\frac{L_{\rm acc}}{4\pi\, R^2\, \sigma} \right]^{1/4} . \tag{5.18}$$

X-ray satellites have identified large numbers (hundreds) of galactic x-ray sources, with typical x-ray luminosities $L_x \sim 10^{37}$ erg s^{-1}. Many of these are observed to be in binary systems; furthermore many are observed to vary on short timescales, implying that the size of the emitting region must be small ($R_e \lesssim c\,\Delta t$). We can then write Eq. 5.18 as

$$T_b \simeq 1.1 \times 10^7 \left(\frac{L_{37}}{R_{10}^2} \right)^{1/4} , \tag{5.19}$$

where $L_{37} = L_{\rm acc}/10^{37}$ erg s^{-1} and $R_{10} = R/10$ km. Thus accretion onto neutron stars (or black holes) naturally provides x-ray temperatures (1 keV $\approx 1.1 \times 10^7$ K) for accretion luminosities comparable to the observed.

Now for black holes there is no solid surface, so what happens? In principle, most of the rest-mass energy can simply be transported across the event horizon. However, spherical accretion is unlikely to occur in general, as infalling gas will have some angular momentum. It will most likely then form an *accretion disk* around the black hole; the gas in the disk slowly loses energy and angular momentum due to viscosity and spirals into the hole. (This will probably happen for neutron stars as well.) The *x-ray binaries* are therefore excellent candidates for compact objects, either neutron stars or black holes. Some are observed to be pulsars, which clearly identies them as neutron stars.

How can we tell whether accretion is occuring onto a neutron star or a black hole? Clearly, if we determine that the mass is greater than the maximum mass of a neutron star, accretion must be occuring onto a black hole. The best candidates for accretion onto a compact object are the so-called single-lined spectroscopic binaries.

What can we say about the mass of a binary system? Take the following figure:

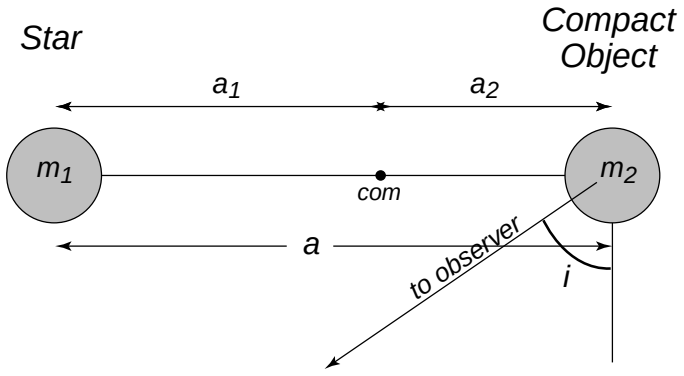


Figure 5.4: Geometry of an X-ray binary.

Here i is the inclination of the orbital plane to our line of sight. We also have

$$\begin{aligned} a &= a_1 + a_2 \\ M_1 a_1 &= M_2 a_2, \end{aligned} \tag{5.20}$$

from the definition of the center of mass.

Any emission line from the star M_1 will be Doppler-shifted by its motion about the center of mass of the system. The amplitude of this Doppler variation (in the non-relativistic limit) is just v_1, the projection of the orbital velocity of M_1 along the line-of-sight:

$$v_1 = \frac{2\pi}{P}\, a_1 \, \sin i, \tag{5.21}$$

where P is the orbital period. By measuring v_1 and P, we determine $a_1 \, \sin i$. From Kepler's laws,

$$\frac{G(M_1 + M_2)}{a^3} = \left(\frac{2\pi}{P}\right)^2, \tag{5.22}$$

while from Eq. 5.20,

$$a = \frac{M_1 + M_2}{M_2}\, a_1. \tag{5.23}$$

Writing Eq. 5.21 as

$$a_1 = \frac{P v_1}{2\pi \sin i}, \tag{5.24}$$

and substituting into Eqs. 5.23 and 5.22:

$$\begin{aligned} \frac{G(M_1 + M_2)}{a^3} &= \frac{G(M_1 + M_2)\, M_2^3}{(M_1 + M_2)^3}\, \frac{(2\pi \sin i)^3}{(P v_1)^3} \\ &= \left(\frac{2\pi}{P}\right)^2 \\ \Longrightarrow \frac{P v_1^3}{2\pi G} &= \frac{(M_2 \sin i)^3}{(M_1 + M_2)^2} \equiv f(M_1, M_2, i). \end{aligned} \tag{5.25}$$

This is the "mass function." (Note that it has dimensions of mass.) If $M_2 \gg M_1$, then $f \simeq M_2 \sin^3 i$; if $M_1 \gg M_2$, $f \approx M_2^3 \sin^3 i / M_1^2$. In either case, f is less than the true mass of the compact object M_2. Without further information, it is impossible to go beyond Eq. 5.25 without making additional assumptions.

For some x-ray binaries, it has been possible to determine the mass function of the optical member of the binary, from Doppler shifts of x-ray lines. If we denote

$$f_1 = \frac{(M_1 \, \sin i)^3}{(M_1 + M_2)^2}, \quad f_2 = \frac{(M_2 \, \sin i)^3}{(M_1 + M_2)^2}, \tag{5.26}$$

then the ratio of these two expressions gives the mass ratio

$$q = \frac{M_2}{M_1}, \tag{5.27}$$

and

$$M_2 = \frac{f_1 q\,(1+q)^2}{\sin^3 i}. \tag{5.28}$$

A unique determination for the mass of the compact object depends on knowing $\sin i$. Limits on $\sin i$ can be derived if, e.g., the x-ray emission is periodically eclipsed by the stellar companion, or for that matter, if the x-ray source is never eclipsed.

There are currently eight stellar-mass black hole candidates known:

Object	Mass Function ($M_\odot$)	Estimated Mass ($M_\odot$)	Comments
Cyg X-1	0.25	~ 10	Steady source w/ high-mass companion
LMC X-1	0.14	6	″
LMC X-3	2.3	9	Transient source w/ low-mass companion
Nova Muscae	3.1	6	″
GS $2023+33$ (V404 Cyg)	6.3	10	″
GRO J$0422+32$	~ 3	4.5	″
A0620	3.2	6	″
GRO J$1655-40$ (Nova Scorpius)	~ 3	4.5	″ Superluminal radio jet source

How can we claim that sources with such small mass functions (e.g., Cyg X-1) are candidates for black holes? We must use other information about these systems. For example, the distance to the Cyg X-1 system is fairly well-determined from observations of the optical component to be $d \sim 2.5$ kpc. With this distance, the luminosity of the optical star is quite high (it is an OB supergiant) and it must be rather massive, with M *at least* 8.5 $M_\odot$ and more probably ~ 20 $M_\odot$ (from stellar structure calculations). We will get a minimum mass if we take $\sin i = 1$, so solving

$$\frac{M_2^3}{(M_1+M_2)^2} = 0.25$$

for $M_1 = 8.5$–20 $M_\odot$ gives $M_2 = 3.3$–5.5 $M_\odot$.

5.4 Massive Black Holes in Galaxies

Several percent of galaxies exhibit *active galactic nuclei*: high luminosity, non-thermal emission which is generated on very small size scales (as determined by light travel-time arguments for varying sources). As with galactic x-ray binaries, the emission extends to x-ray wavelengths; this alone with the high luminosities from small ($r \ll 1$ pc) volumes, suggests that the emission arises from accretion onto a *massive* ($M_{\rm BH} \sim 10^6$–10^9 $M_\odot$) central black hole. The emission from these *AGN* can be comparable to the luminosity from the rest of the galaxy; the most extreme examples are the quasars (for *quasi-stellar radio sources*) in which the AGN is so luminous that it completely dwarfs emission from the host galaxy (if any).

Obviously the techniques used for trying to estimate masses which are used for galactic x-ray binaries are of no use here. How can we try to ascertain the existence of these massive black holes? We must look for gravitational evidence of large masses in very small volumes, or, equivalently, to look for very large *mass densities*.

Observationally this has proved very difficult: even at ~ 10 Mpc (a relatively nearby galaxy), $1'' \approx 50$ pc. The advent of the Hubble Space Telescope (HST) improved the available resolution by quite a bit, but still produced no clear-cut examples, as the mass of stars associated with the galactic nucleus generally swamps the putative black hole mass at the size scale observed.

Ground-based radio observations with the VLBA (Very Long Baseline Array) have now produced ironclad evidence for a massive black hole in a galaxy, NGC 4258 (M 106). Radio observations in the 22 GHz water maser line reveal a warped disk of gas in essentially perfect Keplerian rotation ($v \propto R^{-1/2}$), with $R_{\rm inner} = 0.13$ pc and $R_{\rm outer} = 0.25$ pc. This implies

$$\begin{aligned} M_c &= M_{\rm BH?} = 3.6 \times 10^7\,{\rm M}_\odot \\ \rho_c &> 4 \times 10^9\,{\rm M}_\odot\,{\rm pc}^{-3}. \end{aligned}$$

To understand the significance of this, we have to detour briefly into stellar dynamics.

Suppse we have a system like a normal (no massive black hole) galactic nucleus, made up of a very large number of stars. We will assume that this system is *self-gravitating*, that is, the stars themselves provide the gravitational potential in which they move. According to the *Virial Theorem*, for an isolated system

$$2K + W = 0, \tag{5.29}$$

where K is the total kinetic energy and W is the total potential energy. If the total mass of the stellar system is M, then the kinetic energy is just $K = \frac{1}{2}M\langle v^2\rangle$, where $\langle v^2\rangle$ is the mean-square speed of the stars. If the radius of the system is R, then the potential energy $W \approx -GM^2/R$, so

$$\begin{aligned} 2\left(\frac{1}{2}M\langle v^2\rangle\right) &= \frac{GM^2}{R} \\ \langle v^2\rangle &\approx \frac{GM}{R}. \end{aligned} \tag{5.30}$$

Since the mass of an individual star is much smaller than the mass of the galactic nucleus, each star moves through a rather smooth potential produced by all the other stars in the nucleus. The typical deflection of a star by gravitational encounters with other stars is quite small, and one can show (continuing on from Eqs. 4.65 and 4.66) that the time it takes for the velocity of a star to shift by an amount $|\delta v|$ comparable to its original velocity (the *relaxation time*) is

$$t_R \sim \frac{0.1N}{\ln N} t_{\rm cross}, \tag{5.31}$$

where the crossing time

$$t_{\rm cross} = \frac{R}{v} \tag{5.32}$$

is just the typical time it takes for a star to cross the nucleus.

A typical galactic nucleus has a velocity dispersion $\langle v^2 \rangle^{1/2} \approx 200$ km s^{-1}, which implies $t_{\rm cross} \approx 5 \times 10^5 \, R_{100}$ yr, where the nucleus is $R = 100 \, R_{100}$ pc; from Eq. 5.30, this also implies that $M \sim 10^9$ M$_\odot$ for these values. The relaxation time is basically the timescale in which the system "forgets" the initial conditions.

Another important stellar-dynamical timescale is the *evaporation* timescale. From time to time, an encounter (a gravitational "collision") between stars in the system will give a star enough energy to escape. There is a slow and irreversible "leakage" of stars from the system. With the aid of the Virial Theorem, we can estimate this timescale in terms of the relaxation time.

One of the consequences of Newton's law of gravitation is that the force due to a spherical distribution of matter contained within radius R is the same as a point mass of the same total mass for $r > R$:

$$F(r) = -\frac{GM(r)}{r^2} \tag{5.33}$$

$$M(r) = 4\pi \int_0^r \rho(R) R^2 \, dR. \tag{5.34}$$

The *escape velocity* at radius r is determined by the condition

$$\frac{1}{2} m v^2 + m\Phi(r) > 0, \tag{5.35}$$

where $\Phi(r)$ is the gravitational potential at radius r. Hence

$$v_e^2 = 2|\Phi(r)| = -2\Phi(r), \tag{5.36}$$

where $|\Phi(r)|$ is the magnitude of the potential at r.

The mean-square escape speed from a spherical system is then just the density-weighted average of v_e^2:

$$\begin{aligned} \langle v_e^2 \rangle &= \frac{4\pi \int r^2 \rho(r) v_e^2(r) \, dr}{4\pi \int r^2 \rho(r) dr} \\ &= \frac{4\pi \int r^2 \rho(r) (-2\Phi(r)) \, dr}{M} \\ &= \frac{-8\pi \int r^2 \rho(r) \Phi(r) \, dr}{M}, \end{aligned} \tag{5.37}$$

where M is the total mass of the system.

What is this integral? For a spherically symmetric system, the total gravitational potential energy is

$$\begin{aligned} W &= \frac{1}{2} \int 4\pi r^2 \rho(r) \, \Phi(r) \, dr \\ &= 2\pi \int r^2 \rho(r) \, \Phi(r) \, dr, \end{aligned} \tag{5.38}$$

where the factor of $\frac{1}{2}$ arises because we don't want to count any star more than once, i.e., each star contributes to ρ and Φ equally.

Thus we can write Eq. 5.37 as

$$\langle v_e \rangle = -\frac{4W}{M}. \tag{5.39}$$

Now from the Virial Theorem (Eqs. 5.29 and 5.30), we know that the mean-square speed of the stars is just

$$\langle v^2 \rangle = -\frac{W}{M} \tag{5.40}$$

(since $K = \frac{1}{2}M\langle v^2 \rangle = -W/2$). Thus

$$\langle v_e^2 \rangle = 4\langle v^2 \rangle \tag{5.41}$$

and the root mean-square (RMS) escape velocity $\langle v_e^2 \rangle^{1/2}$ is just twice the RMS stellar velocity $\langle v^2 \rangle^{1/2}$.

Assume that the actual distribution of stellar velocities (more precisely, speeds) is a Maxwellian:

$$f(v) = \left(\frac{2}{\pi}\right)^{1/2} \sigma^{-3} v^2 e^{-v^2/2\sigma^2}, \tag{5.42}$$

where σ is the *velocity dispersion*. This is related to the RMS velocity by

$$\langle v^2 \rangle^{1/2} = v_{\rm RMS} = \sqrt{3}\sigma \tag{5.43}$$

Then the fraction of stars which have velocities exceeding twice the RMS velocity is

$$f_{\rm esc} = \int_{2v_{\rm RMS}}^{\infty} f(v)\, dv = \left(\frac{2}{\pi}\right)^{1/2} \sigma^{-3} \int_{2v_{\rm RMS}}^{\infty} v^2 e^{-v^2/2\sigma^2}\, dv. \tag{5.44}$$

To simplify this, let $u = v/\sqrt{2}\sigma$; then

$$f_{\rm esc} = \frac{2}{\sqrt{\pi}} \cdot 2 \int_{u_{\rm RMS}}^{\infty} u^2 e^{-u^2}\, du. \tag{5.45}$$

We can do the integral by parts:

$$\begin{aligned} 2\int_{u_{\rm RMS}}^{\infty} u^2 e^{-u^2}\, du &= \left[-ue^{-u^2}\right]_{u_{\rm RMS}}^{\infty} + \int_{u_{\rm RMS}}^{\infty} e^{-u^2}\, du \\ &= u_{\rm RMS} e^{-u_{\rm RMS}^2} + \frac{\sqrt{\pi}}{2}\,{\rm Erfc}(u_{\rm RMS}), \end{aligned}$$

where Erfc is the complementary error function. Since $u_{\rm RMS} = 2\sqrt{3}\sigma/\sqrt{2}\sigma = 2\sqrt{3/2}$, evaluation gives

$$f_{\rm esc} = \frac{2}{\sqrt{\pi}} \cdot 6.604 \times 10^{-3} = 7.45 \times 10^{-3}. \tag{5.46}$$

We can approximately account for evaporation by assuming a fraction $f_{\rm esc}$ of the stars are lost every relaxation time (because a Maxwellian distribution will be re-established every relaxation time). If the original number of stars in the system is N, then

$$\frac{dN}{dt} \approx -\frac{f_{\rm esc} N}{t_R} \equiv \frac{N}{t_{\rm evap}}, \tag{5.47}$$

where the evaporation timescale is

$$t_{\rm evap} \equiv \frac{t_R}{f_{\rm esc}} \approx 134 t_R. \tag{5.48}$$

The process of evaporation therefore limits the lifetime of any stellar system to ~ 100 relaxation times.

What does this have to do with NGC 4258? The upper limit to any deviation from Keplerian velocities is $\Delta v \leq 3$ km s^{-1}. If we had purely Keplerian motion (i.e., a point mass like a black hole), then the velocities at the inner and outer radii would be related by

$$\left(\frac{r_{\rm out}}{r_{\rm in}}\right)^{1/2} v_{\rm out} = v_{\rm in}. \tag{5.49}$$

Denote the ratio of the mass contained between radii $r_{\rm in}$ and $r_{\rm out}$ to that contained within $r_{\rm in}$ by

$$\delta M = \frac{M(r_{\rm out}) - M(r_{\rm in})}{M(r_{\rm in})}. \tag{5.50}$$

Similary, define the velocity difference

$$\delta v = \left(\frac{r_{\rm out}}{r_{\rm in}}\right)^{1/2} v_{\rm out} - v_{\rm in}. \tag{5.51}$$

A little algebra shows that

$$\delta M \simeq 2\frac{\delta v}{v_{\rm in}} \quad \text{for } \delta v / v_{\rm in} \ll 1. \tag{5.52}$$

From the observations, $\delta v / v_{\rm in} \lesssim 0.003$, and so $\delta M \lesssim 1\%$.

This immediately says that the mass contained between $r_{\rm in}$ and $r_{\rm out}$ is less than about 1% of the mass within $r_{\rm in}$. If the central mass is not a massive black hole, it must be very sharply cut off within $r_{\rm in}$. Now suppose that we try to replace the central mass with a stellar cluster. Even if we make the (ridiculous) assumption that we can cut it off abruptly at $r_{\rm in} = 0.13$ pc (which maximizes the crossing, relaxation, and evaporation times), we get

$$\begin{aligned} v_{\rm RMS} &\approx 1100 \quad \text{km s}^{-1}{}^{1} \\ t_{\rm cross} &\approx 120 \quad \text{yrs.} \end{aligned}$$

Assuming that the stars have typical masses of 1 $M_\odot$,

$$\begin{aligned} t_R &\approx 2 \times 10^5 \, t_{\rm cross} \approx 2.4 \times 10^7 \quad \text{yrs} \\ t_{\rm evap} &\approx 3 \times 10^9 \text{ yrs.} \end{aligned}$$

This is already disturbingly short, since the age of the galaxy must be greater than 10^9 yrs.

However this model is absurd. Any realistic model for a stellar cluster will be centrally concentrated, and the density will decline as some power of the radius. For example, the density profiles of globular clusters and galactic nuclei can be described reasonably well by profiles of the form

$$\rho = \rho_0 \left[1 + (r/r_c)^2\right]^{-\alpha/2}, \tag{5.53}$$

where ρ_0 is the central density, r_c the core radius, and $\alpha \approx 4$–5 for galactic nuclei and globular clusters.

(Note that for $\alpha = 2$, the cluster mass increases with radius at least as fast as r, so that $\delta M \gtrsim 1$; clearly $\alpha > 2$.)

For the choice $\alpha = 5$, for example, requiring $\delta M < 0.01$ requires $r_c < 0.012$ pc, and $\rho_0 \gtrsim 4.5 \times 10^{12}$ $M_\odot$. In this case the evaporation time is

$$t_{\rm evap} \lesssim 10^8 \text{ yrs.}$$

This timescale isn't very sensitive to the actual parameters used; for any realistic stellar cluster, the evaporation time is *much less* than the age of the galaxy. Although it is possible to have possibly some exotic form of matter be the central mass (e.g., a massive neutrino ball), the most *conservative* assumption is that the central object is a massive black hole.

Chapter 6

Cosmology

Cosmology is fundamentally concerned with the distribution and dynamics of the material which makes up the universe. *All* the information we possess on the rest of the universe (outside the solar system) comes from collecting photons emitted by astrophysical objects. It is straightforward (although often very tedious) to measure the distribution of objects on the sky. But how do we tell how far away they are?

6.1 Cosmic Distance Scale

The first reasonable estimate of the distances to nearby stars was made by Newton. He assumed that all stars have the same brightness as the Sun; this means that the apparently brightest stars are simply the nearest. Since these are about 10^{11} times fainter than the Sun, the inverse square law for flux gives

$$\frac{D_\star}{D_\odot} \approx \left(\frac{f_\odot}{f_\star}\right)^{1/2} = \left(10^{11}\right)^{1/2} \approx 3 \times 10^5 \tag{6.1}$$

$$D_\star \approx 3 \times 10^5 \text{ AU} = 4.5 \times 10^{18} \text{ cm} \approx 1.5 \text{ pc} \tag{6.2}$$

which is pretty close to correct (although Newton actually made a numerical error of a factor of 100).

6.1.1 Parallax

This is based on the apparent shift of nearby objects due to more distant ones as we change position; for astronomical purposes, this shift is due to the orbit of the Earth about the Sun.

$$\begin{aligned} \tan\psi &= d/D \\ \Longrightarrow D &= d/\tan\psi \simeq d/\psi \end{aligned} \tag{6.3}$$

since ψ is a very small angle.

This is also where the term parsec ("parallax-second") arises from: an object at a distance of 1 parsec has a parallax $\psi = 1''$:

$$1\,\text{parsec} = \frac{1\,\text{AU}}{(2\pi/1.296 \times 10^{6''})} = 3.0856 \times 10^{18}\,\text{cm}$$

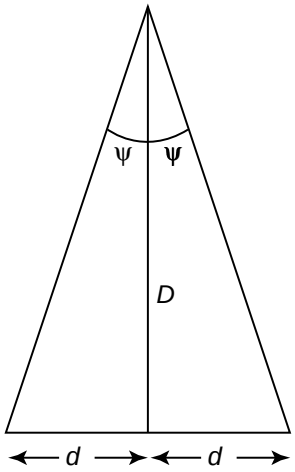


Figure 6.1: Parallax geometry.

The first stellar parallax was measured by Bessel in 1837. The *largest* stellar parallax is $\psi \simeq 0.8''$. This is limited by the smallest parallax which can be reliably measured, and is reliable out to ~ 30 pc.

6.1.2 Standard Candles

There are other techniques which can be used for stars, as discussed in Berry and in Silk, but for the extragalactic distance scale, all of the methods are variants of the technique used by Newton to estimate distances to the stars: they are all based on the idea of a *standard candle*.

If we know the true luminosity of an astrophysical source of radiation, and we measure the flux we receive from it (both in some arbitrary wavelength range), then we can immediately determine its distance:

$$D = \left(\frac{L}{4\pi}\right)^{1/2} \tag{6.4}$$

(assuming of course, that the source is radiating isotropically).

The problem, of course, is knowing what the true luminosity is. Determination of the cosmological distance scale has been an immensely time-consuming and controversial business.

A hotly debated issue in the first quarter of this century was whether the so-called "spiral nebulae"—which we now know to be galaxies like our own—were part of the galaxy or not. The first accurate determination of the distnace to the nearest large galaxy to our own—M31 in Andromeda—was made by Öpik in 1922, using observations of rotation velocities.

Spectroscopic observations had already shown evidence for gas motions in M31; assuming that these represented more or less circular velocities, this gives the mass interior to the

edge of the disk, where the speed is v_c, as

$$\frac{GM}{r^2} = \frac{v_c^2}{r} \Longrightarrow \frac{GM}{(\theta D)^2} = \frac{v_c^2}{\theta D} \tag{6.5}$$

where θD is the angular radius of the disk. since the observed flux f is just

$$f = \frac{L}{4\pi D^2} \tag{6.6}$$

from the galaxy, substituting $D^2 = L/4\pi f$ into Eq. 6.5 and rearranging gives

$$D = \frac{v_c^2 \theta}{4\pi G f} \frac{L}{M} \tag{6.7}$$

Assuming $M/L \sim 3$, and with everything else on the right-hand side of Eq. 6.7 known, Öpik obtained $D \approx 450$ kpc, compared with the modern value $D \approx 770$ pc.

6.1.3 Cepheid Variables

The most important distance indicator to be discovered (which finally settled the debate over spiral nebulae) are the *Cepheid variable* stars. Many stars show regular (i.e., periodic) variations in brightness. In 1912, Henrietta Leavitt showed that there was a linear relation between the apparent magnitude m and the period P for the Cepheid variables in the Small Magellanic Cloud, a small nearby ($D \approx 60$ kpc) galaxy. Since these stars are all at nearly the same distance, she concluded that there is a unique relationship between the absolute magnitude and the period.

Digression on Magnitudes: Traditionally, the brightness of astrophysical objects have been expressed not as fluxes, but as *magnitudes*. This is a logarithmic scale, with the apparent magnitude $m \propto \log f$. If two objects have observed fluxes f_1 and f_2, then

$$m_2 - m_1 = 2.5 \log(f_1/f_2) \tag{6.8}$$

The factor of 2.5 means that a difference in flux of a factor of 100 corresponds to 5 magnitudes.
The *absolute magnitude* M is defined as the magnitude a source would have if it were at a distance of 10 pc. Since $f \propto D^{-2}$,

$$m - M = 5 \log(D/10) \tag{6.9}$$

where D is in parsecs. The absolute magnitude of the Sun is 4.72, while its apparent magnitude is $m_\odot = -26.85$. $m - M$ is called the *distance modulus*.

The discovery of the Cepheid period-luminosity relationship was a major breakthrough, as it made it possible to determine the luminosity simply by observations of the light curve; the distance is then given immediately by the observed flux. In 1923, Hubble discovered Cepheids in M31, thereby establishing its distance unequivocally.

Other "standard candles" include:

- **Novae:** Correlation between magnitude at maximum and fading time.
- **Brightest cluster galaxies:** Assumes that the maximum magnitude is very similar in all clusters. The galaxy luminosity distribution is fit pretty well by the Schecter luminosity function: the number of galaxies with luminosity between L and $L + dL$ per unit volume is

$$\phi(L)\, dL = \phi_\star \left(\frac{L}{L_\star}\right)^\alpha e^{-L/L_\star} \frac{dL}{L_\star} \tag{6.10}$$

  with $\phi_\star \approx 1.2 \times 10^{-2} h^3$ Mpc^{-3}, $\alpha \approx -1.25$, and $L_\star \approx 1.0 \times 10^{10} h^{-2}$ L$_\odot$in the visual.
- **Planetary nebulae luminosity function**
- **Supernovae of Type II**

All these methods have sources of error, but at least they are different for different methods.

6.2 Expansion of the Universe

Using Cepheids to determine distances, and spectra to determine velocities, Hubble in 1929 announced his discovery of the *expansion of the universe*: the velocity of recession of a galaxy from ours is proportional to its distance from us:

$$v = HD \tag{6.11}$$

where H, the Hubble constant, is equal to 75 ± 25 km s^{-1} Mpc^{-1}. As we will see shortly, the precise value of Hubble's constant is of profound cosmological significance, as it is directly related to the age of the universe.

Even before deducing the expansion of the universe, Hubble made another observation of equally profound importance. It was already known that there are many more "spiral nebulae" (galaxies) of small angular extent than large angular extent, as one would expect if the distribution were more or less uniform around us. Hubble quantified this using the following test, which had been devised originally in conjunction with studies of star counts in the Milky Way.

Suppose the universe is static and the geometry is Euclidean, and that all galaxies have the same luminosity L. Due to the geometric dilution of flux, the received flux from each galaxy is $f = L/4\pi D^2$. With all galaxies having the same luminosity, all galaxies with flux $> f$ are at a distance $< D$.

The volume of space per steradian out to a distance D is just $D^3/3$. If the galaxies are distributed uniformly, with average number density n, the average number per steradian brighter than f would be:

$$N(>f) = nV = \frac{nD^3}{3} = \frac{n}{3}\left(\frac{L}{4\pi f}\right)^{3/2} \tag{6.12}$$

That is, $N(>f)$ is proportional to $f^{-3/2}$.

This is (unfortunately) often expressed in magnitudes instead of fluxes; from Eq. 6.8, $f \propto 10^{-0.4m}$, and so Eq. 6.12 becomes

$$N(< m) \propto 10^{0.6m} \tag{6.13}$$

From star counts it was known that the stellar distribution in the Milky Way did *not* obey Eq. 6.13; it is a finite system. However Hubble's galaxy counts closely followed Eq. 6.13.

What does the Hubble relation mean? It is obviously absurd to infer that everything is expanding away from *us*, i.e., that we are at the center of an expanding sphere of galaxies. However, there is an alternate and far more plausible explanation. Take our galaxy as the origin of a system of coordinates.

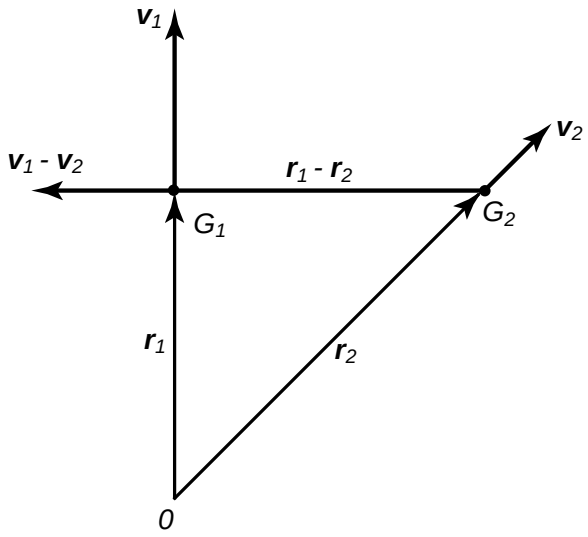


Figure 6.2: Geometry for galaxies in an expanding universe.

Hubble's Law is then $\boldsymbol{v} = H\boldsymbol{r}$, where $\boldsymbol{r}$ is the position vector of a galaxy relative to us. For galaxy G_1 then, its velocity of recession is

$$\boldsymbol{v}_1 = H\boldsymbol{r}_1$$

while for galaxy G_2,

$$\boldsymbol{v}_2 = H\boldsymbol{r}_2.$$

But what is the velocity of galaxy G_1 with respect to G_2? This is just

$$\boldsymbol{v}_1 - \boldsymbol{v}_2 = H\boldsymbol{r}_1 - H\boldsymbol{r}_2 = H(\boldsymbol{r}_1 - \boldsymbol{r}_2)$$

which is just H times the distance of G_1 from G_2. Thus it is clear that each galaxy sees all others receding from it, i.e., there is a uniform expansion of the entire system.

By the end of the 1920s, therefore, observations had established two fundamental facts about the universe:

1. It is expanding.
2. On large enough scales (i.e., volumes big enough to contain reasonable numbers of galaxies), it appears to be fairly *homogeneous* (the same at every point) and isotropic (the same in every direction).

Chapter 7

Theoretical Cosmology

Newton realized that his theory of gravitation raised a cosmological problem. In Newton's time (and for more than two centuries thereafter), it was assumed that the universe was infinite and unchanging, i.e., *static*. The force of gravitation causes a problem for such a cosmos; why doesn't gravity cause the system to collapse?

Newton argued that it was stable because of the fact that it is infinite and uniform: the force on a particle in any direction would be cancelled out by an identical force in the opposite direction. If the universe were *finite* in extent, a distribution of matter initially at rest would inevitably collapse to the center under the influence of its own gravity.

This is an example of a symmetry argument, which frequently occurs in physics, often to great effect. This one, however, is *wrong*. There is a fundamental flaw in Newton's homogeneous infinite universe—it is *unstable*, and the entire universe should collapse in on itself, *unless* it is expanding, i.e., it has enough kinetic energy to overcome (at least temporarily) the attractive force of gravity.

Another flaw with the infinite, static universe of Newton's time is what is commonly referred to as Olber's paradox (although it was first pointed out by Edmund Halley more than a century earlier): why is the night sky dark?

If the universe is infinite and unchanging, then every ray we follow away from the Earth must eventually intercept the surface of a star. Therefore the entire sky should be as bright as the surface of a star!

This is obviously not the case. Olber (1827) proposed that the solution is that there is some material between the stars which absorbs the radiation from them. However this won't work either, because of simple thermodynamics—the absorbing material must eventually become as hot as the radiation it is absorbing (since it is exposed to an infinite bath of radiation) and the problem returns.

7.1 Cosmological Principle

The appearance of Einstein's theory of General Relativity sparked an explosion of interest in theoretical cosmology; in fact, the expansion of the universe could have been predicted

By Phil Maloney.

by Einstein[1] a decade before it was discovered by Hubble.

For simplicity, theoretical models treat the evolution of a smoothed-out version of the universe, i.e., we will ignore (for the moment) the irregularities due to galaxies, clusters, and other local density perturbations. In the first explicit cosmological model based on General Relativity, Einstein (1917) adopted what has since become known as the *Cosmological Principle*: on suitably large scales, the universe is homogeneous and isotropic. This was a major leap of faith at the time, as Hubble's observations of the galaxy distribution were not to appear for another ten years, and star counts had already established that the Milky Way (which observationally was "the universe" at the time) was anything but homogeneous and isotropic. Einstein adopted the Cosmological Principle in the specific context of a closed, "Machian" universe.

The fundamental question we need to answer to study the evolution of this smooth, perfect universe is then: what is the geometry of the spacetime which comprises the universe?

Although General Relativity is necessary to really answer this question, it is possible to restrict the possible forms of the metric dramatically by the adoption of the Cosmological Principle. If the universe is homogeneous and isotropic everywhere, then if we take a spacelike "slice" through spacetime at constant cosmic time t, then these slices must be symmetric about every point in them. Clearly then, the spatial structure of the 3-D *position space* must be one of constant curvature.

7.2 Comoving Coordinates

We discussed the 3-D space of constant curvature earlier (§4.3 starting on page 40). The spatial part of this metric is just

$$dr^2 = \frac{dR^2}{1 - KR^2} + R^2\, d\theta^2 + R^2 \sin^2\theta\, d\phi^2, \tag{4.22}$$

where K is the curvature and R is the radial coordinate. The curvature is of course independent of position, but it may be a function of time.

Now the expansion of the universe means that all galaxies are steadily moving apart from one another. The assumption of homogeneousness and isotropy for the universe means that the expansion (i.e., ignoring any local peculiar velocities due to, say, gravitational interaction of nearby galaxies) cannot alter the relative orientations of galaxies (or other bodies) with respect to one another: that is, it will lead to no rotation or anisotropic stretching.

This means that the proper physical separation between, say, a pair of galaxies, must be

$$d = d_\circ\, a(t), \tag{7.1}$$

where $d_\circ$ is a constant for the pair and $a(t)$ is a universal (the same everywhere) expansion factor. The time derivative of Eq. 7.1 is just the relative velocity of the two galaxies:

$$v = \dot{d} = d_\circ \dot{a} = \frac{\dot{a}}{a} d \equiv H d. \tag{7.2}$$

[1] But was not, in what he later referred to as the greatest blunder of his career.

We therefore see that the value of the coefficient in Hubble's expansion law is determined by the rate of change of the expansion coefficient:

$$H = \frac{\dot{a}}{a}. \tag{7.3}$$

This will, in general, be a function of time; the present-day (observed) value is $H_\circ$.

The Hubble constant also sets an age scale for the universe: if we assume that $\dot{a}$ is constant, then at time

$$\begin{aligned} \frac{a}{\dot{a}} = H_\circ^{-1} &= 3.09 \times 10^{17}\, h \ \text{ s} = 9.8 \times 10^9\, h \ \text{ yrs} \\ \left(\text{where } h \right. &= \left. \frac{H_\circ}{100} \ \text{ km s}^{-1} \text{ Mpc}^{-1}\right) \end{aligned} \tag{7.4}$$

ago, the radius of the universe was zero. (There are numerical factors depending on the time dependence of a, but this sets the scale.)

The fact that the distances between galaxies (or any other objects, or arbitrarily selected points in space) increase uniformly with time due to the expansion suggests that we want to pick our spatial coordinates in a particular way, namely, that we separate out the effect of the expansion. In other words, we want to use the $d_\circ$'s from Eq. 7.1 to describe the relative positions of points.

Imagine for example, that we impose a set of spherical coordinates on the universe at some time, and then let this coordinate grid expand as $a(t)$.[2] The *physical size* of the grid would increase with time, but the *coordinates* would be unchanged: two galaxies with coordinates $(R_1, 0, 0)$ and $(R_2, 0, 0)$ which are unmoving except for their expansion with the universe (generally referred to as the "Hubble flow") will *always* have a *coordinate distance* separation $\Delta R = |R_2 - R_1|$, but their proper *physical separation* will be $d = a(t)\Delta R = a(t)|R_2 - R_1|$. These are referred to as *comoving* coordinates.

7.3 Friedmann-Robertson-Walker Metric

If we take the R-coordinate in Eq. 4.22 to be the comoving radial coordinate, then the spatial part of the metric will just be

$$dr^2 = [a(t)]^2 \left(\frac{dR^2}{1 - KR^2} + R^2\, d\theta^2 + R^2 \sin^2\theta\, d\phi^2 \right). \tag{7.5}$$

The curvature K has not yet been specified (and we will need to return to General Relativity to do so), but it may be positive, negative, or zero, with some arbitrary magnitude if it is non-zero. It will prove convenient to absorb the magnitude of K into the radial coordinate and the scale factor.

To do this, define a new parameter k by

$$K = |K|k. \tag{7.6}$$

[2] And assume $a = 1$ initially, just for convenience.

For $K \neq 0$, so that $k = +1$ or -1 depending on whether K is positive or negative. If we now introduce a rescaled radial coordinate by

$$R^\star = |K|^{1/2} R, \qquad dR^\star = |K|^{1/2}\, dR, \tag{7.7}$$

then the spatial part of the metric (Eq. 7.5) becomes

$$dr^2 = \frac{[a(t)]^2}{|K|}\left(\frac{dR^{\star 2}}{1 - kR^{\star 2}} + R^{\star 2}\, d\theta^2 + R^{\star 2} \sin^2\theta\, d\phi^2\right). \tag{7.8}$$

Here we can define a rescaled expansion factor

$$\begin{aligned} a^\star(t) &= \frac{a(t)}{|K|^{1/2}} \quad K \neq 0, \\ a^\star(t) &= a(t) \qquad K = 0. \end{aligned} \tag{7.9}$$

(Note that this leaves H unaffected.)

Then the spatial part of the metric is

$$dr^2 = [a^\star(r)]^2\left(\frac{dR^{\star 2}}{1 - kR^{\star 2}} + R^{\star 2}\, d\theta^2 + R^{\star 2} \sin^2\theta\, d\phi^2\right). \tag{7.10}$$

What is the time part of the metric? For a comoving observer, with $R^\star$, θ, and ϕ constant, t is just the proper time τ, and so (dropping the $\star$s on the coordinates):

$$d\tau^2 = dt^2 - \frac{[a(t)]^2}{c^2}\left(\frac{dR^2}{1 - kR^2} + R^2\, d\theta^2 + R^2 \sin^2\theta\, d\phi^2\right). \tag{7.11}$$

(More formally, the fact that the spatial coordinates of a comoving particle are constant along a geodesic means that the time coordinate is orthogonal to the spacelike surfaces $t =$ constant.) Here t plays the role of a *cosmic* or *world* time.

Eq. 7.11 is called the Robertson-Walker line element, after the relativists who first showed that it is the most general form for the metric of a spatially homogeneous and isotropic spacetime, independent of General Relativity.

At any given cosmic time t, the geometry of the universe is just given by the spatial part of Eq. 7.11. The *proper distance* between us and some other object, such as a galaxy, at time t is then just

$$D_p = \left[\int_0^R \frac{dR}{1 - kR^2}\right] a(t) = a(t) \times \begin{cases} \sin^{-1} R, & \text{if } k = 1; \\ R, & \text{if } k = 0; \\ \sinh^{-1} R, & \text{if } k = -1 \end{cases} \tag{7.12}$$

(cf. Eqs. 4.23 and 4.26, on p. 42). In writing Eq. 7.12, we have taken $R = 0$ to be us for convenience, as we are free to center our coordinates anywhere we like. The distance is of course proportional to $a(t)$, which changes with time.

We obtain the proper velocity of the galaxy (or whatever) with respect to us by differentiating Eq. 7.12, keeping in mind that the *comoving* radial coordinate R is *constant*:

$$v_p = \dot{D}_p = \dot{a}(t) \int_0^R \frac{dR}{1 - kR^2} = \frac{\dot{a}(t)}{a(t)} D_p. \tag{7.13}$$

This of course is just Eq. 7.2 again, with $H = \dot{a}(t)/a(t)$.

The geometry for $k = 0$ (flat space) is obviously Euclidean; such a universe is termed *open* topologically, as the radius (i.e., distance D_p) and volume increases without limit as the coordinate R increases.

The geometry for $k = -1$ (negative curvature) is not Euclidean; from our earlier discussion of curved spaces, the area of an R-sphere increases as $\sinh^2 D_p$, which for large D_p is much faster than the Euclidean D_p^2 behavior. Such a universe is also *open*.

For $k = 1$, the universe has *closed* geometry: as in our earlier discussion of positively-curved spaces, the area of an R_p-sphere has a maximum, and then decreases with increasing R_p, finally reaching zero again. Such a universe has finite volume but no boundary (just as a 2-dimensional being on the surface of a 2-D sphere will never encounter an edge, but the area of the sphere is finite).

7.4 Redshifts in an Expanding Universe

Again take $R = 0$ to correspond to our position, and consider light reaching us at the present (time $t_\circ$) from a distant galaxy. Two successive wave crests were emitted at times t_e and $t_e + \Delta t_e$, and are received by us at times $t_\circ$ and $t_\circ + \Delta t_\circ$. How are these times related?

The light travels inwards (by definition, since we are at $R = 0$) along a radial null geodesic of the Robertson-Walker metric (Eq. 7.11):

$$0 = dt^2 - \frac{[a(t)]^2}{c^2}\frac{dR^2}{1-kR^2} \tag{7.14}$$

$$\Longrightarrow dt = \pm\frac{a(t)}{c}\frac{dR}{(1-kR^2)^{1/2}} \qquad \begin{cases} + & \text{Receding (outgoing) light ray} \\ - & \text{Approaching (incoming) light ray} \end{cases} \tag{7.15}$$

Let the comoving radius of the distant galaxy be R_e. Then

$$\int_{t_e}^{t_\circ}\frac{dt}{a(t)} = -\frac{1}{c}\int_{R_e}^{0}\frac{dR}{(1-kR^2)^{1/2}} = \frac{1}{c}\int_0^{R_e}\frac{dR}{(1-kR^2)^{1/2}} \equiv \frac{f(R_e)}{c}, \tag{7.16}$$

where

$$f(R_e) = \begin{cases} \sin^{-1} R_e & k = +1 \\ R_e & k = 0 \\ \sinh^{-1} R_e & k = -1 \end{cases} . \tag{7.17}$$

Similarly,

$$\int_{t_e+\Delta t_e}^{t_\circ+\Delta t_\circ}\frac{dt}{a(t)} = \frac{1}{c}\int_0^{R_e}\frac{dR}{(1-kR^2)^{1/2}} = \int_{t_e}^{t_\circ}\frac{dt}{a(t)}, \tag{7.18}$$

since the comoving coordinate R_e is unchanged. Therefore (getting rid of the $1/c$'s),

$$\begin{aligned}\int_{t_e+\Delta t_e}^{t_\circ+\Delta t_\circ}\frac{dt}{a(t)}-\int_{t_e}^{t_\circ}\frac{dt}{a(t)} &= 0\\ &= \int_{t_\circ}^{t_\circ+\Delta t_\circ}\frac{dt}{a(t)}-\int_{t_e}^{t_e+\Delta t_e}\frac{dt}{a(t)}.\end{aligned}$$

Assuming that $a(t)$ does not vary significantly over the intervals $\Delta t_\circ$ and Δt_e, we can remove it from the integrals, and get

$$\frac{\Delta t_\circ}{a(t_\circ)}=\frac{\Delta t_e}{a(t_e)}. \tag{7.19}$$

The emitted and observed wavelengths λ_e and $\lambda_\circ$ are related as usual to the periods by

$$\lambda_e=\frac{c}{\nu_e}=c\Delta t_e,\qquad \lambda_\circ=c\Delta t_\circ.$$

And so the redshift is

$$z=\frac{\lambda_\circ-\lambda_e}{\lambda_e}=\frac{\lambda_\circ}{\lambda_e}-1=\frac{\Delta t_\circ}{\Delta t_e}-1=\frac{a(t_\circ)}{a(t_e)}-1. \tag{7.20}$$

In an expanding universe, $a(t_\circ)>a(t_e)$, i.e., the light takes time to reach us. Therefore the redshift is positive—the light *is really* redshifted. Thus,

$$\Longrightarrow 1+z=\frac{a(t_\circ)}{a(t_e)}.$$

If the redshift is not large, so that the times of emission and reception do not differ by a large amount, let $t_\circ=t_e+dt$. Then

$$\begin{aligned}(1+z)=\frac{a(t_\circ)}{a(t_e)}=\frac{a(t_\circ)}{a(t_\circ-dt)} &\equiv \frac{a(t_\circ)}{a(t_\circ)-\dot a(t_\circ)\,dt}\\ &\simeq 1+\frac{\dot a(t_\circ)}{a(t_\circ)}dt\end{aligned} \tag{7.21}$$

to lowest order in dt. We also have (cf. Eq. 7.16)

$$\int_{t_e}^{t_\circ}\frac{dt}{a(t)}=\int_{t_e}^{t_e+dt}\frac{dt}{a(t)}\simeq\frac{dt}{a(t_e)}=\frac{dt}{a(t_\circ-dt)}\simeq\frac{dt}{a(t_\circ)}. \tag{7.22}$$

Again to lowest order in dt, from Eq. 7.16,

$$\int_{t_e}^{t_\circ}\frac{dt}{a(t)}=\frac{f(R_e)}{c} \tag{7.23}$$

where $f(R)$ is given by Eq. 7.17. For small R, $f(R_e)\approx R_e$ for all three values of k, and so

$$\frac{dt}{a(t_\circ)}\approx\frac{R_e}{c}. \tag{7.24}$$

Combining this with Eq. 7.21,

$$\begin{aligned} 1+z &\approx 1+\dot{a}(t_\circ)\cdot\frac{R_e}{c} \\ z &\approx \frac{\dot{a}(t_\circ)R_e}{c}. \end{aligned} \tag{7.25}$$

Thus for small redshifts, the redshift is directly proportional to distance. Note also that this is independent of $|K|$ (cf. Eqs. 7.7 and 7.9).

7.5 Horizons

When discussing black holes in the Schwarzschild metric, we encountered the concept of an *event horizon* (which in this case was just the Schwarzschild radius): no information (e.g., photons) could ever reach us from inside the event horizon.

In cosmology, there are two important horizons: the *particle horizon*, and the *event horizon*. The particle horizon (also called the object horizon) arises in answering the following question: what is the comoving coordinate R of the most distant object we can see now? This will obviously be the objects which emitted their light at the beginning of the universe, t_B. In some model universes, $t_B = 0$, in others $t_B = -\infty$. (We'll discuss this later, when we consider the dynamical evolution of the universe explicitly.) In either case, we have from the Robertson-Walker metric (cf. Eq. 7.16 again)

$$\int_{t_B}^{t_\circ}\frac{dt}{a(t)} = \frac{1}{c}\int_0^{R_{\rm ph}}\frac{dR}{(1-kR^2)^{1/2}} = \frac{f(R_{\rm ph})}{c} \tag{7.26}$$

$$\Longrightarrow R_{\rm ph} = \begin{cases} \sin\left(c\displaystyle\int_{t_B}^{t_\circ}\frac{dt}{a(t)}\right) & k=+1 \\ c\displaystyle\int_{t_B}^{t_\circ}\frac{dt}{a(t)} & k=0 \\ \sinh\left(c\displaystyle\int_{t_B}^{t_\circ}\frac{dt}{a(t)}\right) & k=-1 \end{cases}. \tag{7.27}$$

To actually evaluate $R_{\rm ph}$ requires a specific form for $a(t)$. No object with $R > R_{\rm ph}$ can be seen by us. However, whatever the value of k, there may not be a particle horizon, depending on the form of $a(t)$; in some model universes, the entire universe is visible to us (in principle!).

The event horizon arises in asking a slighlty different question: what is the coordinate $R_{\rm eh}$ of the most distant event occurring *now* (at time $t_\circ$) which we will *ever* be able to see? This coordinate is the event horizon. The light from an event at the event horizon must reach us by the time the universe ends at time t_E; t_E is often infinite, but as we will see later there are universes which stop expanding at a finite time and recollapse.

Analogous to Eqs. 7.26 and 7.27, we get that the event horizon is given by

$$R_{\text{eh}} = \begin{cases} \sin\left(c\int_{t_\circ}^{t_E}\frac{dt}{a(t)}\right) & k=+1 \\ c\int_{t_\circ}^{t_E}\frac{dt}{a(t)} & k=0 \\ \sinh\left(c\int_{t_\circ}^{t_E}\frac{dt}{a(t)}\right) & k=-1 \end{cases}. \tag{7.28}$$

Depending on the model universe, R_{eh} may be infinite, in which case we eventually get to see all events, or it may be finite; in the latter case, we may *still* get to see all events if the universe is bounded ($k=+1$) and R_{eh} is greater than the maximum radius of the universe.

The particle or object horizon determines the greatest possible distance at which an object can have had any effect on our locality.
The event horizon is the greatest distance at which an object will *eventually* be able to affect our locality.

7.6 Luminosity Distance

We observe the rest of the universe by receiving photons emitted by astrophysical objects. In a static, Euclidean universe, flux is related to luminosity by

$$f = \frac{L}{4\pi D^2}.$$

This will obviously not be true in an expanding, curved space. Define the *luminosity distance* by

$$f = \frac{L}{4\pi D_L^2} \Longrightarrow D_L = \left(\frac{L}{4\pi f}\right)^{1/2} \tag{7.29}$$

where L is the true, intrinsic luminosity of the source. Using the Robertson-Walker metric, we can calculate D_L exactly.

Let the coordinates of the emission event be R_e, t_e. The light reaches us at $t = t_\circ$; at this time, the area of the hyper-spherical wavefront is

$$A = 4\pi R_e^2\, a^2(t_\circ). \tag{7.30}$$

The flux of energy crossing this surface is reduced by two additional factors:

1. **The redshift of the radiation:**

$$\lambda_\circ = (1+z)\lambda_e \quad \rightarrow \quad \nu_\circ = \frac{\nu_e}{1+z}$$
$$E = h\nu \quad \rightarrow \quad E_\circ = \frac{E_e}{1+z}.$$

2. **Time dilation:** Recall Eq. 7.19, $\Delta t_\circ/a(t_\circ) = \Delta t_e/a(t_e)$, which implies

$$\begin{aligned} \Longrightarrow \frac{\Delta t_\circ}{\Delta t_e} &= \frac{a(t_\circ)}{a(t_e)} \\ &= 1+z. \end{aligned}$$

In addition to having lower energy, the *rate of arrival* of photons is therefore reduced by a factor of $1+z$, relative to the rate at which they are emitted at $R = R_e$.

Thus the energy flux arriving at the spherical hyper-surface which includes us is

$$\begin{aligned} f &= \frac{L}{4\pi R_e^2\, a^2(t_\circ)\,(1+z)^2} \\ &= \frac{L\, a^2(t_e)}{4\pi R_e^2\, a^4(t_\circ)}. \end{aligned} \tag{7.31}$$

Thus the luminosity distance is

$$D_L = \frac{R_e\, a^2(t_\circ)}{a(t_e)} = R_e\, a(t_\circ)(1+z). \tag{7.32}$$

Note that this is *not* the same as the proper distance D_p (Eq. 7.12); these two distances are equivalent in general for small R_e (small z).

7.7 Deceleration Parameter $q_\circ$

Most observed objects (unsurprisingly) have small redshifts ($z \ll 1$) and therefore $t_\circ - t_e$ is also small (i.e., $(t_\circ - t_e)/t_\circ \ll 1$). It is therefore useful to expand $a(t)$ in a power series about $t = t_\circ$:

$$\begin{aligned} a(t) &= a(t_\circ) + (t-t_\circ)\,\dot a(t_\circ) + \frac{1}{2}(t-t_\circ)^2\,\ddot a(t_\circ) + \cdots \\ &= a(t_\circ)\left[1 + (t-t_\circ)\frac{\dot a(t_\circ)}{a(t_\circ)} + \frac{1}{2}(t-t_\circ)^2\frac{\ddot a(t_\circ)}{a(t_\circ)} + \cdots\right] \\ &= a(t_\circ)\left[1 + (t-t_\circ)\frac{\dot a(t_\circ)}{a(t_\circ)} + \frac{1}{2}(t-t_\circ)^2\frac{\ddot a(t_\circ)\,a(t_\circ)}{\dot a^2(t_\circ)}\left(\frac{\dot a(t_\circ)}{a(t_\circ)}\right)^2 + \cdots\right] \\ &= a(t_\circ)\left[1 + (t-t_\circ)H_\circ + \frac{1}{2}(t-t_\circ)^2\frac{\ddot a(t_\circ)\,a(t_\circ)}{\dot a^2(t_\circ)}H_\circ^2 + \cdots\right] \\ &= a(t_\circ)\left[1 + H_\circ(t-t_\circ) - \frac{1}{2}q_\circ H_\circ^2(t-t_\circ)^2 + \cdots\right] \end{aligned} \tag{7.33}$$

where the *deceleration parameter* is

$$q_\circ \equiv -\frac{\ddot a(t_\circ)\,a(t_\circ)}{\dot a^2(t_\circ)} = -\frac{\ddot a(t_\circ)}{a(t_\circ)H_\circ^2}. \tag{7.34}$$

Note that $q_\circ$ is positive if $\ddot{a}(t_\circ)$ is negative, i.e., if the expansion is slowing down.

Using Eq. 7.33 in Eq. 7.20 for z yields a power series for z in terms of $(t_\circ - t_e)$:

$$z = H_\circ(t_\circ - t_e) + \left(1 + \frac{1}{2}q_\circ\right) H_\circ^2(t_\circ - t_e)^2 + \cdots \tag{7.35}$$

Inverting this power series gives a series for $t_\circ - t_e$ in terms of z:

$$(t_\circ - t_e) = \frac{1}{H_\circ}\left[z - (1 + \frac{1}{2}q_\circ)z^2 + \cdots\right] \tag{7.36}$$

Now, from Eq. 7.25, we have for small z

$$\begin{aligned} R_e = \frac{cz}{\dot{a}(t_\circ)} \Longrightarrow R_e\, a(t_\circ) &= cz\frac{a(t_\circ)}{\dot{a}(t_\circ)} \\ &= \frac{cz}{H_\circ}. \end{aligned}$$

Substituting this expression into Eq. 7.31

$$\begin{aligned} f &= \frac{L}{4\pi}\frac{a^2(t_e)}{a^2(t_\circ)}[R_e\, a(t_\circ)]^{-2} \\ &= \frac{LH_\circ^2}{4\pi c^2 z^2}\left[\frac{a(t_e)}{a(t_\circ)}\right]^2. \end{aligned}$$

With $a(t_e)/a(t_\circ)$ from Eq. 7.33, and using Eqs. 7.36 for $(t_\circ - t_e)$, we finally get

$$f = \frac{LH_\circ^2}{4\pi c^2 z^2}[1 + (q_\circ - 1)z + \cdots] \tag{7.37}$$

This finally gives us a direct relation between f and z, provided we have a population of standard candles for which we know L.

For galaxies close enough for the $(q_\circ - 1)z$ term to be negligible, the slope of f vs. z^{-2} gives us $H_\circ$. If we can then extend our measurements out to larger z, so that the $(q_\circ - 1)z$ term starts to contribute significantly, then we can determine $q_\circ$ from the deviation of f vs. z^{-2} from a straight line.

As we will see later, $q_\circ$ is directly related to the dynamical state of the universe; if we could determine $q_\circ$ precisely, we would know the ultimate fate of the universe: continued expansion, or halt and recollapse. This is *the* "classical" cosmological test; astronomers (notably Allan Sandage) have been struggling with this problem for decades. It has been plagued by both *selection effects* and *evolutionary effects*.

7.8 Cosmic Dynamics

So far, the form of $a(t)$ has been left indeterminate. We now want to relate $a(t)$ to the mass-energy content of the universe. Although we need General Relativity to do this, we can get a surprisingly long way using Newtonian mechanics. However, we will need two special results from General Relativity:

1. In a spherically symmetric system, the gravitational force (i.e., the acceleration) at a given radius is determined by the mass within that radius. This is known as *Birkhoff's Theorem* (first mentioned back on p. 62), and is the General Relativity generalization of Newton's first theorem.
2. The active gravitational mass density is equal to the sum of the matter density plus $1/c^2$ times the energy density of radiation and relativistic particles (e.g., neutrinos).

Our Newtonian approximation will be valid provided that all velocities are $\ll c$ and the potential $\phi \ll c^2$. For a region of size L, this latter condition amounts to

$$\frac{GM(<L)}{L} \sim \frac{G\rho L^3}{L} \ll c^2. \tag{7.38}$$

As we will see shortly, the value of the Hubble constant $H \sim (G\rho)^{1/2}$, and so

$$\begin{aligned} H^2L^2 &\ll c^2 \\ L &\ll \frac{c}{H}. \end{aligned} \tag{7.39}$$

Since the relative velocity due to the Hubble expansion $v = HL$, Eq. 7.39 also implies that $v \ll c$. The length c/H (the *horizon distance*) is the distance a photon can have travelled in the age of the universe.

With the aid of Birkhoff's Theorem, we can ignore the effect of any matter or energy outside our sphere of radius L. Including the effect of the expansion, write $L = L_\circ\, a(t)$. The acceleration of the surface of the sphere is then given by

$$\frac{d^2L}{dt^2} = -\frac{GM}{L^2}, \tag{7.40}$$

where M is the mass (including relativistic mass) within the sphere. There are *no* pressure forces, since $\nabla P = 0$ by the assumption of homogeneity. At present, the energy density in radiation and relativistic particles is much less than the rest mass energy in matter. However we will see shortly that this was not always true, so we cannot ignore the radiation mass-energy.

The pressure due to both matter and radiation is now unimportant. At early times, when pressure *was* important, the pressure due to radiation greatly exceeded that due to matter. Therefore ignore the matter pressure at all times[3] and use the radiation equation of state:

$$P = \frac{1}{3}\rho_r c^2 \tag{7.41}$$

where the radiation mass energy $\rho_r = u_r/c^2$, where u_r is the radiation energy density (in erg cm^{-3}). Then the gravitational mass density is

$$\rho = \rho_m + \frac{3P}{c^2}, \tag{7.42}$$

[3]In General Relativity, pressure-less matter is technically known as "dust."

and the mass within radius L is just $M = \rho V = \frac{4\pi}{3} L^3 \rho$. Since $L_\circ$ is a constant, Eq. 7.40 can then be written as an equation for the scale factor $a(t)$:

$$\ddot{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) a \tag{7.43}$$

where a, ρ, and P are all functions of time.

In order to integrate this equation, we need to know how ρ and P vary with the scale factor $a(t)$. This is given by the first law of thermodynamics:

$$\frac{d}{dt}(\rho c^2 V) = -P\frac{dV}{dt}. \tag{7.44}$$

This just says that the change in the internal energy $\rho c^2 V$ (noting that $\rho = \rho_m + \rho_r$ includes the rest-mass energy) is equal to minus the work done in the expansion (i.e., $dU + PdV = dS = 0$ since there can be no heat exchange in a homogeneous universe). Using Eq. 7.42 in 7.44, this becomes

$$\begin{aligned} \frac{d}{dt}(\rho c^2 V) + P\frac{dV}{dt} &= 0 \\ \frac{d}{dt}(\rho_m c^2 V + 3PV) + P\frac{dV}{dt} &= 0 \\ &= \frac{d\rho_m}{dt} c^2 V + \rho_m c^2 \frac{dV}{dt} + 3P\frac{dV}{dt} \\ &\quad + 3V\frac{dP}{dt} + P\frac{dV}{dt}. \end{aligned} \tag{7.45}$$

Except for *very* early times (the first few seconds after the Big Bang), the coupling between matter and radiation was very small, in the sense that the energy transfer between matter and radiation $\Delta E \ll \rho_m c^2, \rho_r c^2$. Eq. 7.45 then can be separated into two equations:

$$\frac{d\rho_m}{dt} c^2 V + \rho_m c^2 \frac{dV}{dt} = 0 \tag{7.46}$$

$$4P\frac{dV}{dt} + 3V\frac{dP}{dt} = 0. \tag{7.47}$$

Eq. 7.46 is just:

$$\begin{aligned} \frac{d\rho_m}{\rho_m} &= -\frac{dV}{V} \\ \Longrightarrow \rho_m &\propto V^{-1} \\ \Longrightarrow \rho_m V &= \text{constant}, \end{aligned} \tag{7.48}$$

while Eq. 7.47 is

$$\begin{aligned} \frac{dP}{P} &= -\frac{4}{3}\frac{dV}{V} \\ P &\propto V^{-4/3} \\ \Longrightarrow PV^{4/3} &= \text{constant} \\ \Longrightarrow \rho_r V^{4/3} &= \text{constant}. \end{aligned} \tag{7.49}$$

Since V/a^3 is equal to a constant,

$$\rho_m = \rho_{m\circ}\frac{a_\circ^3}{a^3(t)} \tag{7.50}$$

$$\rho_r = \rho_{r\circ}\frac{a_\circ^4}{a^4(t)}, \tag{7.51}$$

where the zero subscripts denote present-day values. Note that the radiation mass-energy density increases faster (by one power of $a(t)$) than the matter density as we go back in time; hence at some time in the past, the universe will become radiation-dominated, rather than its present matter-dominated state.

Physically this is because although the numbers of particles and photons are both conserved as we run the universe backward, the radiation is also being blueshifted (just the reverse of the expansion-induced redshift). Thus the radiation energy density increases by an additional factor of $a(t_\circ)/a(t)$.

7.9 Friedmann Equations

We can write Eq. 7.44 in the form

$$\begin{aligned} d(\rho c^2 V) + P dV &= 0 \\ d\rho + \rho\frac{dV}{V} + \frac{P}{c^2}\frac{dV}{V} &= 0 \\ d\rho + \left(\rho + \frac{P}{c^2}\right)\frac{dV}{V} &= 0. \end{aligned}$$

Using $V \propto a^3(t)$ and $dV/V = 3da/a$, along with a little algebra, we can write Eq. 7.44 as

$$\begin{aligned} 3\left(\rho + \frac{P}{c^2}\right) &= -\frac{d\rho}{da}\cdot a \\ \frac{P}{c^2} &= -\frac{a}{3}\frac{d\rho}{da} - \rho. \end{aligned} \tag{7.52}$$

We can use this to eliminate P/c^2 from Eq. 7.43:

$$\begin{aligned} \ddot{a} &= -\frac{4\pi G a}{3}\left(\rho + 3\left[-\frac{a}{3}\frac{d\rho}{da} - \rho\right]\right) \\ &= \frac{4\pi G a}{3}\left(2\rho + a\frac{d\rho}{da}\right) \\ &= \frac{4\pi G}{3}\left(2\rho a + a^2\frac{d\rho}{da}\right). \end{aligned} \tag{7.53}$$

Note that

$$\begin{aligned} \frac{d}{dt}(a^2\rho) &= 2a\cdot\dot{a}\rho + a^2\frac{d\rho}{dt} \\ &= 2a\dot{a}\rho + a^2\frac{d\rho}{da}\frac{da}{dt} \\ &= 2a\dot{a}\rho + a^2\dot{a}\frac{d\rho}{da}. \end{aligned}$$

Thus the right-hand side is just

$$\frac{1}{\dot{a}}\frac{d}{dt}(a^2\rho).$$

If we multiply both sides by $\dot{a}$, then the left-hand side is

$$\ddot{a}\dot{a} = \frac{1}{2}\frac{d}{dt}(\dot{a})^2.$$

So Eq. 7.53 can be integrated once:

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}(\dot{a}^2) &= \frac{4\pi G}{3}\frac{d}{dt}(a^2\rho) \\
\frac{\dot{a}^2}{2} &= \frac{4\pi G}{3}\rho a^2 + E \\
\dot{a}^2 &= \frac{8\pi G}{3}\rho a^2 + 2E,
\end{aligned} \tag{7.54}$$

where E is a constant which is related to the energy.

Using the actual field equations of General Relativity, we obtain Eq. 7.54 again, with the constant $2E$ now identified as $-k\,c^2$. Thus

$$\dot{a}^2 = \frac{8\pi G}{3}\rho a^2 - k\,c^2. \tag{7.55}$$

We will include one further complication. As we will see shortly, Eq. 7.55 implies that the universe is *dynamic*, i.e., evolving. In his first model for the universe, Einstein sought a static solution. In order to produce this, he was forced to introduce an additional term into the field equations (and therefore in Eq. 7.55). This is known as the *cosmological constant*.

The reason for this is apparent from Eq. 7.43: in order for $\ddot{a}$ to be zero, $\rho + 3P/c^2 = 0$. In other words, if the density $\rho > 0$, the pressure associated with this matter must be negative, which is impossible for any matter or radiation. Einstein therefore introduced the cosmological constant Λ, which acts like a repulsion term:

$$\ddot{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right)a + \frac{\Lambda}{3}a, \tag{7.56}$$

where the factor of $1/3$ is for convenience. With the inclusion of this term, Eq. 7.55 becomes

$$\dot{a}^2 = \frac{8\pi G}{3}\rho a^2 + \frac{\Lambda}{3}a^2 - k\,c^2. \tag{7.57}$$

this is known as *Friedmann's equation*—or, for the truly GR-inclined, the initial-value equation.

7.9.1 Critical Density

Using the Friedmann equation, we can now explicitly solve for $a(t)$, and therefore determine the histories of model universes, one of which (hopefully) is a good approximation to our own.

For simplicity, let us first set the cosmological constant $\Lambda = 0$, so that Friedmann's equation is

$$\dot{a}^2 + k\,c^2 = \frac{8\pi G}{3}\rho a^2. \tag{7.58}$$

Note that the pressure is not explicitly present in this equation. For the moment, let us also assume that the pressure is also negligible (i.e., ignore the contribution of radiation and relativistic particles to the right-hand side). In this case the density as a function of a is given by Eq. 7.50:

$$\rho = \rho_\circ \frac{a_\circ^3}{a^3} \tag{7.50}$$

(where we have dropped the m subscripts on ρ). Substituting into Eq. 7.58:

$$\dot{a}^2 + k\,c^2 = \frac{8\pi G}{3}\frac{\rho_\circ a_\circ^3}{a}, \tag{7.59}$$

where $a_\circ = a(t_\circ)$. Now, Eq. 7.58 holds at all times, including the present where $t = t_\circ$. We can then write this as an equation for $k\,c^2$:

$$\begin{aligned}\frac{k\,c^2}{a_\circ^2} &= \frac{8\pi G}{3}\rho_\circ - \left(\frac{\dot{a}_\circ}{a_\circ}\right)^2 \\ &= \frac{8\pi G}{3}\rho_\circ - H_\circ^2 \\ &= \frac{8\pi G}{3}\left(\rho_\circ - \frac{3H_\circ^2}{8\pi G}\right).\end{aligned} \tag{7.60}$$

Hence whether k is > 0, equal to 0, or < 0 depends on whether $\rho_\circ > \rho_c$, $\rho_\circ = \rho_c$, or $\rho_\circ < \rho_c$, respectively, where the *critical density* is

$$\rho_c \equiv \frac{3H_\circ^2}{8\pi G}. \tag{7.61}$$

We can also write $q_\circ$ for the zero-pressure case in terms of ρ_c. Recall that

$$q_\circ \equiv -\frac{\ddot{a}_\circ}{a_\circ H_\circ^2}. \tag{7.34}$$

From Eq. 7.43, with $P = 0$,

$$\ddot{a}_\circ = -\frac{4\pi G}{3}\rho_\circ a_\circ. \tag{7.62}$$

Thus

$$\begin{aligned}q_\circ &= \frac{4\pi G}{3}\rho_\circ H_\circ^{-2} \\ &= \frac{1}{2}\frac{\rho_\circ}{\rho_c}.\end{aligned} \tag{7.63}$$

7.9.2 Ω and Flat, Closed, and Open Universes

It is customary to define

$$\Omega \equiv \frac{\rho}{\rho_c}. \tag{7.64}$$

Thus $\Omega = 1$ corresponds to the critical case of a flat, $k = 0$ universe.

We can now determine the evolution of the universe for the three possible zero pressure, zero-Λ models (usually referred to as the *Friedmann models*).

$k = 0$ This is the flat, $\Omega = 1$ universe ($\Omega_c = 1$). For convenience, define $A^2 \equiv (8\pi G/3)\rho_\circ a_\circ^3$, so that the Friedmann equation is

$$\dot{a}^2 + k\,c^2 = \frac{A^2}{a} \tag{7.65}$$

so that in this case,

$$\frac{da}{dt} = \frac{A}{a^{1/2}}. \tag{7.66}$$

This can be trivially integrated:

$$\begin{aligned} a^{1/2}da &= A\,dt \\ a^{3/2} &= \frac{3}{2}At + \not{c} \to 0 \qquad \text{for } a = 0 \text{ at } t = 0 \\ a &= \left(\frac{3A}{2}\right)^{2/3} t^{2/3}. \end{aligned} \tag{7.67}$$

This is known as the Einstein-de Sitter model. Note that if we differentiate Eq. 7.67

$$\dot{a} = \frac{2}{3}\left(\frac{3A}{2}\right)^{2/3} t^{-1/3}, \tag{7.68}$$

which approaches zero as $t \to \infty$. Thus the $\Omega = 1$, $k = 0$ universe slows to a halt, but only as t approaches infinity. Obviously, models with $\Omega > 1$ will stop expanding (and re-collapse) at some finite time, while those with $\Omega < 1$ will always expand (for $\Lambda = 0$).

$k = 1$ This is a closed, $\Omega > 1$ universe. In this case, the Friedmann equation becomes

$$\begin{aligned} \dot{a}^2 + c^2 &= \frac{A}{a} \\ \frac{da}{dt} &= \left(\frac{A^2 - c^2 a}{a}\right)^{1/2}. \end{aligned} \tag{7.69}$$

This can be written as an equation for t:

$$t = \int_0^a \left(\frac{a}{A^2 - c^2 a}\right)^{1/2} da. \tag{7.70}$$

To get this into a more convenient form, define the angle ψ by

$$a \equiv \frac{A^2}{c^2} \sin^2\left(\frac{\psi}{2}\right). \tag{7.71}$$

(We'll see the reason for the factor of $1/c^2$ momentarily.) Then

$$\begin{aligned} da &= \frac{1}{c^2} A^2 \cdot 2 \sin\left(\frac{\psi}{2}\right) \cos\left(\frac{\psi}{2}\right) \cdot \frac{1}{2} d\psi \\ &= \frac{A^2}{c^2} \sin\left(\frac{\psi}{2}\right) \cos\left(\frac{\psi}{2}\right) d\psi \\ \frac{a}{A^2 - c^2 a} &= \frac{c^{-2} A^2 \sin^2\left(\frac{\psi}{2}\right)}{A^2 - A^2 \sin^2\left(\frac{\psi}{2}\right)} \\ &= \frac{1}{c^2} \frac{\sin^2\left(\frac{\psi}{2}\right)}{1 - \sin^2\left(\frac{\psi}{2}\right)} \\ &= \frac{1}{c^2} \frac{\sin^2\left(\frac{\psi}{2}\right)}{\cos^2\left(\frac{\psi}{2}\right)}. \end{aligned}$$

And so Eq. 7.70 becomes

$$\begin{aligned} t &= \frac{A^2}{c^2} \int_0^\psi \sin^2\left(\frac{\psi}{2}\right) d\psi \\ &= \frac{1}{2} \frac{A^2}{c^2} \int_0^\psi (1 - \cos\psi)\, d\psi. \end{aligned}$$

We use the half-angle formula from trigonometry:

$$t = \frac{1}{2} \frac{A^2}{c^2} (\psi - \sin\psi), \tag{7.72}$$

while a is given by (cf. Eq. 7.71)

$$a = \frac{A^2}{c^2} \sin^2\left(\frac{\psi}{2}\right) = \frac{1}{2} \frac{A^2}{c^2} (1 - \cos\psi). \tag{7.73}$$

This gives $a(t)$ parametrically, in terms of $a(\psi)$, while $t(\psi)$ is given by Eq. 7.72. Note that a is a *periodic* function of time; it will be zero whenever ψ is an integer multiple of 2π.

$k = -1$ This is an open, $\Omega < 1$ universe. In this case, the Friedmann equation becomes

$$\frac{da}{dt} = \left(\frac{A^2 + c^2 a}{a}\right)^{1/2} \tag{7.74}$$

which gives the equation for t as

$$t = \int_0^a \left(\frac{a}{A^2 + c^2 a} \right)^{1/2} da. \tag{7.75}$$

In this case, define ψ by

$$a \equiv \frac{A^2}{c^2} \sinh^2 \left(\frac{\psi}{2} \right). \tag{7.76}$$

And so the equation for t (7.75) becomes

$$\begin{aligned} t &= \frac{A^2}{c^3} \int_0^\psi \sinh^2 \left(\frac{\psi}{2} \right) d\psi \\ &= \frac{1}{2} \frac{A^2}{c^3} \int_0^\psi (\cosh \psi - 1)\, d\psi \\ &= \frac{1}{2} \frac{A^2}{c^3} (\sinh \psi - \psi), \end{aligned} \tag{7.77}$$

using the half-angle formulae for hyperbolic sine and cosine. This again gives an expression for $t(\psi)$, while $a(\psi)$ is

$$a = \frac{1}{2} \frac{A^2}{c^2} (\cosh \psi - 1). \tag{7.78}$$

From Eqs. 7.78 and 7.77, we can obtain an expression for $\dot{a}$:

$$\begin{aligned} \frac{da}{dt} = \frac{da}{d\psi} \frac{d\psi}{dt} &= \frac{1}{2} \frac{A^2}{c^2} \sinh \psi \cdot \left[\frac{1}{2} \frac{A^2}{c^3} (\cosh \psi - 1) \right]^{-1} \\ \frac{1}{c} \frac{da}{dt} &= \frac{\sinh \psi}{\cosh \psi - 1} \\ \Longrightarrow \frac{\sinh \psi}{\cosh \psi} &= \tanh \psi = 1 \qquad \text{as } \psi \to \infty. \end{aligned} \tag{7.79}$$

Thus $\dot{a} \to c$ as $t \to \infty$ for $k = -1$ models. (This seems odd, but remember that we folded a factor of $|K|^{1/2}$ into the scale factor for $K \neq 0$.)

In fact, from Eq. 7.74, we see that, since $a \to \infty$ as $t \to \infty$ (from Eq. 7.78), as $t \to \infty$,

$$\frac{da}{dt} \to c.$$

The physical origin of different behavior for the $k = 0, \pm 1$, $P = 0$, and $\Lambda = 0$ Friedmann models is easiest to understand by returning to our Newtonian derivation, which gave

$$\dot{a}^2 = \frac{8\pi G}{3} \rho a^2 + 2E \tag{7.54}$$

where $2E$ was a constant, which the field equation derivation of General Relativity identifies as $-k\, c^2$.

We derived Eq. 7.54 by integrating Eq. 7.53 for $\ddot{a}$. Eq. 7.53, in classical terms, is an *equation of motion* (in this case, for the entire universe!). By integrating this equation, we obtain 7.54, with $2E$ as a constant of integration. Thus $2E$ is an *integral of motion*, in this case, an energy equation.

With our definition of A^2, the equation for $\dot{a}$ can be written (cf. Eq. 7.65)

$$\dot{a}^2 - \frac{A^2}{a} = -k\,c^2. \tag{7.80}$$

Pursuing our Newtonian analogy, we can identify $\dot{a}^2$ as the kinetic energy, $-A^2/a$ as the potential energy, and $-k\,c^2$ as the total energy.

If $k = +1$, then the total energy of the universe is negative, i.e., the total kinetic energy is less than the gravitational potential energy. In this case the expansion must eventually come to a halt, when

$$\begin{aligned} \frac{A^2}{a} &= c^2 \\ \Longrightarrow a &= \frac{A^2}{c^2}, \end{aligned} \tag{7.81}$$

and the universe contracts with increasing time.

If $k = 0$, then the total energy of the universe equals zero, and the kinetic energy is just large enough to allow the universe to keep expanding, but at an ever-decreasing rate ($\dot{a} \to 0$ as $t \to \infty$).

If $k = -1$, then the total energy is positive, and the universe has enough kinetic energy to allow it to expand forever, at an eventually constant rate ($\dot{a} \to c$ as $t \to \infty$). This latter behavior arises because the potential energy term eventually becomes negligible as $a \to \infty$, and so the expansion of the universe "coasts" at constant velocity.

7.9.3 Models with Non-zero Λ

In general, the scale factor as a function of time in these models requires the use of elliptic functions, whose physical meaning is rather opaque. For the special case of a flat ($k = 0$) universe, however, it is possible to get closed-form expressions for $a(t)$.

$k = 0$: For these flat-space models, Friedmann's equation simplifies to

$$\dot{a}^2 = \frac{A^2}{a} + \frac{1}{3}\Lambda a^2. \tag{7.82}$$

We have two cases to consider: $\Lambda > 0$ and $\Lambda < 0$.

$\Lambda > 0$: In the first case, introduce a new variable,

$$u \equiv \frac{2\Lambda}{3A^2}a^3. \tag{7.83}$$

Differentiating this equation gives

$$\dot{u} = \frac{2\Lambda}{A^2}a^2\dot{a}. \tag{7.84}$$

Using Eqs. 7.83 and 7.84 in Eq. 7.82 results in:

$$\begin{aligned}
\frac{A^4}{4\Lambda^2}\dot{u}^2 a^{-4} &= \frac{A^2}{a} + \frac{1}{3}\Lambda a^2 \\
\dot{u}^2 &= \frac{4\Lambda^2}{A^4}\left[A^2 a^3 + \frac{1}{3}\Lambda a^6\right] \\
&= \frac{4\Lambda^2}{A^2}a^3 + \frac{4}{3}\frac{\Lambda^3}{A^4}a^6 \\
&= 6\Lambda u + 3\Lambda u^2 \\
&= 3\Lambda(2u + u^2). \qquad (7.85)
\end{aligned}$$

Therefore,

$$\dot{u} = (3\Lambda)^{1/2}(2u + u^2)^{1/2}, \qquad (7.86)$$

where we have taken the positive square root since both terms on the right-hand side of Eq. 7.82 are positive, and therefore $\dot{a}$ (and hence $\dot{u}$) must be > 0. This also implies that $a = 0$ when $t = 0$; we'll discuss this further shortly.

With this assumption (i.e., a Big Bang cosmology), we can solve Eq. 7.86 as:

$$\int_0^u \frac{du}{(2u + u^2)^{1/2}} = \int_0^t (3\Lambda)^{1/2}\, dt = (3\Lambda)^{1/2} t. \qquad (7.87)$$

To perform the integral over u, first complete the square:

$$\int_0^u = \frac{du}{[(u^2 + 1)^2 - 1]^{1/2}} = \int_1^v \frac{dv}{(v^2 - 1)^{1/2}},$$

where $v \equiv u + 1$. The integral is just $\cosh^{-1} v|_1^v = \cosh^{-1} v$ since $\cosh^{-1} 1 = 0$:

$$\begin{aligned}
\Longrightarrow \cosh^{-1} v &= (3\Lambda)^{1/2} t \\
v &= \cosh\left(\sqrt{3\Lambda}t\right) \\
u &= v - 1 = \cosh\left(\sqrt{3\Lambda}t\right) - 1. \qquad (7.88)
\end{aligned}$$

And so, finally,

$$\begin{aligned}
a^3 &= \frac{3A^2}{2\Lambda}\left[\cosh\left(\sqrt{3\Lambda}t\right) - 1\right] \\
a &= \left(\frac{3A^2}{2\Lambda}\right)^{1/3}\left[\cosh\left(\sqrt{3\Lambda}t\right) - 1\right]^{1/3}. \qquad (7.89)
\end{aligned}$$

This is, of course, an ever-expanding universe.

$\Lambda < 0$: In this case, introducing the new variable u by

$$u \equiv -\frac{2\Lambda}{3A^2}a^3, \tag{7.90}$$

and proceeding as before, we get the result

$$\begin{aligned} a^3 &= \frac{3A^2}{2(-\Lambda)}\left[1-\cos\left(\sqrt{-3\Lambda}t\right)\right] \\ a &= \left(\frac{3A^2}{2(-\Lambda)}\right)^{1/3}\left[1-\cos\left(\sqrt{-3\Lambda}t\right)\right]^{1/3}. \end{aligned} \tag{7.91}$$

This is clearly another *periodic* universe: a will equal zero whenever

$$\begin{aligned} \cos\left(\sqrt{-3\Lambda}t\right) &= 1, \\ \sqrt{-3\Lambda}t &= n\cdot 2\pi \Longrightarrow t = \frac{2\pi n}{\sqrt{-3\Lambda}}. \end{aligned}$$

7.9.4 Classification of Friedmann Universes

In general, solutions require elliptic functions, which are not very enlightening. However, it is possible to classify the dynamical behavior of Friedmann universes much more simply, without solving for the exact behavior of $a(t)$. The acceleration and velocity equations for a in the Friedmann cosmologies are:

$$\ddot{a} = -\frac{1}{2}\frac{A^2}{a^2} + \frac{\Lambda}{3}a \tag{7.92}$$

$$\dot{a}^2 = \frac{A^2}{a} - k\,c^2 + \frac{\Lambda}{3}a. \tag{7.93}$$

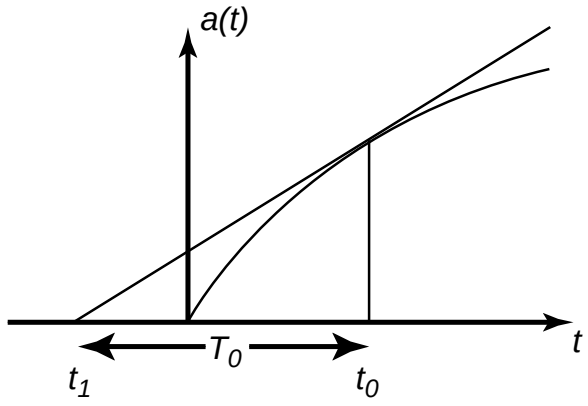


Figure 7.1: Expansion of the universe over time.

Consider first the past histories of Friedmann universes. If $\ddot{a} < 0$ for all t (or equivalently, for all a), then the universe *must* have begun in an infinitely dense *singularity*—the *Big Bang*—less than a Hubble time ($\equiv H_\circ^{-1}$) ago:

$$t_\circ - t_1 = \frac{a(t_\circ)}{\dot{a}(t_\circ)} = T_\circ.$$

The tangent to the $a(t)$ curve at the present, $t = t_\circ$, intercepts the t-axis (i.e., $a = 0$) at some time t_1. $T_\circ$ is thus the time since $a = 0$, if the universe had always been expanding at its current rate. If $\ddot{a} < 0$ always, that means that the rate of expansion $\dot{a}$ has been slowing down ever since the Big Bang; therefore

1. The universe must have begun in a Big Bang singularity.
2. The time since the Big Bang is less than the Hubble time $T_\circ$.

Clearly if $\Lambda \leq 0$, then $\ddot{a}$ is always < 0—the universe must begin in a Big Bang singularity.

What if $\Lambda > 0$? To investigate this, define the right-hand side of Eq. 7.93 to be a function, $f(a)$:

$$f(a) \equiv \frac{A^2}{a} - k\,c^2 + \frac{\Lambda}{3}a^2. \tag{7.94}$$

Clearly from Eq. 7.92, if $\Lambda > 0$ then $\ddot{a}$ is not necessarily always less than zero. Whenever $a(t)$ has a maximum or minimum, then $\dot{a}$ must be zero.

Note that we can write Eq. 7.93 as

$$\dot{a} = \pm\left(f(a)\right)^{1/2}. \tag{7.95}$$

Since $\dot{a}$ must be real, $f(a) \geq 0$ always.

Now, $\dot{a}$ has zeros whenever $f(a)$ has zeros. If $k = 0$ or $k = -1$, all the terms on the right-hand side of Eq. 7.94 are positive, and so $f(a)$ (and therefore $\dot{a}$) can never be zero.

This leaves the case $k = +1$. When will $f(a)$ have zeros in this case? $f(a)$ will have a maximum, minimum, or inflection point whenever $df/da = 0$. This is

$$\begin{aligned}
\frac{df}{da} = -\frac{A^2}{a^2} + \frac{2\Lambda}{3}a &= 0 \\
\frac{2\Lambda}{3}a^3 &= A^2 \\
a_c^3 &= \frac{3A^2}{2\Lambda} \\
a_c &= \left(\frac{3A^2}{2\Lambda}\right)^{1/3}.
\end{aligned} \tag{7.96}$$

So at this critical value of a_c (note there is only one), f has a maximum, minimum, or saddle point. If we take the second derivative,

$$\frac{d^2 f}{da^2} = \frac{2A^2}{a^3} + \frac{2\Lambda}{3},$$

we see that $f(a)$ has a minimum at a_c for $\Lambda \geq 0$, as $d^2f/da^2 > 0$. At this critical value of a, $f(a)$ has the value:

$$\begin{aligned} f(a) &= \frac{A^2}{A^{2/3}}\left(\frac{2\Lambda}{3}\right)^{1/3} - c^2 + \frac{\Lambda}{3}\left(\frac{3}{2\Lambda}\right)^{2/3} A^{4/3} \\ &= \left(\frac{\Lambda}{3}\right)^{1/3} A^{4/3}\left(2^{1/3} + 2^{-2/3}\right) - c^2 \\ &= 2^{-2/3}3\left(\frac{\Lambda}{3}\right)^{1/3} A^{4/3} - c^2 \\ &= \left(\frac{9}{4}\Lambda A^4\right)^{1/3} - c^2. \end{aligned} \tag{7.97}$$

This defines a critical value of Λ:

$$\Lambda_c = \frac{4c^6}{9A^4} \tag{7.98}$$

at which $f(a) = 0$ at $a = a_c$. Clearly, then, for $0 < \Lambda < \Lambda_c$, $f(a)$ goes through zero, so $\dot{a} \to 0$ and then changes sign, since for $0 < \Lambda < \Lambda_c$, $f(a)$ is < 0 at $a = a_c$.

Recalling that $A^2 = 8\pi G\rho_\circ a_\circ^3/3$, we can also write Λ_c as

$$\Lambda_c = \frac{c^6}{(4\pi G\rho_\circ r_\circ^3)^2}. \tag{7.99}$$

Now, we have established that for $0 < \Lambda < \Lambda_c$, $\dot{a}$ must go to zero and then change sign. (The sign change is mandated by the requirement that $f(a) \geq 0$.) Depending on initial conditions, there are then two possible model universes:

1. An oscillating universe, which begins in a singularity, expands to a finite radius, and then collapses back to a singularity.
2. A model which *contracts* initially to a finite minimum size $a_{\min}$, and then re-expands. $a(t)$ is symmetric in time about the minimum value. This is one of only two Friedmann models which do *not* have a Big Bang origin.

Similar considerations apply to the future evolution. If $\Lambda < 0$, then Eq. 7.92 shows that $\ddot{a} < 0$ always. Therefore at some point the expansion must halt and the universe recollapses—ending in the Big Crunch.

If $\Lambda = 0$, $\ddot{a} \to 0$ as $a \to \infty$; as we have already seen, for $k = 0$ or -1, the universe can expand forever, although for $k = +1$, it recollapses.

If $\Lambda > 0$, as we have just seen, the expansion can be halted if $\Lambda < \Lambda_c$; otherwise the expansion must continue forever.

What about $\Lambda = \Lambda_c$? This critical value of Λ results in the original *Einstein universe*, a static model in which $a = a_c$ always. It was, of course, precisely to obtain a static solution that Einstein introduced the cosmological constant in the first place. Both $\dot{a}$ and $\ddot{a}$ must be zero.

However there is a serious flaw with Einstein's static model. It is unstable. Write the equation of motion as:

$$\begin{aligned} \ddot{a} &= -\frac{4\pi G\rho}{3}a + \frac{\Lambda}{3}a \\ &= \frac{a}{3}(\Lambda - 4\pi G\rho) = \frac{a_c}{3}(\Lambda_c - 4\pi G\rho). \end{aligned} \tag{7.100}$$

Right off the bat, this indicates that there is something peculiar about this model: we have set a fundamental constant, Λ, equal to a density.

This equation also immediately indicates the nature of the instability. If we perturb the density to slightly higher or lower densities, then $\ddot{a}$ is non-zero, being negative or positive, respectively. Thus $\dot{a}$ will also become non-zero, and the universe must run away from the static solution, either collapsing or expanding.

This leads to two additional possible models with $\Lambda = \Lambda_c$: one begins as the Einstein static model at early times and eventually expands away from it; the other begins in a singularity and asymptotically approaches the Einstein static universe. The former is known as the Eddington-Lemaître universe. In principle, the universe could have spent an arbitrarily long time in a static state before the expansion began.

The final $k = +1$, $\Lambda > 0$ model is for $\Lambda > \Lambda_c$. This is known as Lemaître's model. In a sense, models with $\Lambda > \Lambda_c$ are a combination of the two non-static $\Lambda = \Lambda_c$ models: the universe expands from a singularity, and $a(t)$ exhibits a pronounced "kink" where the expansion rate slows, so the universe spends an extended period of time in a state in which a is nearly constant, until eventually the Λ-term becomes dominant and the expansion resumes at an accelerating rate. The closer Λ is to Λ_c, the longer the universe remains in a nearly static state.

Finally let us mention one other solution: the *de Sitter universe*, in which k, P, and $\rho = 0$, i.e., this is an empty universe. Since $\rho = 0$ means $A^2 = 0$, Eq. 7.93 simplifies to

$$\begin{aligned} \dot{a}^2 &= \frac{\Lambda}{3}a^2 \\ \Longrightarrow \dot{a} &= \left(\frac{\Lambda}{3}\right)^{1/2} a. \end{aligned} \tag{7.101}$$

This can be trivially integrated:

$$\begin{aligned} \frac{da}{a} &= \left(\frac{\Lambda}{3}\right)^{1/2} dt \\ a &= ce^{(\Lambda/3)^{1/2}t}, \end{aligned} \tag{7.102}$$

where the constant c determines the value of the scale factor at $t = 0$. Thus the universe expands exponentially with time.

This model is of historical interest because it was the first expanding universe solution. It is also the end state of all the continually-expanding $\Lambda > 0$ models:

$$\dot{a}^2 = \frac{A^2}{a} - k\,c^2 + \frac{\Lambda}{3}a^2 \to \frac{\Lambda}{3}a^2 \qquad \text{as } a \to \infty.$$

I.e., at large a, both the matter and curvature terms become negligible compared to Λ, and the universe is driven to exponential expansion.

As we have seen, nearly all of the Friedmann universes begin in a Big Bang singularity. At early times, all of these Big Bang models behave the same way. This is also obvious from the above equation: since $a \to 0$ at the beginning of a Big Bang universe, the matter term A^2/a must dominate at small a even if k and Λ are non-zero. Thus at early times, all Big Bang models behave like the Einstein-de Sitter $k = 0$, $\Lambda = 0$ universe:

$$\begin{aligned} \dot{a} &= \frac{A}{a^{1/2}} \\ \Longrightarrow \quad a &= \left(\frac{3A}{2}\right)^{2/3} t^{2/3} \end{aligned} \quad . \tag{7.67}$$

7.10 The Steady State Model

A radically different picture of the universe is the *Steady State* model, proposed by Bondi and Gold (1948) and Hoyle (1948). This is based on what is generally referred to as the *perfect cosmological principle*: the universe is not only homogeneous but unchanging.

This immediately implies that the universe must be expanding, on thermodynamic grounds: a static, infinitely old universe would reach thermodynamic equilibrium. In a contracting universe, blueshifting of radiation leads to a situation in which radiation dominates over matter, while in an expanding universe redshifting leads to the opposite. Since this is the observed condition of our universe, it must be expanding.

Since the universe is expanding and unchanging, matter must be continually created to keep the mean density constant. Furthermore, this must be a flat, $k = 0$ universe, since expansion of the universe causes the curvature K to decrease. This is most easily seen from the spatial part of the Robertson-Walker metric (Eq. 7.5):

$$dr^2 = [a(t)]^2 \left(\frac{dR^2}{1 - KR^2} + R^2\, d\theta^2 + R^2 \sin^2\theta\, d\phi^2 \right) . \tag{7.5}$$

If we introduce the new coordinate $\rho = a_\circ R$, this can be written as

$$dr^2 = \frac{a^2}{a_\circ^2} \left(\frac{d\rho^2}{1 - K'\rho^2} + \rho^2\, d\theta^2 + \rho^2 \sin^2\theta\, d\phi^2 \right) , \tag{7.103}$$

which aside from an overall scale factor, is unaltered from Eq. 7.5 *except* that the curvature is now $K' = K/a_\circ^2$.

The dynamics of the Steady State universe are very simple. Since the Hubble constant

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

must be the same at all times,

$$\begin{aligned} \frac{\dot{a}}{a} &= \frac{1}{T_\circ} && \text{where } T_\circ \text{ is a constant} = H_\circ \\ \Longrightarrow a &= ce^{t/T_\circ} && \text{where } c \text{ is a constant.} \end{aligned}$$

This is just the de Sitter solution, with an exponentially expanding universe, although this is *not* an empty universe.

The Steady State model also predicts that the deceleration parameter is equal to -1:

$$\begin{aligned} q_\circ &= -\frac{\ddot{a}}{a} H_\circ^{-2} \\ \dot{a} &= \frac{a}{T_\circ} = a H_\circ \\ \ddot{a} &= \dot{a} H_\circ \\ q_\circ &= -\frac{\dot{a}}{a} H_\circ \, H_\circ^{-2} = -H_\circ^2 \, H_\circ^{-2} = -1. \end{aligned}$$

As we will see in the next chapter, the discovery of the cosmic microwave background—generally interpreted as the signature of a hot Big Bang origin for the universe—put an end to interest in Steady State models, except for the truly committed.

7.11 The Effect of Radiation

So far, we have ignored the effect of pressure, i.e., radiation, in the evolution of the universe. Although the contribution of radiation to the dynamics of the universe is now unimportant, at early times this was not true. As we saw earlier, the mass-energy density of radiation scales like a^{-4} vs. a^{-3} for matter (Eqs. 7.50 and 7.51), and thus at sufficiently early epochs radiation dominates. We can quantify this simply: the densities of matter and radiation were equal at a time t_E given by

$$\begin{aligned} \rho_{m_\circ} \frac{a_\circ^3}{a(t_E)} &= \rho_{r_\circ} \frac{a_\circ^4}{a(t_E)} \\ \Longrightarrow \frac{a(t_E)}{a(t_\circ)} &= \frac{\rho_{r_\circ}}{\rho_{m_\circ}}. \end{aligned} \tag{7.104}$$

For times earlier than t_E, radiation dominated the dynamics; hence, $0 < t < t_E$ is known as the radiation-dominated era.

At early times, the radiation mass-energy density dominates over all other terms. If we define

$$B^2 = \frac{8\pi G \rho_{r_\circ} a_\circ^4}{3}, \tag{7.105}$$

then

$$\begin{aligned} \dot{a}^2 &= \frac{B^2}{a^2} \\ a \, da &= B \, dt \\ a &= \sqrt{2} B^{1/2} t^{1/2} \end{aligned} \tag{7.106}$$

(again assuming that $a \to 0$ as $t \to 0$ for a Big Bang universe). Note the difference from the $t^{2/3}$ scaling for the early stages of a radiation-less universe; the universe expands more slowly due to the active gravitational mass associated with the radiation, cf. Eq. 7.43:

$$\ddot{a} = -\frac{4\pi G}{3} \left(\rho + \frac{3P}{c^2} \right) a. \tag{7.43}$$

As we will see very shortly, the cosmic microwave background radiation is one of the key pieces of evidence for a Big Bang origin to the universe. To understand why, let us suppose that at some epoch the universe is filled with radiation, and that the radiation has a thermal (blackbody) distribution, with temperature $T_r = T_r(t)$; i.e.,

$$B_\nu = \frac{2h\nu^3}{c^2}\frac{1}{e^{h\nu/kT_r(t)} - 1} \tag{7.107}$$

is the frequency distribution of the radiation in units of ergs cm^{-2} s^{-1} Hz^{-1} sr^{-1}. The *number* of photons in a volume $v(t)$ with frequencies between ν and $\nu + d\nu$ is then given by

$$\begin{aligned} dN(t) &= \frac{4\pi}{c}\cdot\frac{1}{h\nu}B_\nu & (7.108)\\ &= \frac{8\pi\nu^2}{c^3}\left(e^{h\nu/kT_r(t)} - 1\right)^{-1} v(t)\, d\nu. & (7.109)\end{aligned}$$

Assuming that the number of photons is conserved, the number in the volume does not change with time. Because of the expansion, however, at some new time t' the photons at frequency ν have been redshifted:

$$\nu' = \nu\frac{a(t)}{a(t')}, \qquad\qquad d\nu' = d\nu\frac{a(t)}{a(t')} \tag{7.110}$$

for $t' > t$, while the volume has expanded to

$$v(t') = v(t)\frac{a^3(t')}{a^3(t)}. \tag{7.111}$$

With these two expressions and conservation of the number of photons, we get

$$\begin{aligned} dN(t') = dN(t) &= \frac{\frac{8\pi}{c^3}\left(\nu'\frac{a(t')}{a(t)}\right)^2 v(t')\,\frac{a^3(t)}{a^3(t')}\,d\nu'\,\frac{a(t')}{a(t)}}{\exp\left[h\nu'\,\frac{a(t')}{a(t)}/kT_r(t)\right] - 1} & (7.112)\\ &= \frac{8\pi\nu'^2}{c^3}\,v(t')\,d\nu'\,\left(e^{h\nu'/kT_r(t')} - 1\right)^{-1}, & (7.113)\end{aligned}$$

where $T_r(t') = T_r(t)\,a(t)/a(t')$. Thus the radiation preserves its blackbody spectrum as the universe expands, but the temperature of the blackbody decreases with the expansion, with $T_r \propto (1+z)$.

At any time t, we can obtain the energy density of radiation by integrating Eq. 7.109 over frequency, with $v(t) = 1$:

$$\begin{aligned} u_r(t) &= \int_0^\infty h\nu\, dN(t)\, d\nu = \frac{8\pi}{c^3}\int_0^\infty h\nu^3\,\left(e^{h\nu/kT_r} - 1\right)^{-1} d\nu\\ &= \frac{8\pi}{c^3h^3}\int_0^\infty \epsilon^3\left(e^{\epsilon/kT_r} - 1\right)^{-1} d\epsilon \qquad (\text{where } \epsilon = h\nu)\\ &= \frac{8\pi}{c^3h^3}k_B^4\,T_r^4\int_0^\infty \frac{x^3}{e^x - 1}dx \qquad (\text{where } x = \frac{\epsilon}{k_BT_r}). \end{aligned} \tag{7.114}$$

The integral on the right-hand side of Eq. 7.114 is standard:

$$\int_0^\infty \frac{x^3}{e^x - 1} dx = \frac{\pi^4}{15}.$$

Thus

$$\begin{aligned} u_r(t) &= \frac{8\pi^5}{15c^3h^3} k_B^4 \, T_r^4 \\ &= \frac{4\sigma}{c} T_r^4 \text{ erg cm}^{-3} \end{aligned} \tag{7.115}$$

where $\sigma = 8\pi^5 k_B^4 / 60 c^2 h^3 = 5.67 \times 10^{-5}$ erg cm^{-2} s^{-1} K^{-4} is the Stefan-Boltzmann constant.

Chapter 8

Observational Cosmology

8.1 Cosmic Microwave Background

Direct evidence for a hot Big Bang origin of the universe is provided by the cosmic microwave background (CMB), discovered Penzias and Wilson in 1965 (for which they received the 1978 Nobel Prize in physics). Specifically what they found was an apparently uniform radiation field characterized by a temperature of about 3 K, and apparently with a blackbody spectral shape. Since the temperature is so low, the radiation peaks at $\lambda \sim 1$ mm, i.e., in the microwave region of the spectrum. An enormous amount of effort has gone into improved measurements of the CMB spectrum over the last 30 years, culminating in the COBE mission in 1991.

The best-fit value to the temperature of the CMB is $T = 2.728 \pm 0.004$ K (Fixsen *et al.* 1996). The CMB is almost uniform. The main departure from uniformity is the *dipole* pattern, consistent with the idea that the Local Group of galaxis is moving through the CMB at 627 ± 22 km s^{-1} towards the direction $[l, b] = [276^\circ \pm 3^\circ, 30^\circ \pm 3^\circ]$ (in the constellation of Hydra)[1] We will return to analyzing the fluctuations in the CMB in Ch. 11.

8.1.1 Energy Density of the Universe

With the value of $T = 2.728$ K for the mean CMB temperature, the present-day energy density is

$$u_r(t_0) = 4.24 \times 10^{-13} \text{ erg cm}^{-3}, \tag{8.1}$$

or, in terms of the equivalent mass density,

$$\rho_r(t_0) = \frac{u_r(t_0)}{c^2} = 4.72 \times 10^{-34} \quad \text{g cm}^{-3}. \tag{8.2}$$

By Phil Maloney.

[1] These numbers are arrived at after correcting for the motion of the Solar System around the Milky Way, and the small correction for the motion of the Milky Way with respect to the other galaxies in the Local Group. See Kogut *et al.* 1993, ApJ, 419, 1.

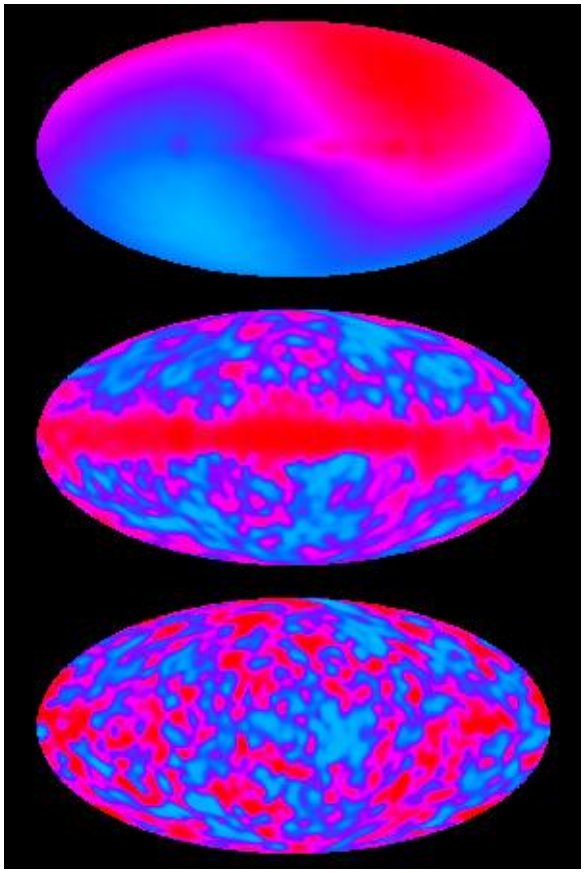


Figure 8.1: CMB fluctuations as seen by the COBE DMR. The top picture shows the dipole anisotrophy as a result of the Local Group's motion with respect to the CMB. The second picture shows emission from the plane of the Milky Way, after the dipole component has been subtracted out. The last picture shows the data after the galactic emission has been removed, showing fluctuations on the order of $\sim 10^{-5}$ of the 2.728 K background.

Recall that the critical density ρ_c, which produces a flat universe, is

$$\begin{aligned} \rho_c &= \frac{3H_0^2}{8\pi G} \\ &= 1.88 \times 10^{-29}\, h^2 \text{ g cm}^{-3}; \text{ with } h = H_0/100 \text{ km s}^{-1} \text{ Mpc}^{-1}. \end{aligned} \tag{8.3}$$

The contribution of radiation to the density parameter, $\Omega = \rho/\rho_c$, is then

$$\Omega_r = \frac{4.72 \times 10^{-34}}{1.88 \times 10^{-29}} = 2.5 \times 10^{-5}. \tag{8.4}$$

So clearly the contribution of radiation to the mass density of the universe is very small at the present.

Since it is extremely difficult to see how such a uniform thermal radiation field could have been produced under conditions similar to present-day, the CMB is evidence for a

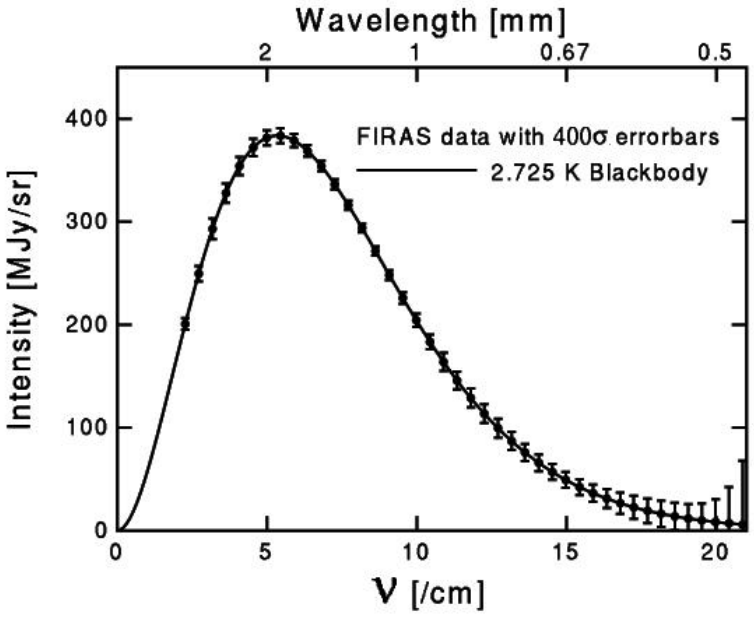


Figure 8.2: The CMB blackbody spectrum as measured by the FIRAS detector onboard COBE. The spectrum was measured at 43 equally spaced points along the curve; the line shows a blackbody fit to the data. The error bars are a tiny fraction of the width of the line in the plot, so they have been multiplied by a 400 to make them visible.

much hotter, denser phase of the universe at early times. To most astronomers, this settled the question of Big Bang vs. Steady State cosmologies.

(There are other background radiation fields, such as the X-ray background, but these are non-thermal and contribute much less to the total energy density.)

8.2 Baryon/Photon Number

If we write the present-day matter density as

$$\rho_{m_0} = \Omega_b \rho_c \qquad (b = \text{baryonic})$$

then the transition from a radiation-dominated universe to a matter-dominated one occurred at a redshift:

$$\begin{aligned} (1+z_{\rm eq})^{-1} &= \frac{a(t_{\rm eq})}{a(t_0)} = \frac{4.72\times10^{-34}}{1.88\times10^{-29}\,\Omega_b h^2} \\ \Longrightarrow z_{\rm eq} &\approx 4\times10^4\,\Omega_b h^2. \end{aligned} \tag{8.5}$$

Here we have explicitly distinguished baryons from other possible forms of matter, such as massive neutrinos, because the way in which they interact with radiation is different.

As we will see shortly, calculations of nucleosynthesis (element formation) during a hot early stage of the universe indicate that $\Omega_b h^2 \approx 0.015$, which indicates that the redshift of matter-radiation equality $z_{\rm eq} \approx 1000$. At this time the temperature of the radiation field was ~ 3000 K.

If we extrapolate back to earlier epochs (larger redshifts), the temperature of the radiation field gets hotter and hotter. At a redshift $z = 10^{10}$, $T \approx 3 \times 10^{10}$ K. Why is this significant?

At this epoch, the characteristic photon energy is:

$$\begin{aligned} h\nu \sim kT &\sim 4 \times 10^{-6} \text{ erg} \\ &\sim 3 \text{ MeV}. \end{aligned}$$

At this energy, the CMB photons are energetic enough to photodisintegrate complex nuclei into neutrons and protons, so that there would be no heavy elements at this redshift. As we will see shortly, it is possible to make detailed predictions about the abundances of the light elements during this phase of the universe.

Why does the radiation have a blackbody spectrum? The heat capacity of the radiation field at constant volume is just

$$c_v^r = \frac{dU_r}{dT} = \frac{16\sigma}{c} T_r^3 \tag{8.6}$$

from Eq. 7.115. Assuming that the matter consists of atomic hydrogen, the heat capacity of the matter is just

$$c_v^m = \frac{3}{2} n k_B = \frac{3}{2} k_B \cdot 1.124 \times 10^{-5}\, \Omega_b h^2 \text{ cm}^{-3}. \tag{8.7}$$

The ratio of these two quantities

$$\frac{c_v^m}{c_v^r} = 4 \times 10^{-9}\, \Omega_b h^2, \tag{8.8}$$

which is independent of redshift. At high redshifts, where the interaction between matter and radiation is strong, the matter relaxes to the radiation temperature since the heat capacity of the radiation is so much greater; in thermal equilibrium, the spectrum remains thermal no matter how strong the interaction between matter and radiation.

The total number of photons per unit volume in the CMB is given by integrating Eq. 7.109 for $dN(t)$ over frequency, with $v(t) = 1$:

$$\begin{aligned} n_\gamma = \int_0^\infty dN(t)\, d\nu &= \frac{8\pi}{c^3 h^3} \int_0^\infty \epsilon^2 \left(e^{\epsilon/kT_r} - 1 \right)^{-1} d\epsilon \\ &= \frac{8\pi}{c^3 h^3} k_B^3\, T_r^3 \int_0^\infty \frac{x^2}{e^x - 1} dx. \end{aligned} \tag{8.9}$$

The integral is:

$$\int_0^\infty \frac{x^2}{e^x - 1} dx = 2\zeta(3) = 2.404$$

where ζ is the Riemann ζ-function. Thus

$$\begin{aligned} n_\gamma &= \frac{16\pi}{c^3h^3}\zeta(3)\,k_B^3\,T_r^3 \\ &= 416(1+z)^3\ \mathrm{cm}^{-3}. \end{aligned} \tag{8.10}$$

What is the entropy in this radiation? For an isolated system at fixed volume,

$$\begin{aligned} dS = \frac{du_r}{T} &= \frac{c_v^r\,dT}{T} \\ &= \frac{16\sigma}{c}T_r^2\,dT_r, \end{aligned} \tag{8.11}$$

which is trivially integrable to

$$\begin{aligned} S_\gamma &= \frac{16\sigma}{3c}T_r^3 \\ &= \frac{16}{3c}\cdot\frac{8\pi^5k_B^4}{60c^2h^3}T_r^3 \\ &= \frac{32}{45}\frac{\pi^5k_B^4}{c^3h^3}T_r^3. \end{aligned} \tag{8.12}$$

This is the entropy per unit volume in blackbody radiation at temperature T_r. The entropy per photon is then

$$\begin{aligned} \frac{S_\gamma}{n_\gamma} &= \frac{\left[\frac{32}{45}\frac{\pi^5k_B^4}{c^3h^3}T_r^3\right]}{\left[\frac{16\pi}{c^3h^3}\zeta(3)\,k_B^3\,T_r^3\right]} \\ &= \frac{2\pi^4k_B}{45}\zeta^{-1}(3) \\ &= 3.6\,k_B. \end{aligned} \tag{8.13}$$

The ratio of the number of photons to the number of baryons is

$$\begin{aligned} \eta^{-1} = \frac{n_\gamma}{n_b} &= \frac{416}{1.124\times10^{-5}\,\Omega_bh^2} \\ &= 3.7\times10^7(\Omega_bh^2)^{-1} \\ \eta = \frac{n_b}{n_\gamma} &= 2.7\times10^{-8}\,\Omega_bh^2. \end{aligned} \tag{8.14}$$

What is the significance of this? We showed earlier (Eq. 8.8) that the heat capacity of the matter is negligible compared to the radiation. Thus all of the entropy is in the radiation. Eq. 8.13 shows that the dimensionless entropy per photon, $S_\gamma/n_\gamma k_B$, is of order unity. Thus the dimensionless entropy per baryon is approximately the ratio of the number of photons to the number of baryons, given as η^{-1} in Eq. 8.14. This is a huge number—in other words, the entropy of the universe is very high compared to its matter content. This has significant implications for the formation of the light elements during the era of nucleosynthesis.

One final point: the formation of the elements, at $z\sim10^{10}$, occurred during the radiation-dominated era (i.e., $z_{\rm EF}\gg z_{\rm eq}$), so the radiation energy density determines the

dynamics—as we saw earlier, both Λ and curvature are negligible during this phase. The solution for the scale factor during this era is Eq. 7.106:

$$\begin{aligned} a &= \sqrt{2Bt}t^{1/2}, \qquad B^2 = 8\pi G\rho_{r_0}a_0^4/3 \\ &= \sqrt{2}(8\pi G\rho_{r_0}/3)^{1/4}a_0. \end{aligned}$$

Since $a/a_0 = (1+z)^{-1} = T_{r_0}/T_r$, we can write this as an equation for t in terms of the radiation density or temperature:

$$\begin{aligned} t &= \left(\frac{3}{32\pi G\rho_{r_0}}\right)^{1/2}\left(\frac{T_{r_0}}{T_r}\right)^2 \\ &= \left(\frac{3}{32\pi G\rho_{r_0}}\left(\frac{T_{r_0}}{T_r}\right)^4\right)^{1/2} \\ &= \left(\frac{3}{32\pi G\rho_r}\right)^{1/2}, \end{aligned} \tag{8.15}$$

since $\rho_r \propto T_r^4$.

As we noted earlier, the formation of the elements occurred when the mean photon energy dropped to about 3 MeV, at $z \sim 10^{10}$, when the radiation temperature $T_r \sim 3 \times 10^{10}$ K. Expressing ρ_r in terms of the temperature using Eqs. 7.115 and 8.2 and evaluating the constants in Eq. 8.15:

$$\begin{aligned} t &= \left(\frac{3c^2}{32\pi Gu_r}\right)^{1/2} \\ &= \left(\frac{3c^3}{128\pi G\sigma T_r^4}\right)^{1/2} = 2.3\left(\frac{T_r}{10^{10}}\right)^{-2}\ \text{s}. \end{aligned} \tag{8.16}$$

Thus nucleosynthesis became possible when the universe was about 2 seconds old.

8.3 Nucleosynthesis in the Big Bang

The rest mass of an electron or positron $m_ec^2 \simeq 0.511$ MeV. For temperatures $T_r \gtrsim 1$ MeV, therefore, there is a "sea" of electron-positron pairs produced by photon-photon interactions:

$$\gamma + \gamma \rightleftharpoons e^- + e^+$$

The free neutrons and protons produced by photo-disintegration of nuclei by the radiation field are thermally coupled to the sea of electron-positron pairs (and a similar sea of neutrino-antineutrino pairs) by the reactions

$$\begin{aligned} e^- + p &\leftrightarrow n + \nu \\ \bar{\nu} + p &\leftrightarrow n + e^+ \\ n &\leftrightarrow p + e^- + \bar{\nu}, \end{aligned} \tag{8.17}$$

where $\bar{\nu}$ denotes an antineutrino. (In Eq. 8.17, the neutrinos are all electron neutrinos.) As long as the reactions in Eq. 8.17 occur at rates much faster than the expansion rate, the proton to neutron ratio will be at the thermal equilibrium value.

This ratio is very important to the outcome of nucleosynthesis, because essentially all of the neutrons get incorporated in ^{4}He. In thermal equilibrium, the neutron-to-proton ratio is just given by the Boltzmann distribution,

$$\frac{n}{p} = e^{-\Delta E/kT} \tag{8.18}$$

where ΔE is the proton-neutron rest-mass energy difference,

$$\Delta E = (m_n - m_p)c^2 = 1.2934 \text{ MeV}. \tag{8.19}$$

There is a high rate of capture of neutrons to form deuterium radiatively:

$$p + n \rightarrow d + \gamma, \tag{8.20}$$

the first step in the formation of the elements. However the reverse photodissociation reaction keeps the deuterium abundance extremely low. This is a direct consequence of the tiny value of η, the baryon to photon ratio.

When the temperature drops to $T \sim 10^{10}$ K, the weak interaction rates for Eqs. 8.17 drop below the expansion rate, and the neutron to proton ratio essentially "freezes out," at

$$\begin{aligned} \frac{n}{p} &= e^{-1.2934 \text{ MeV}/k_B \cdot 10^{10}} \\ &= 0.22. \end{aligned} \tag{8.21}$$

The neutron-to-proton ratio continues to decline beyond this temperature due to neutron decay, with $t_N \approx 17$ minutes ($= 1020$ seconds). This is not a huge effect, however, because nucleosynthesis is essentially complete before the universe is 1000 seconds old.

When the temperature has dropped to $T \sim 0.1$ MeV, there is no longer a significant number of thermal photons more energetic than 2.2 MeV, the binding energy of deuterium. (Note that 1 MeV $\sim 1.16 \times 10^{10}$ K.) This is at $t \sim 170$ seconds. By this time, the electron-positron pairs have annihilated, and the neutron-proton ratio has dropped to about 1/7.

At this point the deuterium abundance has grown large enough for the deuterons to produce helium, via the sequence of reactions:

$$\begin{aligned} d + d &\leftrightarrow t + p \text{ (where } t = \text{tritium)} \\ d + d &\leftrightarrow {}^3\text{He} + n \\ t + d &\leftrightarrow {}^4\text{He} + n \\ {}^3\text{He} + d &\leftrightarrow {}^4\text{He} + p. \end{aligned} \tag{8.22}$$

Essentially all of the neutrons wind up in ^{4}He, the most tightly bound of the light nuclei. If we assume that all of the neutrons form He, then it is trivial to calculate the mass fraction of ^{4}He: 2 neutrons and 2 protons are required to form ^{4}He. The number of ^{4}He nuclei is then 1/2 the number of neutrons at the time of He formation, at $T \sim 0.1$ MeV. Since the

^{4}He nucleus weighs ≈ 4 times as much as a neutron or proton, and ignoring everything except ^{4}He and H:

$$\begin{aligned} x_{^4\mathrm{He}} &\approx \frac{4n_{^4He}}{n_N} \qquad (n_N = n_n + n_p = \text{Total \# of baryons}) \\ &= \frac{4(n_n/2)}{n_n + n_p} = \frac{2(n/p)_{T=0.1}}{1 + (n/p)_{T=0.1}} \\ &\simeq 0.25. \end{aligned}$$

8.3.1 Predictions of Big Bang Nucleosynthesis

Essentially only D, ^{3}He, ^{4}He, and ^{7}Li are produced; this is due to the absence of stable nuclei of masses 5 and 8, so that build-up of heavy elements essentially shuts off at ^{4}He. Trace amounts of ^{7}Li are produced by

$$\begin{aligned} {}^4\mathrm{He} + t &\rightarrow {}^7\mathrm{Li} + \gamma \\ {}^3\mathrm{He} + {}^4\mathrm{He} &\rightarrow {}^7\mathrm{Li} + \gamma. \end{aligned} \tag{8.23}$$

Stars get around the mass barrier through the triple-α process:

$$\begin{aligned} {}^4\mathrm{He} + {}^4\mathrm{He} &\rightarrow {}^8\mathrm{Be} + \gamma \\ {}^3\mathrm{He} + {}^8\mathrm{He} &\rightarrow {}^{12}\mathrm{C} + \gamma. \end{aligned} \tag{8.24}$$

The second step is able to occur in stars *before* the decay of the highly unstable ^{8}Be nucleus because the densities are very high, considerably higher than the density at $T \sim 1$ MeV in the Big Bang.

What are these calculations sensitive to? Most of the nuclear physics input into the calculations (i.e., cross-sections for reactions) is well-determined; the only signifcant exception is for ^{7}Li, for which the predicted uncertainty is $\approx 50\%$.

The predicted ^{4}He abundance depends almost entirely on the value at which the neutron-to-proton ratio freezes out. This ratio depends on the rates of the weak interactions which interconvert protons and neutrons. These same rates determine the lifetime of free neutrons:

$$\tau_{1/2}(n) = 10.5 \pm 0.2 \text{ minutes}. \tag{8.25}$$

It is important to note that the predicted abundances depend only on the ratio of baryons to photons, η, and not on the individual values of n_b and n_γ. The reason for this is that during the epoch of nucleosynthesis, the universe is radiation-dominated, so the radiation density determines the expansion rate. Since the temperature and density of the blackbody radiation field are directly related ($\rho_r \propto T_r^4$), the expansion rate is then simply a function of temperature. The various reaction rates are all proportional to thermally-averaged cross sections, $\langle \sigma v \rangle = \bar{\sigma}(T)$[2] times the number densities of the various species, $n_i \propto \eta n_\gamma = n_i(\eta, T)$, since, as just noted, n_γ and T_r are directly related.

The results of standard Big Bang nucleosynthesis models then depend on one cosmological parameter, η, and two physical ones: $\tau_{1/2}(n)$, and the number of additional relativistic (at the time of Big Bang nucleosynthesis) particles, which contribute to the energy density. How do these affect the calculations?

[2] I.e., averaged over a Maxwellian velocity distribution.

η: If η is larger, then the abundances of ^{3}H, D, and ^{3}He, which depend on η^x where $x > 0$, build up slightly earlier, and this formation of ^{4}He begins earlier, when the neutron-to-proton ratio is larger, leading to a greater abundance of ^{4}He. The amounts of D and ^{3}He which survive therefore also depend on η, in the opposite way: there is more surviving D and ^{3}He for smaller values of η.

$\tau_{1/2}(n)$: All of the weak interaction rates are proportional to $1/\tau_{1/2}(n)$. An increase in $\tau_{1/2}(n)$ therefore leads to a decrease in the weak interaction rates that interconvert neutrons and protons. Thus the neutron-to-proton ratio will freeze-out at a higher temperature, leading to an increase in the predicted ^{4}He abundance.

N_ν: Any relativistic particles add to the energy density at the time of Big Bang nucleosynthesis, and therefore lead to an increase in the expansion rate.[3] We can express this as a constraint on the number of neutrino flavors, N_ν. At present 3 are known: ν_e, ν_μ, and ν_τ. Increasing N_ν leads to freeze-out of n/p earlier, at a higher value.

8.3.2 Comparison with Observations

The tricky part to comparing with observations is trying to determine whether the abundances being measured in any astronomical source are truly primordial (not counting the difficulties in making the observations and determining the abundances!).

Deuterium : The abundance of deuterium (usually expressed as the D/H ratio) has been determined in solar system objects and in the local interstellar medium. For the solar system objects and the molecular interstellar medium (ISM), the measurements are based on observations of deuterated molecules (e.g., DHO vs. H_2O, DCO vs. HCO). The best (?) determination in the solar system is for the atmosphere of Jupiter, with D/H $\simeq 1$–4×10^{-5}; a similar range is found from studies of the local ISM.

Since deuterium is very weakly bound, it is easy to destroy in stars through nuclear burning at $T \gtrsim 5 \times 10^5$ K. It is very difficult to find "contemporary" astrophysical sites for D formation. Thus the present value of the D/H ratio should be taken as a lower limit to the Big Bang nucleosynthesis value. With $(\mathrm{D/H})_p \gtrsim 1 \times 10^{-5}$ (where p stands for "primordial"), this leads to an upper bound on η of about 10^{-9}.

^{3}He : This is very difficult to measure in astronomical sources; the only way to do it (so far) relies on the ^{3}He$^+$ hyperfine line, as measured in H II regions. Its abundance has also been measured in meteorites and the solar wind.

^{3}He is much harder to destroy by nuclear processes than D, and it is hard to do this without also producing heavy elements and/or large amounts of ^{4}He. Of equal importance is the fact that deuterium burning produces ^{3}He, so that measurements of the ^{3}He abundance in the solar wind represent the sum of the Sun's original ^{3}He + D abundances.

The bottom line from stellar nucleosynthesis models is that the ^{3}He abundance has probably been reduced by burning in stars ("astration") by no more than a factor of two,

[3] At a fixed temperature.

and the upper limit to the primordial D+^{3}He abundance is

$$\left(\frac{\mathrm{D} + ^3\mathrm{He}}{\mathrm{H}}\right)_p \lesssim 8 \times 10^{-5}$$

which in turn implies that $\eta \gtrsim 4 \times 10^{-10}$.

^{7}Li : Since stellar nucleosynthesis produces heavy elements such as oxygen and carbon, one criterion for stars of primordial (or nearly) composition is to require that the metallicities are very low. The abundance of ^{7}Li as determined in the local ISM, young stars, and meteorites is $\gtrsim 10$ times the Big Bang nucleosynthesis value. However since ^{7}Li is produced by cosmic ray spallation and some stellar processes, there was little reason for believing this reflected primordial abundances.

^{7}Li determinations have been made in old, metal-poor stars ($z = z_\odot/12$ to $z_\odot/250$) with masses 0.6–1.1 $\mathrm{M}_\odot$. There is a clear correlation of ^{7}Li/H with stellar mass; the highest mass stars show a plateau at $^7\mathrm{Li/H} \approx (1.1 \pm 0.4) \times 10^{-10}$. There is some speculation that this could be the result of deeper surface convective zones in lower mass stars. If we take this to be the primordial abundance, then η is in the range (2–5) $\times 10^{-10}$. With 50% uncertainty, take $\eta = (1\text{–}7) \times 10^{-10}$.

^{4}He : Since ^{4}He *is* produced in stars, the observed ^{4}He abundance will in general not be the primordial value, but will be greater. Again, if we look at very metal-poor objects we expect to find a trend of ^{4}He abundance vs. metallicity. This is another area fraught with controversy, where systematic effects clearly dominate the errors.

The current best estimates suggest the primordial ^{4}He mass fraction is

$$0.22 \lesssim Y_p \lesssim 0.25,$$

where Y_p is the primordial helium abundance. This depends not only on η, but also on N_ν. Taking all of these constraints together, we get

$$\begin{aligned} 4 \times 10^{-10} \lesssim \quad \eta \quad &\lesssim 7 \times 10^{-10} \\ 2 \lesssim \quad N_\nu \quad &\lesssim 4. \end{aligned}$$

Thus, as we have seen here, it is possible to derive a consistent set of values for the abundances of the light elements from Big Bang nucleosynthesis. Furthermore, Big Bang nucleosynthesis makes a direct prediction for the number of neutrino flavors, ≥ 2 and ≤ 4. In fact, an independent determination of the number of neutrino flavors is provided by measuring the decay width of an elementary particle called the Z^0 boson. This has now been found to give the constraint $N_\nu \approx 3 \pm 0.5$.

8.3.3 Omega in Baryons

Since we know the temperature—and therefore the energy density—of the cosmic microwave background quite accurately, combined with the Big Bang nucleosynthesis constraints on η, we get a direct estimate of the density of baryons in the universe:

$$0.010 \leq \Omega_b\, h^2 \leq 0.016. \tag{8.26}$$

This immediately tells us that baryons alone (i.e., normal matter) cannot close the universe, since even the smallest plausible value of h (~ 0.5) gives an upper limit to $\Omega_b \lesssim 0.1$.

How many baryons do we see? We can estimate the contribution from baryons in stars by using the galaxy luminosity function, provided we know how to convert from luminosity to mass. In order to do this, we need dynamical information, i.e., observations of stellar or gas motions to determine $v(r)$. Then we can estimate the mass from

$$\frac{GM(r)}{r^2} = \frac{v_c^2}{r} \tag{8.27}$$

as before, where v_c is the outermost observed velocity/radius.

Most spiral galaxies exhibit flat rotation curves (which in itself has remarkable implications, which we will return to shortly). There is a well-established correlation between the maximum rotation velocity and the luminosity, the so-called *Tully-Fisher relation*:

$$v_c = 220\,(L/L_\star)^{0.22}\ \mathrm{km\ s^{-1}}, \tag{8.28}$$

where $L_\star$ is the characteristic luminosity in the galaxy luminosity function, Eq. 6.10. Application of this relation gives $M/L \sim 12\,h$ with only a very weak dependence on galaxy luminosity. Since most of the luminosity comes from galaxies near $L_\star$, the mass density contributed by galaxies is

$$\begin{aligned} \rho_{\rm gal} &\sim \frac{M}{L} \cdot L_\star \cdot \phi_\star \\ &\approx 12\,h \cdot 1.0 \times 10^{10}\,\mathrm{L_\odot}\,h^{-2} \cdot 1.2 \times 10^{-2}\,h^3\ \mathrm{M_\odot\ Mpc^{-3}} \\ &= 1.4 \times 10^9\,h^2\ \mathrm{M_\odot\ Mpc^{-3}}, \end{aligned} \tag{8.29}$$

which in terms of the critical density is

$$\begin{aligned} \Omega_{\rm gal} = \frac{\rho_{\rm gal}}{\rho_c} &= \frac{9.5 \times 10^{-32}\ \mathrm{g\ cm^{-3}}}{1.88 \times 10^{-29}\ \mathrm{g\ cm^{-3}}} \\ &= 5 \times 10^{-3}. \end{aligned} \tag{8.30}$$

This immediately suggests that there are unobserved baryons, since even for $h = 1$ this is less than the Big Bang nucleosynthesis value by a factor of at least 3. In fact, as we will discuss later, there is direct observational evidence for baryons which have not coalesced into galaxies, in the form of an intergalactic medium, as seen in absorption against high redshift objects such as quasars.

8.4 Cosmic Problems

So the standard Big Bang cosmology can explain the origin of the light elements, requires the observed number of neutrino families, and predicts approximately the observed number of baryons. It also explains the origins of the cosmic microwave background radiation. Can we stop here? *No!*

Although extremely successful in accounting for many properties of the observed universe, the standard Big Bang model has a number of shortcomings.

8.4.1 The Smoothness or Horizon Problem

In deriving the Robertson-Walker metric, we assumed that the universe was homogeneous and isotropic. Why should this be? In general, Einstein's field equations have solutions which are inhomogeneous and/or anisotropic.

The best evidence that the universe *is* homogeneous and isotropic comes from observations of the microwave background. The COBE measurements show that the background is smooth to a few parts in 10^6 on scales above a few degrees. At smaller scales—down to about $20''$—the cosmic microwave background is isotropic to better than one part in 10^4.

To understand the significance of this, we need to digress back to the evolution of the universe for a moment. At early times, the gas and radiation are in thermal equilibrium. Under conditions of thermodynamic equilibrium, the ionization equilibrium is given by the Saha equation: assuming pure hydrogen for implicity, let $n = n_{\rm H^+} + n_{\rm H}$ be the total density of hydrogen nuclei, and $x = n_e/n = n_{\rm H^+}/n$ is the electron fraction. Then

$$\frac{n_e n_p}{n_{\rm H} n} = \frac{x^2}{1-x} = \frac{(2\pi\, m_e\, kT)^{3/2}}{nh^3} e^{-B/kT}, \tag{8.31}$$

where the binding energy (ionization potential) $B = 13.6$ eV.

With $T = 2.736\,(1+z)$ K and $n = 1.12 \times 10^{-5}\,\Omega_b\, h^2 (1+z)^3$ cm^{-3}, and plugging in all the constants, this can be written as

$$\log\left[\frac{x^2}{1-x}\right] = 20.99 - \log\left[\Omega_b\, h^2 (1+z)^{3/2}\right] - \frac{25050}{1+z}. \tag{8.32}$$

The ionization fraction has dropped to 1/2 at a redshift

$$z_{\rm dec} = 1360; \qquad T_{\rm dec} = 3700 \text{ K} \quad \text{for } \Omega_b\, h^2 = 0.013.$$

The equlibrium ionization drops very rapidly; the redshift at which it has declined to 10^{-3} is

$$z_{10^{-3}} = 1030, \qquad T_{10^{-3}} = 2820 \text{ K} \quad \text{for } \Omega_b\, h^2 = 0.013.$$

In fact, the ionization fraction plateaus at a value above the equilibrium value after recombination, due to scattering of Lyman continuum photons.

The epoch at which the universe recombined is also referred to as the *Decoupling Era.* Why? When the universe was highly ionized, the photons and matter were very tightly coupled together, by Thompson scattering off of the free electrons. (We will see that this has very important implications when we talk about structure formation.)

The optical depth at any epoch z is then set by the electron density and the Thompson cross-section, $\sigma_T = 6.65 \times 10^{-25}$ cm^{-2}:

$$\tau \simeq \sigma_T\, n_e\, ct, \tag{8.33}$$

taking the length scale to be $L = ct$, where t is the expansion age of the universe. Assuming that the mass-energy density dominates over curvature and Λ, so that we approach the Einstein-de Sitter solution (cf. Eq. 7.67), and we can write the expansion time as

$$t = \frac{2}{3}\left[H_0\, \Omega^{1/2}\, (1+z)^{3/2}\right]^{-1}. \tag{8.34}$$

And so

$$\tau = 0.046 \times (1+z)^{3/2}\,\Omega_b\,\Omega^{-1/2}\,h. \tag{8.35}$$

(Note the distinction between Ω_b and Ω, which refers to the total density.)

For $x = 1$, this is $\tau \sim 40\,\Omega^{-1/2}$ at $z = 1000$ (~ 170 if $\Omega = \Omega_b$) so that each photon has been scattered many times. Once the universe recombines, however, this rapidly drops to a value $\tau \ll 1$, and the universe becomes transparent to the cosmic microwave background photons. Thus the isotropy of the cosmic microwave background measures the isotropy of the *surface of last scattering*, at $z = z_{\rm dec}$. (Note that this is coincidentally, also about the time $t_{\rm eq}$ of matter-radiation equality.)

If the universe had been substantially inhomogeneous or anisotropic, a significant signature would have been imprinted on the microwave background. *If* the entire observable universe were causally connected at the time of the last scattering of the cosmic microwave background photons, then one could imagine that some microphysical process acted to smooth out any temperature fluctuations to produce a uniform temperature.

However, this is not possible, because at the time of decoupling, the particle horizon size was much smaller than the size of the region which is now observable (i.e., the present particle horizon size): assume for simplicity a flat universe; the distance to the particle horizon (assuming a matter-dominated universe) is

$$R_{\rm ph} = 2\,cH_0^{-1}(1+z)^{-3/2}. \tag{8.36}$$

The present-day particle horizon—our observable universe—is just $2\,cH_0^{-1}$. At some earlier epoch, the size of this region—that is, the size of the region which grows into our presently observable universe—is

$$R^0_{\rm ph}\frac{a(t)}{a(t_0)} = \frac{R^0_{\rm ph}}{1+z} = 2\,cH_0^{-1}(1+z)^{-1} = R^0_{\rm ph}(z). \tag{8.37}$$

Hence at decoupling/recombination at $z \approx 1000$, the ratio of the particle horizon to the present-day observable universe (at the size it was then) was

$$\frac{R_{\rm ph}}{R^0_{\rm ph}} = (1+z)^{-1/2} \approx 3.2 \times 10^{-2}. \tag{8.38}$$

In other words, at the era of decoupling of the radiation field from matter, our universe (as we now observe it) consisted of $\sim 10^5$ causally disconnected regions. For this reason the smoothness problem is also called the *horizon problem*.

8.4.2 Formation of Small-Scale Problem

The tight coupling between matter and radiation prior to decoupling also means that any matter density fluctuations will have imprinted a signature on the cosmic microwave background. The observational limits constrain these fluctuations to very small amplitudes ($\delta\rho/\rho \lesssim 10^{-5}$), but this is not really the problem. The problem is that, in a baryon-only universe, it is very difficult to make structure formation models work on galaxy-mass scales; we will discuss this later on.

8.4.3 Flatness Problem

Models of Big Bang nucleosynthesis and observations of galaxies both suggest that $\Omega \sim 0.01$, i.e., it is *not* equal to 1. But it's not that far from 1. Should we be surprised by this? Maybe!

We earlier defined the present-day value of the critical density to be

$$\rho_c = \frac{3H_0^2}{8\pi G}. \tag{7.61}$$

In general, of course, this is a function of time. We can re-write Friedmann's equation (with a zero cosmological constant) as

$$\begin{aligned} \dot{a}^2 + k\,c^2 &= \frac{8\pi G\rho}{3}a^2 \\ H^2 + \frac{k\,c^2}{a^2} &= \frac{8\pi G\rho}{3} \\ \frac{k\,c^2}{H^2a^2} = \frac{8\pi G\rho}{3H^2} - 1 &\equiv \Omega - 1, \end{aligned} \tag{8.39}$$

where $\rho_c \equiv 3H^2/8\pi G$ and $\Omega = \rho/\rho_c$, as before. We can re-write Eq. 8.39 as an equation for H^2:

$$\begin{aligned} H^2 &= \frac{8\pi G\rho}{3} - \frac{k\,c^2}{a^2} \\ \Longrightarrow \frac{k\,c^2}{H^2a^2} &= \frac{3k}{8\pi G\rho\,a^2 - 3k} = \frac{1}{\frac{8\pi G\rho\,a^2}{3k} - 1}. \end{aligned}$$

And so we get an expression for Ω:

$$\begin{aligned} \Omega &= 1 + \frac{k\,c^2}{H^2a^2} \\ &= \frac{1}{1 - \frac{k\,c^2/a^2}{8\pi G\rho/3}}. \end{aligned} \tag{8.40}$$

The second term on the right-hand side of Eq. 8.40 becomes extremely small at early times; since $\rho \propto a^{-3}$ (for matter dominated) or a^{-4} (for radiation dominated), the second term becomes $\propto a(t)$ or $a^2(t)$ respectively.

This indicates that at very early times, Ω was extremely close to 1, and approaches 1 to arbitrarily good accuracy at small t. If Ω is not *exactly* 1 (i.e., $k = 0$) then why should it be fairly close to 1 now?

Another way of seeing this problem is the following. Write Friedmann's equation, including curvature, as

$$\left(\frac{\dot{a}}{a}\right)^2 = H^2 = \frac{8\pi G\rho}{3} - \frac{k\,c^2}{a^2} + \frac{\Lambda}{3}. \tag{8.41}$$

Each term on the right-hand side of Eq. 8.41 varies with time in a different way. We have seen that the predictions of Big Bang nucleosynthesis, occurring at $z \sim 10^{10}$, are in excellent

agreement with observations, and indicate that the mass-energy density is dominated by radiation and neutrinos (and nothing else!). This term varies like $(1+z)^4$. The curvature term $\propto a^{-2} \propto (1+z)^2$, while the Λ term is constant. Because of these different dependencies, in principle each term can dominate at a different epoch.

At $z \sim 10^{10}$, the radiation mass-energy term is 40 orders of magnitude larger than its current value, the curvature term is 20 orders of magnitude larger, and the Λ term is unaltered. If the curvature term dominates now (i.e., it is $\gg 10^{-2}$, the value of the mass density—now dominated by matter rather than radiation) at $z \sim 10^{10}$ it was down from the mass-energy density term by ~ 16 orders of magnitude.

In other words, although the curvature dominates now, by assumption, at $z \sim 10^{10}$ it was negligible, and the balance between the kinetic energy term ($\dot{a}^2$) and the potential energy term ($\propto G\rho a^2$) held to an accuracy of ~ 1 part in 10^{16}. If the cosmological constant dominates now, then this balance was even tighter (~ 1 part in 10^{36}).

This raises two questions. Why was there such a perfect balance between the kinetic and potential energy terms, and why should we just happen to be here in the epoch when this balance is disappearing? This problem does not arise in the Einstein-de Sitter $k = 0$, $\Omega = 1$ universe, because the ratio of the kinetic and potential energy terms is independent of time:

$$\begin{aligned} \dot{a}^2 &= \frac{8\pi G}{3}\rho a^2 \\ \Longrightarrow \frac{\dot{a}^2}{G\rho a^2} &= \text{constant.} \end{aligned}$$

This is also known as the "Dicke coincidence argument," after R. Dicke, who first pointed it out in the 1960s. No one paid all that much attention to these problems, however, until a solution was proposed!

The theoretical prejudice that $\Omega = 1$ has now become quite firmly entrenched. If we take these arguments and Big Bang nucleosynthesis at face value, we are forced to a remarkable conclusion: the mass of the universe is dominated by something other than normal (baryonic) matter.

8.5 Dark Matter

At about the same time that astronomers began to worry about the problems with the standard Big Bang model of the universe, a related problem cropped up in the field of galaxy dynamics.

From studies of the light distribution in spiral galaxies, it was well established that their surface brightness profiles are typically exponential in radius. Assuming that the mass-to-light ratio is constant, i.e., assuming that the mass of the galaxy is dominated by the stars, then galactic rotation curves should exhibit a Keplerian fall-off at large radius; once we are well outside the radius containing most of the mass, then

$$\begin{aligned} \frac{v_c^2}{R} &\approx \frac{GM}{R^2} \\ v_c &\equiv \left(\frac{GM}{R}\right)^{1/2} \quad \text{with } M = \text{constant, so } v \propto R^{-/2}. \end{aligned}$$

(Since galaxies are flattened, not spherical, there will be a correction to the rotation velocity of order unity, but we can ignore this.)

This is not what is observed in spiral galaxies. Instead, the rotation curves tend to rise to a maximum, and stay at that value, i.e., they are *flat*. A flat rotation curve, with v = constant = $v_{\rm max}$, immediately implies that GM/R = constant, or in other words, $M \propto R$.

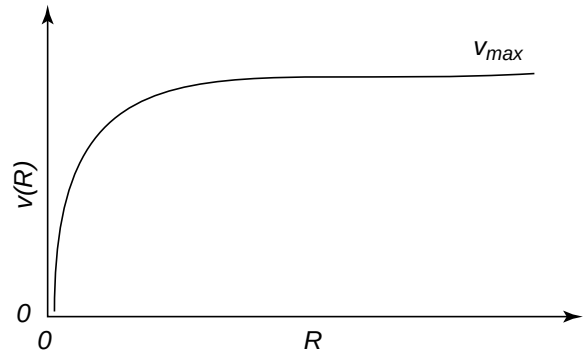


Figure 8.3: A sample rotation for a spiral galaxy, which rises to some maximum value $v_{\rm max}$ at large distances R from the galactic center.

Since this does not at all resemble the distribution of light, this immediately implies that the ratio of mass to luminosity—the mass-to-light ratio—must increase rapidly with radius. This is usually described by saying that the luminous parts of galaxies must be embedded within massive "halos" of dark matter.

If the dark matter distribution is spherical, then the fact that $M \propto R$ implies that (assuming a power-law density profile)

$$\begin{aligned} M &= 4\pi\,\rho_\circ \int_0^R r^2 \left(\frac{r_\circ}{r}\right)^\alpha dr \\ &= 4\pi\,\rho_\circ\, r_\circ^\alpha \int_0^R r^{2-\alpha}\, dr \\ &= \frac{4\pi\,\rho_\circ\, r_\circ^\alpha}{3-\alpha} R^{3-\alpha}, \end{aligned} \tag{8.42}$$

which would imply $\rho_{\rm halo} \propto r^{-2}$. The evidence for spherical dark matter halos is very weak, however.

How big and how massive are galaxies? Our knowledge is limited by the distance to which we can measure rotation curves. The 21 cm hyperfine line of atomic hydrogen is the most useful tracer, since atomic hydrogen distributions in galaxies are usually considerably more extended than the stellar distributions. However, the atomic hydrogen eventually gets ionized by the extragalactic radiation field, which limits how far out it is possible to measure the rotation curves. For those galaxies with extended rotation curves, the total mass at the last measured point is typically $\sim$ 3–4 times the mass in stars.

What is this mass? There are two possibilities: baryonic and non-baryonic. Proposed objects that fall into the baryonic category include normal astrophysical phenomena (such

as low-mass stars, brown dwarfs or Jupiter-sized planets, and stellar remnants, e.g., neutron stars and black holes), as well as more exotic creatures such as super-massive (10^{5-6} $M_\odot$) black holes. Non-baryonic matter includes normal neutrinos, which might have a non-zero mass, as well as exotic particles that may or may not exist, e.g., WIMPs.

8.5.1 Hot Dark Matter: Massive Neutrinos

If neutrinos have mass, they can contribute substantially to the total mass of the universe, because there are so many of them: there is a background sea of neutrinos analogous to the cosmic microwave background photons.

How many neutrinos are there? At very early times, the neutrinos were in equilibrium along with everything else. We can estimate the total number the same way we did for photons. There is a slight difference, however, because neutrinos, with spin 1/2, are fermions, whereas photons have spin zero and are bosons. Recall that for photons, the energy density is

$$\begin{aligned} u_r &= \frac{8\pi}{c^3h^3}\int_0^\infty \epsilon^3(e^{\epsilon/kT_r}-1)^{-1}\,d\epsilon \\ &= \frac{8\pi}{c^3h^3}k_B^4T_r^4\int_0^\infty \frac{x^3}{e^x-1}\,dx \end{aligned} \quad . \tag{7.114}$$

If the temperature is very high ($kT \gg$ neutrino rest mass), then the neutrinos are relativistic, and we can set the neutrino momentum $\rho = \epsilon/c$. (If neutrinos have zero rest mass, of course, this is always true.) In terms of momentum rather than energy, the energy density of a single family of neutrinos is

$$\begin{aligned} u_\nu &= \frac{2}{h^3}\int_0^\infty \frac{4\pi\, p^2\, dp\, pc}{e^{pc/kT_\nu}+1} \\ &= \frac{8\pi}{c^3h^3}k_B^4T_\nu^4\int_0^\infty \frac{x^3}{e^x+1}\,dx, \end{aligned} \tag{8.43}$$

where T_ν is the neutrino temperature. Note e^x+1 in the denominator rather than e^x-1 for photons, which is due to Fermi-Dirac vs. Bose-Einstein statistics. We use

$$\int_0^\infty \frac{x^3}{e^x+1}dx = \frac{7}{8}\int_0^\infty \frac{x^3}{e^x-1}dx,$$

and therefore the energy density in neutrinos is 7/8 the energy density in radiation at the same temperature.

A similar calculation to determine the number density of neutrinos (cf. Eqs. 8.9 and 8.10 for photons) gives

$$n_\nu = \frac{12\pi}{c^3h^3}\,\zeta(3)\,k_B^3T_\nu^3, \tag{8.44}$$

which is 3/4 the number of photons at the same temperature.

The present-day temperature of the neutrinos is slightly lower than that of the photons. Why? The neutrinos are kept in equilibrium by weak interactions. As we saw in discussing

nucleosynthesis, the expansion rate eventually becomes larger than the weak interaction rate, which causes the neutron/proton ratio to "freeze out." This also means that the neutrinos aren't kept in equilibrium anymore. This happens to occur *before* the electron-positron pairs annihilate and dump their energy (and entropy) into the radiation field. The result is:

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_r, \tag{8.45}$$

with $T_r^0 = 2.736$ K and $T_\nu = 1.95$ K. Just like the photons, $T_\nu \propto (1+z)$. Also, the number density varies as $(1+z)^3$. With Eq. 8.45 and the present-day value for the radiation temperature, we get that the present-day number density of neutrinos plus anti-neutrinos in a single family is

$$n_\nu = \frac{3}{11} n_\gamma = 113 \text{ cm}^{-3}. \tag{8.46}$$

If neutrinos in one family have rest mass m_ν, then this family contributes a mass density

$$\rho_\nu = n_\nu m_\nu. \tag{8.47}$$

In terms of the neutrino density relative to the critical density,

$$\begin{aligned} m_\nu = \frac{\rho_\nu}{n_\nu} &= 1.88 \times 10^{-29} \frac{\Omega_\nu h^2}{n_\nu} \text{ g} \\ &= 1.66 \times 10^{-31} \Omega_\nu h^2 \text{ g} \\ m_\nu c^2 &= 1.5 \times 10^{-10} \Omega_\nu h^2 \text{ ergs} \\ &= 93 \Omega_\nu h^2 \text{ eV}. \end{aligned} \tag{8.48}$$

In the Standard Model of particle physics, neutrinos are massless. Most of the attempted extensions of the Standard Model which have massive neutrinos predict that the neutrino masses scale with the mass of their associated leptons, e.g.,

$$(m_{\nu_e} : m_{\nu_\mu} : m_{\nu_\tau}) : (m_e^2 : m_\mu^2 : m_\tau^2).$$

Since $m_e \approx 0.511$ MeV, $m_\mu \approx 0.105$ GeV, and $m_\tau \approx 1.8$ GeV, these extensions of the Standard Model predict that the mass of the τ neutrino is much greater than that of the μ or e neutrino. If some scheme like this is correct, then one family (flavor) of neutrino dominates, and its mass is given by Eq. 8.48. For comparison, the experimental limits to the masses of the individual neutrinos are:

$$m_{\nu_e} \lesssim 8 \text{ eV}, \qquad m_{\nu_\mu} \lesssim 250 \text{ keV}, \qquad m_{\nu_\tau} \lesssim 35 \text{ MeV}.$$

Chapter 9

Formation of Structure in the Universe

So far we have considered the evolution of a smoothed-out version of the universe. However, the universe as we see it today is anything *but* smooth, except on the very largest scales: matter is clustered into stars, galaxies, clusters of galaxies, even clusters of clusters. How did it wind up this way?

In order to address this question, we have to understand how perturbations to the average density of the universe evolve, since it is these perturbations which will evolve into objects such as galaxies.

9.1 Jeans Mass

A crucial concept in the evolution of perturbations in a self-gravitating fluid is the *Jeans mass*. Consider a spherical cloud of radius R_c, mass M_c, and density ρ_c (assumed uniform). A spherical shell within the cloud of radius r and thickness dr has mass

$$dM(r) = 4\pi r^2 \rho \, dr. \tag{9.1}$$

The pressure gradient across the shell must balance the gravitational attraction on the shell, in equilibrium:

$$4\pi r^2 \, dP = -\frac{GM(r)}{r^2} dM(r), \tag{9.2}$$

which can be re-written as

$$3V \, dP = -\frac{GM(r)}{r} dM(r). \tag{9.3}$$

Integrating both sides of this equation from the center of the cloud to the cloud edge gives

$$3 \int_{P_c}^{P_s} V \, dP = \int_0^{R_c} \frac{GM(r)}{r} dM(r), \tag{9.4}$$

By Phil Maloney.

where P_s is the pressure at the cloud surface. The left-hand side can be integrated by parts:

$$\begin{aligned} 3\int_{P_c}^{P_s} V\,dP &= 3\Big[PV\Big]_{\text{center}}^{\text{edge}} - 3\int_0^{V_c} P\,dV \\ &= 3\,P_s\,V_s - 3\int_0^{V_c} P\,dV. \end{aligned}$$

Assume the cloud is isothermal. The energy of the cloud gas (per unit volume) is $\varepsilon = \frac{3}{2}nkT = \frac{3}{2}P$ where n is the particle number density. Thus

$$-\int_0^{V_c} P\,dV = -\frac{2}{3}\int_0^{V_c} \varepsilon\,dV = -\frac{2}{3}\varepsilon\,V_c = -\frac{2}{3}T_k,$$

where T_k is the total kinetic energy of the cloud. Thus

$$3\int_{P_c}^{P_s} V\,dP = 3\,P_s\,V_c - 2\,T_k.$$

The right-hand side of Eq. 9.4 is just the gravitational energy of the cloud:

$$\begin{aligned} \int_0^{R_c} \frac{GM(r)}{r}dM(r) \equiv W &= \frac{16\pi^2}{3}\,\rho_c^2\,G\int_0^{R_c} r^4\,dr \\ &= \frac{16\pi^2}{3}\,\rho_c^2\,G\,\frac{R_c^5}{5} = \frac{3}{5}\frac{GM_c^2}{R_c}. \end{aligned}$$

Therefore

$$3\,P_s\,V_c = 2\,T_k - W. \tag{9.5}$$

This is another form of the Virial Theorem, only now allowing for the effect of an external pressure.

Writing the internal energy as $\varepsilon = \frac{3}{2}P_c = \rho_c\,kT/\mu$ where μ is the mean mass per particle, we can write Eq. 9.5 as

$$4\pi\,R_c^3\,P_s = \frac{3\,M_c\,kT}{\mu} - \frac{3}{5}\frac{GM_c^2}{R_c}. \tag{9.6}$$

If we assume the surface pressure term is negligible, then the condition for equilibrium is

$$\frac{3\,M_c\,kT}{\mu} = \frac{3}{5}\frac{GM_c^2}{R_c}. \tag{9.7}$$

If the right-hand side of this equation is larger than the left-hand side, then the cloud must collapse, since the internal pressure forces are not strong enough to resist gravity. Thus the cloud will collapse if

$$\begin{aligned} &\frac{3}{5}\frac{GM_c^2}{R_c} > \frac{3\,M_c\,kT}{\mu} \\ &\Longrightarrow R_c < \frac{3\,GM_c\,\mu}{15\,kT}. \end{aligned} \tag{9.8}$$

The critical radius for gravitational collapse is then

$$R_{\rm cr} = \frac{3\,GM_c\,\mu}{15\,kT}. \tag{9.9}$$

If we compress a cloud of mass M_c (for fixed μ and T) to a radius smaller than $R_{\rm cr}$, it must collapse.

We can re-write Eq. 9.8 as

$$\begin{aligned} \frac{4\pi\,R_c^3}{3}\frac{G\rho_c}{5R_c} &> \frac{kT}{\mu} \\ \frac{4\pi}{15}G\rho_c &> \frac{kT}{\mu}\frac{1}{R_c^2}. \end{aligned}$$

The sound speed in the gas—the speed at which pressure disturbances travel—is $c_s \simeq (kT/\mu)^{1/2}$. Therefore,

$$\frac{4\pi}{15}G\rho_c > \frac{c_s^2}{R_c^2}. \tag{9.10}$$

The right-hand side of this equation is just $1/t_s^2$, where $t_s = R_c/c_s$ is the time for a sound wave to travel a cloud radius. Thus, we have

$$t_s \geq \left(\frac{15}{4\pi\,G\rho_c}\right)^{1/2}. \tag{9.11}$$

What is the right-hand side? In the absence of pressure, the equation of motion of a shell would be

$$\frac{d^2r}{dt^2} = -\frac{GM(r)}{r^2} = -\frac{4\pi r}{3}\rho G \equiv a_g. \tag{9.12}$$

From simple dimensional arguments, $r \sim a_g t_{\rm ff}^2$, where $t_{\rm ff}$ is the free-fall time scale for the shell to reach the center, which results in

$$t_{\rm ff} \sim \left(\frac{3}{4\pi\,G\rho_c}\right)^{1/2}. \tag{9.13}$$

The equation of motion can actually be solved exactly, giving

$$t_{\rm ff} = \left(\frac{3\pi}{32\,G\rho_c}\right)^{1/2}. \tag{9.14}$$

The right-hand side of Eq. 9.11 can then be written as

$$\left(\frac{15}{4\pi\,G\rho_c}\right)^{1/2} = \frac{2\sqrt{10}}{\pi}t_{\rm ff} \simeq 2\,t_{\rm ff},$$

and therefore,

$$t_{\rm ff} \lesssim \frac{1}{2}t_s. \tag{9.15}$$

Thus the cloud will collapse only if the gravitational free-fall time is less than the sound crossing time. this makes simple physical sense: if $t_{\rm ff} < t_s$, then the cloud collapses before the internal pressure has time to respond to halt the collapse.

We can re-write Eq. 9.15 as an expression for a critical mass. Since $t_s = R/c_s$, and

$$R = \left(\frac{3M_c}{4\pi\,\rho_c}\right)^{1/3},$$

for a fixed cloud mass and density, cubing both sides of Eq. 9.15 leads, after a little algebra, to

$$\begin{aligned} M_J &= \left(\frac{3\pi^5}{32}\right)^{1/2} c_s^3\,\rho_c^{-1/2} G^{-3/2} \\ &= \frac{1}{4}\left(\frac{3}{2}\right)^{1/2} \pi\rho_c \left(\frac{\pi c_s^2}{G\rho_c}\right)^{3/2}. \end{aligned} \tag{9.16}$$

This is known as the Jeans mass. For a given density and temperature (or c_s), objects with masses $M > M_J$ are unstable to gravitational collapse, while objects with $M < M_J$ are stabilized by their internal pressure.

In a static universe, only perturbations with $M > M_J$, i.e., $t_{\rm ff} < t_s$, can collapse. In an expanding universe, there is an additional criterion: a perturbation can only collapse (i.e., grow in amplitude) if the gravitational collapse time $t_{\rm ff}$ is less than the expansion timescale t_E.

Assuming a flat, mass-energy dominated universe, Friedmann's equation (7.57) is just

$$\begin{aligned} \dot{a}^2 &= \frac{8\pi G}{3}\rho a^2 \\ \dot{a} &= \left(\frac{8\pi G}{3}\rho\right)^{1/2} a \end{aligned} \tag{7.57}$$

and the expansion time $a/\dot{a}$ is just

$$t_E = \frac{a}{\dot{a}} = \left(\frac{3}{8\pi\,G\rho}\right)^{1/2}. \tag{9.17}$$

If we are interested in a universe with only one mass-energy component, then this is never a problem, since the same density enters into the equation for the expansion time (Eq. 9.17) and the collapse time (9.14). Suppose that we are interested in a perturbation to a component which is only a minor fraction of the total mass density, however, and the main component is smoothly distributed on the scale of the perturbation. This is the case, for example, for perturbations to the baryons during the radiation-dominated era, when the mass-energy density in radiation dominates by a large factor over that in baryons. In that situtation,

$$t_E \sim \left(\frac{1}{G\rho_R}\right)^{-1/2} \ll t_{\rm ff} \sim \left(\frac{1}{G\rho_B}\right)^{-1/2}, \tag{9.18}$$

and the rapid expansion inhibits collapse.

If the size scale of a perturbation is larger than the horizon d_H, then processes such as pressure support, etc., cannot possibly affect it, as the perturbation is by definition larger than the maximum causally-connected region. What happens in this case?

We can derive the behavior by the following clever argument. Consider a spherical region of radius $\lambda > d_H$, containing matter of mean density ρ_1, embedded in a $k = 0$ universe of density $\rho_\circ$. Let $\rho_1 = \rho_\circ + \delta\rho$, where $\delta\rho$ is small and positive. Thus, this is a positive density perturbation. By spherical symmetry (i.e., Birkhoff's Theorem), the matter outside this region cannot affect the evolution of the perturbation. Since $\rho_1 > \rho_\circ$, and $\rho_\circ$ corresponds to a flat ($k = 0$) universe, the perturbation must behave as a $k = 1$ universe. With negligible cosmological constant, Friedmann's equation is

$$\left(\frac{\dot{a}}{a}\right)^2 = H^2 = \frac{8\pi\, G\rho}{3} - \frac{kc^2}{a^2}. \tag{8.41}$$

Thus these two regions obey the evolution equations:

$$\begin{aligned} H_1^2 &= \frac{8\pi\, G\rho_1}{3} - \frac{c^2}{a_1^2} \\ H_\circ^2 &= \frac{8\pi\, G\rho_\circ}{3}. \end{aligned} \tag{9.19}$$

Note that $H_\circ$ does *not* refer to the present-day value of H! But

$$H_1 = \frac{\dot{a}_1}{a_1}, \qquad H_\circ = \frac{\dot{a}_\circ}{a_\circ}.$$

Now, compare the perturbed universe with the background universe when their expansion rates are the same, i.e., $H_1 = H_\circ$. Then

$$\frac{8\pi G}{3}(\rho_1 - \rho_\circ) = \frac{c^2}{a_1^2}, \tag{9.20}$$

or

$$\left(\frac{\rho_1 - \rho_\circ}{\rho_\circ}\right) = \left(\frac{\delta\rho}{\rho_\circ}\right) = \left(\frac{3c^2}{8\pi\, G\rho_\circ a_1^2}\right). \tag{9.21}$$

In general, if $H_\circ = H_1$ at some time, then $a_\circ \neq a_1$ at that time. But if $\delta\rho/\rho_\circ$ is small, then a_1 and $a_\circ$ will differ by a small amount and we can approximate $a_1 \approx a_\circ$.

Since $\rho_\circ \propto a^{-4}$ in the radiation-dominated phase and $\rho_\circ \propto a^{-3}$ in the matter-dominated phase,

$$\left(\frac{\delta\rho}{\rho}\right) \propto \begin{cases} a^2 & \text{radiation-dominated} \\ a & \text{matter-dominated.} \end{cases}$$

Thus perturbations on scales larger than the horizon size always grow; since $a \propto t^{1/2}$ (radiation-dominated) and $a \propto t^{2/3}$ (matter-dominated).

$$\left(\frac{\delta\rho}{\rho}\right) \propto \begin{cases} t & \text{radiation-dominated era} \\ t^{2/3} & \text{matter-dominated era.} \end{cases}$$

What happens to perturbations that are smaller than the horizon depends on both the epoch and the material involved in the perturbation.

We will consider three potential components to the universe: baryons, radiation, and dark matter (explicitly non-baryonic). This last may be either massive neutrinos, or some more exotic form of elementary particle. Before the universe recombines at $z_{\rm dec}$, the baryons and photons are tightly coupled by Thompson scattering of the photons off the electrons, while the dark matter particles (henceforth denoted as DM) are not; thus the evolution of these two components is quite different.

Since the particle horizon size increase with time, a perturbation which was initially larger than the horizon size $R_{\rm ph}$—and was therefore growing—will eventually enter the horizon. What happens then?

As we have already seen, there are two processes which can prevent it from growing further: pressure (i.e., $M < M_J$) and expansion (if the mass density of the universe is dominated by some species other than the perturbed component). For the DM, there is no "pressure" as such (we can assume the DM to be collision-less); but the analogous support is provided by the velocity dispersion of the DM particles (just as stellar systems such as galaxies are supported by the stellar velocity dispersions). If neither of these two processes are effective, then the perturbations will grow.

Clearly during the radiation-dominated era, no DM or baryon perturbations can grow, as the expansion time (dominated by the radiation density) is too short. Growth of perturbations with $\lambda < R_{\rm ph}$ can only occur once the universe is matter-dominated.

If we define the *Jeans Length* by

$$\frac{4\pi}{3}\lambda_J^3\rho = M_J, \tag{9.22}$$

then from Eq. 9.16,

$$\lambda_J = \left(\frac{3\pi c_s^2}{8G\rho_c}\right)^{1/2}. \tag{9.23}$$

For perturbations with $\lambda \gg \lambda_J$ (i.e., $M \gg M_J$) then pressure effects will be negligible, and the perturbations will grow like super-horizon speed perturbations (assuming $\Omega \approx 1$) with $\delta\rho/\rho \propto a$, or $t^{2/3}$. Perturbations with $\lambda \gtrsim \lambda_J$ also grow, but at a slower rate, due to the effects of pressure.

Consider first the evolution of a DM perturbation. At some high temperature T_D, the DM will have decoupled from the thermal equilibrium, and it will have become non-relativistic ($v \ll c$) once the temperature drops below $T_{\rm nr} \approx m_{\rm DM}c^2$. Assume that the DM is non-relativistic when the perturbation enters the horizon (we will see shortly why this is important).

Since it is collisionless, the DM does no pressure work in the expansion. However, its velocity dispersion (or temperature) decreases as a^{-1} due to redshifting of the particle momentum, just like the redshifting of photon energies. This has an important consequence: if the velocity dispersion of the DM is v, then Eq. 9.23 says that

$$\lambda_J \propto \frac{v}{\rho^{1/2}}, \tag{9.24}$$

where ρ corresponds to the dominant component. For $z > z_{\rm eq}$ ($a < a_{\rm eq}$), this is the radiation, while for $z < z_{\rm eq}$ ($a > a_{\rm eq}$), it is the DM. (In a multi-component medium it is the velocity dispersion—or sound speed—of the *perturbed* component which enters into Eq. 9.23, since it provides the pressure support, but the density of the *dominant* component, since it is the gravitationally dominant component which causes the collapse.) When the DM is relativistic, its velocity dispersion $v \simeq c$ is constant.

Since $\rho = \rho_R$ for $a < a_{\rm eq}$, $\rho^{-1/2} \propto \rho_R^{-1/2} \propto a^2$ for $a < a_{\rm eq}$, while for $a > a_{\rm eq}$, $\rho^{-1/2} = \rho_{\rm DM}^{-1/2} \propto a^{3/2}$. Thus

$$\lambda_J \propto \begin{cases} a^2 & (a < a_{\rm nr}) \\ a & (a_{\rm nr} < a < a_{\rm eq}) \\ a^{1/2} & (a_{\rm eq} < a). \end{cases} \tag{9.25}$$

When the DM is relativistic, $\lambda_J \approx R_{\rm ph}$. (This is just the size of the causally-connected universe.) Once the DM goes non-relativistic, the Jeans length increases at most as fast as the expansion rate. Hence any DM perturbation which is non-relativistic when it enters the horizon will have $\lambda > \lambda_J$.

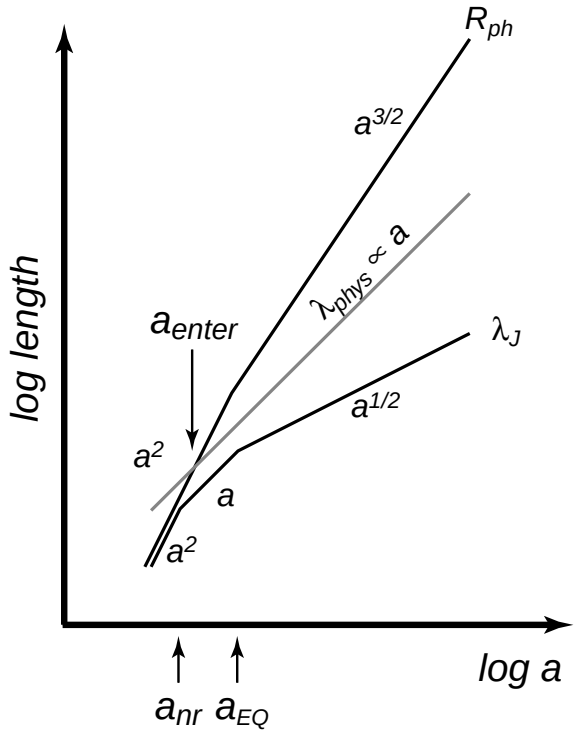


Figure 9.1: Evolution of a dark matter perturbation.

We can then divide the evolution of a DM perturbation into three stages.

1. $\boldsymbol{a < a_{\rm enter}}$. The perturbation wavelength is larger than the horizon; $\delta\rho/\rho \propto a^2$.

2. $\boldsymbol{a_{\rm enter} < a < a_{\rm eq}}$. The perturbation enters the horizon. Since $\lambda > \lambda_J$, pressure forces (velocity dispersion, in this case) cannot halt the collapse. However, $\rho_R \gg \rho_{\rm DM}$, so expansion prevents further growth: $\delta\rho/\rho =$ constant.
3. $\boldsymbol{a_{\rm eq} < a}$. ρ is now dominated by $\rho_{\rm DM}$, and the perturbation grows. If $\lambda \gg \lambda_J$, $\delta\rho/\rho \propto a$.

The scalings for λ_J in Eq. 9.25 give the Jeans mass for the DM component as (with $\rho_{\rm DM} \propto a^{-3}$ in the non-relativistic space)

$$M_J = \frac{4\pi}{3}\lambda_J^3 \rho_{\rm DM} \propto \begin{cases} a^2 & (a < a_{\rm nr}) \\ \text{constant} & (a_{\rm nr} < a < a_{\rm eq}) \\ a^{-3/2} & (a_{\rm eq} < a). \end{cases} \tag{9.26}$$

Thus the DM Jeans mass decreases steadily once the universe reaches the matter-dominated stage.

Plugging in numbers, we get

$$M_J = 3.2 \times 10^{14}\ {\rm M}_\odot\, (\Omega h^2)^{-2} \left(\frac{a}{a_{\rm eq}}\right)^{-3/2} \qquad \text{for } a > a_{\rm eq}, \tag{9.27}$$

since

$$\frac{a(t_\circ)}{a(t_{\rm eq})} = 4 \times 10^4\, \Omega h^2, \tag{8.5}$$

so therefore

$$M_J(t_\circ) = 4 \times 10^7\ {\rm M}_\odot\, (\Omega h^2)^{-7/2}.$$

This is much smaller than, say, the mass of a galaxy ($\sim 10^{11}$ M$_\odot$).

For baryons, the behavior is somewhat different because of their coupling to the photons. For $a < a_{\rm dec}$ the baryons and photons are in pressure equilibrium. After decoupling, the matter temperature drops faster than the radiation, because the matter does work in the expansion (adiabatic expansion) and so $T_m \propto a^{-2}$, rather than a^{-1} for T_r.

When the photons and baryons are tightly coupled, the pressure and density are dominated by the radiation, and the characteristic value of the velocity dispersion $v^2 \approx \frac{1}{3}c^2$ (with the factor of 1/3 arising from the one-dimensional velocity dispersion). (For the allowed range of $\Omega_b h^2$ from Big Bang nucleosynthesis, $z_{\rm dec}$ and the redshift of equality of mass-energy density in *baryons* and photons are very similar.) After decoupling, the characteristic velocity dispersion drops from $\frac{1}{3}c^2$ to

$$v^2 \approx \frac{5}{3}\frac{kT_\circ}{m_p}(1 + z_{\rm dec}), \tag{9.28}$$

which is a factor $\approx 2\times10^{-9}$. The Jeans mass is $\propto v^3$, so this drops by a factor $\simeq 8.3\times10^{-14}$. This is an enormous change: just before and after decoupling,

$$\begin{aligned} M_J(t \lesssim t_{\rm dec})_{\rm baryon} &= 3.1 \times 10^{16}\ {\rm M}_\odot \left(\frac{\Omega_b}{\Omega}\right)(\Omega h^2)^{-1/2} \\ M_J(t \gtrsim t_{\rm dec})_{\rm baryon} &= 2.5 \times 10^{3}\ {\rm M}_\odot \left(\frac{\Omega_b}{\Omega}\right)(\Omega h^2)^{-1/2}, \end{aligned} \tag{9.29}$$

inconsequence of the sudden, huge pressure drop.

As for the DM perturbations, baryon perturbations evolve in 3 stages:

$$\left(\frac{\delta\rho}{\rho}\right) \propto \begin{cases} a^2 & (a < a_{\rm enter}) \\ \text{constant} & (a_{\rm enter} < a < a_{\rm eq}) \\ a & (a_{\rm dec} < a). \end{cases}$$

(This last assumes $\lambda > \lambda_{J,\rm baryon}$.)

There is an important distinction here from the DM case. Note that DM perturbations grow once $a > a_{\rm eq}$, while baryon perturbations grow only for $a > a_{\rm dec}$. If $\Omega_{\rm DM} \simeq 1$, then $z_{\rm eq} \simeq 4 \times 10^4\, h^2$, while $z_{\rm dec} \approx 1000$. Hence DM perturbations can begin growth well before baryonic perturbations; from the time $t_{\rm eq}$ to $t_{\rm dec}$, the DM perturbations grow by a factor $a_{\rm dec}/a_{\rm eq} \approx 21\,\Omega h^2$. Thus when the baryons decouple from the photons, the DM perturbations are much more important, and the baryons will rapidly fall into the DM potential wells.

9.2 Spectrum of Perturbations

What are these perturbations? Presumably they represent primordial fluctuations in the universe; we do not presently have a good understanding of their origins.

However, what we observe in the universe will *not* in general be a direct reflection of some primordial spectrum of fluctuations. (By the spectrum we basically mean $d\rho/\rho$ as a function of mass scale.) The reason is that there are processes which basically *filter* the spectrum, one of which affects relativistic matter, the other of which affects baryons.

1. **Free-streaming.** Suppose we have a DM perturbation, and the DM is still relativistic when it enters the horizon. Because the DM particles do not interact collisionally, each particle is free to move along a geodesic in spacetime. This means that on sufficiently small scales, any perturbations will be wiped out, because the DM particles are free to stream from an overdense region to an underdense region, thereby eliminating the perturbation.

   What is this scale? When the DM is relativistic, it essentially has velocity c. The proper distance traveled by a particle in time t can be written as

   $$L = a(t) \int_0^t \frac{v(t')}{a(t')}\, dt' \tag{9.30}$$

   (since the proper velocity $v = a\, dl/dt$). In time $t_{\rm nr}$, the particle will travel a distance $2\, ct_{\rm nr}$ (where the extra factor of 2 coming from the integration over the expansion rate). At present, this free-streaming scale is $L_{\rm FS} = \frac{a_\circ}{a_{\rm nr}} \cdot 2\, ct_{\rm nr}$.

   The epoch $t_{\rm nr}$ depends on the mass of the DM particle:

   $$L_{\rm FS} = 0.5\ {\rm Mpc}\, (\Omega_{\rm DM} h^2)^{1/3} \left(\frac{m}{1\ {\rm keV}}\right)^{-4/3}. \tag{9.31}$$

For neutrinos, for example, with $\Omega_\nu h^2 = m_\nu/93$ eV,

$$L_{\rm FS} = 28 \text{ Mpc} \left(\frac{m_\nu}{30 \text{ eV}}\right)^{-1} \tag{9.32}$$

$$M_{\rm FS} = 4 \times 10^{15} \text{ M}_\odot \left(\frac{m_\nu}{30 \text{ eV}}\right)^{-2}. \tag{9.33}$$

Any perturbations on mass scales smaller than this will be wiped out by free-streaming.

DM thus gets divided into two varieties: hot dark matter (relativistic at the epoch of horizon crossing) and cold dark matter (non-relativistic). The former also corresponds to a "top-down" scenario for structure formation (all the power in density fluctuations is on large scales), while cold dark matter is a "bottom-up" scenario.

2. **Photon viscosity.** At $t \ll t_{\rm dec}$, the photons and baryons are very tightly coupled, and any fluctuations on scales smaller than the photon mean free path

$$l_{\rm MFP} = \frac{1}{x_e n_e \sigma} \simeq 1.3 \times 10^{29} x_e^{-1} (1+z)^{-3} (\Omega_b h^2)^{-1} \text{ cm} \tag{9.34}$$

will be obliterated.

However, this works on even larger scales: the photons can diffuse out of over-dense regions, and they drag the matter with them. This actually has its largest effect just as the universe is recombining, as the photon mean free path increases. This scale turns out to be

$$M_s \simeq 6.2 \times 10^{12} \left(\frac{\Omega}{\Omega_b}\right)^{3/2} (\Omega h^2)^{-5/4} \text{ M}_\odot. \tag{9.35}$$

This process is known as "Silk damping," after Silk (1968). Since the limits from Big Bang nucleosynthesis restrict $\Omega_b/\Omega \lesssim 0.1$, the scale mass must be $M_s \gtrsim 10^{14}$ M$_\odot$ in either a baryon-only or a flat ($\Omega = 1$) universe.

Thus in either baryon-only or hot dark matter-dominated universes, the minimum mass scale of perturbations available to form galaxies, clusters, etc., once growth resumes is much greater than the mass of a galaxy, or even a cluster, and so galactic mass objects must form by the collapse and subsequent fragmentation of these very large (both in mass and size) scales.

In a cold dark matter-dominated universe, however, this is not the case: from Eq. 9.31, if the mass of the dark matter particle $m_{\rm DM}c^2 \gg 1$ keV, then $L_{\rm FS} \ll 0.5$ Mpc. For reference, assuming that the mass of a typical galaxy is $M_{\rm gal} \simeq 10^{11}$–$10^{12}$ M$_\odot$, then with $\Omega = 1 \Longrightarrow \rho = \rho_{\rm crit} = 1.88 \times 10^{-29} h^2$ g cm^{-3}, the radius corresponding to a galaxy mass is

$$M_{\rm gal} = \frac{4\pi}{3} \lambda^3 \rho_{\rm crit} \Longrightarrow \lambda = (0.44\text{–}0.95) h^{-2/3} \text{ Mpc}$$

for $M_{\rm gal} = 10^{11}$–10^{12} M$_\odot$.

Thus the mass of a typical galaxy corresponds to a region ~ 1 Mpc in size today. In other words, to form a galaxy we needed to collapse all the matter in a region which today would be ~ 1 Mpc in size.

Since $L_{\rm FS} \ll 0.5$ Mpc for $M_{\rm DM}c^2 \gg 1$ keV, in cold dark matter scenarios perturbations on scales much smaller than galactic mass scales can survive until growth resumes at $z = z_{\rm eq}$.

9.2.1 Linear/Non-Linear Perturbations

As we have seen, density fluctuations grow as $\delta\rho/\rho \propto a$ in the matter-dominated era (only after recomination for baryonic perturbations). For a flat ($\Omega = 1$, $k = 0$) universe,

$$\frac{\delta\rho}{\rho} \propto t^{2/3}$$

so the density perturbations grow as power-laws in time. The flat $\Omega = 1$ universe gives the optimum case for forming structure gravitationally. For $\Omega < 1$ growth is shut off by the expansion of the universe, while for $\Omega > 1$ the time available for the growth of density fluctuations is less.

In deriving this result for the growth of density perturbations, we assumed that $\delta\rho \ll \rho$; this is known as the *linear* regime. Once $\delta\rho/\rho$ reaches ~ 1, the perturbation becomes *nonlinear*: the gravitational attraction due to the perturbation (its self-gravity) becomes dominant (locally) over the expansion. While the perturbation was in the linear phase, its physical size continued to increase as the universe expanded (but not as fast as the scale factor a) while the density contrast $\delta\rho/\rho$ grew; once $\delta\rho/\rho$ reaches ~ 1, however, and its self-gravity becomes dominant, the perturbation breaks away from the expansion: it reaches its maximum size at this point, then collapses and forms a bound object.

Thus in order to have formed a bound object which is no longer influenced by the expansion of the universe (e.g., a galaxy), the corresponding density fluctuation $\delta\rho/\rho$ on the relevant size/mass scale must have reached 1 before today.

For a flat universe, the age as a function of redshift is

$$t = 2 \times 10^{17} (1+z)^{-3/2}\, h^{-1} \text{ seconds}. \tag{9.36}$$

For baryon perturbations, as we have seen, growth does not start until the decoupling matter and radiation, at $z_{\rm dec} \approx 1500$, when the universe was

$$t_{\rm dec} \approx 3.4 \times 10^{12}\, h^{-1} \text{ seconds}$$

old (about 110,000 years). Thus baryon perturbations can only have grown by

$$\left(\frac{t_\circ}{t_{\rm dec}}\right)^{2/3} = \left(\frac{2 \times 10^{17}}{3.4 \times 10^{12}}\right)^{2/3} \approx 1.5 \times 10^3. \tag{9.37}$$

Thus at $z_{\rm dec}$, the magnitude of the density fluctuations must have been at *least* $(\delta\rho/\rho)_{z_{\rm dec}} \gtrsim 6 \times 10^{-4}$. In reality, this is an underestimate, because galaxies formed at $z > 0$, so a more realistic value is $(\delta\rho/\rho)_{z_{\rm dec}} \gtrsim 10^{-3}$.

Note that at $z_{\rm dec}$, a region corresponding to a galactic mass—$\lambda \sim 1$ Mpc—was ~ 1 kpc in size. This corresponds to an angular size scale $\theta \sim 30\,\Omega\,h$ arcsec on the sky today.

This is a big problem for baryon-only universes—the limits on $\delta T/T = \delta\rho/\rho \lesssim 10^{-4}$ on these size scales. Thus, in order to have grown to nonlinearity and collapse, baryon-only

perturbations have to be ~ 10 times larger than allowed by limits to fluctuations in the microwave background.

For dark matter perturbations, growth begins substantially earlier, at $z = z_{\rm eq}$, rather than $z = z_{\rm dec}$. Since $z_{\rm eq} \simeq 4 \times 10^4\,\Omega h^2$, in a flat $\Omega = 1$ universe:

$$\begin{aligned} t_{\rm eq} &\simeq 2 \times 10^{17} (4 \times 10^4 h^2)^{-3/2} h^{-1} \text{ seconds} \\ &\approx 2.5 \times 10^{10} h^{-4} \text{ seconds.} \end{aligned}$$

And so dark matter perturbations will have grown by

$$\left(\frac{t_\circ}{t_{\rm eq}} \right)^{2/3} \simeq z_{\rm eq} = 4 \times 10^4 h^2, \tag{9.38}$$

so that

$$\left(\frac{\delta\rho}{\rho} \right)_{z_{\rm eq}} ({\rm DM}) \lesssim 10^{-4} \quad \text{for the allowed range of } h.$$

Of course, at $z_{\rm dec}$ these perturbations will have grown to the same amplitude as the baryon perturbations would have to be; however since the baryons will not respond (i.e., fall into) the dark matter perturbations until *after* decoupling, this is not a problem.

9.3 Primordial Spectrum of Perturbations

The usual convention for specifying the initial (i.e., primordial) psectrum of perturbations to the density of the universe is to specify the amplitude (that is, $\delta\rho/\rho$) at the time they enter the horizon. This eliminates some General Relativistic ambiguities which arise in dealing with super-horizon-size perturbations. (These ambiguities can be resolved, but this is not something we want to get into.) It is usually assumed that the primordial fluctuation spectrum is a featureless power-law (i.e., no low-mass or high-mass cutoffs):

$$\left(\frac{\delta\rho}{\rho} \right)_{\rm hor} (M) = A\, M^{-\alpha}, \tag{9.39}$$

where A is a normalization constant (setting the actual amplitude), and α is the slope. The amplitude can be fixed by comparison with observations (e.g., the COBE results, which measure the amplitude of density perturbations on very large scales, which are still in the linear regime).

Can we say anything *a priori* about the slope, α?

We have seen that on galaxy scales ($M \sim 10^{11}$–10^{12} $\mathrm{M}_\odot$, $\lambda \sim 1$ Mpc), the amplitude of perturbations to the density $\delta\rho/\rho \sim 10^{-4}$ at $z_{\rm eq}$. In fact, the size of the horizon ($R_{\rm ph} \sim 11\ {\rm Mpc}\,(\Omega h^2)^{-1}$) was only a few times larger than the galaxy mass scale at $z_{\rm eq}$, so we must have had

$$\left(\frac{\delta\rho}{\rho} \right)_{\rm hor} (M_{\rm gal}) \sim 10^{-4}$$

to order of magnitude.

The COBE results for the amplitude of fluctuations on much larger size/mass scales (the size of the present-day horizon, corresponding to $M \sim 10^{22}\,\Omega h^2$ M$_\odot$) indicate that

$$\left(\frac{\delta\rho}{\rho}\right)_{\rm hor,\,now} (M \sim 10^{22}\ {\rm M}_\odot) \sim 10^{-5}.$$

If the initial spectrum of perturbations was not cut off in some fashion at long wavelengths (large masses), then the fact that $(\delta\rho/\rho)_{\rm hor}$ is constrained to be very similar on mass scales that differ by 10 or 11 orders of magnitude implies that α must be very small, i.e., not very different from zero.

What about masses much less than a galaxy mass? If $\alpha > 0$, then $(\delta\rho/\rho)_{\rm hor}$ increases with decreasing mass.

Any perturbation with $(\delta\rho/\rho)_{\rm hor} \sim 1$ will collapse into a black hole: as we saw earlier on p. 135, any perturbation in non-relativistic matter has $M > M_J$ when it crosses the horizon. If $(\delta\rho/\rho)_{\rm hor} \sim 1$, the perturbation immediately breaks away from the expansion, and acts like a small closed universe: since $M > M_J$, pressure forces will be unable to resist the collapse, and the result will be collapse to a black hole. Black holes with masses $< 10^{15}$ g will have evaporated by now via Hawking radiation, but larger mass ones will still be around, while those with $M \simeq 10^{15}$ g will now be evaporating.

Such a scenario will produce too much of a γ-ray background (due to the evaporating black holes) and produce far too much mass in primordial black holes (which would, for example, have non-trivial consequences for the dynamics of stellar systems). Thus scales much smaller than galaxy masses also require that α is very small (at least if $\alpha > 0$).

The combination of these two arguments favors $\alpha = 0$ or 1, so that $(\delta\rho/\rho)_{\rm hor}$ is independent of mass. One particularly important model has $\alpha = 2/3$, which is known as the *Harrison-Zel'dovich* spectrum (Harrison 1970, Zel'dovich 1972). We will discuss more about the Harrison-Zel'dovich spectrum in Ch. 11.

9.4 Structure Formation: The Virial Theorem

At $z < z_{\rm dec}$, baryons are free to fall into the dark matter potential, as they are no longer tied to the radiation field. What happens next? This depends on the depth of the potential well.

Recall from our discussion of the Virial Theorem with no external pressure, in equilibrium we must have (Eq. 9.7)

$$\frac{3MkT}{\mu} = \frac{3}{5}\frac{GM^2}{R}$$

(where μ is the mean mass per particle). We can use this to define the *virial temperature* for an object to be supported by thermal motions against its own gravity:

$$\begin{aligned} T_{\rm vir} &= \frac{Gm_\mu}{5kR} \\ &= 10^6 \left(\frac{M/10^{11}\ {\rm M}_\odot}{R/10\ {\rm kpc}}\right)\ {\rm K}. \end{aligned} \tag{9.40}$$

(We have taken $\mu = 1.67 \times 10^{-24}$ g in calculating this, i.e., we have assumed an atomic gas.) At $z_{\rm dec} \sim 1000$, the baryons and radiation will have equal temperatures, $T_b \sim 4000$ K.

> Dark matter potentials with $T_{\rm vir} \lesssim T_b$ will have very little effect on the baryons, as these DM fluctuations are not sufficiently "deep."
>
> Dark matter potentials with $T_{\rm vir} \gg T_b$ will have a large effect on the nearby baryons, which will fall in with no initial pressure support (because the baryons are too "cold" compared to the temperature $T_{\rm vir}$ characterizing the depth of the DM potential).

What happens to the baryons after they fall in? For a dissipation-less system (such as the dark matter perturbations) which breaks away from the expansion when it has a radius $R_{\rm max}$, the subsequent collapse to a new equilibrium (a process known as virialization for a self-gravitating system) leads to a decrease in size by a factor of two. This is easy to show using the Virial Theorem. When the DM perturbation reaches its maximum radius (when $\delta\rho/\rho \sim 1$), then its kinetic energy is zero (because the velocity goes to zero as it goes from expansion to collapse). Thus its total energy E just equals its gravitational potential energy:

$$\begin{aligned} E &= W = -\frac{GM^2}{R_{\rm max}} \\ T &= 0. \end{aligned} \tag{9.41}$$

After the perturbation collapses and virializes at its new radius $R_\circ$, it obeys the virial theorem, so $2T - W = 0$.

$$\Longrightarrow v^2 = \frac{GM}{R_\circ},$$

where v is the velocity dispersion of the system; this is just twice the kinetic energy per unit mass:

$$\frac{T}{M} = \frac{1}{2}v^2 = \frac{1}{2}\frac{GM}{R_\circ}. \tag{9.42}$$

The gravitational potential energy is $-GM^2/R_\circ$, and so the total energy is

$$E = T + W = \frac{1}{2}\frac{GM^2}{R_\circ} - \frac{GM^2}{R_\circ} = -\frac{1}{2}\frac{GM^2}{R_\circ}. \tag{9.43}$$

However, since we have *assumed* that this system is dissipation-less, there can be no loss of energy, and so the energy in Eqs. 9.41 and 9.43 is the same:

$$\begin{aligned} -\frac{1}{2}\frac{GM^2}{R_\circ} &= -\frac{GM^2}{R_{\rm max}} \\ \Longrightarrow R_\circ &= \frac{1}{2}R_{\rm max}. \end{aligned} \tag{9.44}$$

Thus the equilibrium radius after virialization is $\frac{1}{2}$ the radius at maximum expansion (also known as "turnaround"); the mean density therefore increases by a factor of 8.

9.5 Cooling of Baryonic Gas

Baryons are *not* dissipation-less, however: as the baryons fall into the potential well, they are gravitationally accelerated until they collide in the center of the DM potential well. At this point, collisions between the baryons lead to shock heating of the baryons up to the virial temperature (because this is the temperature corresponding to the kinetic energy they have acquired by falling into the potential well). What happens to the baryons next depends on whether they can cool, or simply remain at $T_{\rm vir}$.

There are 3 main processes at work in the early universe, when the matter consists essentially of just H and He.

1. **Compton cooling:** This occurs when cosmic microwave background (CMB) photons scatter inelastically off of electrons. (This is the same process which imprints fluctuations on the CMB.)

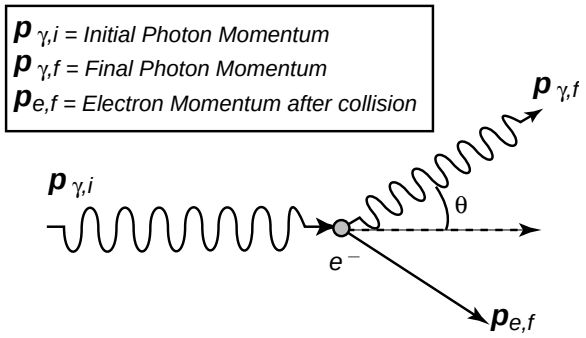


Figure 9.2: Scattering of CMB photons by free electrons.

   If the electrons have more energy than the photons, the photons gain energy from the electrons; if the photons have more energy, the reverse is true (this is known as inverse Compton scattering). This process acts to drive the photons and electrons to the same temperature. The cooling rate per unit volume (ergs cm^{-3} s^{-1}) is proportional to $n_\gamma n_e$.

   Because the magnitude of Compton cooling depends on the energy density in the CMB (since it depends not only on the number density of photons, $\propto (1+z)^3$, but also their momenta $\propto T_r \propto 1+z$), it is very dependent on redshift, scaling as $(1+z)^4$. For the redshifts at which galaxy-mass perturbations go non-linear ($z \lesssim 6$), Compton cooling is unimportant.

2. **Recombination:** When a proton and electron recombine, a photon is emitted:

$$e^- + p^+ \rightarrow {\rm H} + \gamma,$$

   where the photon energy $E \gtrsim 13.6$ eV. This energy is therefore lost from the gas (the electrons) and goes into the radiation field, which is no longer coupled to the matter for $z < z_{\rm dec}$. The cooling rate per unit volume is

proportional to $n_e n_p T^{-1/2} \propto n_e^2 T^{-1/2}$ (since, neglecting helium, $n_e = n_p$); the $T^{-1/2}$ dependence comes from averaging over a Maxwellian velocity distribution for the electrons. The *faster* the electron passes the proton, the less likely it is to recombine, given the shorter encounter time between the two. Since the typical velocity is porportional to $T^{1/2}$ for a Maxwellian velocity distribution, this leads to the $T^{-1/2}$ dependence.

3. **Bremsstrahlung (Free-Free):** Even if they do not recombine, an electron which passes sufficiently close to a proton will be accelerated by the electromagnetic attraction between the two; since an accelerated charged particle radiates, this leads to radiation, and again loss of energy from the gas; the magnitude of the energy loss is set by the charge-to-mass ratio of the electron. The cooling rate per unit volume for this process is proportional to $n_e n_p T^{1/2} \propto n_e^2 T^{1/2}$. The $T^{1/2}$ scaling again arises from the velocity scaling; it is positive in this case because the number of encounters between protons and electrons increases with T.

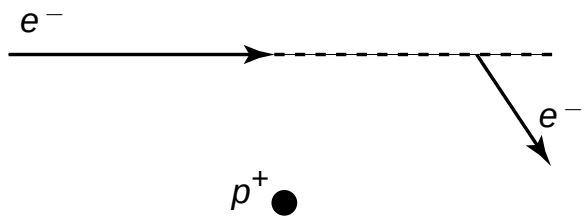


Figure 9.3: An electron accelerated by a proton and subsequently emitting radiation via the Bremsstrahlung effect.

The cooling rate per unit volume is usually written as

$$\dot{E} = n^2 \Lambda(T) \text{ ergs cm}^{-3} \text{ s}^{-1}, \tag{9.45}$$

where $\Lambda(T)$ is the *cooling function* (not to be confused with the cosmological constant!). For recombination and Bremsstrahlung cooling, $\Lambda(T)$ is independent of density. (This is true for most cooling processes, which is why Eq. 9.45 is defined this way.)

The cooling timescale is then

$$t_{\rm cool} = \frac{E}{\dot{E}} = \frac{3}{2}\frac{nkT}{n^2\Lambda(T)} = \frac{3}{2}\frac{kT}{n\Lambda(T)}. \tag{9.46}$$

The evolution of baryons after they fall into the dark matter potential wells and virialize then depends on the values of 3 timescales: the cooling time $t_{\rm cool}$, the dynamical time $t_{\rm dyn} (= t_{\rm ff})$, and the expansion time $t_E = H^{-1}$.

If $t_{\rm cool} > t_E$, then the baryons can have evolved very little since they fell into the DM perturbation. If $t_{\rm cool} > H_\circ^{-1}$, the present value of the expansion timescale, then the cooling time is greater than the age of the universe.

If $t_{\rm cool} < t_E$ but $t_{\rm cool} > t_{\rm dyn}$, where the dynamical (free-fall) time is (Eq. 9.14)

$$t_{\rm dyn} = \left(\frac{3\pi}{32G\rho}\right)^{1/2} = \frac{\pi}{2}\left(\frac{2GM}{R^3}\right)^{-1/2}. \tag{9.47}$$

(Note that, since $\Omega_{\rm DM} \gg \Omega_{\rm baryon}$ for interesting dark matter, it is the dark matter density which determines the dynamical time.) Then the baryons can cool, but they evolve quasi-statically (i.e., through a series of near-steady states) because the pressure has time to keep up with the dynamical evolution; thus the baryon distribution evolves on the timescale $t_{\rm cool}$.

If $t_{\rm cool} < t_{\rm dyn} < t_E$, then the baryons cool and collapse rapidly, on a dynamical timescale. Numerically,

$$t_{\rm cool} \sim 8\times10^6 \text{ yr} \left(\frac{n}{1 \text{ cm}^{-3}}\right)^{-1} \left[\underbrace{\left(\frac{T}{10^6 \text{ K}}\right)^{-1/2}}_{\text{Bremsstrahlung term}} + 1.5 \underbrace{\left(\frac{T}{10^6 \text{ K}}\right)^{-3/2}}_{\text{Recombination term}}\right]^{-1}. \tag{9.48}$$

Important point: $T_{\rm vir} \propto M/R \propto R^2\rho$, while $t_{\rm dyn} \propto \rho^{-1/2}$, $t_{\rm cool} \propto \rho^{-1}$. Thus the cooling and dynamical times depend only on ρ, but $T_{\rm vir}$ depends on both R and ρ.

9.6 Galaxy Formation

There are two ways to get the gas to cool on a dynamical timescale:

1. Keep $T_{\rm vir}$ low, so that $\Lambda(T)$ is large (i.e., make the total cooling rate large by making $\Lambda(T)$ large).
2. Make the density large (make the total cooling rate large by increasing ρ since $t_{\rm cool}$ drops faster with increasing ρ ($\propto \rho^{-1}$) than $t_{\rm dyn}$ ($\propto \rho^{-1/2}$).

Using the approximate cooling function (Eq. 9.48) and Eq. 9.40 for $T_{\rm vir}$, we find that the baryons will be able to cool on a dynamical timescale provided that:

1. $M \lesssim 10^{12}$ M$_\odot$ ($T_{\rm vir} \lesssim 10^6$ K)

or

2. $R \lesssim 100$ kpc ($T_{\rm vir} \gtrsim 10^6$ K).

Thus galactic masses and dimensions are naturally picked out by the cooling timescale. However this is only true if we have density perturbations on these mass scales! As we have seen, this is true of cold dark matter models, but not hot dark matter or baryon-only universes.

The evolution of structure is qualitatively different in cold dark matter (CDM) and hot dark matter (HDM)-dominated universes. In CDM models, density perturbations exist over a very wide range of mass scales, from $M \sim 10^{14}$ M$_\odot$ down to $M \sim 10^5$ M$_\odot$ (the lower limit depends on how "cold"—i.e., how massive—the CDM particles actually are), with comparable amplitude. As perturbations on small scales evolve, they also merge together

as part of larger scale perturbations. This is known as "hierarchical clustering," which simply means that structures build up from smaller scale (smaller mass) structures.

In the HDM scenario (and similarly in a baryon-only universe), the initial perturbations at $z = z_{\rm eq}$ peak at $M = M_{\rm FS} \approx 10^{14}$ M$_\odot$. Thus the initially forming structures all have masses around this value.

Because this corresponds to such a large size scale ($\sim$ tens of Mpc), the assumption of spherical symmetry which we made in discussing the dynamical times and virialization is poor for describing the evolution of these size scales. Basically, these perturbations have to collapse such a long way that any initial asymmetries in the shape and velocities is magnified during collapse. The most likely result is that collapse occurs first along one dimension, leading to a flattened, 2-dimensional structure known as a "Zel'dovich pancake," after the Russian physicist who first proved this result. Objects on smaller scales then form by further collapse and fragmentation of the pancake.

9.7 Correlation Function

Clearly the resulting distribution of objects—galaxies, clusters, and super-clusters—must be very different in these two scenarios. this is apparent from visual inspection of simulations of the evolution of CDM and HDM universes.

How can we quantify this? The most commonly-used statistical measure of the spatial distribution (and probably the most powerful) is the spatial *two-point correlation function.*

Suppose we have a spatial distribution of galaxies with average number density n per Mpc^3. Rigorously, when we say that the number density is n, we mean that the probability of finding a galaxy in a volume element dV is

$$dP = n\, dV. \tag{9.49}$$

If galaxies are randomly distributed with respect to one another, then the probability of finding galaxies within two volumes dV_1 and dV_2 simultaneously (the *joint* probability) is obtained from Eq. 9.49 in the usual way:

$$dP_2 = n\, dV_1\, n\, dV_2 = n^2\, dV_1\, dV_2. \tag{9.50}$$

Because we have assumed the positions of galaxies are uncorrelated, this depends only on the size of the two volume elements, and not on their spatial separation.

To make allowance for the possibility that galaxies actually are spatially correlated, re-write Eq. 9.50 for the joint probability with an additional term:

$$dP_2 = n^2\, dV_1\, dV_2\, [1 + \xi(r)], \tag{9.51}$$

where Eq. 9.51 is now the joint probability that galaxies are found in the two volumes dV_1 and dV_2, *separated by distance* r. $\xi(r)$ is by definition, the two-point correlation function. Obviously if $\xi(r) = 0$ for all r, then Eq. 9.51 reduces to the result for the random distribution Eq. 9.50.

We can define $\xi(r)$ in another, equivalent way: if we start with a position centered on a galaxy, then the *conditional probability* that we find another galaxy in a volume element

dV at distance r is

$$dP_c = n\,dV\,[1 + \xi(r)], \tag{9.52}$$

which reduces to Eq. 9.49 if $\xi(r) = 0$.

Observations show that $\xi(r) \neq 0$ for galaxies; the galaxy-galaxy correlation function

$$\xi_{\rm gg}(r) = \left(\frac{r_\circ}{r}\right)^{1.8}, \quad r_\circ = (5.4 \pm 1)\,h^{-2}\ {\rm Mpc}. \tag{9.53}$$

Eq. 9.53 holds over the range 10 kpc $\leq hr \leq$ 10 Mpc. 9.53 shows that the galaxy-galaxy correlation function falls to unity at $r = r_\circ = 5.4\,h^{-1}$ Mpc; on smaller scales, galaxies are highly correlated.

In addition, clusters of galaxies are also correlated with one another; the cluster-cluster correlation function is

$$\xi_{\rm cc}(r) = \left(\frac{r_\circ}{r}\right)^{1.8}, \quad r_\circ \approx 18\,h^{-1}\ {\rm Mpc}. \tag{9.54}$$

Under the assumption that only gravity is important in forming structure, it is straightforward (if computationally expensive) to follow the evolution of structure in CDM and HDM scenarios and compare them with observations. The results are summarized in the following two sub-sections.

9.8 Hot Dark Matter Models

Because of the elimination of all perturbations on scales less than a few tens of megaparsecs due to free-streaming of the HDM (usually assumed to be neutrinos), formation of galaxies depends on large scales going non-linear and fragmenting.

Because of this "top-down" mode of structure evolution, HDM models run into trouble when compared to the real universe. This is because scales much larger than galaxies go non-linear first; if we require that galaxy formation took place at some reasonable amount of time before the present (current observations suggest that the redshift of galaxy formation $z_f \gtrsim 1$), then the HDM models produce too much clustering. (The slope of the 2-point correlation function is too steep, and its amplitude is too high.) This is true even if one makes a distinction between the "neutrino" distribution (the total matter distribution) and the "galaxy" distribution (mass concentrations which have already collapsed).

In addition, one of the attractive features of an HDM universe—the formation of large sheets and filaments of galaxies, as seen in the real universe—is only a transient stage, as the Zel'dovich pancakes formed early on continue to collapse in the remaining two dimensions, forming first filaments and then quasi-spherical "clusters." The latter are far too massive to correspond to anything in the observed universe, so that some unknown mechanism would have to act to prevent visible (baryonic) matter from falling into these mass concentrations. For these reasons, hot dark matter models are generally not regarded as viable at present.

9.9 Cold Dark Matter Models

In this scenario there is plenty of power in density fluctuations on galactic and sub-galactic mass scales, so we do not run into the problem of having to wait for fragmentation on larger scales, as in HDM models.

However matching both the slope *and* the amplitude of the galaxy-galaxy correlation function requries $\Omega h \lesssim 0.2$. Thus either $h \lesssim 0.2$ for $\Omega = 1$, which is an unacceptably low value for the Hubble constant, or else $\Omega \sim 0.2$ for $h \sim 1$, which suffers from the usual arguments against $\Omega < 1$ (in particular, $\Omega = 1$ is predicted by inflation, which we will discuss in the next chapter). Equivalently, the predicted peculiar (non-Hubble flow) velocities of galaxies in this scenario are too large compared with observations. This is not surprising, since observations of the velocity dispersions in groups and clusters of galaxies give mass estimates corresponding to $\Omega \sim 0.1$ (discussed further below); thus if galaxies accurately trace the mass distribution, then observations show that $\Omega < 1$, and we do not live in a critical-density universe. The remedy to this problem is to assume that galaxies do *not* trace the mass distribution accurately, but only a subset of it.

The idea behind this is as follows. A point we have ignored so far is that, at a given time and a given size (or mass) scale, all fluctuations in density will not have the same amplitude; instead, there will be a range of values. In most models, this spread in the actual amplitudes of density fluctuations around the mean value is Gaussian in shape. In "biased" galaxy formation, it is assumed that only density peaks with amplitudes greater than some minimum threshold times the mean value actually become galaxies. Physically, this could be because only exceptionally high density peaks can capture *and* cool sufficient mass in baryons to form visible galaxies; in the absence of actual hydrodynamic simulations with realistic treatment of heating and cooling, this is only a plausibility argument.

In these CDM models with biased galaxy formation, the galaxies represent a more highly clustered subset of the total mass distribution. (This is because an exceptionally high-density peak in the density distribution is statistically more likely to occur in a region of higher than average density.) These biased CDM models can produce a good match to the galaxy-galaxy correlation function for $\Omega = 1$ and reasonable values of the Hubble constant.

However they still have difficulties in producing enough power on large scales ($M \gg M_{\rm gal}$) compared to galactic-mass scales. This is reflected in a too-small value of the cluster-cluster correlation function. A similar problem has arisen as a result of the COBE detection of fluctuations in the microwave background (on very big scales) at the 10^{-5} level: if the amplitudes of these fluctuations are used to set the normalization constant A in Eq. 9.39, then everybody's favorite $\alpha = 2/3$ (Harrison-Zel'dovich) CDM model produces too much power on small scales. Matching the normalization on large scales produces over-clustering on the galaxy-mass level. CDM models also tend to have trouble producing voids as large as those seen in the real galaxy distribution.

A truly realistic assessment of either HDM or CDM models requires an accurate inclusion of gas dynamical processes, heating and cooling, and feedback effects (e.g., the injection of energy in the form of radiation and supernovae once massive stars begin forming). Work along these lines is still very preliminary.

9.10 Variations on CDM

Recent trends in structure formation models include the following:

1. **Open Cold Dark Matter (OCDM):** These models can generally reproduce the sponge-like distributions of matter seen in surveys. The difference between normal CDM and OCDM is that for the latter, the epoch of matter and radiation equilibrium occurs later, and hence the growth of structure occurs over a smaller range in z. The fluctuation spectrum has less power at smaller scales. The only problem with these models is that they conflict with inflationary theory which predicts $\Omega = 1$.

2. **ΛCDM:** These have become very trendy, usually with $\Omega_{\rm CDM} \approx 0.2$ and the rest provided by Ω_Λ so that $\Omega_{\rm CDM} + \Omega_\Lambda = 1$. They give qualitatively similar dynamics as the OCDM models. They do have the advantage of producing an older universe for a given value of $H_\circ$, because the cosmological constant stretches out the expansion timescale. This avoids the over-clustering at late times which troubles $\Omega_{\rm CDM} = 1$ models. They suffer from horrible fine-tuning problems, however.

3. **Mixed Dark Matter (HCDM):** These usually have $\Omega_{\rm HDM} \approx 0.2$ and $\Omega_{\rm CDM} \approx 0.8$. The HDM provides the large-scale power and the CDM provides the small-scale power. These are computationally very difficult, because of the interaction between the hot and cold dark matter.

4. **Cold Dark Matter with Decaying Neutrinos (τCDM):** The objective of this class of models is to increase the radiation to matter energy densities so that $z_{\rm eq}$ for matter-radiation equilibrium occurs later (i.e., at smaller values of z), similar to the OCDM models. However if you have more families of relativistic particles, Big Bang nucleosynthesis would result in excessive helium production. The τCDM models therefore propose relativistic particles which decay away before the epoch of nucleosynthesis. Whether such particles really do exist is of course another question altogether.

Chapter 10

Inflation

10.1 A Solution for Problems with the Big Bang?

As discussed previously in §8.4 (pp. 119–123), although the Big Bang model is highly successful (especially in predicting the light element abundances), there are a number of problems that the standard model does not address. In fact, the traditional Big Bang theory can be thought of as really addressing what happens to the Universe *after* the Big Bang, and how it evolves from its originally hot, dense state. It does not say anything about why the Universe started expanding from this initial condition, what was it that expanded, or what might have preceded it. Although the concept of inflation was proposed to solve several of the problems in the standard Big Bang model, some theorists believe that it may also contain clues to these more fundamental questions about the origins of the Universe.

The central idea is that the Universe underwent a phase transition very early in its history, at $t \sim 10^{-35}$ seconds, when the temperature dropped below the point where it was possible to maintain symmetry between the strong nuclear force and the electroweak force. This phase transition lasted about 10^{-34} seconds.

Why was this important? Consider a more familiar example, the freezing of water to ice. If we take a container of water at room temperature, say, and steadily remove heat from it, the temperature of the water will drop until it reaches the freezing point. At this point, although we continue to remove heat, the temperature remains constant while the fraction of ice increases. Only *after* all the liquid water in the container has frozen does the temperature continue to drop. The latent heat of crystallization of the ice keeps the temperature constant until the phase transition is complete.

A constant energy density behaves just like a positive cosmological constant; when this term is dominant, the scale factor expands exponentially,

$$a \propto e^{(\Lambda/3)^{1/3} t}$$

(cf. Eq. 7.102). Since the epoch of inflation lasts for a factor of 100 in age, the expansion of the Universe can be huge: in typical inflation models, a grows from as little as $\sim e^{70} \sim 10^{30}$ to as large as $\sim e^{100} \sim 10^{43}$. This enormous rapid expansion solves some of the biggest difficulties with the standard Big Bang model:

By Ka Chun Yu and Phil Maloney.

1. **The horizon (or smoothness) problem:** prior to the epoch of inflation, our presently observable Universe *was* a causally connected region. Particles inside this region therefore have the opportunity to attain a uniform, homogeneous state. It is only *after* inflation then that the particles are driven so far apart from the exponential expansion that they cannot communicate with or affect each with light signals. So it is not difficult to understand why the Universe appears to be so homogeneous and isotropic, as shown by the microwave background. Although all of the volume enclosed by the observable CMB was not causally connected at the time of recombination, it was so even earlier in time, before inflation.
2. **The flatness problem:** Whatever the curvature of the Universe *before* inflation, the enormous expansion of the Universe during the inflationary era will have made it asymptotically flat. Of course, it is not possible to transform a $k \neq 0$ Universe into a $k = 0$ Universe, and at late enough times, when the particle horizon exceeds the size of the inflated region, we would be able to tell that $k \neq 0$, so it does not solve this problem forever. It does, however, solve it for a *very* long time (orders of magnitude longer than the current age of the Universe, for typical inflation models).
3. **Hidden relics problem:** Many Grand Unified Theories (GUT) predict the creation of magnetic monopoles, gravitinos, and other exotic particles after the phase transition when the strong nuclear force breaks from the electroweak force. We should expect to see enormous numbers of these exotic particles, as well as topological defects like cosmic strings and domain walls in the present-day Universe. In fact, the density of magnetic monopoles should exceed $\Omega = 1$. Inflation however expands the bulk of the exotic particles out of our particle horizon, so there is at most one magnetic monopole inside the present horizon.

(We will cover other problems that inflation addresses in §10.2.1.)

The horizon problem returns in a slightly different form in discussing the origin of perturbations. A perturbation of physical size $\lambda_\circ$ today had a proper length $\lambda_\circ(a(t)/a_\circ)$ in the past; $a(t)/a_\circ \propto t^n$, where $n < 1$. For the radiation-dominated phase, $n = 1/2$, for a matter-dominated Universe, $n = 2/3$.

The size of the horizon is

$$R_H \approx \frac{c}{H(t)} = c\left(\frac{a}{\dot{a}}\right) = \frac{ct}{n}$$

Since this scales as t, the ratio

$$\frac{\lambda(t)}{c\,H^{-1}(t)} \propto t^{n-1}$$

Since $n - 1 < 0$, this ratio increases with decreasing t; at early enough times, any perturbation will be bigger than the horizon. As we have discussed earlier, with every size scale λ we can associate a mass scale:

$$M(\lambda) = \frac{4\pi}{3}\bar{\rho}(t)\,\lambda^3(t) = 1.5 \times 10^{11}\ \mathrm{M}_\odot\,\Omega h^2 \left(\frac{\lambda}{0.5\ \mathrm{Mpc}}\right)^3$$

where $\bar{\rho}(t)$ is the mean density of the Universe. This mass scale is independent of redshift since $\bar{\rho} \propto a^{-3}$ while $\lambda \propto a$.

The size scale of a perturbation of a given mass will be bigger than the horizon at all redshifts $z > z_{\rm enter}(M)$, where

$$\begin{aligned} z_{\rm enter}(M) &= 1.41 \times 10^5 (\Omega h^2)^{1/3} \left(\frac{M}{10^{12}\ {\rm M}_\odot} \right)^{-1/3}, \\ & \qquad M < M_{\rm eq} \approx 3.2 \times 10^{14}\ {\rm M}_\odot\, (\Omega h^2)^{-2} \left(\frac{T_{R_\circ}}{2.75\ {\rm K}} \right)^6 \\ &= 1.1 \times 10^6 (\Omega h^2)^{-1/3} \left(\frac{M}{10^{12}\ {\rm M}_\odot} \right)^{-2/3}, \\ & \qquad M > M_{\rm eq} \end{aligned}$$

$M_{\rm eq}$ is the mass corresponding to

$$\lambda_{\rm eq} \approx 13\ {\rm Mpc}\, (\Omega h^2)^{-1} \left(\frac{T_{R_\circ}}{2.75\ {\rm K}} \right)^2,$$

the size scale which crosses (enters) the horizon at $t = t_{\rm eq}$. Perturbations on scales with $\lambda < \lambda_{\rm eq}$ enter the horizon at $t < t_{\rm eq}$ ($z > z_{\rm eq}$).

From the above estimate, perturbations on $\sim$ galaxy-mass scales were bigger than the horizon for $z \gtrsim 10^6$. How can causal processes have operated on super-horizon sized scales?

Inflation provides a way out of this problem. Suppose the Universe is radiation-dominated up to some time t_i, but expanded exponentially in the interval $t_i \leq t \leq t_f$; after t_f, the Universe returns to radiation-dominated, and then (eventually) matter-dominated. Thus,

$$a(t) = a_i\, e^{H(t-t_i)}, \quad t_i \leq t \leq t_f$$

where $a_i = a(t_i)$.

This has drastic consequences. All physical size scales increase exponentially, but the horizon size is unchanged:

$$H(t) = \frac{\dot{a}}{a} = \frac{H\, a_i\, e^{H(t-t_i)}}{a_i e^{H(t-t_i)}} = H,$$

so the horizon size remains constant during inflation. This means that a given length scale can actually cross the horizon twice in inflationary models.

For inflation that is driven by the GUT phase transition, the start of the inflationary period is at $t_i \approx 10^{-35}$ seconds, while it ends at $t_f \approx 10^{-33}$ seconds. For example, consider a scale $\lambda_\circ \sim 1$ Mpc ($M \sim 1.2 \times 10^{12}\ {\rm M}_\odot\ \Omega h^2$), i.e., a typical bright galaxy mass. At the end of inflation, this size scale was

$$\lambda(t_f) = \lambda_\circ \frac{a(t_f)}{a(t_\circ)} \simeq 1.8 \times 10^{-2}\ {\rm cm}.$$

This is much bigger than the horizon size at t_f:

$$\frac{c}{H(t_f)} \simeq 1.4 \times 10^{-24}\ {\rm cm}$$

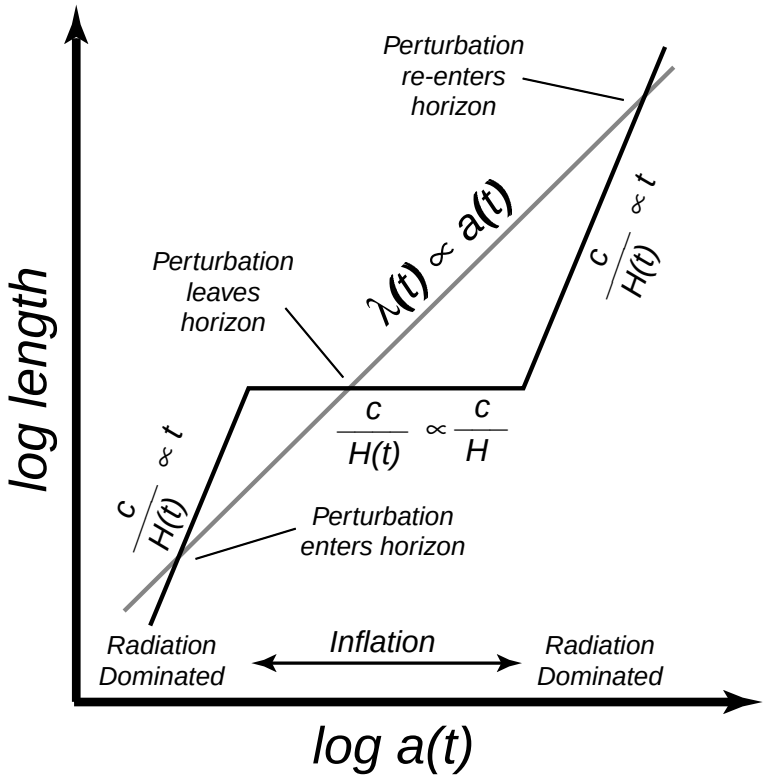


Figure 10.1: A perturbation in a Universe undergoing inflation.

However, before inflation, this size scale was smaller by a factor of

$$e^{-Ht_f} \simeq e^{-70} \approx 4 \times 10^{-31}$$

so that at t_i,

$$\lambda(t_i) \simeq 7 \times 10^{-33} \text{ cm}$$

which is much *smaller* than the horizon size at $t = t_i$.

10.2 The Origin of Inflation

We have explored how inflation can resolve many problems with the basic Big Bang model. However we have not discussed in any detail about the underlying physics behind the expansion. We basically need a short-term cosmological constant that starts the expansion period. Looking at the Friedmann equation for a flat ($k = 0$) Universe,

$$\dot{a}^2 = \frac{8\pi G}{3}\rho a^2 + \frac{\Lambda}{3}a^2, \tag{7.57}$$

notice that even in an empty Universe with $\rho = 0$, a positive cosmological constant Λ can still result in a net force acting on a test particle. Since this can occur without any matter or energy density, we attribute the repulsive force that results in expansion to the vacuum.

One can think of the vacuum as not just empty space, but as the ground state for any physical theory. The ground state is the lowest energy state, but it should also appear to be the same in all coordinate systems; i.e., *the vacuum is Lorentz invariant.* Instead of describing it as a vector field (as with electromagnetic theory) or with tensors (as in General Relativity), the vacuum must be described by a *scalar field.* It has been shown by Zel'dovich (1968) that a scalar field follows the following equation of state:

$$p_{\rm vac} = -\rho_{\rm vac} c^2 \tag{10.1}$$

This relationship follows from the 1st Law of Thermodynamics with the proviso that mass-energy density $\rho_{\rm vac}$ must be constant if the vacuum expands or is compressed:

$$dE = dU + p\,dV = \rho_{\rm vac} c^2\,dV - \rho_{\rm vac} c^2\,dV = 0. \tag{10.2}$$

Classically the vacuum has the lowest energy state $\rho_{\rm vac} = 0$. However things are more complicated in quantum mechanics. Note that the simplest energy system in quantum mechanics is the harmonic oscillator, which has the potential

$$V(x) = \frac{1}{2} m\omega^2 x^2, \tag{10.3}$$

where we are following a particle with mass m that oscillates along the x direction. However, the energy that a particle can have is quantized, so that the possible energies are

$$E_n = \frac{1}{2}\hbar\omega + n\hbar\omega, \tag{10.4}$$

with $n = 0, 1, 2, \ldots$. The lowest possible energy is therefore

$$E_0 = \frac{1}{2}\hbar\omega. \tag{10.5}$$

This is the *zero-point energy* of the vacuum. Alternatively, this can be thought of simply as a consequence of Heisenberg's Uncertainty Principle, where at the level of "empty" space, virtual particle-antiparticle pairs appear and disappear. In quantum field theory, the vacuum field can be interpreted as a collection of harmonic oscillators of all frequencies. The vacuum energy is a sum over all possible contributing modes:

$$E_0 = \sum_j \frac{1}{2}\hbar\omega_j, \tag{10.6}$$

where the total energy is computed by putting the system in a box with volume L^3, letting $L \to \infty$, and summing up over all modes. Standard periodic boundary conditions are set up so that we add only the allowed wavenumbers

$$k_i = \frac{2\pi}{\lambda_i} = \frac{2\pi n_i}{L}, \tag{10.7}$$

where n_i is an integer. If we add up all possible contributing wavelengths, the vacuum energy E_0 diverges as $k_i \to \infty$. However we have good reason to believe that quantum mechanics starts to break down at large enough energies, so we do not have to integrate out to infinite wavenumbers.

10.2.1 The Planck Era

We can guess when quantum mechanics as we know it breaks down. As we saw earlier in classifying the possible Friedmann Universes (§ 7.9 starting on p. 93), using General Relativity to evolve the Universe backward in time, almost all model Universes begin in an initial singularity, with $a = 0$, $\rho = \infty$. The precise nature of the singularity depends on the geometry of the Universe.

$k = +1$ (Closed Universe): A *finite* amount of matter is packed into *zero* proper volume. *All of spacetime* is packed into this "point" singularity; there is nothing outside it.

$k = 0$ or -1 (Open Universe): An *infinite* amount of matter is packed into *infinite* proper volume, so the singularity is "everywhere." This is a consequence of the unchanged topology of the Universe: an infinite (open) Universe is *always* infinite, even initially.

An initial singularity arises as we try to extrapolate back to $t = 0$ using General Relativity. As $t \to 0$, radiation dominates over matter, and so the dynamics is just that of the radiation-dominated era, the expansion rate is

$$H = \frac{\dot{a}(t)}{a(t)} = \frac{1}{2t} \tag{10.8}$$

When the temperature of the radiation is T, the characteristic *radiation frequency* $\omega = 2\pi\nu$ is given by

$$\begin{aligned} h\nu &= \hbar\omega = kT \\ \Longrightarrow \omega &= \frac{kT_r}{\hbar} = \frac{k}{\hbar}\left(\frac{3c^2}{32\pi G\sigma}\right)^{1/4} t^{-1/2} \end{aligned} \tag{10.9}$$

where we have used Eq. 8.16 to relate t and T in the radiation-dominated era; σ is the Stefan-Boltzmann constant.

The expansion rate *cannot* exceed the frequency ω without quantum effects becoming important. This restriction $H(t) < \omega$ means that General Relativity breaks down for times earlier than

$$t_{\rm Pl} \approx \frac{\hbar^2}{4k^2}\left(\frac{32\pi G\sigma}{3c^2}\right)^{1/2} = \pi\left(\frac{hG}{45c^5}\right) \sim 10^{-43}\ \text{s}, \tag{10.10}$$

which is roughly the *Planck time*. At $t_{\rm Pl}$, the temperature of the Universe is $T \sim 5 \times 10^{31}$ K.

Alternatively we can think of this as the time when quantum effects and gravitational effects become equally important. Remember that in quantum mechanics, we can always associate a *Compton wavelength* λ with a particle of mass m by the relationship

$$\lambda_{\rm C} = \frac{\hbar}{mc} \tag{10.11}$$

Gravitational effects become extremely important at distances on the order of the Schwarzschild radius. Equating these two fundamental length scales from quantum mechanics and General Relativity,

$$\left(\lambda_{\rm C} = \frac{\hbar}{mc}\right) = \left(r_{\rm Schw} = \frac{2Gm}{c^2}\right), \tag{10.12}$$

we can define a Planck mass:

$$m_{\rm Planck} = \sqrt{\frac{\hbar c}{G}} = 1.22 \times 10^{19}\ {\rm GeV}\, c^{-2}, \tag{10.13}$$

a corresponding length scale, the Planck length:

$$l_{\rm Planck} = \frac{\hbar}{m_{\rm Planck} c} = \sqrt{\frac{\hbar G}{c^3}} = 1.62 \times 10^{-33}\ {\rm cm}, \tag{10.14}$$

and a Planck time scale:

$$t_{\rm Planck} = \frac{l_{\rm Planck}}{c} = \sqrt{\frac{\hbar G}{c^5}} \sim 5.31 \times 10^{-44}\ {\rm cm}. \tag{10.15}$$

Our knowledge of physics breaks down when we reach these mass, length, or time scales. Until we can come up with a theory of quantum gravity, we cannot describe any event prior to the Planck time, $\sim 10^{-43}$ sec, after the Big Bang.

10.2.2 Inflationary Phase Transition

The cosmological phase transition that drives inflation is thought to come as the result of new hypothetical particles that exist during the GUT era. This is analogous to the *Higgs field*, which was introduced to remove singularities in electroweak theory, and to give the $W^{\pm}$ and Z^0 bosons mass. Measurements at CERN of the masses of these particles have confirmed electroweak theory to a high degree of precision. The Higgs field is predicted to be a scalar field, which is exactly the type of force necessary to create a vacuum-driven expansion.

The basic idea is to postulate a particle, the "inflaton," which is important during the GUT era before the split between the strong and electroweak forces occurred. Following the work by Alan Guth, the founder of inflationary theory, let us suppose that the inflaton has a temperature-dependent quantum field φ and potential $V(\varphi, T)$. At a high temperature greater than $T_{\rm crit}$, the potential has a minimum at $\varphi = 0$. The Universe would settle into this minimum value, and $\varphi = 0$ would gradually pervade all of spacetime.

Suppose that the potential at $\varphi = 0$ is non-zero, so that $V(0, T > T_{\rm crit}) > 0$. All observers would observe this same value for the scalar field in all reference frames, so we may think of $V(\varphi, T)$ as being a property of the vacuum. However the field would actually fluctuate due to the Heisenberg Uncertainty principle around its *vacuum expectation value*

$$\langle \varphi_{\rm crit} \rangle = 0,$$

while the potential energy fluctuates around the mean *vacuum energy*

$$\langle V(0, T_{\rm crit}) \rangle > 0.$$

The value of this vacuum energy can be equated with Λ in Friedmann's equation, and will thus give the repulsive force that drives an exponential cosmological expansion.

This expansion cannot go on forever (or else our Universe today would look very different!). The key here is that the potential depends on temperature, and the expansion can drive the energy density of the Universe low enough so that $V(\varphi, T)$ will significantly change its shape. It is assumed then that in the hot GUT era, the potential was symmetric about $\varphi = 0$, and as the Universe cooled and went through its GUT phase transition, the potential developed additional minima at $\varphi = \pm\phi_{\rm min}$. If the potential is at all like the Higgs potential, it can be described as

$$V(\varphi, T) = -(\mu^2 - aT^2)\,\varphi^2 + \lambda\phi^4, \tag{10.16}$$

where μ is the mass of the field, and a and λ are constant. The minima for V will occur at

$$\varphi_{\rm min} = \begin{cases} 0 & \text{for } T > T_{\rm crit} \\ \sqrt{\frac{\mu^2 - aT^2}{2\lambda}} & \text{for } T < T_{\rm crit}, \end{cases} \tag{10.17}$$

with the critical temperature

$$T_{\rm crit} = \frac{\mu}{\sqrt{a}}, \tag{10.18}$$

and

$$V_{\rm min} = \begin{cases} 0 & \text{for } T > T_{\rm crit} \\ -\frac{(\mu^2 - aT^2)^2}{4\lambda} & \text{for } T < T_{\rm crit}. \end{cases} \tag{10.19}$$

If the inflaton field is like the Higgs field in electroweak theory, then the scalar field is complex so it is defined over the φ_1 and φ_2 plane. The original symmetric $V(\varphi_1, \varphi_2, T)$ is paraboloidal in shape, while the new post-GUT potential has a 'Mexican hat' shape so that it has rotational symmetry. The "true" vacuum is found at the circle along the bottom of the hat, while the minimum in the original symmetric potential is a "false vacuum."

Note that the critical temperature when the phase transition occurs is $T_{\rm crit} \propto \mu$. Thus inflation will set in at the mass scale of the scalar field that produces inflation. For GUTs, this is roughly $mc^2 \sim 10^{15}$ GeV. This characteristic energy is expected to be prevalent about 10^{-34} sec after the Big Bang, and is often refered to as the GUT energy scale. In typical inflationary scenarios, the exponential expansion period lasts for about 100 times as long as this. Thus the scale factor of the Universe increased by a factor of roughly $e^{100} \approx 10^{43}$. The horizon scale at the start of inflation was only $r = ct \approx 3 \times 10^{-24}$ cm and so was inflated to a size of 3×10^{19} cm at the end of inflation. After inflation, the Universe would expand at much slower rates, so that this original horizon scale would now be 3×10^{44} cm, which is much larger than the present size of the observable Universe ($\sim 10^{28}$ cm).

By originating as a result of a GUT phase transition, inflation also manages to solve a number of other problems.

4. **The expansion of the Universe:** Inflation naturally explains why the Universe is expanding. The idea that the Universe is expanding because it has always expanded in the past is not very satisfactory. Inflation however forces the initial quantum fluctuation that is the beginning of the Universe into expansion.

5. **Structure formation in the Universe:** Inflation not only predicts the extremely uniform cosmic microwave background radiation that we observe today, but it also makes predictions about fluctuations in this smooth background. More precisely, quantum fluctuations in the inflationary period can create the seeds of the density perturbations which eventually evolve into the structures we have today in the present Universe. Although the specific details for how this works depends on the exact scalar potential used, a wide range of theories predict a Harrison-Zel'dovich type spectrum.
6. **Baryon-antibaryon asymmetry:** Sakharov (1967) outlined three rules that must hold for baryons to outnumber antibaryons:
   (a) The baryon number must be violated, so that more particles are formed than antiparticles.
   (b) C (charge conjugation) and CP (charge conjugation and parity) must be violated. This means particles and antiparticles must behave differently in certain reactions.
   (c) The matter-antimatter asymmetry must be created under non-equilibrium conditions. Similar arguments to the ones in § 8.3 for the primordial nucleosynthesis of the light elements are used. Since the masses of a proton and antiproton are expected to be exactly the same, thermal equilibrium would imply that they are created in equal numbers.

   These rules may be satisfied by observations in particle physics as well as conditions in the early Universe. The first rule comes naturally out of the symmetry breaking in GUTs (with a secondary prediction that the proton must be unstable, with lower limits on the decay time of $\gtrsim 10^{32}$ yrs). C and CP violation has been observed in the decay of neutral K^0 and $\bar{K}^0$ mesons, with a slight imbalance of one extra matter particle for every 1000 decays—much greater than the $\sim 10^{-8}$ asymmetry necessary for baryogenesis. The details by which the baryons and antibaryons reach their final asymmetry are not entirely understood, although the typical theory involves a hypothetical massive boson and antiboson which are involved with the unification of the strong and electroweak forces. After symmetry breaking and the GUT phase transition, they decay into the final baryon states with the matter-antimatter asymmetry.

Many people have worked on inflation theory since its initial proposal by Guth (1981). The main difficulty is choosing the right potential $V(\varphi, T)$. Usually what people have done is to work backwards from what type of inflation is necessary to explain observations, and then to derive the scalar field and reconstruct the inflation potential from these requirements. Problems with Guth's classical inflation model, involving quantum tunneling through the potential, resulted in non-stop inflation and a Universe that would continue to expand exponentially forever. Thus other variants have been proposed over the years, including Andre Linde's chaotic inflation which doesn't fix the initial minimum of the false vacuum at $\varphi = 0$, but instead assigns it a random, fluctuating starting value φ_a. However no single model has been completely satisfactory, and so new versions of inflation continue to be

developed, often in what seems to be a rather ad hoc manner, although they all have basic foundations in modern quantum and gravitation theory.

Note then that there is not a single theory of inflation, but actually a class of theories. Although inflation solves many fundamental problems of the Big Bang, this feature does not guarantee that inflation actually occurred. The very ad hoc nature of the many variants of inflation may even argue against the theory as a whole. However many physicists would agree that the inflation model solves more problems than it introduces, and some variant of inflation is here to stay.

10.3 The Earliest Phase of the Universe

We have seen in previous chapters that the earlier we look back in time after the start of the Big Bang, the higher the temperatures and densities are observed. Let's summarize the most significant events in the early universe:

$\sim 10^4$ K, 1 eV, $\sim 10^5$ yr after BB:

Formation of atoms; decoupling of matter and radiation.

$\sim 10^5$ K, 10 eV, $\sim 10^4$ yr after BB:

Domination of matter energy over radiation energy densities.

10^9 K, $\sim$ 90 keV, $\sim$ 3 min after BB:

Neutron decay becomes important; nucleosynthesis starts at this time and ends ~ 30 min later.

10^{10} K, $\sim$ 1 MeV, $\sim$ 1 sec after BB:

Neutrinos decouple; e^--e^+ pairs annihilate.

3×10^{10} K, $\sim$ 3 MeV, $\sim$ 0.1 sec after BB:

The weak interactions that interconvert n and p become unimportant, so that unequal numbers of n and p freeze out.

$\sim 10^{13}$ K, 1 GeV, $\sim 10^{-5}$ sec after BB:

Quark/hadron transition: quarks are confined into baryons and mesons.

$\sim 10^{15}$ K, $\sim$ 1 TeV, $\sim 10^{-12}$ sec after BB:

End of electroweak unification; the electromagnetic and weak nuclear forces split.

$\sim 10^{26}$ K, $\sim 10^{14}$ GeV, $\sim 10^{-33}$ sec after BB:

End of Grand Unification of the strong and electroweak force; origin of matter-antimatter asymmetry; creation of magnetic monopoles; era of inflation.

$\gtrsim 10^{32}$ K, $\gtrsim 10^{19}$ GeV, $\lesssim 10^{-43}$ sec after BB:

Unification of all forces including gravity: supergravity?, superstrings?, supersymmetry?, M-brane theory?

The limit of known physics occurs at times less than the Planck time $\lesssim 10^{43}$ sec. Hawking and Penrose (1988, 1989) have shown that at $t = 0$, singularities are unavoidable with the current theory of gravity, and the field equations of General Relativity break down. Hence, quantum gravity is absolutely necessary for understanding physics shorter than the Planck time. However this has not stopped physicists from speculating what might have occurred *before* this time, and even wondering what might have started off the Big Bang to begin with. These ideas are of course highly speculative, and there is no guarantee that any of them might become testable in the near (or far) future. Many of these ideas usually involve the Universe beginning as an acausal quantum fluctuation.

Remember that in quantum theory, a vacuum is never truly empty. Virtual particle-antiparticle pairs appear and disappear, "borrowing" energy from the vacuum, and then giving up this energy a short time later in an act of mutual annihilation, with the only requirement that the Heisenberg Uncertainty principle must be obeyed:

$$\Delta E \, \Delta t \gtrsim h \tag{10.20}$$

Before the Big Bang, spacetime may not exist, but the laws of quantum mechanics are thought to still operate. Thus mini-universes may be popping in and out of existence in a kind of quantum "foam." For one of these virtual universes to actually grow into a "real" universe requires inflation to take hold. The probability for inflation to start is irrelevant; it only has to take hold once, in all of eternity, after which it inflates the Universe.

Here is a possible variant origin for the Universe based on the above ideas. Guth (2001) has recently developed a model for "eternal inflation." Inflation ends once the Universe cools to the point where the phase transition ends, and the GUT symmetry is broken. However the rate at which this occurs can be thought of as quantum-like probability, that can be expressed with a half-life. The scalar potential can decay at slightly different rates, and hence end its phase transitions in different parts of the Universe at different times. Thus one part of the expanding spacetime bubble may end its inflationary phase and continue expanding normally, forming a *pocket universe.* The other parts continue their exponential expansion, until these inflationary pockets also subdivide themselves into other pocket universes that are no longer inflationary, and inflationary pockets. This continues *ad infinitum* until you get an infinite tree of reproducing Universes, all separate from each other (since inflation has expanded them all outside of each others' horizons).

Chapter 11

Analyzing the Cosmic Microwave Background

After subtracting of the dipole anisotropy, residual variations in the CMB are found at the $\delta T = 29 \pm 1$ μK level, or

$$\frac{\delta T}{T} = 1.06 \times 10^{-5} \tag{11.1}$$

level. Thus the CMB is amazingly isotropic, although it is not completely smooth. We can describe the level of fluctuations by *spherical harmonics.* Spherical harmonics can be thought of as the modes of vibration of a sphere, where each harmonic produces a pure "note" of a definite frequency.

The normal procedure in describing the CMB fluctuations is to examine the changes in temperature from point to point in the sky. This distribution of T is then expanded as a sum of spherical harmonics:

$$\frac{\delta T(\theta, \phi)}{T} = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=+l} a_{lm} Y_{lm}(\theta, \phi), \tag{11.2}$$

where θ and ϕ are the usual spherical angles, and the normalized functions Y_{lm} are defined by

$$Y_{lm}(\theta, \phi) = \left[\frac{2l+1}{4\pi} \frac{(l-|m|)!}{(l+|m|)!} \right]^{1/2} P_{lm}(\cos\theta) e^{im\phi} \times \begin{cases} -1^m & \text{for } m \geq 0, \\ 1 & \text{for } m < 0, \end{cases} \tag{11.3}$$

where $P_{lm}(\cos\theta)$ are the associated Legendre polynomials of order l. This expansion into spherical harmonics is analogous to a Fourier decomposition of a wave into plane wave elements. For each wave number l, there are $m = 2l + 1$ separate modes producing the same "note," where $-l \leq m \leq +l$. Thus for the quadrupolar mode of $l = 2$, there are five different spherical harmonics functions: $Y_{2,-2}, Y_{2,-1}, Y_{2,0}, Y_{2,1}, Y_{2,2}$.

By Ka Chun Yu

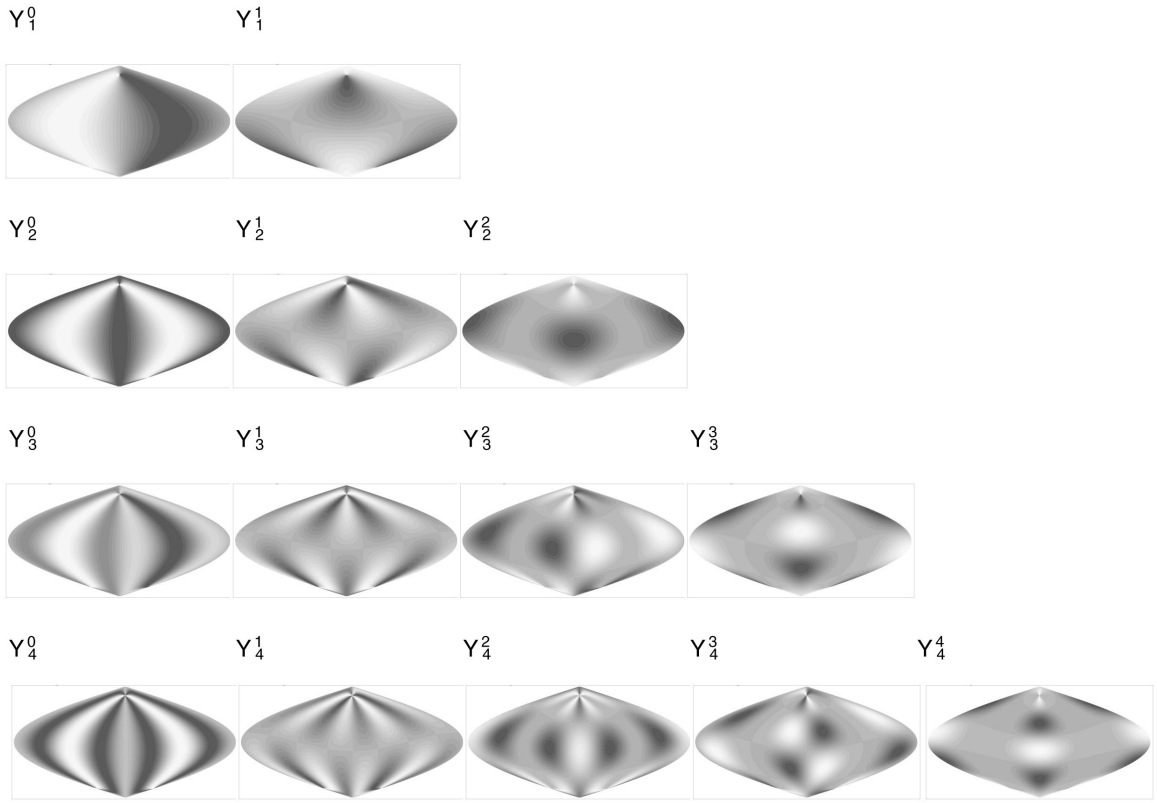


Figure 11.1: A diagram showing the amplitudes of spherical harmonics with $l = 1, 2, 3, 4$ and positive m. Each function is defined over the entire surface of a sphere, and hence is shown in its entirety via a sinusoidal projection, with the poles at the top and bottom. The $l = 1$ modes are dipoles; the $l = 2$ modes are quadrupolar modes. The $l = 0$ mode is not shown since it is just an amplitude without any angular dependences.

The Y_{lm} (Fig. 11.1) are a complete orthonormal set of functions on a sphere, so that

$$\int_{4\pi} Y^*_{lm} Y_{l'm'} d\Omega = \delta_{ll'}\delta_{mm'}, \tag{11.4}$$

where the asterisk means a complex conjugate, the integral is taken over the whole sky (i.e., for a spherical element of solid angle, $d\Omega = \sin d\theta\, d\phi$), and δ_{ij} is the Kronecker δ-function (which is $= 1$ if $l = l'$ and $m = m'$, and $= 0$ otherwise). If we use the orthogonality condition Eq. 11.4, the values for a_{lm} can be found by multiplying the temperature distribution over the sphere by Y^*_{lm} and integrating over the sphere:

$$a_{lm} = \int_{4\pi} \frac{\delta T}{T}(\theta, \phi) Y^*_{lm} d\Omega. \tag{11.5}$$

The a_{lm} are generally complex and follow the condition,

$$\langle a^*_{l'm'} a_{lm} \rangle = C_l \delta_{ll'}\delta_{mm'}, \tag{11.6}$$

where C_l is the *angular power spectrum*:

$$C_l \equiv \frac{1}{2l+1} \sum_m a_{lm} a_{lm}^* = \langle |a_{lm}|^2 \rangle. \tag{11.7}$$

Thus the power spectrum measures the mean square temperature fluctuation in the $2l+1$ spherical harmonic modes at each l. The analyses start with the quadrupole mode $l = 2$ since the $l = 0$ monopole mode is just the mean temperature over the observed part of the sky, and the $l = 1$ mode corresponds to the dipole anisotropy. Higher multipoles correspond to fluctuations on angular scales

$$\vartheta \simeq \frac{60^\circ}{l}. \tag{11.8}$$

Thus for higher angular resolution observations, more terms of high l must be included to describe the power spectrum.

Do we expect fluctuations in the CMB? We see today that matter in the universe is not homogeneously distributed, but it is collected into galaxies, groups and clusters of galaxies, and super-clusters, with large voids in between. We would expect the CMB to contain lumpy seeds of the cosmic structures that we see today.

If we were to follow photons from the surface of last scattering at recombination, photons climbing out of a gravitational well (in regions of higher density) would be redshifted. Those photons coming from a region of low density "rolls down" a gravitational potential and hence are blueshifted. While the photons traverse across the universe to reach us, they may run across additional pockets of matter, but the blueshift going in is compensated by a redshift climbing out (unless the gravitational potential changes during the traverse), so the photons' frequencies should not be affected after recombination. Photons also suffer time dilation compared to unshifted photons.

CMB photons thus preserve a "memory" of the density fluctuations from emission from the surface of last scatter (LS). The combination of gravitational redshift and time dilation is known as the *Sachs-Wolfe effect*, where both effects contribute to $\delta T/T$ in a way that is linearly dependent on $\delta\rho/\rho$. We found in Eq. 3.38 that the shift in frequency for photons climbing out of a gravitational potential well is:

$$\frac{\delta\nu}{\nu} = \frac{\delta T}{T} = \frac{G\delta M}{dc^2} \approx \frac{\delta\phi}{c^2},$$

where $\delta\phi$ is the Newtonian gravitational potential and d is the size of the perturbation. For the time dilation term, we use Eq. 7.20 that tells us the cosmic scale factor a was smaller in the past when the radiation was emitted:

$$\frac{\delta T}{T} = -\frac{\delta a}{a}. \tag{11.9}$$

In the standard models in the matter-dominated era, density fluctuations grow as $\delta\rho/\rho \propto a$. For a flat universe, $a \propto t^{2/3}$ (Eq. 7.67), or $\delta\rho/\rho \propto t^{2/3}$. Thus the cosmic scale factor will

incrementally change with time as

$$\begin{aligned} \delta a &\propto \frac{2}{3} t^{-1/3}\, \delta t \\ \implies \frac{\delta a}{a} &= \frac{\frac{2}{3} t^{-1/3}\, \delta t}{t^{2/3}} \\ &= \frac{2}{3} \frac{\delta t}{t}. \end{aligned} \tag{11.10}$$

As shown in Eq. 3.38 though, $\delta\nu/\nu = -\delta t/t$ is just the Newtonian gravitational redshift. The net result of the gravitational redshift and the time dilation contributions is therefore

$$\frac{\delta T}{T} = \frac{\delta\phi}{c^2} - \frac{2}{3}\frac{\delta\phi}{c^2} = \frac{1}{3}\frac{\delta\phi}{c^2}. \tag{11.11}$$

The physical size of the perturbation d at redshift z corresponds to a physical size today of $d_\circ = d/(1+z)$. From $1 + z = a_\circ/a$ and $\rho \propto a^{-3} \propto (1+z)^3$, we know that

$$\frac{\delta\rho}{\rho} \propto a \propto \frac{1}{1+z}.$$

This can then be used to give us the size of the density perturbation at the time of decoupling:

$$\begin{aligned} \frac{\delta\rho}{\rho} &= \frac{\delta\rho_\circ}{\rho_\circ}\frac{1}{1+z} \\ \implies \delta\rho &= \frac{\delta\rho_\circ}{1+z}\frac{\rho}{\rho_\circ} \\ &= \delta\rho_\circ \frac{(1+z)^3}{1+z} \\ &= \delta\rho_\circ\, (1+z)^2. \end{aligned} \tag{11.12}$$

Since $\delta M \approx \delta\rho\, d^3$, it follows that

$$\delta\phi \approx \frac{G\,\delta M}{d} \approx G\delta\rho_\circ\, d_\circ^2. \tag{11.13}$$

The gravitational potential in Eq. 11.13 does not have any dependence on z; thus $\delta\phi$ at any redshift is the same as that estimated for the perturbation once it has evolved linearly to the present epoch.

In §9.3, we saw that there are a number of arguments which suggest that the primordial fluctuation spectrum,

$$\frac{\delta\rho}{\rho} \propto M^{-\alpha}, \tag{9.39}$$

is likely to be nearly flat, i.e., $\alpha = 0$ or $\alpha = 2/3$ for the Harrison-Zel'dovich spectrum. Along with $M \approx \rho_\circ\, d_\circ^3$, Eq. 9.39 will give

$$\begin{aligned} \delta\rho_\circ &\propto \rho_\circ M^{-\alpha} \\ &\propto d_\circ^{-3\alpha} \\ \implies \delta\phi &\approx G\delta\rho_\circ\, d_\circ^2 \\ &\propto d_\circ^{2-3\alpha}. \end{aligned} \tag{11.14}$$

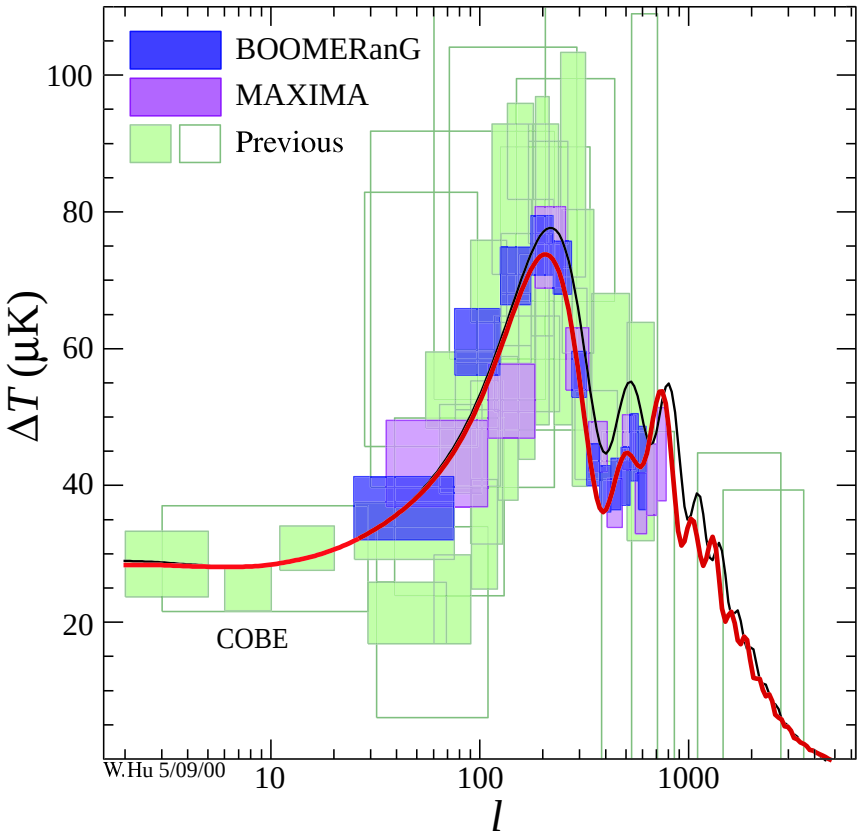


Figure 11.2: A summary of observations of the CMB power spectrum including the recent results from BOOMERanG and MAXIMA (2000), shown as blue and purple boxes. These two balloon-borne experiments measured patches of the sky (instead of the entire sky *á la* COBE) and are the best results to date for data that covers the angular range in the first acoustic peak of the power spectrum. The boxes show the size of the error bars as well as the width of the data bins. The solid lines represent different universe models.

If the perturbation now subtends an angle $\theta = d_\circ/D$ in the sky, then

$$\begin{aligned} \delta\phi &\propto \theta^{2-3\alpha} \\ \Longrightarrow \frac{\delta T}{T} &\propto \theta^{2-3\alpha} \propto \theta^0 = 1 \qquad \text{for } \alpha = 2/3. \end{aligned} \tag{11.15}$$

Thus the amplitude of the fluctuations due to the Sachs-Wolfe effect is independent of angular scale. (The Harrison-Zel'dovich perturbation spectrum is thus known as a *scale-invariant* spectrum.)

Fig. 11.2 shows a current compilation of measurements of the power spectrum. At $l < 30$ (i.e., large angular scales) in the points associated with COBE measurements, the spectrum is flat as predicted by the Sachs-Wolfe effect. In fact one can make a direct estimate of the perturbation power spectrum index α by analyzing the COBE data. This

has been done by a number of authors who derive $\alpha \approx 0.68$, which is remarkably close to the Harrison-Zel'dovich index of 2/3 (Bennett *et al.* 1996, Hancock *et al.* 1997).

Near $l = 200$, the power spectrum peaks in the first of a series of *Sakharov or acoustic peaks*.[1] The wavelengths associated with these higher order multipoles are smaller than the horizon at last scattering. Thus this region looks fundamentally different from $l \lesssim 100$ because different physical processes start to effect the inhomogeneities within the LS horizon.

The acoustic peaks are the result of density perturbations in the baryon and photon fluid which exists before recombination. The radiation pressure from the photons resists the gravitational compression due to the inhomogeneities in the fluid. This pressure reverses the compression until the perturbation overshoots its original size, making it more rarefied than the surrounding medium. The pressure from outside the perturbation then acts to reverse the motion of the perturbation. This cycle continues in a cycle of oscillating compressions and rarefactions. These oscillations due to the photon pressure can then be thought of as acoustic standing waves inside the LS surface horizon. They oscillate about regions of low and high energy density, with the shorter wavelengths of the potential fluctuation resulting in faster fluid oscillations, since the frequency of oscillation is given by

$$\omega_i = k_i c_s, \tag{11.16}$$

where k_i is the wavenumber of the oscillation, and c_s is the sound speed in the medium. The size of the mode k_1 is inversely related to the distance the sound wave can travel by recombination. Thus for a perturbation with a *sound horizon* at recombination of d_s,

$$k_1 = \frac{\pi}{d_s}.$$

A mode with twice the wavenumber (or half the wavelength) oscillates with twice the frequency, and hence this mode $k_2 = 2k_1$ can compress and then rarefy before recombination. Similarly the mode $k_3 = 3k_1$ can compress, rarefy, and compress again. These first three modes correspond to the first, second, and third acoustic peaks in the power spectrum.

At recombination, the photons begin to free-stream and will not be available to provide the pressure in the fluid. Thus the photon-baryon fluid stops oscillating. The modes that are frozen in an extremum of their oscillation will have enhanced temperature fluctuations. Waves that have completed a half-integral number of oscillations (every other mode) by the time of LS will be at maximum amplitude. The rest of the wave modes which were at mid-phase at LS will have smaller amplitudes. Thus variations in the oscillation phase with wavelengths rsults in a series of peaks as a funciton of l.

The exact form of the power spectrum depends on virtually every important parameter of the universe, including Ω and $H_\circ$. Thus if we were able to measure the power spectrum accurately out to high l, we would be able to constrain these parameters much more accurately. So how exactly do physical parameters of the universe effect the power spectrum?

In a flat universe, the dominant angular scale for CMB fluctuations, the angle subtended by the sonic horizon at the surface of last scatter is roughly 1°. In our temperature fluctuation spectrum, this corresponds to $l = 180$. If the universe is open, photons move on

[1] These peaks are also called "Doppler" peaks, although they do not have anything to do with Doppler shifts.

more rapidly diverging paths in negatively curved space. Due to this effect, the dominant angular scale for microwave shifts to smaller angular scales (and hence larger l). This result leads to Fig. 11.3 where the CMB power spectrum shifts to higher l.

The position of the first acoustic peak in the CMB power spectrum is most sensitive to the value of Ω. The peak moves to the right in proportion to $1/\sqrt{\Omega}$ for large values of Ω, with only a weak dependence on the Hubble constant $H_\circ$, Ω_b, and other cosmological parameters. Decreasing Ω to ~ 0.2 has the dramatic consequence of shifting the first acoustic peak from $l \approx 200$ to $l \approx 600$. The BOOMERanG and MAXIMA results thus strongly imply that $\Omega \approx 1$. Note however that the first acoustic peak *does not* tell you the actual breakdown of Ω into the various energy and mass densities, be it radiation, baryonic matter, CDM, HDM, or Λ. However measurements of dark matter in galaxy clusters imply that there isn't enough cold dark matter to get Ω up to 1; the upper limits to neutrino mass, as we have seen, also can contribute no more than a few percent. Any shortfall would have to be made up by a cosmological constant (or *dark energy* as it is popularly being called currently). This is supported by the independent line of evidence of an accelerating expansion from supernovae light curves.

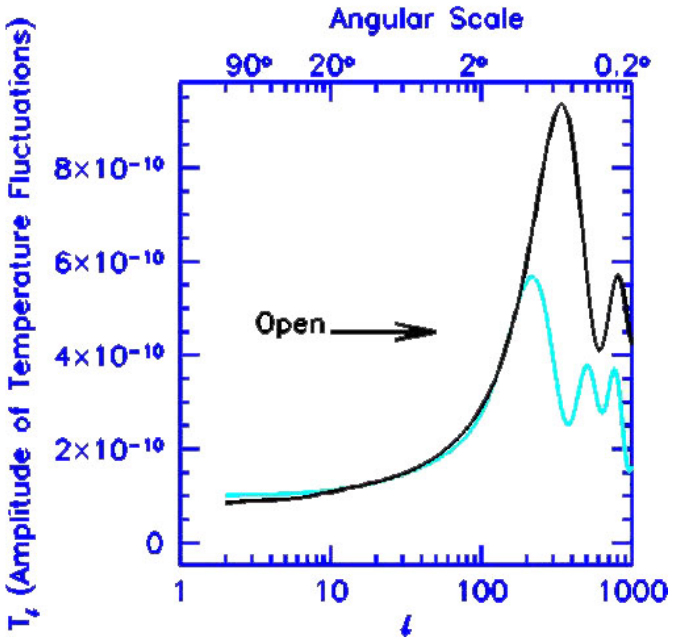


Figure 11.3: A comparison of power spectra from an open (black) and a flat (cyan) universe, a standard cold dark matter model.

The behavior of the early universe fluid will also depend upon the number densities of baryons relative to photons. Baryons increase the effective mass of the baryon-photon fluid that is oscillating. The greater gravitational potential leads to a larger compression of the fluid in the potential well. This is similar to a larger mass on a spring that results in a wider swing in oscillations. This increase in the amplitude of the oscillation does not change the

location of the zero point. The first, third, fifth, ... acoustic peaks are associated with how far the plasma is compressed or "falls" into the potential wells. Thus increasing the baryon density ρ_b will enhance the odd-numbered peaks. The even acoustic peaks however trace the movement of fluid outwards, or how much the plasma rarefies. Thus the even peaks will be relatively suppressed when the baryon-photon ratio goes up. By measuring the relative heights of the peaks in the fluctuation spectrum, we should be able to determine the density of protons and electrons relative to that of the radiation.

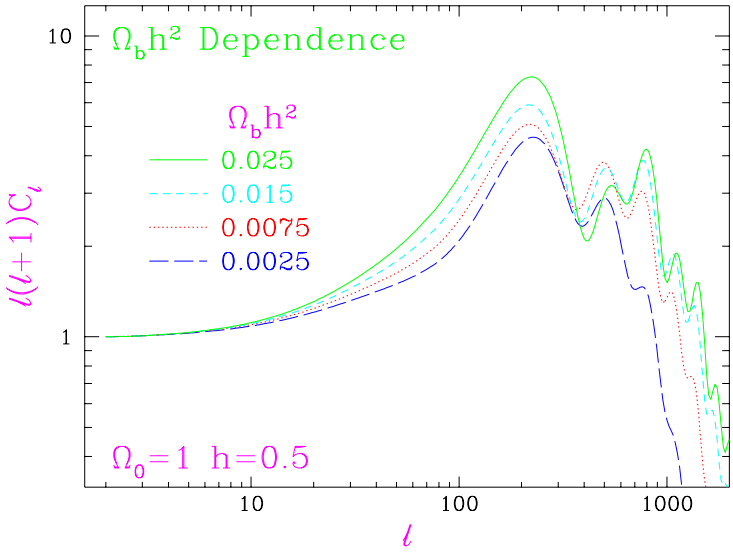


Figure 11.4: Increasing the ratio of baryon to photons results in different enhancements of the odd acoustic peaks with respect to the even peaks.

If there is a cosmological constant, then the depth of gravitational potential wells decay with time. Thus, a photon which falls into a deep potential well gets to climb out of a slightly shallower well. The net effect leads to a slight increase in photon energy along this path. Another photon which travels through a low density region (which produces a potential "hill") will lose energy as it gets to descend down a shallower hill than it climbed up. Because of this effect, a model with a cosmological constant will have additional fluctuations on large angular scales. Large angular scale measurements are most sensitive to variations in the gravitational potential at low redshift.

Additionally the higher acoustic peaks can also be used to measure the ratio of dark matter energy density to radiation density. This has to do with what happens to modes in the radiation-dominated versus the matter-dominated eras. Density fluctuations can damp easily during the radiation epoch as a result of photon viscosity, leaving the gravitational potential to decay away. The fluid bouncing back from a compression sees no potential to climb out of, and hence, the amplitude of the oscillation increases dramatically. Since the shorter wavelength modes start oscillating first, it is the larger l modes that will be driven in this manner. A comparison of the amplitudes of the high multipole peaks, most importantly the first three peaks, should give us the matter-energy density ratio ρ_m/ρ_r.

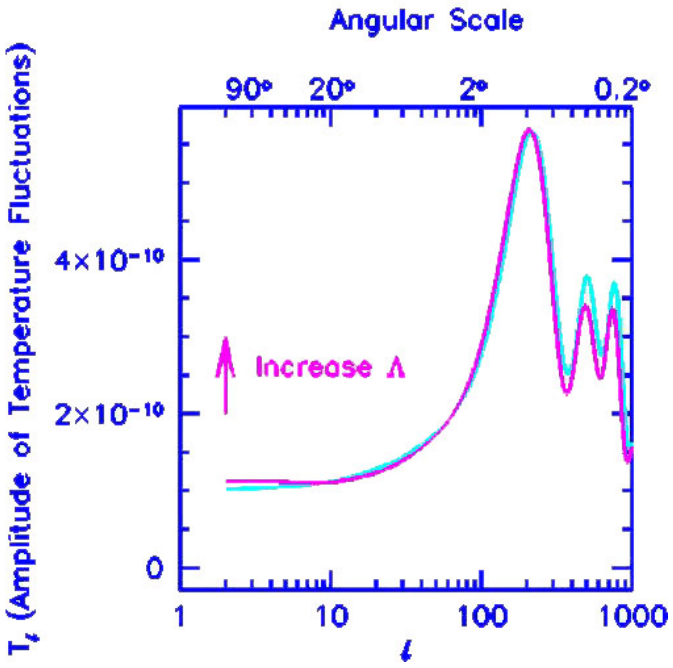


Figure 11.5: A cosmological constant reduces the depth of gravitational potential wells over time, resulting in additional fluctuations at large angular scales.

Finally the amplitudes of the highest wavenumber modes decrease quickly in the power spectrum as a damping tail. The fluctuations associated with these small angular scales have sizes that are comparable to the distance that photons travel during recombination. Note that recombination does not occur instantly, and so photons do have a chance to scatter for a bit longer. If photons random walk a distance equal to the size of the fluctuation before they start free-streaming, then hot and cold photons can mix and average out, resulting in a damping of the amplitude of these high order peaks. Increasing ρ_b will decrease the mean free path of the photons, and shift the damping tail to higher l. Increasing the matter density ρ_m will increase the relative age of the universe at recombination, and move the damping tail the other direction towards larger scales. Finally the curvature of the universe will also affect the location of these peaks in the damping tail. Measuring high multipole peaks and determining the location of the damping tail can therefore give a consistency check to the determinations of the baryon and matter densities, and the curvature from the lower l peaks.

Appendix A

Gravitational Lensing

A.1 Basics of Gravitational Lensing

We have seen in Ch. 4 that the theory of General Relativity predicts the deflection of light by a gravitating body. The amount of deflection was given by:

$$\alpha = -\frac{4GM}{bc^2}, \tag{ref113}$$

where b is the impact parameter, and M is the mass of the deflector. For small deflection angles, b is amost exactly the closest approach distance of the light rays to the deflector.

If the deflector is precisely aligned with the background source, then the light would be equally deflected above, below, and off to the side of the deflector. The observer would therefore see a circular ring of emission. The idea for this effect originated first with Chwolson (1924), and subsequently by Einsten (1936), although today, they are known as *Einstein rings*.

It is easy to calculate the size of the *Einstein ring radius*. Let D_d be the distance from the observer to the deflector, D_s be the distance to the source, and D_{ds} be the distance between the deflector and the source. If θ_E is the angular radius of the Einstein ring, then by simple geometry,

$$\theta_E = \alpha \left(\frac{D_{ds}}{D_s} \right), \tag{A.1}$$

where α is given by Eq. 4.75. Eq. A.1 is thus

$$\theta_E - \frac{4GM}{bc^2} \left(\frac{D_{ds}}{D_s} \right). \tag{A.2}$$

Since $b = \theta_E D_d$,

$$\theta_E^2 = \frac{4GM}{c^2} \left(\frac{D_{ds}}{D_s D_d} \right) = \frac{4GM}{c^2} \frac{1}{D}, \tag{A.3}$$

By Ka Chun Yu.

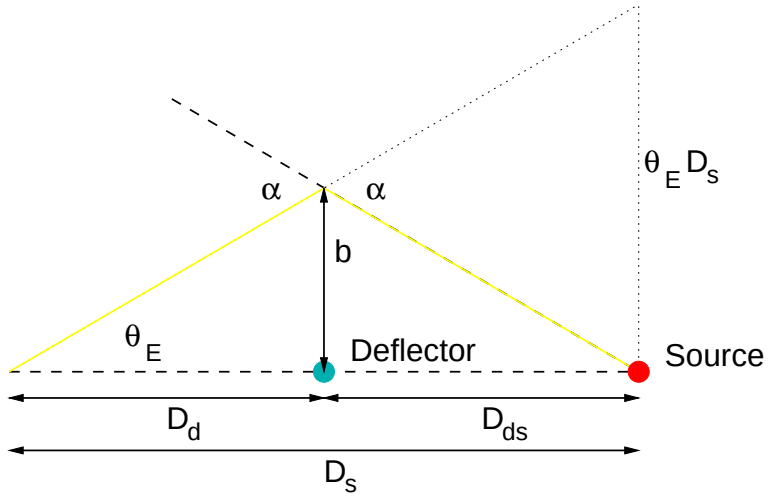


Figure A.1: The gravitational lensing geometry when the deflector and background source are aligned.

where $D = D_s D_d / D_{ds}$. The *Einstein angle* θ_E is therefore

$$\theta_E = \left(\frac{4GM}{c^2} \right)^{1/2} D^{-1/2}. \tag{A.4}$$

We have worked out this derivation assuming Euclidean geometry. However the relation in Eq. A.4 also works at cosmological distances if the D terms are *angular diameter distances*. This is a term often used in the literature and is defined as

$$D_A = \frac{D_m}{1+z}. \tag{A.5}$$

The usefulness of the above definition comes about in the expression for the angular size of an object derived in a Robertson-Walker metric. Here the *distance measure* is $D_m = |k|^{-1/2} \sin k^{1/2} R$ where k is the curvature constant in Eq. 7.11, and R is the spatial coordinate in the Robertson-Walker metric. The angular size of an object at cosmological distances is therefore

$$\Delta\theta = \frac{d}{D_A}, \tag{A.6}$$

where d is the proper length of the object. For small redshifts, $z \ll 1$, and Eq. A.6 turns into the Euclidean expression $d = r\Delta\theta$.

Eq. A.2 can be re-written so that the mass M is in solar masses and the distance D in 10^9 pc ($= 3.056 \times 10^{22}$ km):

$$\theta_E = 3 \times 10^{-6} \left(\frac{M}{\mathrm{M}_\odot} \right)^{1/2} \left(\frac{D}{10^9 \text{ pc}} \right)^{-1/2} \text{ arcsec.} \tag{A.7}$$

Thus clusters of galaxies with $M_{\rm cluster} \sim 10^{15}$ $M_\odot$, and located at cosmological distances, can have Einstein rings with sizes of tens of arcseconds across, which is easily observed.

The Einstein ring will appear if the background source and deflecting mass are precisely aligned. However as the below sequence shows, the appearance of the source changes in appearance depending on the angular separation between the two. Note that the following example is for a point source lens and a compact background object.

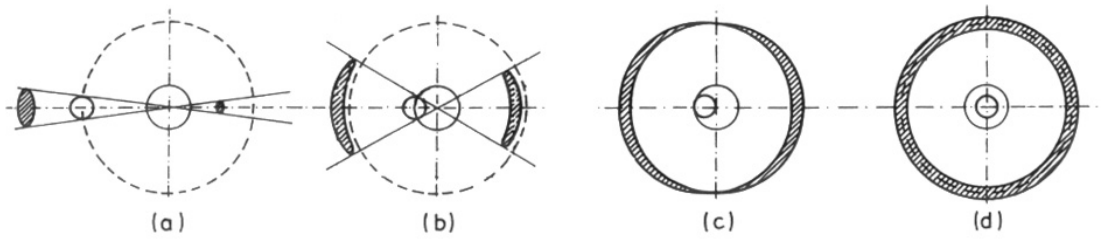


Figure A.2: The changing appearance of the background source as it passes behind a point mass. An Einstein ring is formed when the two sources are perfectly aligned.

Here the dashed circle is the location of the Einstein ring. The large and small solid circles denote the deflector and background source, respectively. As the angular separation of the two objects approaches the Einstein radius, a second image of the source appears on the opposite side of the deflector. Both images grow into arcs that merge into a circle when the objects are exactly aligned.

Gravitationally lensed arcs in galaxy clusters were first reported by Soucail *et al.* (1987) and Petrosian (1986). One of the most famous examples is the cluster Abel 2218 observed with the Hubble Space Telescope by Kneib *et al.* (1996). The rings seen there are incomplete and elliptical, as oppose to being exaclty circular. this suggests that the gravitational potential of the cluster is not precisely spherically symmetric, and the background galaxy is not aligned with the cluster center.

In general, for a compact object that is lensed by an extended source, the gravitational pontential is not spherically symmetric. The lensing effect is also not true lensing in the sense of geometric optics. The total delay in time from a lensed image is a combination of both a geometric delay as well as a gravitational time delay from time dilation. The appearance of the (multiple) image(s) of the background source can merge and/or split when the background object passes through *caustics*. The images formed as a result fo this orientationk tend to lie along *critical lines* or *curves*.

A.2 Microlensing

Gravitational lensing can occur for objects within our galaxy as well. For instance, let's say we observe a star near the center of our galaxy ($D_s \sim 8$ kpc) being lensed by a foreground star located halfway to the Galactic center ($D_d \sim 4$ kpc). Assuming a deflector mass

equivalent to the mass of the Sun, from Eq. A.4, the Einstein angle is

$$\theta_E = \left(\frac{4GM}{c^2}\right)^{1/2} D^{-1/2}\Bigg|_{M=2\times10^{33}\ \mathrm{g}} \approx 4\times10^{-4}\ \mathrm{arcsec}. \tag{A.8}$$

Any arcs or other examples of multiple images will be at this size scale, and will thus be too small to observe. However, there will still be a cumulative brightening of the background star as it passes close to the deflecting star. In fact, if the closest approach between the point mass lens and the source is $b \leq \theta_E$, the peak amplification of the source is

$$\mu_{\max} \gtrsim 1.34. \tag{A.9}$$

This corresponds to a brightening of 0.34 mag, which is easily observed.

The probability that any one star will pass close enough in angular separation to another star is small. As a result, observers must look at many stars (100,000s to millions) over a length of time to catch any chance lensing. Paczyński (1986) first proposed the monitoring of millions of stars in the Large Magellanic Cloud to look for lensing by stars within the halo of our Galaxy. In this way, one could map out the distribution of stellar-mass objects in our halo.

Because monitoring millions of stars will inevitably lead to the detections of variables, one has to separate stars with variability from stars underoing *microlensing*. Fortunately the light curves of stars that are being lensed should brighten and fade symmetrically with time, and the brightening should be achromatic, i.e., the same light curve should be seen at different wavelengths.

For a lens with a relative velocity (in the plane of the sky) v with respect to the source, the time-scale for microlensing light curve variation is:

$$t_\circ = 0.214\ \mathrm{yr} \times \left(\frac{M}{\mathrm{M}_\odot}\right)^{1/2} \left(\frac{D_d}{10\ \mathrm{kpc}}\right)^{1/2} \left(\frac{D_{ds}}{D_s}\right)^{1/2} \left(\frac{200\ \mathrm{km\ s^{-1}}}{v}\right). \tag{A.10}$$

The ratio $(D_{ds}/D_s)^{-1/2}$ is close to 1 if the lenses are in the Galactic halo and the sources in the LMC. There are several attempts currently to find MACHOs (Massive Compact Halo Objects) via microlensing. If the light curves are sampled anywhere from time intervals of an hour to a year, MACHOs in the mass range 10^{-6} $\mathrm{M}_\odot$ to 10^2 $\mathrm{M}_\odot$ can be potentially detected.

Appendix B

Weighing the Universe

With the data we currently possess, what do we actually estimate for Ω? We can subdivide the total Ω into two components:

$$\Omega = \Omega_m + \Omega_\Lambda, \tag{B.1}$$

where Ω_m is the contribution from matter and energy that has gravity, and Ω_Λ is from the vacuum energy. The former includes radiation (which we can estimate from the CMB), hot dark matter such as neutrinos, cold dark matter, and ordinary matter in the form of baryons. Recall Eq. 8.4 which gives the radiation component:

$$\Omega_r = \frac{4.72 \times 10^{-34}}{1.88 \times 10^{-29}} = 2.5 \times 10^{-5}.$$

Summing up all three neutrino families gives:

$$\Omega_\nu = 0.68\Omega_r. \tag{B.2}$$

The best evidence for the baryon density comes from Big Bang nucleosynthesis models. Burles, Nollett, & Turner (1999) give

$$\Omega_b h^2 = 0.020 \pm 0.002, \tag{B.3}$$

at the 95% confidence level. Actually attempting to measure the luminous and dark mass in galaxies requires a number of different methods. Adding up all the stars that we see gives just a tiny fraction:

$$\Omega_\star \approx 0.005 \pm 0.002. \tag{B.4}$$

Observations of spiral galaxies however suggest that there are extensive dark matter haloes that keep the galaxies' rotation curves flat out past the edge of the visible matter. The mass-to-light ratios for stars versus galaxies is:

$$\left.\frac{M}{L}\right|_\star = 1\text{–}3\ \frac{\mathrm{M}_\odot}{\mathrm{L}_\odot} \tag{B.5}$$

$$\left.\frac{M}{L}\right|_{\mathrm{galaxy}} = 10\text{–}20\ \frac{\mathrm{M}_\odot}{\mathrm{L}_\odot}. \tag{B.6}$$

By Ka Chun Yu and Phil Maloney.

Thus rotation curves suggests that there is about ten times as much dark matter as there is ordinary baryonic matter.

Similar conclusions are found from other independent techniques. Using the virial theorem along with observed sizes of the clusters and velocity dispersions of the cluster members, Merritt (1987) finds a mass-to-light ratio in galaxy clusters of

$$\left.\frac{M}{L}\right|_{\rm cl,vir} \approx 350 h^{-1} \frac{\rm M_\odot}{\rm L_\odot}, \tag{B.7}$$

which again suggests that the dark matter density is many times the baryonic density.

Studies of the dynamics of galaxies at small cosmological distances (say $r \lesssim 10\,h^{-1}$ Mpc) give

$$\Omega_{r<10} \sim 0.05\text{–}0.2. \tag{B.8}$$

This value is based on statistical analyses of the relative velocities of galaxies as a function of separation (statistical because the redshift only gives one component of the relative velocity of galaxies), on studies of the dynamics of loose groups of galaxies, and estimates of the true mass-to-light ratios of galaxies with flat rotation curves (i.e., including the contribution of the dark matter to the estimate of $\Omega_\star$).

Cluster masses can also be determined by several other techniques. The binding mass can be estimated from observations of temperature and density profiles of hot, x-ray emitting gas in rich clusters, under the assumption that the gas is in hydrostatic equilibrium in the cluster potential. (Note that for masses of 10^{13}–10^{14} $\rm M_\odot$ and radii of ~ 1 Mpc, as is typical of rich clusters, Eq. 9.40 gives $T_{\rm vir} \sim 10^6$–10^7 K, at which temperature the gas will radiate in x-rays.) This gives a density ratio in the range $\Omega_{\rm cl,xray} \sim 0.10$–$0.4$.

For rich clusters that are fortuitous enough to be aligned with background galaxies, the size of the lensing arcs can be related to the cluster mass via Eq. A.7. Studies in these cases give results that are consistent with the x-ray determinations. Including both baryonic and dark matter in the sum total, the density ratio of the gravitating matter is:

$$\Omega_m \sim 0.25, \tag{B.9}$$

of which roughly 1% of this is in the form of baryons.

At larger distances $r \gtrsim 10\,h^{-1}$ Mpc, dynamical studies show

$$\Omega_{r>10} \sim 0.05\text{–}1.0. \tag{B.10}$$

These results are based on attempts to measure peculiar (non-Hubble flow) velocities on these scales. Perhaps the most important (and certainly the least controversial) is that due to the CMB dipole anisotropy. The microwave background is observed to be anisotropic at the $\sim 10^{-3}$ level, being hotter in one direction and cooler in the opposite direction (hence a "dipole" term). This dipole anisotropy indicates that our local patch of the universe is moving with a velocity of ~ 600 km s^{-1} with respect to the frame in which the CMB appears uniform, i.e., the dipole anisotropy is just due to the Doppler effect. The standard interpretation is that this peculiar motion is due to acceleration caused by large-scale density inhomogeneities.

This last estimate gives the first direct indication that $\Omega \sim 1$. However as we have seen, if galaxies are more clustered than the matter distribution, then only observations on very big scales will probe the true value of Ω.

If we believe that $\Omega_{\rm total} = 1$, and $\Omega_m \sim 0.25$, then this implies that $\Omega_\Lambda \sim 0.7$. Two independent lines of reasoning support the conclusion that $\Omega_\Lambda > 0$. First are the results of groups finding Type Ia supernovae at high z and measuring their light curves. Type Ia SNs are believed to be good standard candles. Therefore they are good objects to observe to test the luminosity-redshift relationship (cf. Eq. 7.32). Deviations from $D_L \propto z$ for observations of a standard candle would hint that a cosmological constant Λ is at work. Two separate research groups (Perlmutter *et al.* 1999, Ries *et al.*1998) find similar results showing an acceleration in the Hubble expansion. The best fit from the Perlmutter *et al.* Supernova Cosmology Project team gives

$$\Omega_m = 0.28^{+0.09}_{-0.08}. \tag{B.11}$$

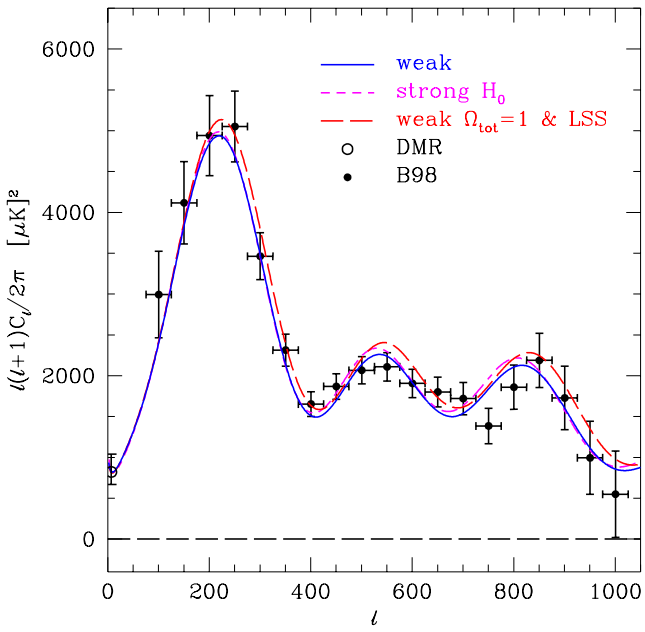


Figure B.1: A plot from Netterfield *et al.* (2001) showing the CMB angular power spectrum out to $l \sim 1000$, plotted with best fit models.

Yet another independent check comes from the BOOMERanG and MAXIMA experiments to measure and analyze the angular power spectrum from the Cosmic Microwave Background. Preliminary results from the BOOMERanG team (de Bernardis *et al.* 2000) show a spectrum with a strong acoustic peak at $l \approx 200$ which strongly suggests a $\Omega = 1$ universe (see Ch. 11). New results using more data that sampled the power spectrum out to $l \sim 1000$ were recently presented by Netterfield *et al.* (2001). They used nine different

fits to extract the values of various cosmological parameters. The median values of these fit results are:

$$\begin{aligned} \Omega_{\text{total}} &\approx 1.00 \\ \Omega_m &\approx 0.38 \\ \Omega_b h^2 &\approx 0.022 \\ \Omega_{\text{CDM}} h^2 &\approx 0.13 \\ \Omega_\Lambda &\approx 0.62. \end{aligned}$$

There is thus growing evidence to suggest that

$$\Omega_m \sim 0.3, \tag{B.12}$$

$$\Omega_\Lambda \sim 0.7, \tag{B.13}$$

$$\Omega_b \sim 0.02. \tag{B.14}$$

Thus not only is baryonic matter—the type we are most familiar with—just a fraction of the total matter density, but the dominant form of energy density in the universe is a cosmological constant (or "dark energy"), something that we have just a bare minimal understanding of.

ASTR 3740: Homework #1

Due: Friday, January 26, 2001

1. **Distribution of globular clusters** [10 pts.]: Early this century, the American astronomer Harlow Shapley studied stellar clusters, including a number of *globular clusters*, spherical distributions of 100,000s to millions of stars. He was able to determine the distance to 69 of them by measuring the periods of *Cepheid variable stars* located in the closest clusters, and then bootstrapping his way to more distant clusters by assuming all globular clusters have roughly the same angular diameter. (Cepheid variables have a well known period-luminosity relationship: if you can determine the periodicity of their variability in brightness, you know how intrinsically bright they are. From this intrinsic brightness, you can estimate how far any Cepheid must be to appear as faint as it does in your observations.)

   Shapley's observations were surprising. They revealed that globular clusters were randomly distributed in the z direction with respect to the plane of the Milky Way: there were just as many above the Galactic plane as there were below it. However they were *not randomly distributed azimuthally*, but appeared to be found in a prefered direction. This was one of the first indications that the Copernican principle applied to our Sun with respect to the Galaxy. Our Sun is not in a privileged location—i.e., the center of the Galaxy—but is instead somewhere in the outskirts of the Galactic disk.

   The table on the next page is a compilation of Shapley's results, showing the name of the globular cluster, its *right ascension* coordinate, its *declination* coordinate, and its radial distance in *parsecs* (= 3.26 light years). The right ascension and declination pinpoint the location of the object in the sky, and are the equivalent to longitude and latitude on the surface of the Earth. Although declinations are similar to latitudes, by varying from 0° to $\pm 90^\circ$, from the celestial equator to the poles, the azimuthal right ascension is measured in *hours*, *minutes* and *seconds*, whereby $24^{\rm h} = 360^\circ$, $60^{\rm m} = 1^{\rm h}$, and $60^{\rm s} = 1^{\rm m}$.

   For this problem, you will follow in Shapley's footsteps by calculating:

   (a) *In what direction* is the distribution of globular clusters centered (i.e., what is the right ascension and declination)? And ...

   (b) *How far* from the Sun is the center of the distribution?

   Hint: One obvious way to solve this problem is to take an average of each set of numbers. However this does not necessarily give you the most correct results—in fact doing this will result in an answer in position that is about 30° away from the true Galactic center (which we now know to be at $17^{\rm h}42^{\rm m}30^{\rm s}, -28^\circ 55' 00''$).

   The standard transformations from spherical coordinates to Cartesian coordinates (which you might or might not need) is:

$$\begin{aligned} x &= r \sin\alpha \cos(\pi/2 - \delta) \\ y &= r \sin\alpha \sin(\pi/2 - \delta) \\ z &= r \cos\alpha \end{aligned}$$

Globular Cluster	R.A.	Decl.	Radial Distance (in 100s of parsecs)
NGC 104	$0^h19.^m6$	$-72°38'$	68
NGC 288	$0^h47.^m8$	$-27°08'$	189
NGC 362	$0^h58.^m9$	$-71°23'$	152
NGC 1261	$3^h09.^m5$	$-55°36'$	256
NGC 1851	$5^h10.^m8$	$-40°09'$	172
NGC 1904	$5^h20.^m1$	$-24°37'$	256
NGC 2298	$6^h45.^m4$	$-35°54'$	244
NGC 2808	$9^h10.^m0$	$-64°27'$	170
NGC 3201	$10^h13.^m5$	$-45°54'$	147
NGC 4147	$12^h05.^m0$	$+19°06'$	526
NGC 4372	$12^h20.^m1$	$-72°07'$	114
NGC 4590	$12^h34.^m2$	$-26°12'$	161
NGC 4833	$12^h52.^m7$	$-70°20'$	164
NGC 5024	$13^h08.^m0$	$+18°42'$	189
NGC 5139	$13^h20.^m8$	$-46°47'$	65
NGC 5272	$13^h37.^m6$	$+28°53'$	139
NGC 5286	$13^h40.^m1$	$-50°52'$	196
NGC 5634	$14^h24.^m4$	$-\ 5°32'$	303
NGC 5897	$15^h11.^m7$	$-20°39'$	149
NGC 5904	$15^h13.^m5$	$+\ 2°27'$	125
NGC 5986	$15^h39.^m5$	$-37°27'$	208
NGC 6093	$16^h11.^m1$	$-22°44'$	200
NGC 6101	$16^h14.^m4$	$-71°58'$	213
NGC 6121	$16^h17.^m5$	$-26°17'$	114
NGC 6144	$16^h21.^m2$	$-25°49'$	244
NGC 6171	$16^h26.^m9$	$-12°50'$	161
NGC 6205	$16^h38.^m1$	$+36°39'$	111
NGC 6218	$16^h42.^m0$	$-\ 1°46'$	123
NGC 6229	$16^h44.^m2$	$+47°42'$	435
NGC 6235	$16^h47.^m4$	$-22°01'$	500
NGC 6254	$16^h51.^m9$	$-\ 3°57'$	120
NGC 6266	$16^h54.^m8$	$-29°58'$	152
NGC 6273	$16^h56.^m4$	$-26°07'$	159
NGC 6284	$16^h58.^m4$	$-24°37'$	370
NGC 6287	$16^h59.^m1$	$-22°34'$	435
NGC 6293	$17^h04.^m0$	$-26°26'$	263
NGC 6304	$17^h08.^m2$	$-29°20'$	322
NGC 6316	$17^h10.^m3$	$-28°01'$	526
NGC 6333	$17^h13.^m3$	$-18°25'$	250
NGC 6341	$17^h14.^m1$	$+43°15'$	123
NGC 6352	$17^h17.^m8$	$-48°19'$	227
NGC 6356	$17^h17.^m8$	$-17°43'$	385
NGC 6362	$17^h21.^m5$	$-66°58'$	130
NGC 6388	$17^h29.^m0$	$-44°40'$	276
NGC 6397	$17^h32.^m5$	$-53°37'$	81
NGC 6402	$17^h32.^m4$	$-\ 3°11'$	233
NGC 6441	$17^h43.^m4$	$-37°01'$	453
NGC 6541	$18^h00.^m8$	$-43°44'$	144

NGC 6584	$18^{\rm h}10\overset{\rm m}{.}6$	$-52^\circ 15'$	253
NGC 6624	$18^{\rm h}17\overset{\rm m}{.}3$	$-30^\circ 24'$	283
NGC 6626	$18^{\rm h}18\overset{\rm m}{.}4$	$-24^\circ 55'$	185
NGC 6637	$18^{\rm h}24\overset{\rm m}{.}8$	$-32^\circ 25'$	209
NGC 6638	$18^{\rm h}24\overset{\rm m}{.}8$	$-25^\circ 34'$	345
NGC 6642	$18^{\rm h}25\overset{\rm m}{.}8$	$-23^\circ 32'$	385
NGC 6652	$18^{\rm h}29\overset{\rm m}{.}2$	$-33^\circ 04'$	305
NGC 6656	$18^{\rm h}30\overset{\rm m}{.}3$	$-23^\circ 59'$	85
NGC 6681	$18^{\rm h}36\overset{\rm m}{.}7$	$-32^\circ 23'$	177
NGC 6712	$18^{\rm h}47\overset{\rm m}{.}6$	$-\ 8^\circ 50'$	312
NGC 6715	$18^{\rm h}48\overset{\rm m}{.}7$	$-30^\circ 36'$	155
NGC 6723	$18^{\rm h}52\overset{\rm m}{.}8$	$-36^\circ 46'$	121
NGC 6752	$19^{\rm h}02\overset{\rm m}{.}0$	$-60^\circ 08'$	79
NGC 6779	$19^{\rm h}12\overset{\rm m}{.}7$	$+30^\circ 00'$	250
NGC 6809	$19^{\rm h}33\overset{\rm m}{.}7$	$-31^\circ 10'$	91
NGC 6864	$20^{\rm h}00\overset{\rm m}{.}2$	$-22^\circ 12'$	455
NGC 6934	$20^{\rm h}29\overset{\rm m}{.}3$	$+\ 7^\circ 04'$	333
NGC 6981	$20^{\rm h}48\overset{\rm m}{.}0$	$-12^\circ 55'$	294
NGC 7006	$20^{\rm h}56\overset{\rm m}{.}8$	$+15^\circ 48'$	626
NGC 7078	$21^{\rm h}25\overset{\rm m}{.}2$	$+11^\circ 44'$	147
NGC 7089	$21^{\rm h}28\overset{\rm m}{.}3$	$-\ 1^\circ 16'$	156
NGC 7099	$21^{\rm h}34\overset{\rm m}{.}7$	$-23^\circ 38'$	172

Here, we assume r is the distance along the radial direction, α is the right ascension angle in the azimuthal direction, and δ is the declination angle. You might (or might not) also wish to have the inverse transforms:

$$\begin{aligned} r &= \sqrt{x^2+y^2+z^2} \\ \delta &= \frac{\pi}{2} - \tan^{-1}\left(\frac{y}{x}\right) \\ \alpha &= \cos^{-1}\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right) \end{aligned}$$

2. **Radioactive decay of relativistic neutrons** [10 pts.]: Free neutrons (those not bound inside an atomic nucleus) undergo radioactive decay with a half-life of only 17 minutes. However if the neutrons are traveling at relativistic speeds, their "internal clocks" are slowed down, so the decay rate measured by an observer at rest is slower. For this problem, you will calculate this effect for neutrons produced in solar flares. Assume that the neutron decay law is:

$$N(t) = N(0)\, e^{-t/17 \text{ minutes}},$$

where t is time measured in the rest frame of the neutrons, $N(0)$ is the number of neutrons at time $t = 0$, and $N(t)$ is the number of neutrons at time t.

Now assume that neutrons leave the Sun, as soon as they are created, at speed $0.5c$. Using a Sun-Earth distance of 1 astronomical unit $= 1.5 \times 10^8$ km, calculate

(a) The fraction of neutrons expected to arrive at the Earth without including special relativistic effects, and

(b) The fraction of neutrons arriving with Special Relativity.

ASTR 3740: Homework #2

Due: Friday, February 2, 2001

1. **The relativity of simultaneity:** (10 pts.) We showed in class that if a reference frame S' moves with velocity $V\hat{\boldsymbol{x}}$ with respect to a frame S, then the coordinates x' and t' are related to x and t by the Lorentz transformation:

$$x' = \frac{x - Vt}{\sqrt{1 - V^2/c^2}}, \qquad t' = \frac{t - Vx/c^2}{\sqrt{1 - V^2/c^2}}$$

   (a) Derive the inverse transformation, i.e., express x and t in terms of x' and t'.

   (b) Consider two events (x_1, t_1) and (x_2, t_2) which have a timelike separation. Show that there is a reference frame S' in which these two events have the same space coordinate, i.e., ${x_1}' = {x_2}'$. Find the velocity of this frame.

   (c) Redo part (b) for events with a spacelike separation, and in this case, find a frame in which these two events are simultaneous, ${t_1}' = {t_2}'$. Can the events in part (b) be made to have a spacelike separation by a Lorentz transformation?

2. **Past and future light cones:** (5 pts.) Draw a spacetime diagram with the horizontal and vertical axes labeled, respectively, x and ct. Label a point P (an event) somewhere on the diagram away from the origin. Draw the path followed by a light ray emitted from the event and propagating to the right, and the path of another light ray propagating to the left. This is called the *future light cone* of P. Then draw the paths of all light rays emitted in the past which arrive at P. This is the *past light cone* of P. Now shade in the part of the diagram occupied by the worldlines of observers which pass through P, and the part of the diagram occupied by events which can be connected to P by some form of signal.

3. **Spacetime diagrams:** (10 pts.) For this exercise, you will get to draw more spacetime diagrams. Start with the standard diagram with axes ct and x for an observer O. Now draw the following:

   (a) The worldline of O's clock at $x = 1$ m.

   (b) The worldline of a particle moving with velocity $v = 0.1c$, and which is at $x = 0.5$ m when $t = 0$ s.

   (c) The ct' and x' axes of an observer O' who moves with velocity $v = 0.5c$ in the positive x direction relative to O and whose origin $(x', t') = (0, 0)$ coincides with that of O.

   (d) The calibration tick along the coordinate axes of O' for intervals of $ct' = 1$ m and $x' = 1$ m.

   (e) The locus of events, all of which occur at the time $ct = 2$ m (i.e., simultaneous as seen by O).

   (f) The locus of events, all of which occur at the time $ct' = 2$ m (i.e., simultaneous as seen by O').

   (g) The event which occurs at $ct' = 0$ m and $x' = 0.5$ m.

   (h) The locus of events at $x' = 1$ m.

(i) The worldline of a photon that is emitted from an event at $ct = -1$ m, $x = 0$ m, travels in the negative x direction, is reflected when it encounters a mirror located at $x' = -1$ m, and is absorbed when it encounters a detector located at $x = 0.75$ m.

(j) The locus of events whose interval Δs^2 from the origin is -1 m^2.

4. **Gravitational redshifts:** (10 pts.) Radiation emitted from the surface of a body of mass M and radius R with wavelength λ_e is observed far from the body to be redshifted to wavelength λ_0. As we saw in class, the redshift z is

$$z \equiv \frac{\lambda_0}{\lambda_e} - 1 \approx \frac{GM}{c^2\, R}$$

(a) Consider a white dwarf star with radius equal to ~ 0.1 that of the Earth's radius ($R_\oplus = 6.052 \times 10^3$ km) and with a mass equal to that of our Sun ($M_\odot = 2 \times 10^{33}$ g). What is the redshift of radiation emitted from such an object?

(b) Suppose you set up a spectrograph on the surface of this white dwarf and measure the wavelength of radiation from singly ionized Sulfur atoms (i.e., S II) located in a distant nebula in interstellar space. The wavelengths of two emission lines measured near the atoms themselves are 6716.4 Å and 6730.8 Å. What would be the wavelengths measured on the white dwarf?

ASTR 3740: Homework #3

Due: Friday, February 9, 2001

1. **The Twin Paradox:** (4 pts.) Sam and Sarah are fraternal twins. On their 21st birthday, Sarah leaves her brother and goes off in a spaceship headed in the $+x$ direction for four years ($= 1.26 \times 10^8$ sec) of *her* time at a speed of $0.75c$. She then stops, reverses course, and travels back at $0.75c$ taking another four years of her time. Assume that she instantly accelerates and decelerates (and manages to survive the tremendous g forces!).

   (a) What is Sarah's age when she returns back to Earth?

   (b) What is Sam's age?

2. (10 pts.) Here is a possible paradox: Sam sees Sarah moving away at $0.75c$ and therefore sees her clocks slowed. However Sarah sees Sam moving in the opposite direction at $0.75c$ and views his clocks as being slowed. Time dilation implies that one person should be older than the other, but is it Sarah or Sam? We will try to reconcile their two different viewpoints in the rest of this problem.

   (a) Draw a space-time diagram from the vantage point of stationary Sam. Now draw the worldline for Sarah on her journey. According to Sarah, what time does her calendar read when she turns around? According to Sam, what time does his calendar read when Sarah stops and reverses direction?

   (b) Let's assume that Sarah sends light pulses back to Sam at regular intervals of once a year (in her time). How many of these signals will Sam have received by the time Sarah has stopped and turned around? How many signals will Sam receive during Sarah's entire return trip? If Sam was also sending light pulses at yearly intervals (in *his* time), would Sarah receive more signals during her outbound trip or her inbound trip?

   (c) If Sam was watching Sarah's spaceship with a telescope the entire time of her journey, describe how the clocks aboard the spaceship would appear to him.

   (d) Has Sarah been in an inertial reference frame during her entire trip? Explain why or why not.

3. **The Bus/Garage Paradox:** (12 pts.) Suppose you have access to an atomic-powered bus which at rest is 20 m in length.[1] You manage to accelerate it up to a speed of $0.8c$. Before the bus and the ground and air around it heats up to the point where everything evaporates in a large explosion, you aim it towards a garage which is 15 m long. Your friend remains at rest by the garage door, ready to slam shut the door as soon as the bus is all the way inside.

   (a) How long does your friend measure the length of the bus to be as it approaches the garage?

   (b) The garage door is initially open and immediately after the bus is entirely inside the building, your friend shuts the door. How long after the door is shut does

[1] See for instance http://bcn.boulder.co.us/campuspress/1995/nov301995/bigbus113095.html.

the front of the bus hit the other end of the garage as measured by the friend? Compute the interval between the events of shutting the barn door and hitting the wall. Is it spacelike, timelike, or null?

(c) In the reference frame of you, the bus driver, what are the lengths of the garage door and your bus?

(d) Do you, as the driver, believe that the bus is entirely inside the garage when its front hits the far wall of the garage? Can you explain why?

(e) After the collision, the bus comes to rest relative to the garage. From your friend's point of view, the 20 m bus now has fit inside a 15 m garage, since the door was shut before the bus stopped. How is this possible? Alternatively, from your point of view, the collision should have stopped the bus *before* the door closed, so the door could not close at all. Was the door closed with the bus inside? Or not?

(f) Draw a spacetime diagram from your friend's POV. Use it to illustrate and justify your conclusions.

(g) Assume the total mass of the bus is 10 tons, or roughly 10^7 g. Now assume all of the bus' kinetic energy is released instantaneously when it stops. What is the equivalent megatonnage of the resulting explosion? (Use for conversion, 9×10^{20} erg $\approx$ 20 kilotons of TNT, from p. 24 of the notes.)

ASTR 3740: Homework #4

Due: Friday, February 16, 2001

1. **Curvature of space by a massive body:** (20 pts.) In this problem, we will examine the curvature of space by a static spherical body with mass M. The proper time interval is given by

$$d\tau^2 = \left(1 - \frac{r_s}{r}\right) dt^2 - \frac{1}{c^2}\left[\frac{dr^2}{1 - r_s/r} + r^2\, d\theta^2 + r^2\, \sin^2\theta\, d\phi^2\right], \tag{1}$$

where $r_s = 2GM/c^2$ is the Schwarzschild radius. Eq. 1 is also known as the *Schwarzschild line element* since it can be used to determine distances between points in the curved spacetime.

   (a) Suppose you wish to measure the radial distance between two radii r_1 and r_2, both located outside r_s. You can calculate the proper length L_{12} between the radii by setting $dt = d\theta = d\phi = 0$ in Eq. 1. The length interval is then

$$L_{12} = \int_{r_1}^{r_2} \frac{dr}{\sqrt{1 - \frac{r_s}{r}}}. \tag{2}$$

   If r/r_s is very large, the integral can be simplified by using the Taylor expansion:

$$\frac{1}{\sqrt{1 - \frac{r_s}{r}}} \approx 1 + \frac{r_s}{2r} \tag{3}$$

   You may do either the full integral or the simplified integral.

   (b) Consider light emitted at r_1 at t_1, and traveling outward to r_2, where it is received by a detector at t_2. Using the fact that $d\tau = 0$ for a light ray, and $d\theta = d\phi = 0$ for propagation in the radial direction, calculate $t_2 - t_1$ in terms of r_2 and r_1. **Hint:** You will again need to integrate over r. You may use the same approximation as in part (a).

   (c) Now suppose two successive crests of a light wave are emitted from r_1 at times t_1 and $t_1 + \delta t_1$, and that they are received at r_2 at times t_2 and $t_2 + \delta t_2$. Relate the emitted frequency ν_e to the *proper time interval* $\delta\tau_1$ by $\delta\tau_1 = 2\pi/\nu_e$. What is the observed frequency $\nu_\circ$ or $2\pi/\delta\tau_2$? Work this out by calculating $\delta\tau_2$ in terms of $\delta\tau_1$, r_1, and r_2. This is the gravitational redshift. **Hint:** From Eq. 1, the proper time interval is related to the coordinate time interval by $\delta\tau = \delta t\sqrt{1 - r_s/r}$ when $dr = d\theta = d\phi = 0$.

(d) When electrons and positrons annihilate, they release γ-ray photons of energy 5.11×10^5 electron volts (eV). A burst of γ-rays of energy 4.2×10^5 eV was detected on March 5, 1979, and was thought to be electron-positron annihilation photons produced near the neutron star SGR 0526–66 located in the Large Magellanic Cloud. Since it originates from a compact, massive object, the radiation should hence be gravitationally redshifted. If this explanation is correct, and the mass of the neutron star is 1 $M_\odot$, how far from the center of the star were the γ-rays produced? **Hint:** Let $r_2 \to \infty$.

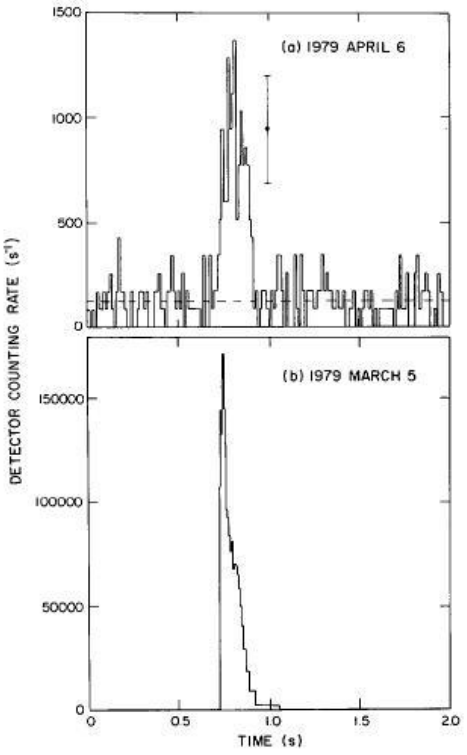


FIG. 1.—Time histories of the 1979 April 6 and March 5 events. (*a*) The time resolution is 11.7 ms, and the background count rate is indicated by a dashed line. (*b*) The time resolution varies from 6.0 ms to 187.5 ms, depending on the count rate.

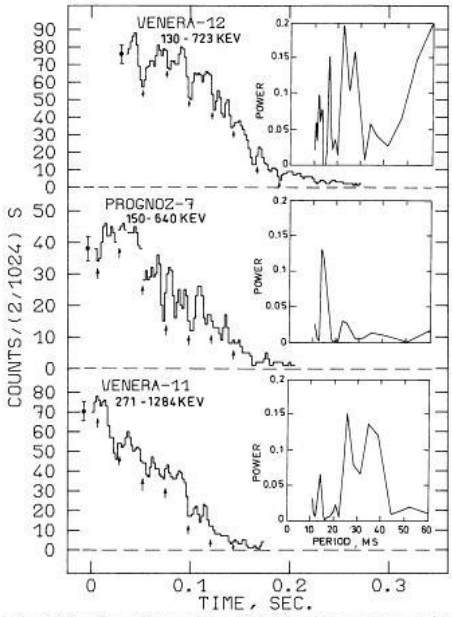


Fig. 1. The first 200 ms of the 1979 March 5 gamma ray burst, observed with 2/1024 s time resolution by identical detectors aboard the Prognoz 7, Venera 11, and Venera 12 spacecraft. A 3-point running average has been used to smooth the data. The first 41 ms of the Venera 12 data are severely affected by dead time losses, and are not shown; similarly, three data points in the Prognoz 7 time history could not be reconstructed from the data, and have been omitted. Dashed lines indicate background levels; arrows indicate a 23 ms period. *Insets:* spectral power in counts^2/Hz as a function of period. These spectra result from a power spectral analysis of the residuals of the raw data, as discribed in the text

ASTR 3740: Homework #5

Due: Friday, February 23, 2001

1. **The binary pulsar:** (25 pts.) In 1975 Russell Hulse (a graduate student) and Joseph Taylor (a professor) discovered a pair of neutron stars (one of which is a pulsar) in orbit around each other. These stars are so close together, and have such strong gravitational fields, that both Special and General Relativistic effects are important in determining their orbits. Thus this system has been an important laboratory for testing the theory of relativity. Hulse and Taylor were awarded the Nobel Prize for Physics in 1992 for their discovery of this binary system. In the following problems, we will examine some aspects of the double pulsar.

   (a) **The classical orbit:** According to Kepler's laws, these neutron stars have an elliptical relative orbit with semimajor axis a and period P. If the two stars have masses M_1 and M_2, then

   $$G(M_1 + M_2)P^2 = 4\pi^2 a^3.$$

   Suppose that $M_1 = M_2 = 1\ \mathrm{M}_\odot$, where the solar mass $1\ \mathrm{M}_\odot = 2 \times 10^{33}$ gm. The period is measured to be 2.79×10^4 s. Calculate a and express it in astronomical units (where $1\ \mathrm{AU} = 1.496 \times 10^{13}$ cm). [**Hint:** If you are clever, you won't need to know G nor the size of an AU.]

   (b) **Orbital velocity:** Assuming each orbit is circular, calculate the velocity of each star using the relation

   $$V^2 = \frac{G(M_1 + M_2)}{a},$$

   and express it as a fraction of the speed of light.

   (c) **Doppler shift:** Suppose the orbits were perpendicular to the plane of the sky. The pulsar period measured in its own rest frame is 0.059 sec. What are the maximum and minimum periods measured, due to the Doppler shift, as the pulsar moves along its orbit?

   (d) **General Relativistic precession:** As discussed in class, GR effects cause the perihelion of an elliptical orbit with semimajor axis a and eccentricity e to advance by an angle $\Delta\phi$ over each orbit, given approximately by

   $$\Delta\phi \approx \frac{6\pi GM}{c^2 a(1 - e^2)}.$$

   Although in the first three parts of this problem, we assumed the orbit to be circular, it actually is not. Let a be the value you calculated before, $e = 0.6$, and assume $M = 2\ \mathrm{M}_\odot$. Calculate the precession angle per orbit.

   (e) Again use an orbital period of 2.79×10^4 sec. Calculate how far the perihelion has precessed in one year. Compare this with the precession of the perihelion of Mercury.

ASTR 3740: Homework #6

Due: Friday, March 2, 2001

1. **Capture by black holes:** (25 pts.) We saw in class that the effective potential per unit mass $V_{\rm eff}(r)$ of a particle with angular momentum per unit mass L moving under the influence of a black hole of mass M is

$$V_{\rm eff}(r) = \frac{L^2}{2r^2}\left(1 - \frac{r_s}{r}\right) - \frac{GM}{r},$$

where r_s is the Schwarzschild radius. This potential allows particles to be captured from infinity, which the Newtonian potential does not allow. The purpose of this problem is to work out some of the conditions for capture.

   (a) The locations at which $V_{\rm eff}$ has maxima or minima are the locations at which

$$\frac{dV_{\rm eff}}{dr} = 0.$$

   Calculate $dV_{\rm eff}/dr$.

   (b) Use the expression you obtained in part (a) to show that maxima and minima are roots of the quadratic equation

$$r^2 - \frac{L^2}{GM}r + \frac{3}{2}\frac{L^2}{GM}r_s = 0.$$

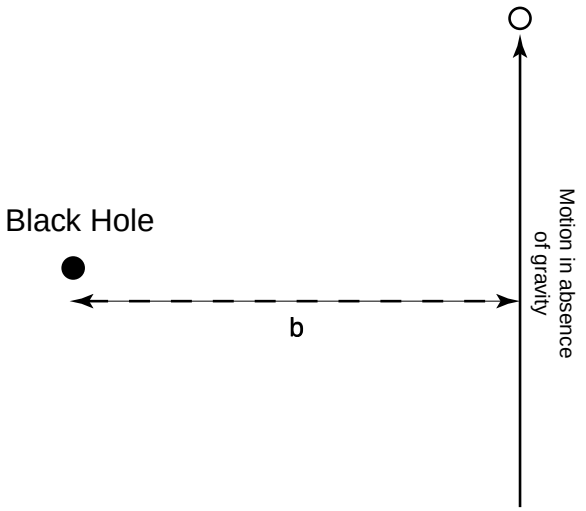
   (c) Show that if

$$\frac{L^2}{GM} < 6r_s$$

   then $V_{\rm eff}$ has no maxima or minima, unlike the case shown in the notes and in class. Sketch $V_{\rm eff}(r)$ in this case. Use your sketch to argue that if a particle has $L < \sqrt{6GMr_s}$, it must fall into the black hole.

   (d) A leading model for quasars is that they are powered by gas falling onto massive black holes at the centers of galaxies. What is the critical angular momentum $\sqrt{6GMr_s}$ for capture onto such a black hole? Assume the mass of the black hole is 10^8 $M_\odot$.

(e) If the speed of a particle very far from the black hole is V, and its distance of closest approach from the hole *in the absence of gravity* would be b, then $L = Vb$. Suppose the average velocity of gas is $V = 300$ km s^{-1}, and that the gas is moving randomly in all directions. What is the maximum b for capture onto the black hole? Compare this with the size of the galactic nucleus, $D_{\rm nucleus} \approx 100$ pc $= 3 \times 10^{15}$ km.



ASTR 3740: Homework #7

Due: Friday, April 6, 2001

Newton's Cosmology (20 pts. total) The first attempt at a physical cosmology was that of Isaac Newton in 1692. Newton argued that an infinite, homogeneous, static universe, while it might have *local* regions of gravitational instability, would be *globally* stable, as there would be equal gravitational forces acting in every direction. There is a fatal flaw in this argument however. The purpose of this problem set is to identify this flaw and determine the fate of Newton's universe.

1. **The force within a spherical shell** (5 pts.) In order to understand the flaw in Newton's cosmology, we need (appropriately enough) ***Newton's First Theorem:*** *A body located within a spherical shell of matter experiences no net gravitational force from that shell.* Suppose we take Newton's infinite, static, homogeneous universe and remove a spherical volume of matter (i.e., remove all the galaxies, which we will assume to be uniformly distributed on average) about some point, which we will call $\boldsymbol{A}$. What does Newton's First Theorem say about the gravitational force on any particle within the empty sphere about $\boldsymbol{A}$ due to the rest of the universe?

2. **Matter inside the spherical volume** (5 pts.) Now suppose we put matter back into the spherical volume around $\boldsymbol{A}$; in fact, put back exactly the same galaxies we removed initially. In keeping with the assumptions of Newton's cosmology, assume all of the galaxies are initially at rest when we put them back. (This is suppose to be a *static* universe, after all.) What are the implications of part (1) for the force exerted on these galaxies by the rest of the universe? If they are initially at rest (zero velocity), what must happen to the galaxies within this volume?

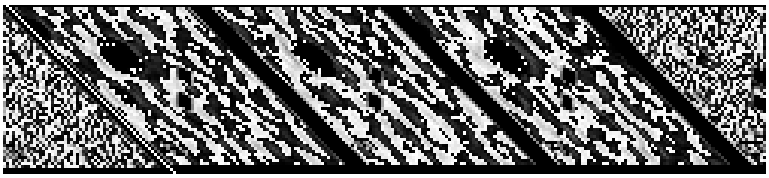
3. **A spherical shell around a second point** (5 pts.) Draw another spherical shell around another point, which we will call $\boldsymbol{B}$. Assume the galaxies within the sphere around $\boldsymbol{B}$ are also at rest initially. From parts (1) and (2), what must happen to the galaxies within this volume?

4. (5 pts.) Now consider a spherical shell which is large enough to contain the shells around both $\boldsymbol{A}$ and $\boldsymbol{B}$. What must happen to the galaxies (initially at rest) within this volume? What are the implications of this result for the fate of Newton's universe?

ASTR 3740: Homework #8

Due: Friday, April 13, 2001

1. **Radioactivity and Cosmology** (5 pts.) We saw in class that a time interval Δt measured within an object at cosmological redshift z is measured by us as a time interval $\Delta t(1+z)$. The purpose of this problem is to apply this to observations of Type II supernovae in distant galaxies. The visible light curves of these supernovae are thought to be powered by the decay of a radioactive isotope of nickel, Ni^{56}, which has a half-life of 6.4 days. That is, the time over which the supernova luminosity declines is proportional to the half life of Ni^{56}. Recently a galaxy was reported with a redshift of $z = 6.68$ (Chen, Lanzetta, & Pascarelle, 2000, Nature, 408, 562):



This object is so distant that even the image on the far right, taken with STIS onboard the Hubble Space Telescope, shows only a faint smudge. However suppose that in the future, a supernova is discovered in this galaxy using the next generation of 100 m class ground-based telescopes. Predict the time over which its luminosity would appear to decay.

2. **Searching for Distant Galaxies** (5 pts.) The light from galaxies appears to be dominated by G and K type giant stars. Assume that this light is blackbody radiation with a temperature $T = 4000$ K. From Wien's law, the wavelength of maximum intensity is

$$\lambda_{\max} = \frac{0.3 \text{ cm}}{T}.$$

Calculate $\lambda_{\max}$ for galaxies. In what part of the spectrum does it fall (i.e., visible, radio, etc.)? Now, using the redshift principle, what should be $\lambda_{\max}$ for a galaxy of redshift z? What type of telescope would be best for detecting galaxies at redshift 6.68—optical, ultraviolet, infrared, or radio?

3. **Particle Horizons** (10 pts.) Can we see every part of the universe? The maximum distance to which we can see (if there is one) is called the particle horizon. The purpose of this problem is to calculate the location of the horizon, $r_{\max}$, in a flat universe ($k = 0$). We determine $r_{\max}$ by requiring that the light emitted from $r_{\max}$ at the beginning of the universe, $t = 0$, be reaching us now. That is,

$$c \int_0^{t_\circ} \frac{dt}{R(t)} = \int_0^{r_{\max}} dr.$$

In a flat universe, we can write

$$R(t) = R(t_\circ) \left(\frac{t}{t_\circ} \right)^{2/3}.$$

Calculate $r_{\max}$ as a function of time $t_\circ$. Can we see more of the universe as time goes on, or less? Explain why.

ASTR 3740: Homework #9

Due: Friday, April 20, 2001

The Coupling of matter and radiation: (30 pts) The purpose of this problem is to study how cosmic background radiation (CBR) photons interact with matter now and in the past.

(a) Assume that space is flat, so that the cosmic scale factor $a(t) \propto t^{2/3}$. Let the present time be $t_\circ$. Show that the radiation which we observe now to have redshift z was emitted at a time t_1 which can be expressed in terms of $t_\circ$ and z by

$$t_1 = \frac{t_\circ}{(1+z)^{3/2}}.$$

This formula is often used to express time in terms of redshift.

(b) Recall that in a flat universe, the present value of the Hubble parameter $H_\circ$ is related to $t_\circ$ by $H_\circ = 2/(3t_\circ)$. Writing $H_\circ$ as usual as $100\,h$ km s^{-1} Mpc^{-1}, and using your solution from (1), what was the age of the universe at the time of recombination, at about $z = 1500$?

(c) Show that the density of (non-relativsitic) matter scales in a flat universe as $(1+z)^3$. Assume (consistent with the abundances of light elements made in the Big Bang), that $\Omega_b h^2 = 0.01$, or the present average density of baryons in the universe, $n_b(t_\circ)$, is $\sim 10^{-7}$ cm^{-3}. What was the average baryon density at the time of recombination?

(d) The Thompson cross-section, i.e., the cross-section for scattering of photons by electrons, has the value $\sigma_T = 6.65 \times 10^{-25}$ cm^2. What is the average distance a photon travels before being scattered both *now* and at $z = 1500$?

(e) What is the average time a photon travels before being scattered in these two cases?

(f) Compare the time for a photon to scatter with the age of the universe, now and also at $z = 1500$. What are your conclusions about the importance of scattering now and at the earlier time?

ASTR 3740: Homework #10

Due: Friday, April 27, 2001

1. **Massive neutrinos** (5 pts) Traditionally, neutrinos (ν) have been thought to be massless. Whether they really are is an open question. If the mass $m_\nu \neq 0$ the cosmological consequences could be major, because the density of ν produced in the early universe is quite high—the present number density n_ν is about 100 cm^{-3}. A recent experiment suggests $m_\nu c^2 \approx 2.5$ eV (1 eV $\approx 1.6 \times 10^{-12}$ erg). Assume this is correct and calculate Ω_ν.

2. **Dark matter in the solar system** (5 pts) Suppose that the universe contains enough invisible matter in the form of elementary particles that $\Omega = 1$, i.e, $\rho = \rho_{\rm crit} \approx 1.88 \times 10^{-29} h^2$ gm cm^{-3}. This dark matter must also fill the solar system. How much of an effect does it have? Compute the mass of dark matter in a sphere 1 AU in radius and compare with the mass of the Sun. What are the prospects for detecting the dark matter through its gravitational effect?

3. **Galactic halos** (10 pts) A variety of observations suggest that galaxies have extended "halos" of dark matter. How big would these halos have to be to close the universe? To answer this question, assume that each halo is a sphere with mass density ρ_h and radius r_h Then the mass of each halo is

$$M_h = \frac{4\pi \rho_h r_h^3}{3}.$$

   If there are n_g galaxies per unit volume with such halos, then the *mean mass density* $\langle \rho_h \rangle$ in galactic halos is

$$\langle \rho_h \rangle = M_h n_g.$$

   Assume that $\rho_h \approx 4 \times 10^{-23}$ gm cm^{-3}, which is about the value in the solar neighborhood. Calculate the radius r_h that galactic halos must have in order that $\langle \rho_h \rangle = \rho_{crit}$. (Assume that $n_g = 0.01 h^3$ Mpc^{-3}.)

ASTR 3740: Review for the Midterm

Note: Midterm is March 12, 2001

The Format: The midterm will be a 50 minute closed-book exam. You will need to bring a pencil and eraser; a calculator, ruler, and extra scratch paper are optional.

What Is Covered: The midterm will test *all* material that has been covered in the lectures up through March 7, 2001. This therefore includes all of Special Relativity, General Relativity, black holes, and the beginning lectures on cosmology.

What To Study: Review *all* of the homework problems. Make sure you understand the solutions completely. (If you are missing a solution, please ask me for one.) Look over and make sure you are familiar with all notes that you might have taken or were handed out. Going through and making sure you understand the official lecture notes should be helpful too!

Some General Study Questions:

1. What is the difference between inertial and gravitational mass? Why are they important in relativity?

2. Work out the expression for the Doppler effect in the classical case and in the relativistic case. Do not just memorize the equations; be sure you know how to derive them. In what instances would you use the relativistic Doppler expression to tell you how fast moving clocks appear move with respect to your (at rest) clocks? How is this different from using the time dilation formula?

3. What are the postulates of General Relativity?

4. What are some of the earthbound laboratory tests of Special Relativity? Of General Relativity? What about astronomical observations that support relativity?

5. Explain the meaning of the Equivalence Principle. What are some of its physical consequences?

6. What is a geodesic? What are some of the differences between a metric for measuring distances on a surface, versus a spacetime metric? What happens when your path deviates from a geodesic in spacetime? What about in a purely spatial metric? If given the equation for a metric, how would you use it to make measurements of distances in a spacetime described by this metric?

7. What is the meaning of the proper time τ? What is the coordinate time t that appears in the Schwarzschild metric?

8. In a spacetime diagram, what past events can influence you? What future events will you be able to send signals to? How can events that are spacelike appear to be simultaneous?

9. Roughly what form does the effective potential for a black hole take? How is it different from the Newtonian effective potential? Under what conditions will particles fall into the black hole versus orbiting it in bound or unbound orbits?

10. What is a *standard candle* in astronomy? Why are they important in the study of cosmology?

11. What is the Cosmological Principle?

12. In the Robertson-Walker metric, what is the parameter k? What do different values of k signify?

ASTR 3740: Review for the Final

Note: Midterm is Wednesday, May 9, 2001; 7:30–10:30 AM

The Format: The midterm will be a closed-book exam. You will need to bring a pencil and eraser; a calculator, ruler, and extra scratch paper are optional.

What Is Covered: The midterm will test *all* material that has been covered in Chapters 6–11 in the notes, including all class lectures and handouts pertaining to these sections.

What To Study: Review *all* of the appropriate homework problems. Make sure you understand the solutions completely. (If you are missing a solution, please ask me for one.) Look over and make sure you are familiar with all notes that you might have taken or were handed out.

Some General Study Questions:

1. What are the three key pieces of evidence that provide observational support for the Big Bang? Be sure you can explain each of these in detail.

2. Derive the redshift relationship in Eq. 7.20. You should not just merely memorize the formula, but be able to obtain this from first principles.

3. The classical expression for the Doppler redshift of a moving object is:

$$z = \frac{\lambda}{\lambda_\circ} - 1 = \frac{v}{c}$$

   The maximum redshift must be $z < 2$ since $v < c$. How is the relativistic Doppler redshift different that allows z to be greater than this? Now compare both of these to the cosmological redshift.

4. What is the particle horizon? Contrast this with the cosmological event horizon (which is not the same as the event horizon around a black hole).

5. Starting with the Friedmann equation, derive the critical density ρ_c.

6. Starting with the Friedmann equation, derive the dependence of the scale factor on time for a flat universe.

7. Using the Friedmann equation and the defintion of the critical density, derive an expression for how Ω in terms of the Hubble constant and the scale factor. If $\Omega = 1$ now (i.e., a flat universe), what does this imply for the value of Ω in the past?

8. Describe the principles behind the Steady State model. What must be occuring in the Universe for this model to be true? What observations do most people agree have falsified Steady State?

9. Show how a radiation field with a blackbody Planck spectrum remains a Planck spectrum in an expanding universe. What does change between the spectra during the expansion?

10. What are the sources and sinks of deuterium in the Universe? How will these affect the attempts to measure the primordial deuterium abundance?

11. Describe in succinct detail the concepts behind the following problems with Big Bang:

    (a) The horizon/smoothness problem.

    (b) The flatness problem.

    (c) The problem of hidden relics (e.g., magnetic monopoles).

(d) The problem of the formation of small-scale structure.

12. How does an inflation model for the early universe solve the above problems?

13. What is the evidence for dark matter in individual galaxies? What about in clusters of galaxies? Be able to explain by what observational means does one infer its existence (including at least one method for dark matter in galaxies and at least two different methods for clusters).

14. Dark matter has been suggested to be in baryonic form, such as black holes, brown dwarfs, and planets, all of which are difficult to detect. Explain why many believe that there *must* be a non-baryonic component to dark matter as well, such as massive neutrinos or exotic particles. You should be able to argue from both a nucleosynthesis viewpoint as well as from a structure formation viewpoint.

15. What is the Jeans mass? What equilibrium conditions are assumed in determining it? What is the Jeans length? Why is the sound speed a useful parameter in determining whether a cloud collapses?

16. Given an expansion time scale for the universe,

$$t_E = \frac{a}{\dot{a}} = \left(\frac{3}{8\pi G\rho}\right)^{1/2}; \tag{1}$$

a free-fall time for a perturbation in the early universe to collapse:

$$t_{\rm ff} = \left(\frac{3\pi}{32G\rho}\right)^{1/2}; \tag{2}$$

and the time for a sound wave to cross the perturbation:

$$t_s = \left(\frac{15}{4\pi G\rho}\right)^{1/2}; \tag{3}$$

what happens to the perturbation if:

(a) $t_s < t_{\rm ff}$?

(b) $t_E < t_{\rm ff}$?

(c) $t_{\rm ff} < t_s < t_E$?

17. Free-streaming and photon viscosity act as filters to damp small fluctuations. What does this imply for how structure in the Universe forms?

18. Perturbations of size $\delta\rho$ grow linearly if $\delta\rho \ll \rho$, where ρ is the background density in the Universe. The perturbation is said to go nonlinear if $\delta\rho \sim \rho$. Describe what happens to the perturbation at this point.

19. What is the difference between the growth of baryonic matter and dark matter perturbations? What is the importance of the scale factors $a_{\rm EQ}$ and $a_{\rm dec}$ (or the corresponding redshifts $z_{\rm EQ}$ and $z_{\rm dec}$) in describing the growth of the perturbations?

20. How is heating of a piece of iron past its Curie point, and then letting it cool akin to the phase transition that is thought to have started inflation?

ASTR 3740: Term Paper

Term papers are due on the last day of class, Friday, May 4. They should be 12-15 pages, should include a bibliography, and may include equations, figures, etc. The topic of the paper should be a scientific paper that you've found, that is on some topic in relativity and/or cosmology. although you may use secondary resources (e.g., popular magazines, websites, textbooks, encyclopediae) to help write this report, the main goal is to write about a scientific result that appears in a scientific paper.

Topic: You should have a paper picked out, preferably by March 16, but no later than the last day of classes before Spring Break (March 23). You should turn into me bibliographic information for the paper (title, authers, journal, volume number, etc.) and the abstract. *It is also highly recommended that you consult with me in person or via email before making a final decision on what to write about.* Because of the technical nature of this assignment, you might find it helpful to consult with me regularly even after you've picked your topic and have started writing. *Do not be afraid to ask questions.* I can make suggestions on finding additional references, and determine whether the scope of the paper is appropriate. It is all too easy to pick a paper that is *too* difficult. In such cases, I would rather you pick a new paper/topic rather than spend all your time on this one assignment.

Here is a list of some possible topics (you may of course choose something not on this list):

- Laboratory tests of relativity
- Solar system tests of relativity
- Outside the solar system tests of relativity
- Rotating (Kerr) black holes
- Astronomical evidence for black holes
- Gravitational radiation
- Gravitational lensing
- The extragalactic distance scale
- Determinations of the Hubble parameter
- Evidence for either an open or closed universe
- Candidates for dark matter
- Big Bang nucleosynthesis
- Distribution of galaxies and/or galaxy clusters; large-scale structure
- Galaxy formation
- The cosmic microwave background radiation

Paper Resources: The astronomical literature is vast. The list of journals that most astronomers publish in however is short. If you limit yourself to just the first four publications, you will have covered $\gtrsim 99\%$ of all important peer-reviewed astronomical papers. Following each title in parentheses is the common abbreviation for each journal that you will see in bibliographies; the Library of Congress catalog number, and the library on campus where you can find the journal (MATHPHYS for the Lester Math-Physics library in Duane and SCIENCE for the Science stacks in Norlin).

Astrophysical Journal and ***Astrophysical Journal Letters*** (ApJ and ApJL; QB1 .A9 MATHPHYS)

Astronomy & Astrophysics (A&A; QB1 .A83 MATHPHYS)

Monthly Notices of the Royal Astronomical Society (MNRAS; 520.6 R81m MATHPHYS)

Nature (Q1 .N2 SCIENCE)

Publications of the Astronomical Society of the Pacific (PASP; QB1 .A423 MATHPHYS)

Annual Reviews of Astronomy & Astrophysics (ARAA; QB1 .A2884 MATHPHYS)

Web Resources: You could spend days sorting through journals looking for something interesting. But there are useful Web resources out there which you may use, that can pinpoint papers for you quickly. These are the online *abstract and paper* services:

- **ADS:** [`http://http://adsabs.harvard.edu/default_service.html`] This service allows you to search for paper abstracts using the authors' names, publication date, words in the title, and words in the abstract. ADS keeps an online archive of scans of pages from papers older than a few months (including papers dating back to the 19th century!) If you are on a CU machine, you can also download the latest papers directly from the journal websites.

- **Astro-ph:** [`http://xxx.lanl.gov/archive/astro-ph`] This abstract and paper service allows you to search for papers that have been submitted but not yet published. Hence you can find the most current and latest work from researchers at this site. However since there is no guarantee that the papers submitted here will actually be accepted for publication, "crank" papers can sometimes slip in under the site managers' noses.

Other Resources: Other references which you might find useful can be found in the Lester Math-Physics library:

Encyclopedia of astronomy and astrophysics , 1989, Meyers, Robert A., Ed. (San Diego: Academic Press). [MATHPHYS REF QB14 .E53 1989] For scientists in one sub-field to keep abreast of topics in other sub-fields, but is generally written at a level that is accessible to upper-division undergraduates.

Scientific American [T1 .S5 MATHPHYS for current year, SCIENCE for past years] Excellent popular magazine with well written and accurate articles written by the scientists who are actually doing research on the topics they write about.

Sky and Telescope [QB1 .S536 MATHPHYS] A popular astronomy magazine, that by nature is non-technical, but it still has many useful feature articles which might be starting points for research.

Astronomy [QB1 .A7998 MATHPHYS] Another popular astronomy magazine, but (in my opinion) even less technically-oriented than *Sky and Telescope*. Its emphasis is more on "pretty pictures."