Astronomy 140 Lecture Notes, Spring 2006

Astronomy 140 is a course about Stellar Systems and Cosmology. The stellar systems we will study are large clusters like globular clusters, and galaxies which are even larger assemblies of stars. The part of the course will involve the calculation of stellar orbits in arbitrary potentials, and the calculation of the potentials generated by arbitrary mass distributions.

The second half of the course deals with cosmology. An important aspect of our study of cosmology is an understanding of the evidence that points toward the Big Bang model of the Universe, and that rules out various alternatives like the Steady State model and tired light models.

Thus the first week of lectures will be devoted to an observational overview. This will discuss the ways that astrophysicists can determine the properties of objects by remote sensing using photometry and spectroscopy.

1. Photometry

Magnitudes: the apparent magnitude is $m = -2.5\log(F/F_\circ)$. [Note that log is the common logarithm (base 10) in astronomical usage. Thus a 1% change in flux is a change of 0.0108 in magnitude.] The standard flux for 0^{th} magnitude is tabulated for various photometric bands in different photometric systems, but in practice one observes standard stars with known magnitudes along with program objects, in order to convert an observed flux ratio into a magnitude difference, which when combined with the known standard star magnitude gives the measured magnitude of the program object. The standard fluxes are usually given in erg/cm^2/sec/Hz for F_ν or W/cm^2/sec/μm for F_λ, or in *Janskies*: 1 Jy $= 10^{-26}$ W/m^2/Hz $= 10^{-23}$ erg/cm^2/sec/Hz. To actually determine the 0^{th} magnitude flux $F_\circ$ is quite difficult, since it requires the comparison of a star with a laboratory standard light source. Stars are much fainter and hotter than lab standards, and are also much further away and observed through the atmosphere. Correcting for all these problems still leaves an uncertainty of a few percent. However, it is possible to compare stars to an accuracy of 0.1%, so magnitudes can be measured to $\Delta m = 0.001$ even though the physical flux level is not known to 0.1% accuracy.

The most common photometric system is the Johnson UBV system which has been extended into the infrared: U (ultraviolet = 360 nm), B (blue = 440 nm), V (visual = 550 nm), R (red = 700 nm), I (infrared = 900 nm), J (1.25 μm), H (1.6 μm), K (2.2 μm), L (3.5 μm), M (4.9 μm), N (10 μm), and Q (20 μm). In this system the 0^{th} magnitude fluxes follow the brightness of an A0V star like Vega. Such a star is approximately a 10^4 K blackbody, but this approximation breaks down for the ultraviolet U band, because the absorption of hydrogen in the $n = 2$ level for $\lambda < 365$ nm causes a large depression in the flux. $F_\circ$'s are $10^{-11.37}$ W/cm^2/μm (1900 Jy) at U, $10^{-11.18}$ W/cm^2/μm (4300 Jy) at B, $10^{-11.42}$ W/cm^2/μm (3800 Jy) at V, $10^{-11.76}$ W/cm^2/μm (2800 Jy) at R, $10^{-12.08}$ W/cm^2/μm (2250 Jy) at I, and $10^{-13.40}$ W/cm^2/μm (650 Jy) at K. The AB magnitude system has the same $F_\circ = 3631$ Jy at all frequencies.

Color: By measuring the magnitude of a source at two different frequencies, you determine a measurement of the color of a source. An example of a color is B-V. Since the magnitudes get bigger when the flux gets smaller, a positive B-V means that the object has *less* blue flux and is thus *redder*.

Bolometric quantities: the total power received over all frequencies per unit area is the bolometric flux:

$$F_{bol} = \int F_\nu d\nu = \int F_\lambda d\lambda \tag{1}$$

Inverse square law: the flux of an object varies as the inverse square of the distance. For objects which radiate uniformly in all directions we can write

$$F_\nu = \frac{L_\nu}{4\pi D^2} \tag{2}$$

where L_ν is the luminosity per unit frequency of the source. Of course, $F_\lambda = L_\lambda/4\pi D^2$ and $F_{bol} = L_{bol}/4\pi D^2$.

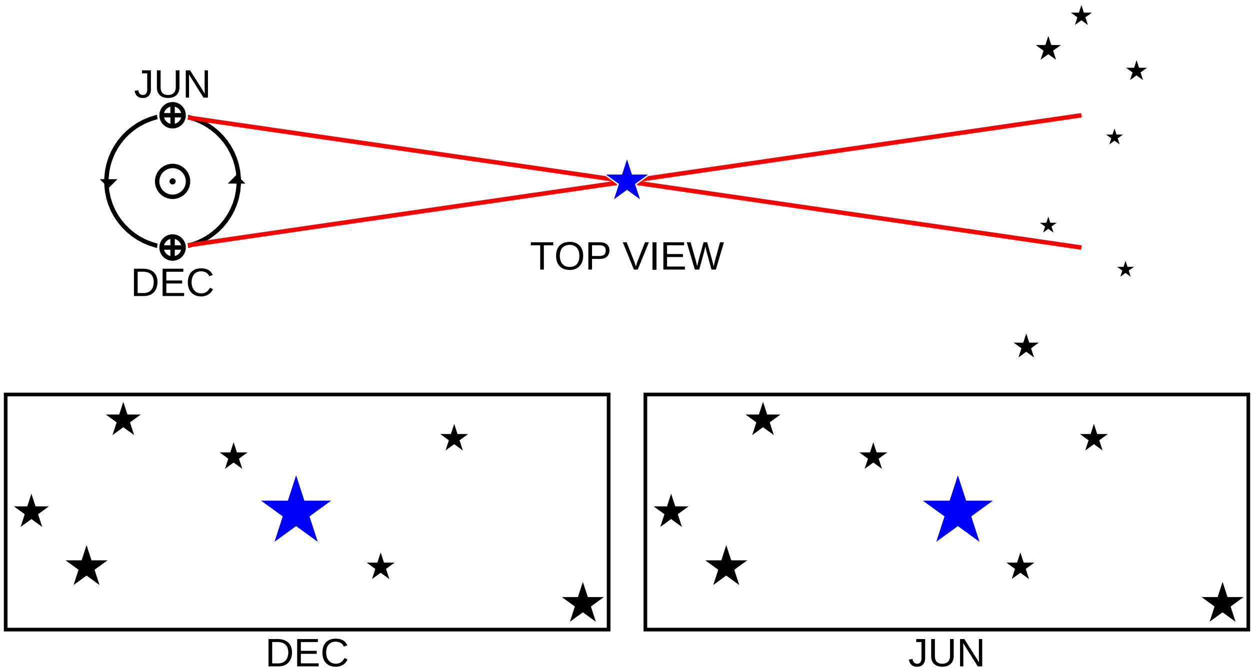

Fig. 1.— Annual parallax diagram. By using the angular motion of the nearby star relative to the background stars over the course of a year, the distance to the nearby star can be determined. The lower two images can be combined stereoscopically by crossing ones eyes, giving a 3-D view.

The bolometric luminosity L_{bol} is the total energy output of the source, and is often given in solar luminosities, $L_{\odot} = 3.826 \times 10^{33}$ erg/sec.

Absolute magnitudes: astronomers usually denote the luminosity of objects by their absolute magnitudes, which are the magnitudes that the object would have at a standard distance, chosen to be 10 parsecs. The absolute magnitude is usually denoted M while the apparent magnitude is m. Their difference is the *distance modulus* or DM:

$$DM = m - M = 5 \log(D/[10 \text{ pc}]). \tag{3}$$

Thus if the distance modulus of the Large Magellanic Cloud is 18.5 then the LMC has a distance of $D = 10^{18.5/5} \times [10 \text{ pc}] = 50.119$ kpc.

2. Astrometry

Proper motions: Stars have individual velocities, that lead to *proper motions.* These aren't "correct" motions, but are rather the property of individual stars. The proper motion is given by $\mu = v_{\perp}/D$ or with v in km/sec and D in parsecs, $\mu\ ''/\text{yr} = 0.21 v_{\perp}/D$. Superimposed on this constant velocity motion is the annual parallax due to the motion of the Earth around the Sun.

A *parsec* is the distance at which an object would have an annual parallax of $1''$. It is thus

$180 \times 60 \times 60/\pi$ au. The astronomical unit (au) is the semi-major axis of the orbit of a test body around the Sun with a period of exactly one sidereal year (31558149.984 sec). The au has been measured by planetary radar to great precision (1.495979×10^{13} cm), but it used to be determined using the diurnal parallax of nearby planets and asteroids, caused by the rotation of an obervatory around the Earth's axis, or by measuring transits of Venus from "the ends of the Earth". Thus 1 pc = 3.085678×10^{18} cm = 3.2616 light-years.

If a star has a parallax of p arcseconds, its actual motion during the year is $\pm p''$, and its distance is $D = 1/p$ pc. Parallaxes can be measured to $\pm 0.003''$ so stars within 100 pc have distances determined by trigonometric parallaxes to an accuracy of better than 30%. The HIPPARCOS satellite measured parallaxes for about 10^5 stars with accuracies $< 0.001''$.

3. Spectroscopy

By measuring the spectrum of a star, one can determine its temperature by comparing the line strengths of lines with highly excited lower levels to the strengths of lines with low excitation lower levels. Since the populations in states with energy E is proportional to $\exp(-E/kT)$ (Boltzmann), this ratio gives T. The spectral type of a star indicates its surface temperature, from O (hottest) through BAFGK to M (coldest). These types are subdivided with a digit: O5, O6... O9, B0... B9, A0..., M10. Two new classes, L & T, have been added recently for very cool *brown dwarf* stars. These stars never burn hydrogen into helium because their masses are less than about 0.08 $M_\odot$. By comparing line ratios from different ions of the same element, one can determine the electron density n_e, since the ratio of ions to neutrals is given by $n_e n_{ion}/n_{atom} \propto \exp(-E_{ion}/kT)$ (Saha). The electron pressure is determined by the opacity and the surface gravity of the star, so it is possible to determine g from a spectrum. The surface gravity of a star is denoted by a roman numeral, from I (lowest) to V (highest). Since low surface gravity implies a large radius through $g = GM/R^2$, and large radii imply large luminosity, these are known as luminosity classes. I is a supergiant, III is a giant, and V is a dwarf or main sequence star.

The spectrum of a star also gives its radial velocity through the Doppler shift. A positive radial velocity is moving away from us: $v_{rad} = dD/dt$.

4. Miscellaneous Parallaxes

Moving cluster method: The top part of Figure 2 shows the space motion of a cluster of stars. Notice that the velocity vectors are parallel so the cluster is neither expanding nor contracting. But when we look at the motions of the stars projected on the sky we see them converging because of perspective effects. The angle to the convergent point is θ. If the cluster is moving towards us the convergent point is really a "divergent" point. The mean proper motion of the cluster, μ, gives $d\theta/dt$. We also need the radial velocity v_r of the cluster measured using the Doppler shift. The

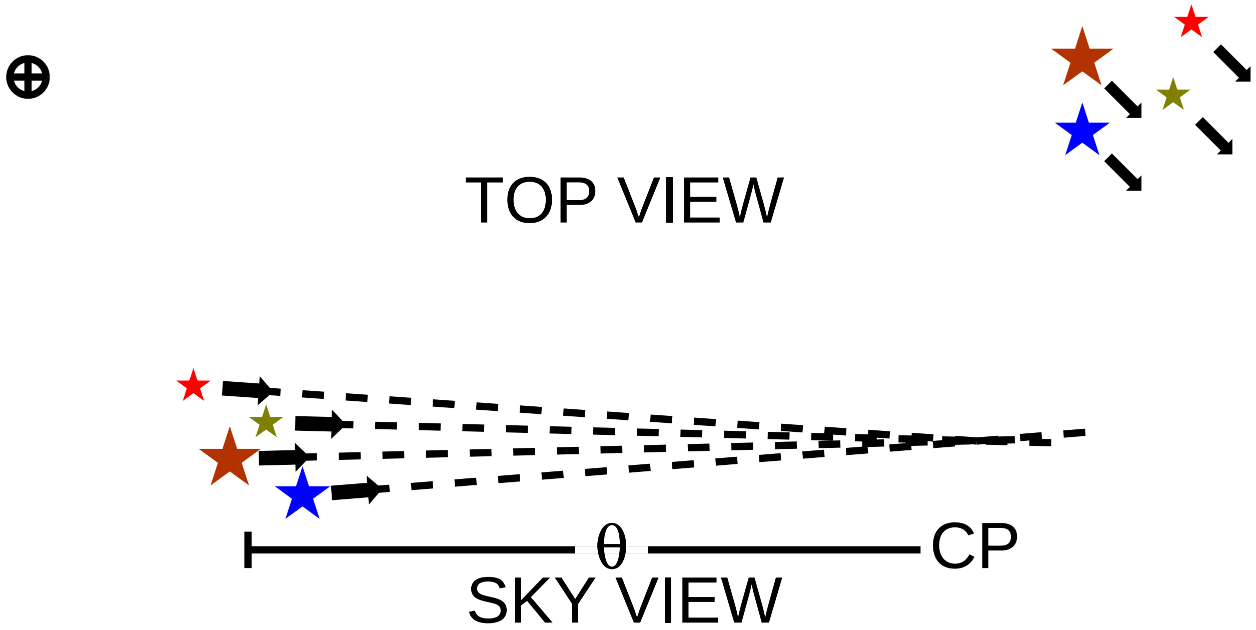

Fig. 2.— Moving cluster parallax. A cluster of stars with fixed physical size has a measured radial velocity and proper motion vectors that seem to converge at the convergent point (CP). These data can be combined to give an accurate distance.

transverse velocity, v_t, of the cluster can be found using $v_t/v_r = \tan\theta$. The distance of the cluster is then

$$\begin{aligned} D &= \frac{v_t}{d\theta/dt} \\ D\ [\mathrm{pc}] &= \left(\frac{v_r}{4.74\ \mathrm{km/sec}}\right)\left(\frac{\tan\theta}{\mu\ [''/\mathrm{yr}]}\right) \end{aligned} \tag{4}$$

The odd constant 4.74 km/sec is one au/year.

Spectroscopic parallax: stars of a given spectral type tend to have the same luminosity, so if the spectral type and flux are known, the distance of a star can be estimated. This is much more reliable when it is done with a cluster of stars, all at the same distance. One can then identify the main sequences stars from colors alone, and get a distance to the cluster by main sequence fitting.

Baade-Wesselink distances: when a star has a variable size, then a comparison of its radial velocity to the size variation can be used to find a distance. This method requires a calibration of the surface brightness or *intensity* as a function of the temperature. Intensity I_ν is the power per unit area per unit frequency per unit solid angle. Typical units for I_ν are erg/cm^2/sec/sr/Hz or Jy/sr. A blackbody has an intensity $I_\nu = B_\nu(T)$ where

$$B_\nu(T) = \frac{2h\nu(\nu/c)^2}{\exp(h\nu/kT) - 1} \tag{5}$$

is the Planck function. The integral of the Planck function over frequency is the total power from

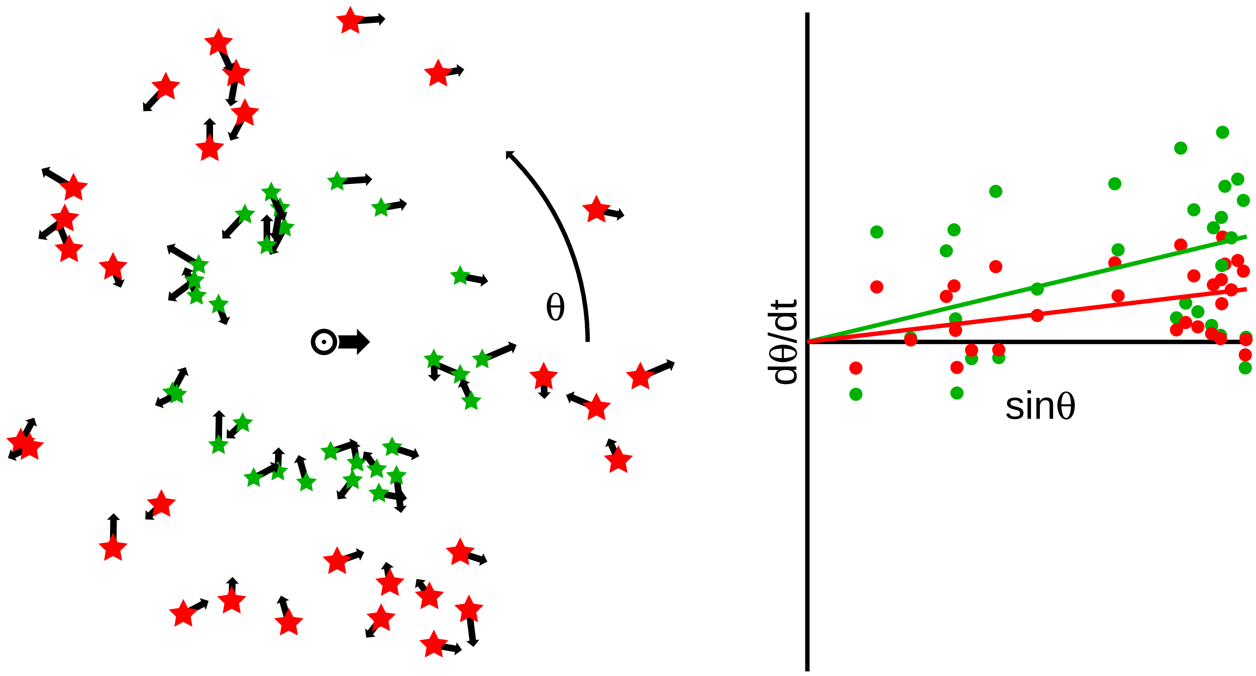

Fig. 3.— Secular and statistical parallaxes: the mean proper motion toward the anti-apex of the solar motion is inversely proportional to the distance, giving the secular parallax. The scatter in the proper motion is also inversely proportional to the distance, giving the statistical parallax.

a blackbody:

$$I_{bol} = \int B_\nu(T)d\nu = \frac{2\pi^4 k^4}{15h^3c^2}T^4 = \frac{\sigma_{SB}}{\pi}T^4 \tag{6}$$

The flux is $F = \int I(\theta,\phi)\cos\theta d\Omega$ but for any star all the angles to the line of sight θ are so small that $\cos\theta \approx 1$. Thus the bolometric flux is $F_{bol} = \theta^2\sigma_{SB}T^4$ where θ is the angular radius of the star. While θ is too small to be measurable except for a few nearby stars, we can solve for

$$\theta(t) = \sqrt{\frac{F_{bol}(t)}{\sigma_{SB}T(t)^4}} \tag{7}$$

Comparing this time variation to the radial velocity variation using $\Delta v_{rad} = Dd\theta/dt$ then gives the distance. This has been applied to pulsating stars, RR Lyra stars and Cepheids, and to Type II supernovae. Note that we had to know the surface brightness vs temperature of the star. We assumed a blackbody above, but if we assume the star had $I_\nu = \epsilon B_\nu(T)$, with emissivity $\epsilon < 1$, then the derived distance would be decreased by the factor $\sqrt{\epsilon}$.

Statistical parallax: If a population of stars has an isotropic velocity distribution, then the scatter in measured radial velocities can be compared to the observed scatter in proper motions to determine a distance for the set of stars.

Secular parallax: the Sun is moving at 19.5 km/sec toward right ascension $\alpha = 18^h4^m$ and declination $\delta = 30°$. This direction the Sun is moving is called the *apex*. The motion is with respect to the average velocity of the stars near the Sun, known as the Local Standard of Rest (LSR). This speed is 4.14 au/yr. Thus a star at rest with respect to the LSR, at a distance of D pc, will have a proper motion of $4.14''/D$ per year if it is perpendicular to the motion of the Sun. We cannot know that any single star is at rest with respect to the LSR, but if we have a population of stars all at the same distance D pc from the Sun they will show an average proper motion toward the anti-apex of $4.14 \sin\theta/D\ ''/\mathrm{yr}$ where θ is the angle between the star and the apex.

5. Interstellar reddening and extinction

The flux of a distant star is reduced not just by the inverse square law, but also by *interstellar extinction*. The apparent magnitude is given by

$$V = M_V + 5\log(D/10) + A_V \tag{8}$$

where A_V is the visual extinction in magnitudes. This extinction is caused by interstellar dust grains that absorb and scatter light. These grains are small compared to the wavelength of light so the extinction increases at shorter wavelengths. Thus A_B is larger than A_V. Since

$$\begin{aligned} B &= M_B + 5\log(D/10) + A_B \\ B - V &= M_B - M_V + A_B - A_V \\ &= (B - V)_\circ + E(B - V) \end{aligned} \tag{9}$$

with the intrinsic color of the star being $(B-V)_\circ = M_B - M_V$ and the color excess begin $E(B-V) = A_B - A_V$. Since the color excess is positive, the stars get redder, and this is called reddening. The extinction A_V is approximately proportional to the column density of interstellar gas, and for a typical density of 1 atom of H per cc the extinction is about 1 magnitude per kpc. The ratio of the color excess to the extinction is fairly constant at $R = A_V/E(B-V) = 3.1$ but there are variations in this ratio. Furthermore, the ratio of color excesses is fairly constant at $E(U-B)/E(B-V) = 0.72$. This means that the color combination $Q = (U-B) - 0.72(B-V)$ is reddening free – not affected by reddening. The magnitude $V' = V - 3.1(B-V)$ is also reddening free so the V', Q diagram can be used for main sequence fitting without worrying about extinction and reddening.

The infrared photometry bands such as K are only slightly affected by reddening. For example, $A_K/A_V = 0.09$. However, in some regions like the galactic center, where $A_V = 30$ magnitudes, the extinction at K is still important.

6. Stellar Evolution

The Hertzsprung-Russell (H-R) diagram plots the luminosity of stars versus their temperatures. The majority of stars in a volume limited sample are arranged in a one dimensional track known as the main sequence. These are the stars that are burning H $\rightarrow$ He in their cores. After about

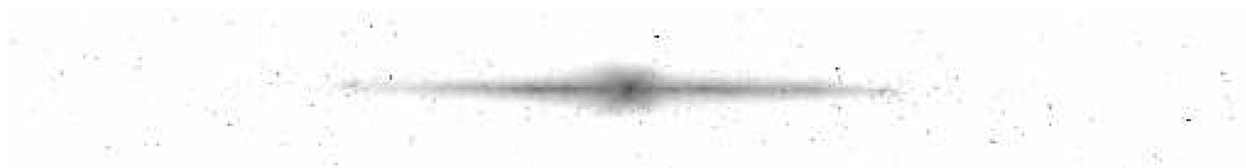

Fig. 4.— The Milky Way in the infrared at 3.5 μm. This picture covers $360° \times 45°$.

10% of the total mass of the star has been converted from hydrogen to helium, the star expands to become a red giant which has a helium core and a hydrogen buring shell. Since the luminosity of a main sequence star is a strong function of its mass, $L \propto M^4$ near 1 $M_\odot$, the lifetime of stars decreases quickly with mass. Thus if a cluster of stars has a large number of stars of different mass but equal ages (an isochrone), then the main sequence will extend from low mass faint red dwarfs up to a main sequence turnoff (MSTO) at a mass given by $M \propto t^{-1/3}$ with $L \propto t^{-4/3}$. Thus the age of cluster can be determined from the luminosity of the stars at the MSTO. But this requires an accurate distance to the cluster.

After stars with $M < 8$ $M_\odot$ become red giants, the helium core grows until central helium burning ignites, and the star heats up and moves along the horizontal branch in the H-R diagram, then cools off as the core burning stops. The once again red star then goes to higher luminosity on the asymptotic giant brach (AGB). Finally they eject their envelopes and become planetary nebulae. The gas dissipates and only the central star, which was the core of the star, remains. This just cools off as a white dwarf with a mass $\leq 1.4 M_\odot$ and a radius close to $R_\oplus$. The oldest, coldest white dwarfs can be used to estimate the age of the disk of the Milky Way. Using the Hubble Space Telescope, Hansen *et al.* (2002) were able to see the oldest, coldest white dwarfs in the globular cluster M4, and derived an age of 12.7 ± 0.7 Gyr for this cluster.

7. Milky Way

The Milky Way, our own galaxy, is a very flat disk of stars. The thickness of the disk is only about 1% of the radius out from the center. In the visible bands, interstellar extinction only allows us to see a few kpc, so the Milky Way appears as a band about ten degrees wide, but in the infrared we can see most of the way across the galaxy so the thinness of the disk is more apparent.

We define a set of angular coordinates (l, b) based on the Milky Way, known as galactic coordinates. The pole of this coordinate system ($b = 90°$) is perpendicular to the disk of the Milky Way, and the zero of galactic longitude l points toward the galactic center (GC).

The surface brightness of the disk of the Milky Way varies like $I = I_\circ \exp(-R/R_d)$ where the disk scale length is $R_d = 3.5 \pm 0.5$ kpc. The location of the Sun is $R_\circ = 8.5 \pm 1$ kpc from the GC. Because the Sun is 2.6 scale lengths out from the GC, 70% of the light from the disk is emitted from radii interior to the Sun's position.

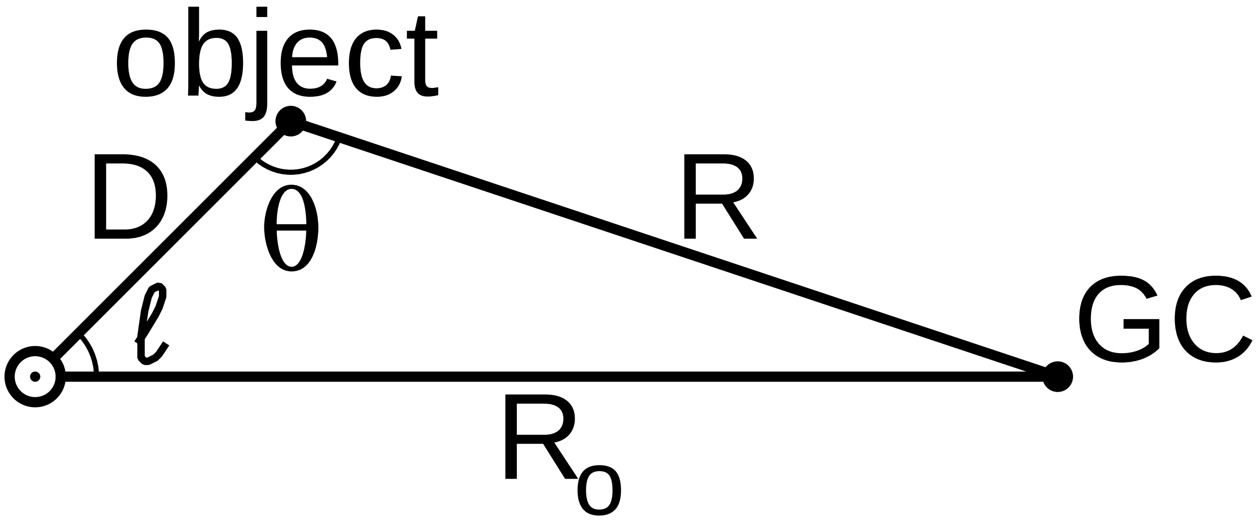

Fig. 5.— Diagram showing the triangle made by the Sun, the Galactic Center and an object at $l = 45°$ and a distance $D = 3$ kpc. The angle θ at the object is needed to convert circular velocities into radial and tangential components.

This thin disk of stars is rotating with a mean circular velocity $v_c(R)$. This rotation is in the sense that the local standard of rest is moving toward $l = 90°,\ b = 0°$. The circular velocity at the Sun's distance from the GC is $v_c(R_\circ) = 220 \pm 15$ km/sec. The rotation of the disk is differential, so the angular speed $\Omega(R) = v_c(R)/R$ decreases with radius. Observations of the radial velocities of nearby stars can be used to determine the amount of differential rotation, since the radial velocity is given by

$$v_r = \sin\theta\, R\ [\Omega(R) - \Omega(R_\circ)] \tag{10}$$

where R is the distance of the object from the GC, and θ is the angle between the GC and the Earth measured at the object as shown in Figure 5. By the law of sines,

$$\frac{\sin\theta}{R_\circ} = \frac{\sin l}{R} \tag{11}$$

so we can write

$$v_r = \sin l\, R_\circ\ [\Omega(R) - \Omega(R_\circ)] \tag{12}$$

Finally we can use $R \approx R_\circ - D\cos l$ where D is the distance from the Earth to the source, as long as $D << R_\circ$. Then

$$v_r = -\sin l \cos l\, R_\circ\, D \frac{\partial\Omega}{\partial R} \tag{13}$$

Then with

$$\frac{\partial\Omega}{\partial R} = \frac{1}{R}\frac{\partial v_c}{\partial R} - \frac{v_c}{R^2} \tag{14}$$

we get

$$\begin{aligned} v_r &= -D\cos l \sin l\ R_\circ \frac{\partial}{\partial R}\left(\frac{v_c}{R}\right) \\ &= D\sin(2l)\,\frac{1}{2}\left(\frac{v_c}{R} - \frac{\partial v_c}{\partial R}\right) \end{aligned} \tag{15}$$

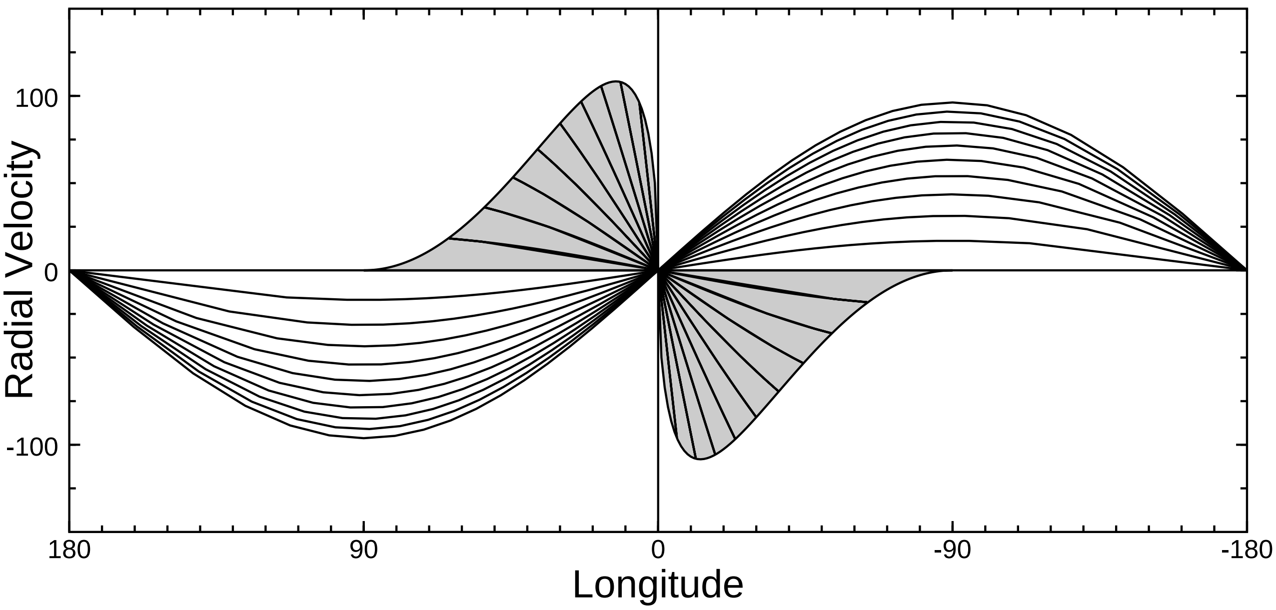

Fig. 6.— Radial velocity *vs.* galactic longitude for a differentially rotating disk. The shaded area is corresponds to emission from inside the solar circle ($R < R_\circ$) while the curves show emission from material in rings at $R = 1, 2, \ldots, 20$ kpc.

The Oort constant

$$A = \frac{1}{2}\left(\frac{v_c}{R} - \frac{\partial v_c}{\partial R}\right) = 14 \pm 1.5 \text{ km/sec/kpc} \tag{16}$$

is thus well determined from radial velocities. In order to determine the absolute value of the angular speed, and thus the length of the "galactic year", we need to measure proper motions with an accuracy much better than (220/8.5) km/sec/kpc $= 0.005''$/yr. The tangential velocity is

$$v_t = v_c(R)\cos\theta + v_c(R_\circ)\cos l \tag{17}$$

The law of cosines gives

$$\begin{aligned} \cos\theta &= \frac{R_\circ^2 - D^2 - R^2}{-2DR} \\ &= \frac{R_\circ^2 - D^2 - (D^2 + R_\circ^2 - 2DR_\circ\cos l))}{-2DR} \\ &= \frac{-2D^2 + 2DR_\circ\cos l}{-2DR} \\ &= \frac{D}{R} - \frac{R_\circ}{R}\cos l \end{aligned} \tag{18}$$

Therefore the tangential velocity is

$$\begin{aligned} v_t &= \frac{D}{R}v_c(R) + \cos l\left(v_c(R_\circ) - \frac{R_\circ}{R}v_c(R)\right) \\ &= D\frac{v_c(R)}{R} + \cos l\, R_\circ\left(\frac{v_c(R_\circ)}{R_\circ} - \frac{v_c(R)}{R}\right) \\ &\approx D\left(\Omega(R) + R_\circ\cos^2 l\frac{\partial\Omega(R)}{\partial R}\right) \end{aligned} \tag{19}$$

The proper motion dl/dt is $-v_t/D$ since the galaxy rotates clockwise as seen from the North Pole while positive angles are counterclockwise. This gives

$$\begin{aligned}
\frac{dl}{dt} &= -\left(\Omega(R) + R_\circ \frac{1+\cos(2l)}{2}\frac{\partial\Omega(R)}{\partial R}\right) \\
&= \left(-\frac{1}{2}\left[\frac{v_c}{R} + \frac{\partial v_c}{\partial R}\right]\right) + \cos(2l)\left(-\frac{1}{2}\left[\frac{\partial v_c}{\partial R} - \frac{v_c}{R}\right]\right) \\
&= B + A\cos(2l) \qquad (20)
\end{aligned}$$

where the Oort constant B is given by

$$B = -\frac{1}{2}\left[\frac{v_c}{R} + \frac{\partial v_c}{\partial R}\right]. \qquad (21)$$

Another way to derive B is to note that the proper motion due to the rotation at the solar circle is just $dl/dt = -v_c(R_\circ)/R_\circ$. There is another contribution from differential rotation, giving $-\cos\theta[\Omega(R) - \Omega(R_\circ)]R/D$ The total proper motion is then

$$\frac{dl}{dt} = \Omega(R_\circ) - \frac{\cos\theta\, R\,[\Omega(R) - \Omega(R_\circ)]}{D} \qquad (22)$$

For small D, $\cos\theta = -\cos l$ since the object-Earth angle at the Galactic Center is close to zero, so

$$\begin{aligned}
\frac{dl}{dt} &\approx -\Omega(R_\circ) - \cos^2 l\, R\frac{\partial\Omega}{\partial R} \\
&= -\frac{v_c}{R} - \cos^2 l\left(\frac{\partial v_c}{\partial R} - \frac{v_c}{R}\right) \\
&= -\frac{v_c}{R} - \frac{1+\cos(2l)}{2}\left(\frac{\partial v_c}{\partial R} - \frac{v_c}{R}\right) \\
&= B + A\cos(2l) \qquad (23)
\end{aligned}$$

Thus the Oort constant

$$B = -\frac{1}{2}\left(\frac{v_c}{R} + \frac{\partial v_c}{\partial R}\right) = -12 \pm 3 \text{ km/sec/kpc} \qquad (24)$$

can only be found from proper motions and is not as well determined as A. Note that the rotation curve of the Milky Way is almost flat, since $\partial v_c/\partial R = -(A+B) = 2 \pm 3$ km/sec/kpc which is consistent with zero.

One interesting fact is that the proper motion of the radio source Sgr A* in the GC is $dl/dt = -0.0059 \pm 0.0004$ ″/yr (Reid *et al.*, astro-ph/9905075). This proper motion is due to the motion of the Sun around the GC, plus any motion of Sgr A* with respect to the true center of mass of the Milky Way. Ignoring the latter gives

$$\frac{v_c(R_\circ) + 15.3 \text{ km/sec}}{R_\circ \text{ [kpc]}} = 28.1 \pm 1.9 \text{ km/sec/kpc} \qquad (25)$$

where 15.3 km/sec is the tangential component of the Sun's motion with respect to the LSR. For $R_\circ = 8.5$ kpc this implies $v_c(R_\circ) = 224 \pm 16$ km/sec which is consistent with other determinations. But this result is quite precise indicating the power of radio VLBI observations.

The random velocities of stars in the disk are about ± 20 km/sec, similar to the Sun's motion with respect to the LSR. As a result, 99% of the kinetic energy of the disk is due to the circular velocity of rotation.

7.1. Spheroid

In addition to the disk, the Milky Way also has a spheroid of stars that are in a much more extended, almost spherical distribution. The ages of these halo stars are typically larger than the ages of stars in the disk, and their chemical compositions are much less metal rich. These stars are known as Population II stars, while the stars in the disk are known as Population I stars. The part of this spheroid close to the GC is bright enough to be easily seen, and is called the bulge. Recent observations have shown that the bulge is not an oblate ellipsoid, but rather is triaxial with its longest axis pointing toward $l = 135^\circ$ from the GC. Thus when we look toward $l = 20^\circ$ we are looking at the closer end of this bar shaped bulge and this part of the bulge extends to higher galactic latitudes b than the part seen at $l = 340^\circ$ which is further away from us.

The halo is not rotating very much, but the random velocities of stars in the halo are similar in magnitude to the rotation speed of the disk. Thus most of the kinetic energy in the halo is due to random motion.

7.2. Globular clusters

Globular clusters are the easiest halo objects to detect. The density of globular clusters seems to fall off as $R^{-3.5}$ or so, but the mass density of must fall off more slowly (R^{-2}) in order to give the observed flat rotation curve. Globular clusters have 10^6 stars, are nearly spherical, and have a core radius of about 1.5 pc, with a tidal radius of 50 pc. The stars in a globular cluster have random velocities of ± 7 km/sec, and negligible net rotation.

8. Other Galaxies

The morphologies of galaxies are classified into two main types: ellipticals (E) and spirals (S). Spiral galaxies are thin disks with bulges like the Milky Way. The spiral patterns are generally logarithmic spirals where the radius of an arm follows the law $r \propto \exp(\alpha\theta)$. The arms can be tightly wound with a low value of α, or very open with a high value of α. α is the pitch angle of the arms. Some spirals have bars: straight segments through the center that connect to the spiral arm pattern. These barred spirals are denoted SB while the unbarred spirals are just S. It now appears that the Milky Way is a barred spiral. Spiral galaxies have bulges that dominate the central part of the disk. The ratio of bulge to disk and the pitch angle α are correlated, with bright bulges

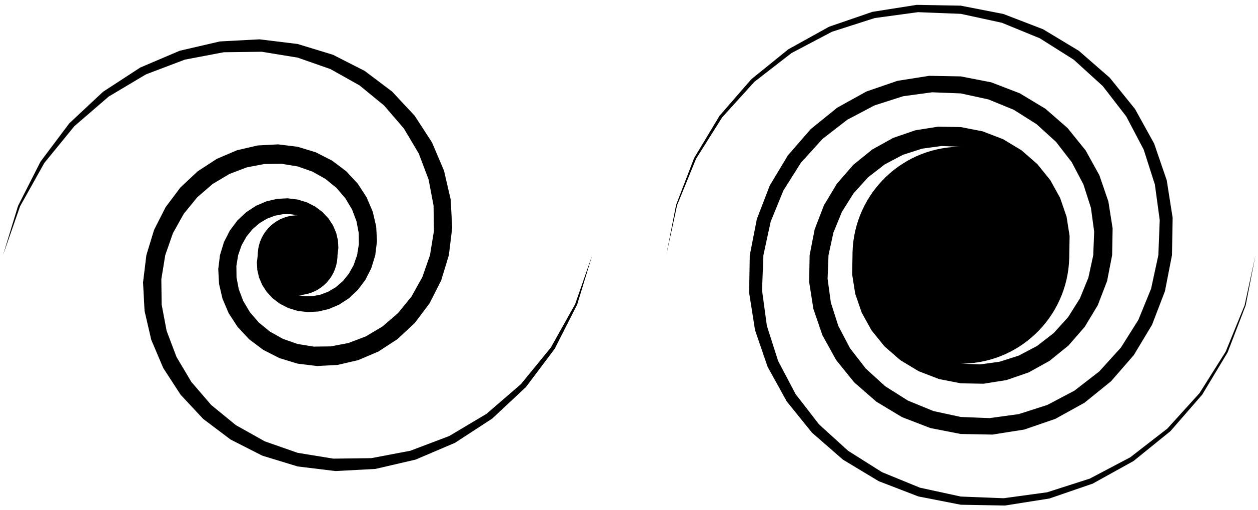

Fig. 7.— Spiral galaxy patterns with different arm pitch angles: $\alpha = 0.23$ on the left and $\alpha = 0.12$ on the right.

going with tightly wound arms. Galaxies with these characteristics are known as Sa (or SBa for the barred version.) Galaxies with very weak bulges tend to have loose arms with high α. These galaxies are Sd or SBd. Sd and SBd galaxies are known as late type galaxies, and tend to have lots of gas and ongoing star formation. Sa and SBa galaxies have little gas, while E galaxies have very little gas or star formation, and are known as early type galaxies.

The light from the disk of spirals follows an exponential law like the Milky Way. The light from the bulge follows laws appropriate for elliptical galaxies.

Elliptical galaxies have elliptical shapes with axes b and a. The ratio of the axes determines the subtype of an E galaxy though $10(1 - b/a)$. Thus a circular galaxy with $b = a$ is an E0 galaxy, while the most elliptical E galaxies have $b/a = 0.3$ and are thus $E7$ galaxies.

Elliptical galaxies are not elongated because of rotation. When a spectrograph slit is laid along the major axis of the elliptical shape, there is very little gradient in the velocity of the the stars. There is, however, a great broadening of the spectral lines, which are several hundred km/sec wide as compared to stellar lines that are usually much narrower. (The lines from the Sun are about 3 km/sec wide.) Thus over 90% of the kinetic energy of the stars in elliptical galaxies is due to random motion.

Some galaxies clearly show a disk but have no gas or ongoing star formation. These galaxies are known as S0 ("ess zero") galaxies. Without young stars marking out spiral arms, these are easily confused with ellipticals when seen face-on.

The diagram below shows the Hubble classification of galaxies in the traditional "tuning fork" diagram. This is *not* an evolutionary sequence! In fact, many astronomers believe ellipticals are the result of mergers between two or more spirals, so the evolution would go the other way.

$$
\begin{tikzcd}
 & & & \mathrm{Sa} \arrow[r] & \mathrm{Sb} \arrow[r] & \mathrm{Sc} \arrow[r] & \mathrm{Sd} \\
\mathrm{E0} \arrow[r] & \mathrm{E7} \arrow[r] & \mathrm{S0} \arrow[ur, bend left] \arrow[dr, bend right] & & & & \\
 & & & \mathrm{SBa} \arrow[r] & \mathrm{SBb} \arrow[r] & \mathrm{SBc} \arrow[r] & \mathrm{SBd} \\
\text{Early} \arrow[rrrrrr] & & & & & & \text{Late}
\end{tikzcd}
$$

There are two empirical laws used for the distribution of brightness with radius in E galaxies: the Hubble law $I = I_\circ/(1 + r/a)^2$ and the de Vaucouleurs law $I = I_\circ \exp(-\alpha r^{1/4})$. The de Vaucouleurs law integrates to a total flux of

$$F = 2\pi \int rI(r)dr = 8\pi\alpha^{-8} \int y^7 e^{-y} dy = 8!\pi\alpha^{-8} I_\circ \tag{26}$$

with $y = \alpha r^{1/4}$. The Hubble law integrates to an infinite flux so it was modified by Abell giving

$$I = \begin{cases} I_\circ/(1+r/a)^2, & r/a < 21.4; \\ 22.4\, I_\circ/(1+r/a)^3, & r/a > 21.4. \end{cases} \tag{27}$$

Note that this law is quite sharply peaked at the center and does not have a continuous second derivative there. The de Vaucouleurs law has the same property. Thus E galaxies have very bright centers. The total flux for this law is

$$F = 2\pi a^2 I_\circ \left(\ln 22.4 - \frac{21.4}{22.4} + \left[1 - \frac{1}{2 \times 22.4} \right] \right) = 6.26\pi a^2 I_\circ \tag{28}$$

Because of the diffuse outer regions of galaxies, it is difficult to measure the total flux, and thus the radius including half the total flux is hard to find. An easier radius to find is the radius at which the encircled flux is growing as the first power of radius:

$$2\pi r_1^2 I(r_1) = \int_0^{r_1} 2\pi I(r) r dr \tag{29}$$

because this depends only on the brighter inner regions. For the exponential disk we have to solve $x^2 = e^x - (1+x)$ to find this radius which gives $r_1 = 1.7933 R_D$. This radius includes 54% of the total flux. For the Hubble-Abell law, we have to solve $x^2/(1+x)^2 = \int_0^x [y/(1+y)^2] dy = \ln(1+x) - x/(1+x)$ which occurs at $x = r_1/a = 2.1626$. This radius includes 15% of the total flux. For the de Vaulcouleurs law we have to solve $x^2 \exp(-x^{1/4}) = \int x \exp(-x^{1/4}) dx$ or $y^8/e^y = 4 \int y^7 e^{-y} dy$. This occurs at $y = 4.8827$ and this radius includes 12% of the total flux. Thus the Hubble-Abell law and the de Vaucouleurs law give similar profiles for elliptical galaxies, and both give much more diffuse emission than the exponential disk.

8.1. Luminosity Function

Galaxies do not all have the same luminosity. Dwarf galaxies like the LMC and M32 are very common, while giant elliptical galaxies like M87 can be seen to great distances. The number of galaxies per unit volume in different luminosity ranges in known as the luminosity function. A very useful analytic approximation to the luminosity function is the Schechter (1976) luminosity function:

$$n(L)dL = \phi_*(L/L_*)^\alpha \exp(-L/L_*)dL/L_* \tag{30}$$

Parameters derived from different galaxy surveys are slightly different, but one set from Loveday *et al.* (1992, ApJ, 390, 338) is $\alpha = -0.96$, $\phi_* = 0.0124$ Mpc^{-3}, and $M(B_J) = -19.67$. (B_J is a photometric band between B and V defined by using a IIIaJ photographic plate and a green glass filter.) These values were derived assuming a Hubble constant $H_\circ = 100$ km/sec/Mpc. If $H_\circ = 50$ is used instead, all the galaxies are twice as far away so the density ϕ_* is 8 times lower and the luminosity L_* is 4 times higher. Loveday *et al.* also found that E and S0 galaxies made up 27% of the sample and that a separate fit to E/S0 galaxies alone gave $\alpha = -0.07$. Thus most of the faint galaxies are spirals or irregulars while more of the bright galaxies are ellipticals.

8.2. Velocity-Luminosity Laws

Both elliptical and spiral galaxies have empirical laws relating their luminosity to their internal velocity. For spirals, this law is known as the Tully-Fisher law:

$$L \propto (\Delta v)^4_{20} \tag{31}$$

where $(\Delta v)_{20}$ is the full width of the 21 cm line from the galaxy at the 20% of peak power level. This width is basically twice the peak circular velocity.

For elliptical galaxies, the Faber-Jackson law relates the velocity dispersion σ measured from the line broadening mentioned earlier and the luminosity. This law is

$$L = L_* \left(\frac{\sigma}{220\ \mathrm{km/sec}}\right)^4 \tag{32}$$

Thus both spirals and ellipticals have luminosities that follow the fourth power of a typical velocity.

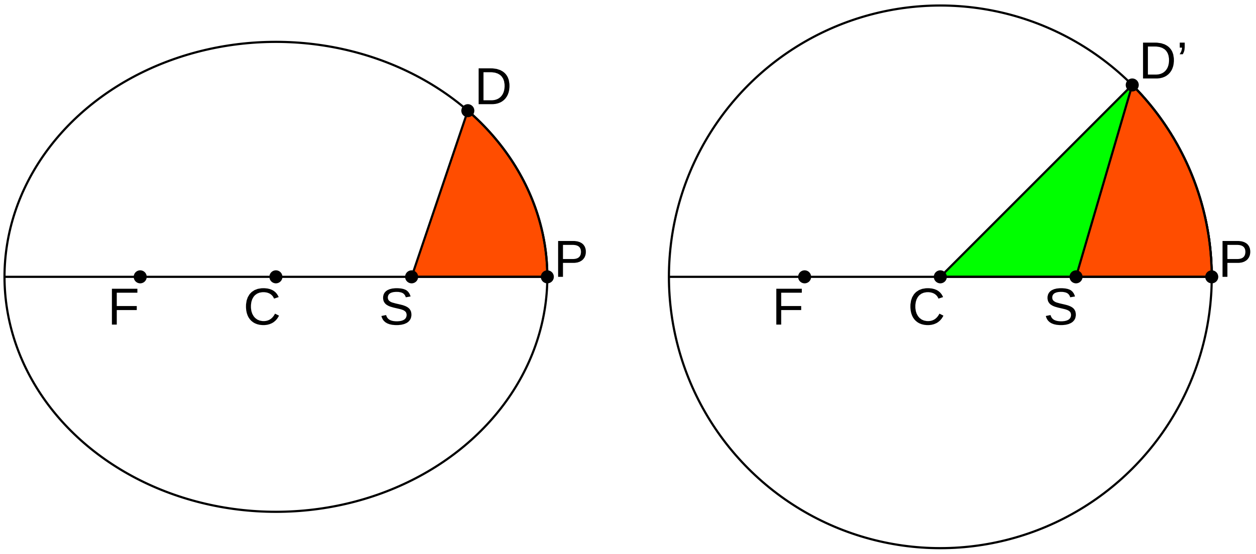

Fig. 8.— Derivation of Kepler's Equation: the area of the elliptical sector SPD on the left is hard to compute, but by stretching the ellipse into a circle this area can be found as the difference between the area of the sector CPD' and the triangle CSD'.

9. Kepler and all that

We will make use of the solutions of the two-body problem several times in this course, so I will review them here. Consider the 2-body orbit of two particles with mass m_1 and m_2. Each particle experiences a force with magnitude $F = Gm_1m_2/r^2$ where $\vec{r} = \vec{r}_1 - \vec{r}_2$ is the interparticle separation and thus the accelerations of the particles around their mutual center of mass are $\ddot{\vec{r}}_1 = -Gm_2\hat{r}/r^2$ and $\ddot{\vec{r}}_2 = +Gm_1\hat{r}/r^2$. Therefore $\ddot{\vec{r}} = -G(m_1 + m_2)\hat{r}/r^2$ which is the same as the equation for a test particle ($m \to 0$) moving in the potential of a mass $M = m_1 + m_2$. Kepler's First Law states that the orbit of a planet is an ellipse with the Sun at one focus. An ellipse centered at $x = 0$ and $y = 0$ has the equation $x^2/a^2 + y^2/b^2 = 1$ where a is the *semi-major axis* and b is the *semi-minor axis* [note that $a > b$]. But the foci are at $x = \pm ae$ where e is the eccentricity of the orbit. The sum of the distances between a point on the ellipse and the two foci is a constant. For the point at $x = a$ and $y = 0$ these distances are $a(1-e)$ and $a(1+e)$ so the constant sum of the distances has to be $2a$. Therefore the semi-minor axis is found using $2\sqrt{b^2 + a^2e^2} = 2a$ so $b = a\sqrt{1-e^2}$.

Kepler's Second Law states that equal areas are swept out by the line between the occupied focus and the orbiting object. The curvilinear "triangle" with corners at the occupied focus S, the periapsis P, and the object D has an area that increases linearly with time, and after one full period it is equal to the area of the full ellipse πab. However, the area SPD is hard to compute. But if we "stretch" the ellipse in the y direction by a factor a/b so it becomes a circle, then the area SPD' is easily found as the difference between the area of the sector CPD' and the area of the triangle CSD'. The area of the sector is $(a^2/2)E$, where the *eccentric anomaly* E is the angle PCD'. The area of the triangle CSD' is given by one-half the base times the height, and the base is ae and the

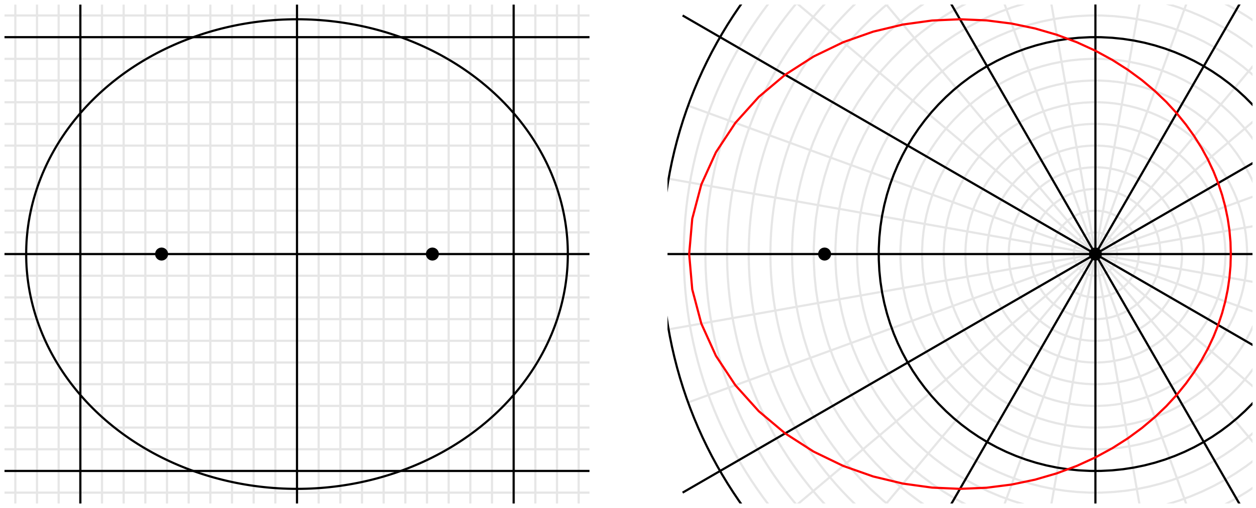

Fig. 9.— Ellipse with $a = 125$ and $e = 0.5$ drawn in Cartesian coordinates on the left and in polar coordinates on the right.

height is $a \sin E$. Scaling these areas by b/a compensates for the previous stretch, so we get

$$\pi ab \frac{t - t_p}{\text{Period}} = \frac{ab}{2}(E - e \sin E) \tag{33}$$

where t_p is the time of periapsis passage. We now define the *mean anomaly*,

$$M = 2\pi(t - t_p)/\text{Period} = \sqrt{\frac{GM}{a^3}}(t - t_p) \tag{34}$$

and have Kepler's Equation:

$$M = E - e \sin E \tag{35}$$

This equation must be solved iteratively. Once E is known it is trivial to find $x = a \cos E$ and $y = b \sin E$. But normally we want coordinates relative to the occupied focus of the ellipse instead of the center of the ellipse, giving $x = a(\cos E - e)$ and $y = b \sin E$. These can be resolved into polar coordinates giving

$$\begin{aligned} \theta &= \tan^{-1}\left(\frac{b \sin E}{a(\cos E - e)}\right) \\ r &= \sqrt{a^2(\cos E - e)^2 + b^2 \sin^2 E} = a(1 - e \cos E) \end{aligned} \tag{36}$$

The angle θ is the angle PSD between the periapsis and the object measured from the occupied focus, and is known as the *true anomaly*. the relation between the radius and the true anomaly is

$$r = \frac{a(1 - e^2)}{1 + e \cos \theta} \tag{37}$$

Both this polar form and the Cartesian form $x^2/a^2 + y^2/b^2 = 1$ of the ellipse are shown in Figure 9.

Kepler's Third Law states that the period squared is proportional to the semi-major axis cubed. In terms of the masses, we have the generalized third law:

$$\frac{4\pi^2}{P^2} = \frac{GM}{a^3} \tag{38}$$

Thus the size of the orbit specifies the period if the masses are known, or the orbit size and period specify the masses.

We can prove that all of Kepler's Laws follow from Newton's Laws using a method from Jacobi (1842, as translated by Gauthier, 2004, AmJPhys, 72, 381). If we define a coordinate system so that the z axis is perpendicular to the initial separation and initial relative velocity, then we have two equations for the x and y components. Since the initial z and $\dot{z}$ are zero, and the force is a central force so $\ddot{z}$ is zero, we have for the third component $z = 0$ at all times. Define $r = \sqrt{x^2 + y^2}$ and ϕ such that $x = r\cos\phi$ and $y = r\sin\phi$. Also define $k^2 = G(m_1 + m_2)$. Then the two remaining equations are

$$\begin{aligned} \frac{d^2x}{dt^2} &= \frac{dv_x}{dt} = \frac{-k^2 x}{r^3} \\ \frac{d^2y}{dt^2} &= \frac{dv_y}{dt} = \frac{-k^2 y}{r^3} \end{aligned} \tag{39}$$

Now let us look at the angular momentum per unit mass of our test particle, which is $\alpha = x\dot{y} - y\dot{x} = r^2\dot{\phi}$. This is also twice the rate at which the area is swept. We see that

$$\frac{d\alpha}{dt} = \dot{x}\dot{y} + x\ddot{y} - \dot{y}\dot{x} - y\ddot{x} = 0 \tag{40}$$

so angular momentum is conserved because this is a central force. This proves Kepler's Second Law, the Law of Areas.

Now let us change variables from t to ϕ. This gives

$$\begin{aligned} \frac{dv_x}{d\phi} &= \frac{dv_x/dt}{d\phi/dt} = \frac{-k^2 x}{r^3}\frac{r^2}{\alpha} = -\frac{k^2}{\alpha}\cos\phi \\ \frac{dv_y}{d\phi} &= \frac{dv_y/dt}{d\phi/dt} = \frac{-k^2 y}{r^3}\frac{r^2}{\alpha} = -\frac{k^2}{\alpha}\sin\phi \end{aligned} \tag{41}$$

We can easily integrate these equations, and see that the velocity vector moves in a circle of radius k^2/α in the v_x, v_y plane, but the circle does not have to be centered at the origin. Thus the velocity vector as a function ϕ is given by

$$\begin{aligned} v_x &= -\frac{k^2}{\alpha}\sin\phi + \beta \\ v_y &= \frac{k^2}{\alpha}\cos\phi + \gamma \end{aligned} \tag{42}$$

We can now find x and y as a function of ϕ:

$$\begin{aligned} \frac{dx}{d\phi} &= \frac{v_x}{d\phi/dt} = \frac{r^2}{\alpha}\left[-\frac{k^2 r^2}{\alpha}\sin\phi + \beta\right] \\ \frac{dy}{d\phi} &= \frac{v_y}{d\phi/dt} = \frac{r^2}{\alpha}\left[\frac{k^2}{\alpha}\cos\phi + \gamma\right] \end{aligned} \tag{43}$$

Remembering that $x = r\cos\phi$ and $y = r\sin\phi$ we can now compute

$$xdy - ydx = r^2 d\phi = \frac{r^3}{\alpha}\left[\frac{k^2}{\alpha}(\cos^2\phi + \sin^2\phi) + \gamma\cos\phi - \beta\sin\phi\right] d\phi \tag{44}$$

We can cancel the $d\phi$ from both sides and get an algebraic equation for r *vs.* ϕ:

$$r = \frac{\alpha^2}{k^2}\left[1 + \frac{\alpha\gamma}{k^2}\cos\phi - \frac{\alpha\beta}{k^2}\sin\phi\right]^{-1} \tag{45}$$

This is the equation for a conic section with one focus at the origin. The eccentricity is $e = (\alpha/k^2)\sqrt{\beta^2+\gamma^2}$. The true anomaly is $\theta = \phi + \tan^{-1}(\beta/\gamma)$. If the eccentricity is less than 1 then we have an ellipse. This is Kepler's First Law. The semi-major axis is given by $a = (r_{max} + r_{min})/2 = \alpha^2/[(1-e^2)k^2]$. The areal rate constant is given by $\alpha = 2\pi ab/P = 2\pi a^2\sqrt{1-e^2}/P$ so the period is deterimined from

$$\begin{aligned} a &= 4\pi^2 a^4/[P^2k^2] \\ \frac{a^3}{P^2} &= \frac{k^2}{4\pi^2} = \frac{G(m_1+m_2)}{4\pi^2} \end{aligned} \tag{46}$$

This is Kepler's Third Law. The book "Feynman's Lost Lecture" by Goodstein & Goodstein contains a derivation of these results that does not use calculus, but what takes a book without calculus is only a couple of pages with calculus.

9.1. Spectroscopic Binaries

In a spectroscopic binary we can measure the velocity amplitude which is given by $K = k^2/\alpha$. Usually we can only measure one component, typically the more massive and thus more luminous component. Furthermore, we don't know the inclination, so what we actually measure is

$$K_1 = \frac{m_2\sin i}{m_1+m_2}\,\frac{G(m_1+m_2)P}{2\pi a^2\sqrt{1-e^2}} \tag{47}$$

We expect $K_1 P/2\pi$ to be a radius, and of course K_1 is a velocity, so the mass of the system should be proportional to K^3P/G. In fact

$$K_1^3(1-e^2)^{3/2}P/(2\pi G) = \frac{G^2 m_2^3\sin^3 i P^4}{(2\pi)^4 a^6} = \frac{m_2^3\sin^3 i}{(m_1+m_2)^2} = m_2\frac{\sin^3 i}{(1+m_2/m_1)^2} \tag{48}$$

is called the *mass function*. Clearly the mass function is less than the companion mass m_2 because $\sin i \le 1$ and $1 + m_2/m_1 \ge 1$.

In a double-lined spectroscopic binary we also know the velocity amplitude of the companions K_2 which gives $m_2/m_1 = K_1/K_2$. And in an eclipsing binary we can measure the inclination i which is always close to $90°$ so $\sin i$ is very close to 1 and well-determined. Thus a double-lined eclipsing binary can be solved completely yielding the two stellar masses.

9.2. Visual Binaries

In a visual binary we can measure the angular separation θ. There are two axes projected on the sky, and observation of the visual orbit can determine the inclination i, the eccentricity e, and the period. An eccentric but not inclined orbit will have an off-center ellipse, while an inclined circular orbit will have a centered ellipse on the sky. For a distance D we get a mass estimate

$$m_1 + m_2 = \frac{4\pi^2\theta^3 D^3}{P^2 G} \tag{49}$$

Unfortunately the distance needs to known quite well in order to get a good mass. On the other hand a fairly poor mass estimate can give a good estimate of the distance.

9.3. Hyperbolic orbits

Note that the semi-major axis of a test particle is related to the total energy per unit mass via

$$\text{Energy} = \frac{v^2}{2} - \frac{GM}{r} = -\frac{GM}{2a} \tag{50}$$

When a particle has so much kinetic energy that its total energy is positive, it is on an unbound hyperbolic orbit. By the energy relation in Eq(50) this implies that the semi-major axis a is negative. The equation for a hyperbola is $x^2/a^2 - y^2/b^2 = 1$. The minus sign on the y term comes from the relation for ellipses $b^2 = (1-e^2)a^2$ but with $e > 1$ instead of $e < 1$. By convention the minus sign is put in the equation instead of using a negative b^2, so we have $b^2 = (e^2-1)a^2$ for hyperbolic orbits. Note that b can be less than or greater than $|a|$.

The polar form for ellipses in Eq(37) works perfectly well for hyperbolae, as seen in Figure 10. Both a and $(1-e^2)$ are negative, so r is positive for $\cos\theta > -1/e$. When $\cos\theta < -1/e$, then r is negative and the other branch of the hyperbola is traced. the singularities when $\cos\theta = -1/e$ give the straight lines making an X through the center. These are the *asymptotes* of the hyperbola.

Kepler's Equation for hyperbolae is modified to

$$\sqrt{\frac{GM}{|a|^3}}(t - t_p) = M = e\sinh E - E \tag{51}$$

and the relation between E and the radius is given by

$$\begin{aligned} x &= |a|(\cosh E - e) \\ y &= b\sinh E \\ r &= |a|(e\cosh E - 1) \end{aligned} \tag{52}$$

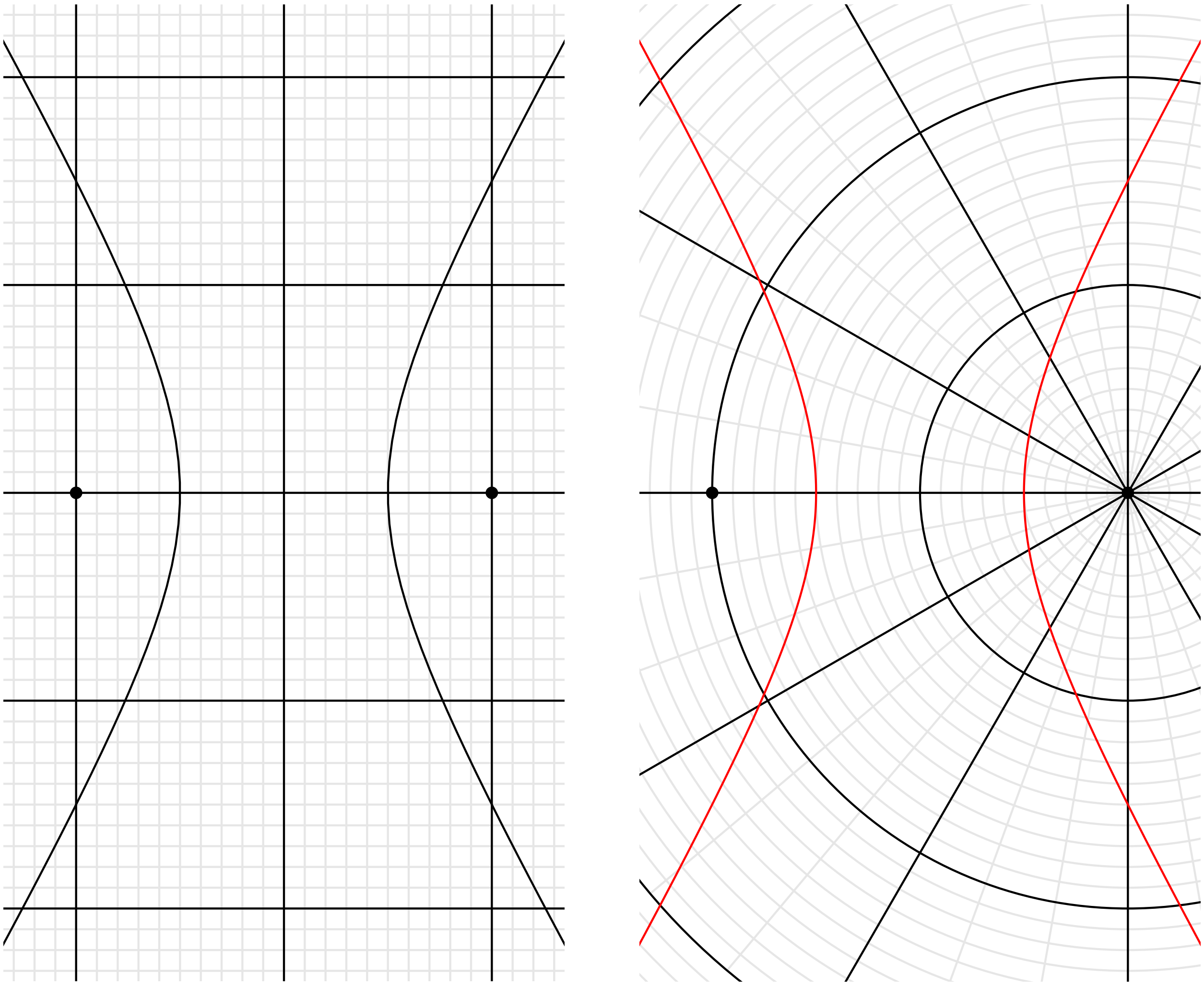

Fig. 10.— Cartesian form of a hyperbola with $a = -50$ and $e = 2$ on the left, and the polar form on the right.

9.4. From elements to positions

Often one wants to compute the position of an object, such as a newly discovered asteroid, from the orbital elements. The elements usually include a, e and the either mean anomaly M at some time or the time of the last periapsis passage. Given these three elements it is easy to compute the x and y coordinates of the object in the coordinate system where the $+x$ direction is toward the periapsis, and the $+y$ direction is where the object moves after periapsis. One just computes M at the desired times, solves Kepler's Equation for E, and then finds x and y. But there are three other orbital elements that specify the orientation of the orbit plane in a standard coordinate system. For objects orbiting the Sun the standard system is ecliptic, while for Earth satellites the celestial coordinate system is used. The orientational elements are: ω, the *argument of the perihelion*, which is the angle between the ascending node and the perihelion measured in the direction of the orbit; i, the *inclination*, which is the angle between the orbit plane and the equator of the standard coordinate system; and Ω, the *longitude of the ascending node*, which is the angle between the zero point of the longitude (or right ascension) and the point where the orbit plane crosses the equator of the standard coordinate system in the upward or North-going direction. Sometimes $\varpi = \Omega + \omega$, the *longitude of the perihelion*, is given instead of ω. The X, Y and Z coordinates in the standard coordinate system are given by

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \cos\Omega & -\sin\Omega & 0 \\ \sin\Omega & \cos\Omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos i & -\sin i \\ 0 & \sin i & \cos i \end{pmatrix} \begin{pmatrix} \cos\omega & -\sin\omega & 0 \\ \sin\omega & \cos\omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \tag{53}$$

9.5. From position and velocity into elements

If the position $\vec{r} = (X, Y, Z)$ and velocity $\vec{v} = (\dot{X}, \dot{Y}, \dot{Z})$ are known, then the elements must be computed. The semi-major axis is found from the energy (per unit mass) of the object:

$$\begin{aligned} -\frac{G(m_1+m_2)}{a} &= 2\left(-\frac{G(m_1+m_2)}{r} + \frac{v^2}{2}\right) \\ a &= \frac{r}{2 - rv^2/GM} \end{aligned} \tag{54}$$

where $M = m_1 + m_2$. The eccentricity and the eccentric anomaly are then found using

$$\begin{aligned} e\cos E &= 1 - \frac{r}{a} \\ e\sin E &= \frac{\vec{r}\cdot\vec{v}}{\sqrt{GMa}} \end{aligned} \tag{55}$$

E should be found using the ATAN2 function in FORTRAN or its equivalent - a two argument arctangent routine that takes both the sine and cosine components and returns an angle covering the entire 0 to 2π range. In general the $\tan^{-1}$ function used in the equations below should be the ATAN2 routine.

The angular momentum per unit mass can also give the eccentricity: a circular orbit has the largest angular momentum while a radial orbit with $e = 1$ has zero angular momentum. Write $L\hat{n} = \vec{r}\times\vec{v}$. Then $L = \sqrt{GMa}\sqrt{1-e^2}$. This formula is not very accurate for nearly circular orbits. But the direction of the angular momentum vector $\hat{n}$ gives the inclination and ascending node of the orbit:

$$\begin{aligned} n_x &= \sin i \sin\Omega \\ n_y &= -\sin i \cos\Omega \\ n_z &= \cos i \end{aligned} \tag{56}$$

which have the solution

$$\begin{aligned} i &= \tan^{-1}\left(\frac{\sqrt{n_x^2 + n_y^2}}{n_z}\right) \\ \Omega &= \tan^{-1}\left(\frac{n_x}{-n_y}\right) \end{aligned} \tag{57}$$

The final element needed is the argument of the perihelion. Define a unit vector toward the ascending node, $\hat{a} = (\cos\Omega, \sin\Omega, 0)$, and a unit vector 90° around the orbit toward the northern part, $\hat{b} = (-\cos i \sin\Omega, \cos i \cos\Omega, \sin i)$. Then the angle from the ascending node to the current position of the object is $\tan^{-1}(\hat{b}\cdot\vec{r}/\hat{a}\cdot\vec{r})$. The true anomaly is $\theta = \tan^{-1}(\sqrt{1-e^2}\sin E/(\cos E - e))$, and the argument of the perihelion is the difference of these two angles:

$$\omega = \tan^{-1}\left(\frac{\hat{b}\cdot\vec{r}}{\hat{a}\cdot\vec{r}}\right) - \tan^{-1}\left(\frac{\sqrt{1-e^2}\sin E}{\cos E - e}\right) \tag{58}$$

9.6. From observations to position and velocity

The most entertaining orbit calculations involve newly discovered objects. Here is a brief run through of Lagrange's method for finding orbits from three observations. Of course Gauss is famous for developing a better orbit determination method and using it to prevent the first discovered asteroid, Ceres, from being lost when it went behind the Sun. But I find Lagrange's method to be easier to understand, and both methods require differential correction to improve the initial orbit.

Consider the position of the object relative to the Earth: $\vec{\rho} = \vec{r} - \vec{R} = \rho\hat{\rho}$, where $\vec{R}$ is the position of the Earth relative to the Sun. If we have three observations of the angular position of the object, $\hat{\rho}$, at three different times, then we can fit an interpolating polynomial to these three points and compute $\hat{\rho}$ and its first and second time derivatives, $\dot{\hat{\rho}}$ and $\ddot{\hat{\rho}}$, for the middle of the observed arc. With these angular derivatives, we can write the second time derivative of the vector:

$$\ddot{\vec{\rho}} = \rho\ddot{\hat{\rho}} + 2\dot{\rho}\dot{\hat{\rho}} + \ddot{\rho}\hat{\rho} \tag{59}$$

But we can also write this acceleration using Newton's laws:

$$\ddot{\vec{\rho}} = \frac{GM\vec{R}}{R^3} - \frac{GM\vec{r}}{r^3} \tag{60}$$

If we equate these two expressions we get a vector equation with 3 components to solve for the three unknowns ρ, $\dot{\rho}$, and $\ddot{\rho}$. If we take the dot product of both sides of the equation with the cross product of the angular position and angular velocity we get

$$\rho[\ddot{\hat{\rho}}\cdot(\hat{\rho}\times\dot{\hat{\rho}})] = GM\left(\frac{[\vec{R}\cdot(\hat{\rho}\times\dot{\hat{\rho}})]}{R^3} - \frac{[(\rho\hat{\rho}+\vec{R})\cdot(\hat{\rho}\times\dot{\hat{\rho}})]}{|\rho\hat{\rho}+\vec{R}|^3}\right) \tag{61}$$

which only involves the distance ρ. We can simplify this to

$$\rho = GM\frac{[\vec{R}\cdot(\hat{\rho}\times\dot{\hat{\rho}})]}{\ddot{\hat{\rho}}\cdot(\hat{\rho}\times\dot{\hat{\rho}})}\left(\frac{1}{R^3} - \frac{1}{|\rho\hat{\rho}+\vec{R}|^3}\right) \tag{62}$$

This needs to be solved iteratively since ρ appears on both sides. Note that this method fails for objects moving in a straight line on the sky (objects with zero inclination do this) because the triple vector product in the denominator vanishes, and for objects 1 au from the Sun because the right hand term vanishes. In these cases one needs more than three observations. Once ρ is found, the radial velocity is

$$\dot{\rho} = GM \frac{[\vec{R} \cdot (\hat{\rho} \times \ddot{\hat{\rho}})]}{2[\dot{\hat{\rho}} \cdot (\hat{\rho} \times \ddot{\hat{\rho}})]} \left(\frac{1}{R^3} - \frac{1}{|\rho\hat{\rho} + \vec{R}|^3} \right) \tag{63}$$

And once $\dot{\rho}$ and ρ are known, the position $\vec{r} = \vec{R} + \rho\hat{\rho}$ and velocity $\vec{v} = \dot{\vec{R}} + \dot{\rho}\hat{\rho} + \rho\dot{\hat{\rho}}$ are known, and can be used to compute the orbital elements.

To illustrate, consider three observations of the position of the "killer" asteroid 1997 XF11. Actually I will use three predicted positions spaced an even 30 days apart. The unit vector $\hat{\rho}$ is given by $(\cos\alpha\cos\delta, \sin\alpha\cos\delta, \sin\delta)$ in the celestial coordinate system with right ascension α and declination δ, and $(\cos\lambda\cos\beta, \sin\lambda\cos\beta, \sin\beta)$ in ecliptic coordinates with ecliptic longitude λ and ecliptic latitude β. In ecliptic coordinates the 1997 XF11 positions are $\hat{\rho}_{-1} = (-0.066032, 0.993301, -0.094834)$ on 8 March 1998, $\hat{\rho} = (-0.177731, 0.981148, -0.075901)$ on 7 April 1998, and $\hat{\rho}_{+1} = (-0.342582, 0.937378, -0.062936)$ on 7 May 1998. I use the standard numerical derivative formulae to give $\dot{\hat{\rho}} \approx (\hat{\rho}_{+1} - \hat{\rho}_{-1})/60 = (-0.00460917, -0.00093205, 0.00053163)$ and $\ddot{\hat{\rho}} \approx (\hat{\rho}_{+1} - 2\hat{\rho} + \hat{\rho}_{-1})/900 = (-0.0000590567, -0.0000351301, -0.0000066318)$. The position of the Earth on 7 April 1998 is $\vec{R} = (-0.9572541, -0.2923791, 0.0000113)$ and the triple vector products are

$$\begin{aligned} \ddot{\hat{\rho}} \cdot (\hat{\rho} \times \dot{\hat{\rho}}) &= -7.3325154 \times 10^{-8} \\ \vec{R} \cdot (\hat{\rho} \times \dot{\hat{\rho}}) &= -5.6145032 \times 10^{-4} \end{aligned} \tag{64}$$

The product GM is known much better than either G or M and is basically the square of the angular frequency of the Earth's orbit: $GM = 2.9591221 \times 10^{-4}$ when units of au's and days are used. Solving Eq(62) iteratively gives $\rho = 2.035$ compared to the exact value 1.988. One gets $a = 1.456(1.442)$, $e = 0.523(0.484)$, $\Omega = 212.5°(214.1)$, $i = 4.0°(4.1)$, $\omega = 104.5°(102.5)$, where the values in parentheses are the true values. These values are pretty good, but the 1% difference in a leads to a 1.5% difference in periods, so after one period the position would be off by 5.4°. But the inclination and ascending node are well determined even by these few observations. I "discovered" an asteroid, 1979 DC, and followed it for two weeks, but Brian Marsden showed that my arc of positions matched an earlier short arc of observations so I was not the real discoverer of this asteroid.

Differential corrections are computed after an initial orbit estimate is found. Given the position $\vec{r}$ and velocity $\vec{v}$ at time $t_\circ$ close to the middle of the observed arc, one can compute the predicted angular position at any time: $\hat{\rho}_c(t, X, Y, Z, \dot{X}, \dot{Y}, \dot{Z})$. Let $\hat{\rho}_{i,c}$ be the predicted position at the time of the i^{th} observation, and let $\hat{\rho}_{i,o}$ be the i^{th} observation, with accuracy σ_i. Then the sum of the weighted squares of errors

$$\chi^2 = \sum_i \frac{|\hat{\rho}_{i,o} - \hat{\rho}_{i,c}|^2}{\sigma_i^2} \tag{65}$$

must be minimized by making slight adjustments to the six parameters $X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}$ until the derivative of χ^2 with respect to each of the six parameters vanishes. The χ^2 statistic has an expected value of n_{df}, where $n_{df} = n_{data} - n_p$ is the *number of degrees of freedom*, $n_p = 6$ is the number of parameters in the fit (the orbital elements), and the number of data points n_{data} is twice the number of positional observations since each observation gives two coordinates on the sky. If the minimized χ^2 is $>>$ n_{df}, then either the observational errors are really much worse than the σ_i's, or else the calculated orbit is really bad.

10. The N-body problem for large N

This material is covered in Chapter 4 of Binney & Tremaine, "Galactic Dynamics".

The problems addressed in astrophysical gravitational dynamics calculations deal with large number of stars. The stars have such small radii (typically 10^{-8} of their separation) that we can almost always treat them as point masses. This mean that we have to solve $3N$ 2^{nd} order coupled non-linear differential equations to find the trajectories of N stars in 3 dimensional space:

$$\ddot{\vec{x}}_i = \sum_{j \neq i} G m_j \frac{\vec{x}_j - \vec{x}_i}{|\vec{x}_j - \vec{x}_i|^3} \tag{66}$$

There is nothing wrong with directly solving these equations for several hundred or even thousands of stars, but galaxies have billions of stars. Or consider the Sun, which has 10^{57} nuclei in it. Should we compute the configuration of the Sun by solving Eqn(66) with $N = 10^{57}$? this is not a very profitable way to proceed. Instead, it is better to break the problem into pieces by first finding the gravitational potential and then finding the orbits of the stars.

The gravitational potential is given by

$$\phi(x) = -\sum_j \frac{G m_j}{|\vec{x}_j - \vec{x}|} \tag{67}$$

and the resulting orbits of the stars are given by

$$\begin{aligned} \dot{\vec{x}}_i &= \vec{v}_i \\ \dot{\vec{v}}_i &= -[\vec{\nabla}\phi](\vec{x}_i) \end{aligned} \tag{68}$$

When evaluating this equation we have to define the gradient of $1/x$ to be zero at $x = 0$.

10.1. Distribution function

So far we haven't simplified anything, but if we have a really large N, then we should be able to find the average potential from the distribution function of the particles. Now we have to ask what we are averaging over. Fundamentally, we are going to average over an ensemble of stellar system, each with N stars, but with random initial conditions. These random initial conditions are constrained to match things we can actually measure, such as core radii and velocity dispersions, but the actual positions and velocities of stars are not fixed. We are actually not very interested in having an ephemeris that lists the positions vs time for all the stars in a globular cluster, but we are interested in knowing the velocity dispersions and mass densities. So we can define a distribution function $f(\vec{x}, \vec{v}, t)$ that tells the density of stars in phase space at $(\vec{x}, \vec{v})$. That is, the expected number of stars is a region of size $d^3\vec{x}$ in space and $d^3\vec{v}$ is velocity space is $n = f(\vec{x}, \vec{v}, t) d^3\vec{x} d^3\vec{v}$. If we have several types of stars with different masses, we can define several distribution functions $f_k(\vec{x}, \vec{v}, t)$ giving the phase space density of the k^{th} type of stars, which have mass m_k.

10.2. Average Potential

We can now define the mean mass density as

$$\langle\rho(\vec{x},t)\rangle = \sum_k m_k \int f(\vec{x},\vec{v},t)d^3\vec{v} \tag{69}$$

and with this we can find the mean potential

$$\overline{\phi}(\vec{x},t) = -G\int \frac{\langle\rho(\vec{x}')\rangle}{|\vec{x}-\vec{x}'|}d^3\vec{x}' \tag{70}$$

With this average potential we can now solve a *different* problem: what are the orbits of the stars in the average potential? These orbits are given by

$$\begin{aligned} \dot{\vec{x}}_i &= \vec{v}_i \\ \dot{\vec{v}}_i &= -\vec{\nabla}\overline{\phi}(\vec{x}_i) \end{aligned} \tag{71}$$

and these equations are much easier to solve. If the resulting orbits are such that the distribution function is the same as the one used to compute the density and potential, then we have found a consistent solution for the average motion in the average potential.

Let us consider a one dimensional problem. The number of stars is the box from x to $x+dx$ and v to $v+dv$ is $f(x,v,t)dxdv$. The flow out of the box at $x+dx$ is $v\,f(x+dx,v,t)dv$ while the flow into the box at x is $v\,f(x,v,t)dv$. The flow in at v is $g(x)\,f(x,v,t)dx$ while the flow out at $v+dv$ is $g(x)\,f(x,v+dx,t)dx$. The net change per unit time is

$$\begin{aligned} \frac{\partial}{\partial t}\left[f(x,v,t)dxdv\right] &= \\ -\left[v\,f(x+dx,v,t)dv - v\,f(x,v,t)dv + g(x)\,f(x,v+dv,t)dx - g(x)\,f(x,v,t)dx\right] &= \\ -dxdv\left[v\frac{\partial f}{\partial x} + g(x)\frac{\partial f}{\partial v}\right] + \ldots \end{aligned} \tag{72}$$

or

$$\frac{\partial f}{\partial t} = -v\frac{\partial f}{\partial x} - g(x)\frac{\partial f}{\partial v} \tag{73}$$

Since we are not considering star formation or destruction, the Lagrangian time derivative of the distribution function Df/dt is equal to zero. The Lagrangian time derivative is taken while "going with the flow", unlike the Eulerian time derivative which is taken at a fixed $(\vec{x},\vec{v})$. Thus

$$Df/dt = \lim_{\Delta t\to 0}\frac{f(\vec{x}+\vec{v}\Delta t, \vec{v}-\vec{\nabla}\overline{\phi}\Delta t, t+\Delta t) - f(\vec{x},\vec{v},t)}{\Delta t} = 0 \tag{74}$$

which gives an equation

$$\vec{v}\vec{\nabla}_x f - \vec{\nabla}\overline{\phi}\vec{\nabla}_v f + \frac{\partial f}{\partial t} = 0 \tag{75}$$

By $\vec{\nabla}_v f$ we mean the gradient of f with respect to its velocity argument. This is just the 3-dimensional version of the 1-D equation we just derived. This equation is known as the "collisionless Boltzmann" equation or the Vlasov equation. Note that it can describe time dependent situations like the spiral arms in a disk galaxy, but often we set $\partial f/\partial t = 0$ to find equilibrium situations. This gives the time-independent Vlasov equation

$$\vec{v}\vec{\nabla}_x f + \vec{g}\vec{\nabla}_v f = 0 \tag{76}$$

with $\vec{g} = -\vec{\nabla}\overline{\phi}$.

We can write these equations in terms of the one particle Hamiltonian of the system:

$$H = \frac{1}{2}mv^2 + m\overline{\phi} \tag{77}$$

Then with $p = mv$ and $q = x$,

$$\frac{\partial H}{\partial p} = v = \frac{\partial q}{\partial t} \tag{78}$$

and

$$\frac{\partial H}{\partial q} = m\vec{\nabla}\overline{\phi} = -\frac{\partial p}{\partial t} \tag{79}$$

Then the collisionless Boltzmann equation is

$$\frac{\partial f}{\partial t} = \frac{\partial H}{\partial q}\frac{\partial f}{\partial p} - \frac{\partial H}{\partial p}\frac{\partial f}{\partial q} = -\{H, f\} \tag{80}$$

where $\{,\}$ is the Poisson bracket. The equivalent in quantum mechanics for the expectation value $\langle A\rangle$ of an operator A is

$$i\hbar\frac{\partial\langle A\rangle}{\partial t} = -\langle[H, A]\rangle$$

where $[,]$ is the commutator of H and A.

In advanced classical mechanics, one learns about transformations of variables, and if the system can be transformed into *action-angle* variables then the time evolution is very simple. In terms of the actions J_i and the angles θ_i the Hamiltonian is

$$H = \mathrm{const} + \sum_i \omega_i J_i \tag{81}$$

so the actions are conserved quantities because $\partial H/\partial q = 0$ and the angles advance at the constant angular rates ω_i. For example, if a system has axial symmetry then the angular momentum around the symmetry axis is a conserved action and the associated angle is like the mean anomaly in the Kepler problem. If the three frequencies ω_i are all different, then the only way to have both $Df/dt = 0$ (which is always true) and $df/dt = 0$ (which is true for an equilibrium model) is to have the distribution function f depend only on the conserved actions J_i.

11. Potential-Density Pairs

When looking for solutions of the Vlasov equation, one needs to have a large library of density distributions with known potentials. One finds in general that the potentials are more spread out than the density distribution, and the potential is more spherical than the density distribution. For example, the density distribution for a point mass is a delta function, while the potential $\phi = -GM/r$ is quite extended. The relations between the potential and the density are

$$\phi(\vec{r}) = -G \int \frac{\rho(\vec{r}')d^3\vec{r}'}{|\vec{r} - \vec{r}'|} \tag{82}$$

and Poisson's equation

$$\nabla^2\phi = \left(\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2}\right) = 4\pi G\rho \tag{83}$$

Note that ∇^2 is the Laplacian operator. Thus the potential is given by an integral over the density, and the density is given by a differential operator applied to the potential.

An arbitrary constant can be added to the potential which will not change the density because the Laplacian of a constant is zero, and it will not change the dynamics because the gravitational acceleration is $\vec{g} = -\vec{\nabla}\phi$ which is zero for a constant.

Consider the potential of a homogeneous sphere with mass $M = (4\pi/3)\rho b^3$, radius b and density ρ. The function $r^2 = x^2 + y^2 + z^2$ has the Laplacian $\nabla^2(r^2) = 6$, so the potential inside the sphere is given by

$$\phi = \frac{4\pi G\rho r^2}{6} + \text{const} \tag{84}$$

For $r > b$ one is outside the sphere and the gravitational field is just that of a point mass M, $\phi = -GM/r$. Matching at $r = b$ fixes the const:

$$\frac{4\pi G\rho b^2}{6} + \text{const} = -\frac{GM}{b} = -\frac{4\pi G\rho b^2}{3} \qquad \rightarrow \qquad \text{const} = -2\pi G\rho b^2 = -\frac{3GM}{2b} \tag{85}$$

Thus the final result is

$$\phi = \begin{cases} -(GM/b)(1.5 - 0.5(r/b)^2), & \text{for } r < b \\ -(GM/r) & \text{for } r > b. \end{cases} \tag{86}$$

Another useful spherically symmetric potential is the Plummer potential given by

$$\phi = -\frac{GM}{\sqrt{r^2 + b^2}} \tag{87}$$

Because this potential approaches $-GM/r$ as $r \to \infty$, the total mass of the object producing this potential is M. Applying the Laplacian operator in spherical coordinates gives

$$\nabla^2\phi = \frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\phi}{dr}\right) = \frac{3GMb^2}{(r^2 + b^2)^{5/2}} = 4\pi G\rho \tag{88}$$

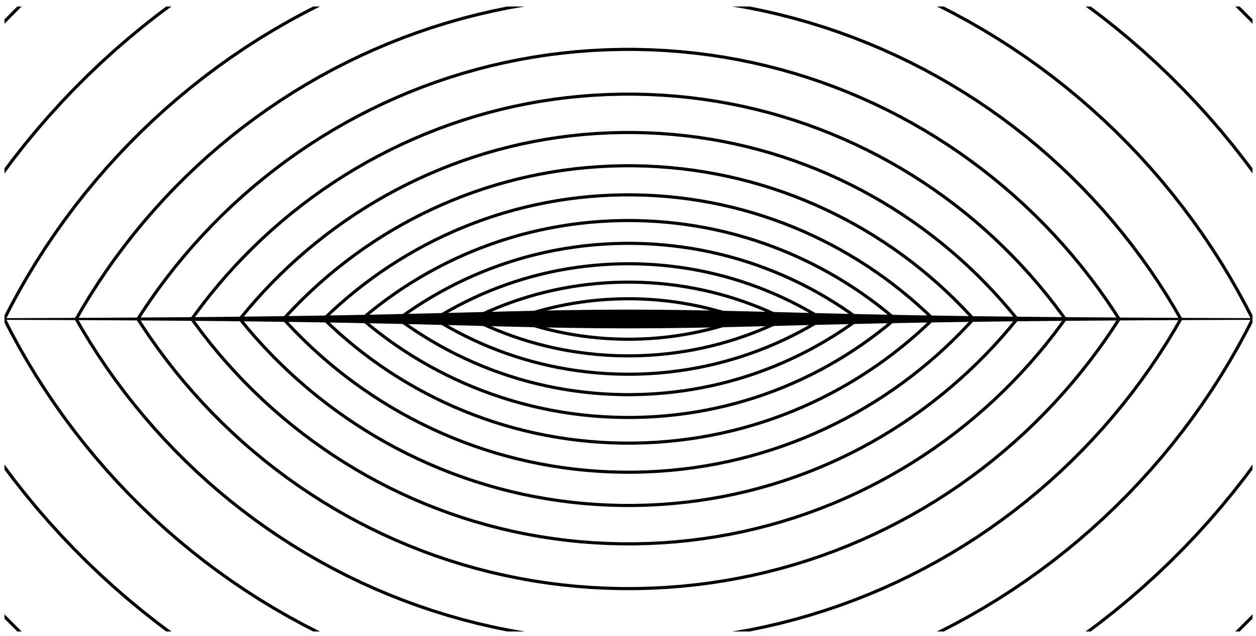

Fig. 11.— Mass density and isopotentials for a Kuzmin disk.

so

$$\rho = \frac{3M}{4\pi b^3}\left(1+\frac{r^2}{b^2}\right)^{-5/2} \tag{89}$$

Note that the central density is the density of a homogeneous sphere with radius b having the mass M, but the actual density distribution extends to infinity varying like r^{-5}.

An important fact about spherically symmetric matter distributions is that the acceleration of gravity at a point (the gradient of the potential) is just that due to the mass contained within a centered sphere having radius equal to the distance of the point from the center of symmetry:

$$|\vec{g}| = \frac{\partial\phi}{\partial r} = \frac{GM(<r)}{r^2} \tag{90}$$

where

$$M(<R) = 4\pi\int_0^R \rho(r)r^2 dr \tag{91}$$

But *this does not imply* that $\phi = -GM(<r)/r$!

If one guesses at a functional form for the potential, one often finds that the Laplacian goes negative in some places. Since the density has to be non-negative, such potentials are not allowed. For power law densities or potentials we get

$$\begin{array}{lllllll}
\phi & = & (2\pi/3)G\rho r^2 & \Longleftrightarrow & & \rho & \\
\phi & = & \sigma^2(r/b) & \Longleftrightarrow & \rho & = & \frac{2\sigma^2}{2\pi Gb}r^{-1} \\
\phi & = & Ar^0 & \Longleftrightarrow & \rho & = & 0 \\
\phi & = & -\frac{GM}{r} & \Longleftrightarrow & \rho & = & M\delta^3(\vec{r}) \\
\phi & = & -Ar^{-2} & \Longleftrightarrow & \rho & < & 0
\end{array} \tag{92}$$

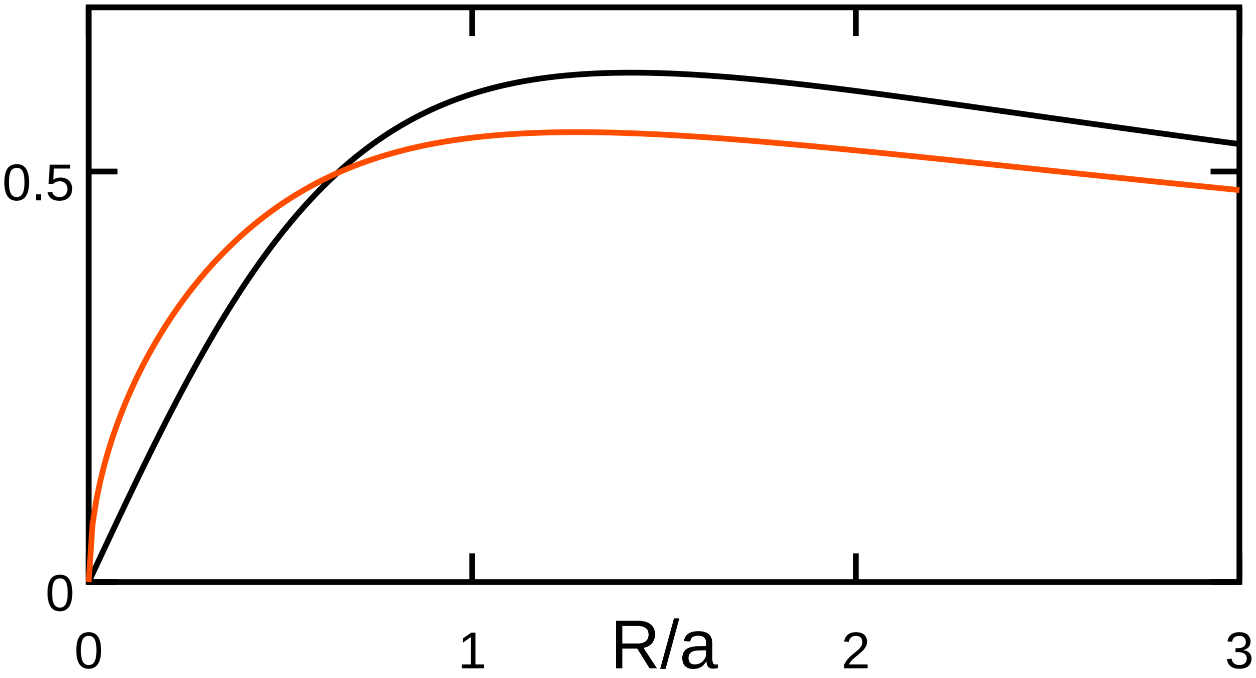

Fig. 12.— Rotation curve of a Kuzmin disk in units of $\sqrt{GM/a}$. Black curve is the correct $v_c = \sqrt{R\partial\phi/\partial R}$ while the orange (or grey) curve shows the formula $v_c = \sqrt{GM(<R)/R}$ which is exact for spherical distributions.

Simulations of the dark matter distribution in halos of galaxies and clusters of galaxies suggest that a density distribution of

$$\rho = \frac{1500\rho_\circ}{x(1+5x)^2} \tag{93}$$

is fairly universal (Navarro, Frenk & White, 1995, MNRAS, 275, 720), where $x = r/r_{200}$, $\rho_\circ$ is the average density of the Universe, and r_{200} is the radius within which the average overdensity is 200. For this density law the potential is

$$\phi = -240\pi G\rho_\circ r_{200}^2 \frac{\ln(1+5x)}{x} \tag{94}$$

For disk galaxies, we need the potential from a thin sheet. A uniform sheet of mass at $z = 0$ with surface density $\Sigma(x,y) = \Sigma_\circ$ gives a potential

$$\phi = 2\pi G\Sigma_\circ |z| \tag{95}$$

For a non-uniform sheet the computations are more complicated. A very simple non-uniform disk model was found by Kuzmin, with a potential for $z > 0$ given by the potential of a point mass M at $x = 0$, $y = 0$ and $z = -a$; but for $z < 0$ one switches to the potential of a point mass M at $x = 0$, $y = 0$ and $z = +a$. This potential is thus

$$\phi = -\frac{GM}{\sqrt{x^2+y^2+(|z|+a)^2}} \tag{96}$$

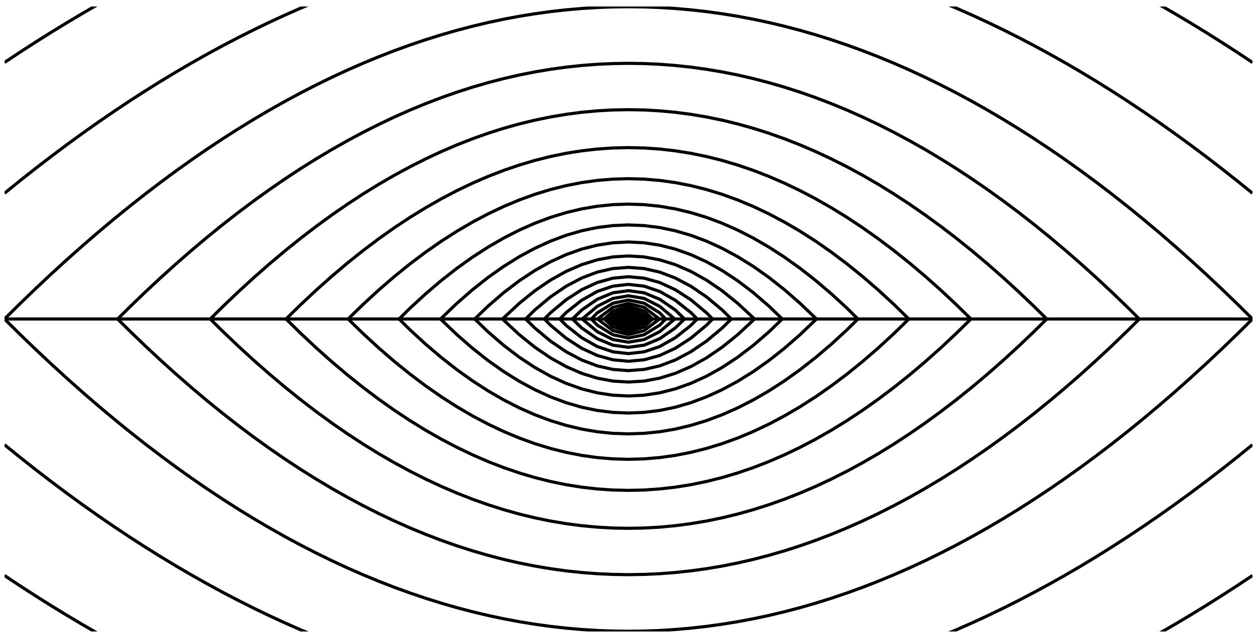

Fig. 13.— Isopotentials for a Mestel disk which has a flat rotation curve $v_c = const.$

For any $z \neq 0$ this is the potential of a point mass at a position different from the position of the mass, and thus the density is zero unless $z = 0$. At $z = \pm\epsilon$, $\partial\phi/\partial z = \pm GMa/(x^2+y^2+a^2)^{3/2}$ so the surface density of the mass sheet at $z = 0$ is

$$\Sigma(x,y) = \frac{Ma}{2\pi(x^2+y^2+a^2)^{3/2}} \tag{97}$$

Note that the mass contained within radius R is

$$M(<R) = \int_0^R \frac{Mar\,dr}{(r^2+a^2)^{3/2}} = \int_0^{R^2} \frac{Ma\,ds}{2(s+a^2)^{3/2}} = M\left(1 - (1+(R/a)^2)^{-1/2}\right) \tag{98}$$

but the velocity of circular orbits is *not* given by

$$v_c = \sqrt{\frac{GM(<R)}{R}}, \tag{99}$$

a formula that applies to spherically symmetric systems. To find the circular velocity for orbits around a Kuzmin disk, we must use

$$\frac{v_c^2}{R} = |g| = \frac{\partial\phi}{\partial R} = \frac{GMR}{(R^2+a^2)^{3/2}} \tag{100}$$

Note that the peak $v_{c,pk} = 0.62\sqrt{GM/a}$ occurs when $R = \sqrt{2}\,a$.

One can write the mass of the disk in terms of the maximum circular velocity and the central surface density $\Sigma(0) = M/(2\pi a^2)$. This gives

$$M = \frac{27}{8\pi G^2}\frac{v_{c,pk}^4}{\Sigma(0)} \tag{101}$$

which is a partial explanation of the empirical Tully-Fisher law relating the luminosity and maximum circular velocity for spiral galaxies. One gets $L \propto v^4$ as long as $(M/L)\Sigma(0)$ is constant. Generally low surface brightness galaxies (with low central surface brightness $I(0)$) have high values of mass to luminosity ratio (M/L). If $(M/L)^2 I_\circ$ is about constant then the both the low surface brightness and the normal surface brightness spirals will follow the same Tully-Fisher relation.

Actually there is one disk mass density for which the $v_c = \sqrt{GM(<R)/R}$ does work: the Mestel disk with $\Sigma(R) = \Sigma_\circ R_\circ / R$. The circular velocity is given by $v_c = \sqrt{2\pi G \Sigma_\circ R_\circ}$ which is independent of R: a *flat rotation curve*. To find the properties of the Mestel disk we use the trick employed for the Kuzmin disk: first note that the potential of an infinite line with mass per unit length λ is

$$\phi = 2G\lambda \ln R \tag{102}$$

so the centripetal force is

$$\frac{v_c^2}{R} = \frac{2G\lambda}{R} \tag{103}$$

and thus $\lambda = v_c^2/(2G)$. Now we want to have all the mass in a flat disk at $z = 0$, so consider the potential given by the following rule: for $z > 0$ use the potential of a linear mass with mass per unit length 2λ running from $z = 0$ to $z = -\infty$, while for $z < 0$ use the potential of a linear mass with mass per unit length 2λ running from $z = 0$ to $z = +\infty$. Thus for the point at (x, y, z) with $R = \sqrt{x^2 + y^2}$ we get

$$\begin{aligned} \phi(x,y,z) &= \int_0^\infty \frac{-2G\lambda dz'}{\sqrt{(|z| + z')^2 + (x^2 + y^2)}} \\ &= -2G\lambda \int_{|z|}^\infty \frac{d\zeta}{\sqrt{\zeta^2 + R^2}} \\ &= -2G\lambda \ln(\zeta + \sqrt{\zeta^2 + R^2})\Big|_{|z|}^\infty \end{aligned} \tag{104}$$

This integral diverges logarithmically as $\zeta = |z| + z' \to \infty$, but we can absorb the divergent evaluation at the upper end of the interval into a constant which can be dropped since adding a constant to a potential does not affect the physics. This gives the potential of the Mestel disk as

$$\phi(R,z) = v_c^2 \ln(|z| + \sqrt{z^2 + R^2}) \tag{105}$$

The isopotentials of this disk are paraboloids all having their foci at $R = 0$ and $z = 0$. For $z > 0$ this is the potential from a mass distribution that vanishes for $z > 0$, so the density vanishes for $z > 0$. Similarly, the density vanishes for $z < 0$. Thus any non-zero density must be confined to the plane $z = 0$. At $z = \pm\epsilon$, $\partial\phi/\partial z = \pm v_c^2/R$ so the surface density of the mass sheet at $z = 0$ is

$$\Sigma(R) = \frac{v_c^2}{2\pi G R} \tag{106}$$

Thus the mass contained within R is given by

$$M(<R) = \int_0^R \frac{v_c^2}{2\pi G R'} 2\pi R' dR' = \frac{v_c^2}{GR} \tag{107}$$

which just happens to agree with the formula for spherically symmetric systems, but this is just a happy coincidence that only works for this case.

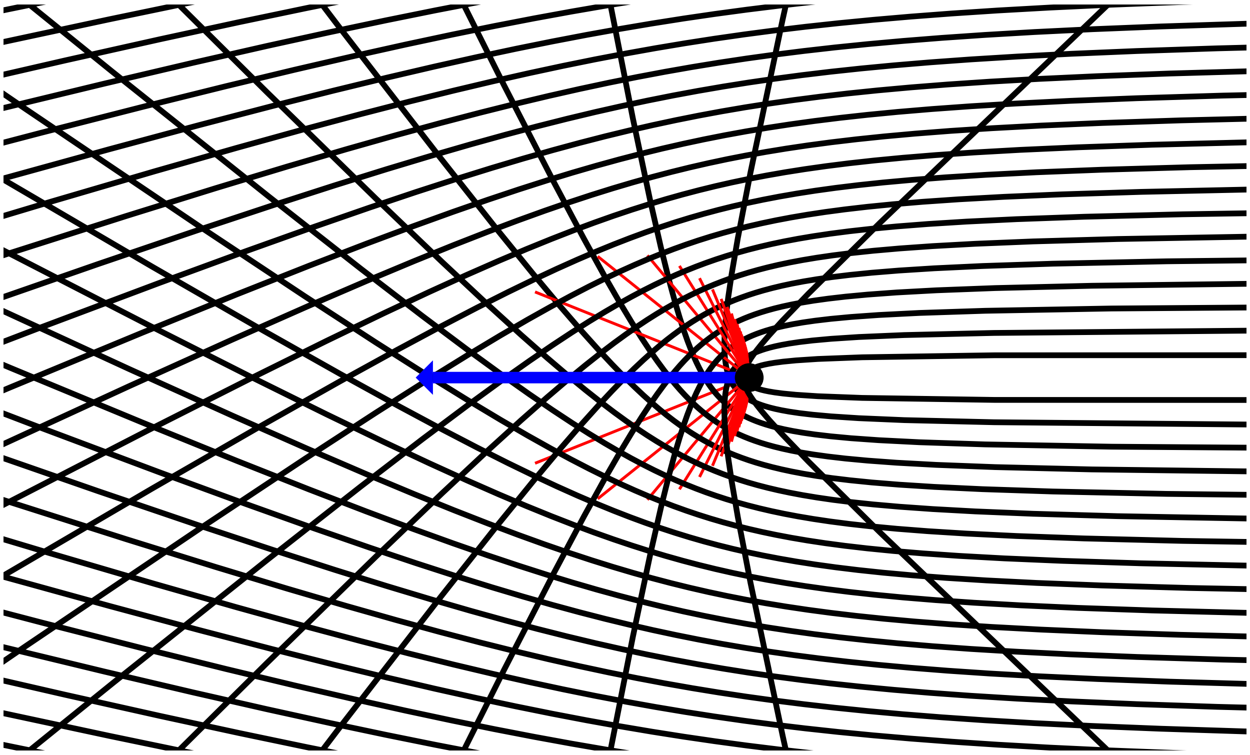

Fig. 14.— Many small bodies moving past a massive body. Each hyperbolic orbit imparts a momentum change to the big body shown by a thin red line. The sum of all the momentum changes is shown by a blue arrow: a net force proportional to the square of the big mass: dynamical friction.

12. Two Body Effects

But there are two "errors" in the collisionless Boltzmann equation that have physical meaning:

1. The passage of a massive body produces a "wake" behind it, which gives a density enhancement proportional to the mass of the passing body, leading to a drag force proportional to the square of the mass. This effect is *dynamical friction.*

2. The actual density of a cluster is bumpy instead of smooth, so a particle's orbit scatters off the bumps leading to a diffusion in phase space at a rate which is independent of the mass of the particle but inversely proportional to the number of particles in the cluster, an effect known as *relaxation.*

12.1. Dynamical Friction

In a two-body interaction, the speed of the incoming particle does not change, but the velocity does change because the direction changes. The velocity change perpendicular to the original track will be randomly positive and negative, averaging to zero (but leading to relaxation), while the velocity change parallel to the original track is always opposite to the original velocity leading to a systematic slowing down known as dynamical friction.

Consider the 2-body orbit of two particles with mass m_1 and m_2. Each particle experiences a force with magnitude $F = Gm_1m_2/r^2$ where $\vec{r} = \vec{r}_1 - \vec{r}_2$ is the interparticle separation and thus the accelerations of the particles around their mutual center of mass are $\ddot{\vec{r}}_1 = -Gm_2\hat{r}/r^2$ and $\ddot{\vec{r}}_2 = +Gm_1\hat{r}/r^2$. Therefore $\ddot{\vec{r}} = -G(m_1+m_2)\hat{r}/r^2$ which is the same as the equation for a test particle moving in the potential of a mass $m_1 + m_2$. We are interested in *unbound* orbits, which are hyperbolic orbits, so we use the equation for a conic section with eccentricity $e > 1$:

$$r = \frac{a(1-e^2)}{1+e\cos\theta} \tag{108}$$

This describes a particle that comes in from infinity starting at angle $\theta_i = -\cos^{-1}(-1/e)$, comes closest at $\theta = 0$, and then departs to infinity at angle $\theta_f = \cos^{-1}(-1/e)$. Thus the total deflection is $\Delta\theta = 2\cos^{-1}(-1/e) - \pi$. The change in the velocity parallel to the original track is $\Delta v = v(\cos(\Delta\theta) - 1) = -2v/e^2$. Now we need to find the semi-major axis a and the eccentricity in terms of the initial impact parameter b and the initial relative velocity v. The speed at periapsis is given by conservation of energy as $v_p = \sqrt{v^2 + 2G(m_1+m_2)/[a(1-e)]}$, and the conserved angular momentum per unit mass is $L = bv = a(1-e)v_p$. Thus

$$b^2v^2 = a^2(1-e)^2v^2 + 2G(m_1+m_2)a(1-e) \tag{109}$$

But the semi-major axis is determined from $2E = -GM/a$ as

$$a = -\frac{G(m_1+m_2)}{v^2} \tag{110}$$

giving

$$b^2v^2 = \frac{[G(m_1+m_2)]^2}{v^2}(1-e)^2 - 2\frac{[G(m_1+m_2)]^2}{v^2}(1-e) \tag{111}$$

or

$$(1-e)^2 - 2(1-e) - \left(\frac{bv^2}{G(m_1+m_2)}\right)^2 = 0 \tag{112}$$

and thus

$$e = \sqrt{1 + \left(\frac{bv^2}{G(m_1+m_2)}\right)^2} \tag{113}$$

and the net change in velocity parallel to the track is

$$\Delta v = \frac{-2v}{1 + [(bv^2)/(G(m_1+m_2))]^2} \tag{114}$$

Now we need to integrate this over all the collisions that occur per unit time. This number is $2\pi nvbdb$. But we also need to convert the change in the relative velocity Δv into the change in the velocity of m_1, which is $\Delta v_1 = (m_2/(m_1+m_2))\Delta v$. We get

$$\begin{aligned} \frac{dv_1}{dt} &= \frac{-4\pi n m_2 v^2}{m_1+m_2}\int_0^R \frac{bdb}{1+[(bv^2)/(G(m_1+m_2))]^2} \\ &\approx \frac{-4\pi n m_2 G^2(m_1+m_2)}{v^2}\ln\left(\frac{Rv^2}{G(m_1+m_2)}\right) \end{aligned} \tag{115}$$

where R, the size of the system, must be specified to end the logarithmic divergence of the integral. Note that the effective minimum impact parameter is $b_{min} = G(m_1 + m_2)/v^2$. Note that the acceleration of m_1 is opposite to v_1 and inversely proportional to the relative velocity. Because this is a central, inverse square law force, the integral over the distribution of v_2 will behave just like the integral of the force of gravity over an extended object. In particular, if the velocities v_2 are spherically symmetric, only the masses with $|v_2| < |v_1|$ will contribute to the net acceleration. Then the net acceleration is

$$\frac{dv_1}{dt} = \frac{-4\pi G^2 \rho(m_1 + m_2)F(< v_1)}{v_1^2} \ln\left(\frac{R\overline{v}^2}{G(m_1 + m_2)}\right) \tag{116}$$

The function $F(< v_1)$ gives the fraction of the objects m_2 that are moving slower than v_1, and $\overline{v}$ is a typical relative velocity. This effect is most important for heavy objects with $m_1 >> m_2$ that are moving at moderate speeds. Very fast objects do not decelerate much. The most important facts about *dynamical friction* in the large m_1 limit are that

1. The deceleration only depends on the product nm_2: the mass density of the objects being scattered.
2. The deceleration is proportional of the mass of the object m_1, so the force is proportional to m_1^2.

12.2. Relaxation Time

The actual mass density for any element in our ensemble is a sum of delta function spikes at the locations of the stars, and thus the actual potential has a forest of $1/x$ cusps. But these cusps are missing in the average potential, which is a much smoother function. Figure 15 shows the effect of these cusps when a smooth density model is used to represent various particle numbers. Thus when we find orbits using Eqn(71) we make an error by neglecting $\epsilon = \phi - \overline{\phi}$ which a bumpy function with $1/x$ cusps. This potential is generated by the density $\sum_j m_j \delta^3(\vec{x} - \vec{x}_j) - \langle\rho(\vec{x})\rangle$. Because this density averages to zero by definition, ϵ is small when averaged over large scales, but on small scales it has bumps. Particles moving through a bumpy potential field will be scattered when they hit a bump and this will lead to a diffusion of the phase space density. The rate of this diffusion will depend on the size of the bumps, which is given by the mass of the individual stars relative to the total mass of the system, or in other words, $1/N$. Thus a system with very many stars will be well described by a solution of the Vlasov equation, but a system with a small number of stars will not.

We can make a quantitative estimate of the length of time the collisionless Boltzmann solution will be valid. Consider a test body moving past a point mass with velocity v and impact parameter b. This body will experience an acceleration in the direction perpendicular to its track of

$$g_y(t) = \frac{Gmb}{(b^2 + v^2t^2)^{3/2}} \tag{117}$$

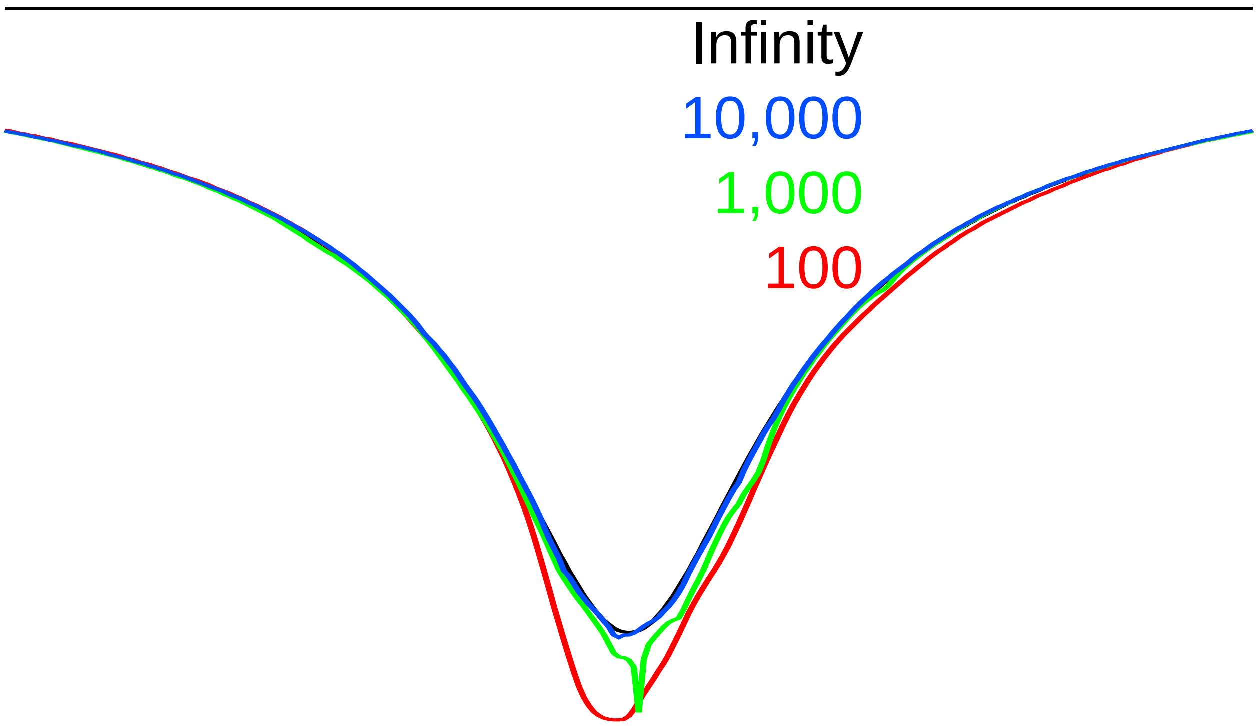

Fig. 15.— The potential from N mass points distributed following the Plummer law for $N = \infty$, 10^4, 10^4 and 100. The irregularities in the potential for small N scatter particles off of their smooth orbits leading to relaxation.

The integral of this acceleration is

$$\Delta v = \int g_y(t)dt = \int \frac{Gmb}{(b^2+v^2t^2)^{3/2}}dt = \frac{Gm}{bv}\int \frac{dx}{(1+x^2)^{3/2}} = \frac{2Gm}{bv} \tag{118}$$

Now the Δv's from different collisions will be in different directions, so we should not add the Δv's, but we can sum $(\Delta v)^2$. The number of collision per unit time is $2\pi bnvdb$ so

$$(\Delta v)^2 = \int 2\pi bnv\left(\frac{2Gm}{bv}\right)^2 dbdt = \frac{8\pi nG^2m^2dt}{v}\int\frac{db}{b} \tag{119}$$

When $(\Delta v)^2 = v^2$ the velocities have been scrambled by scattering, and this takes a time given by the *relaxation time*,

$$t_{relax} = \frac{v^3}{8\pi nG^2m^2\ln\Lambda} \tag{120}$$

where $\Lambda = b_{max}/b_{min}$. But if we let $N = (4\pi/3)r_c^3 n$, and $r_c = \sqrt{9\sigma^2/(4\pi Gnm)}$, $v = \sqrt{3}\sigma$, and $t_{cross} = 2r_c/v$, we get

$$\begin{aligned} t_{relax} &= t_{cross}\frac{v^4}{16\pi nG^2m^2r_c\ln\Lambda} \\ &= t_{cross}\frac{9\sigma^4}{16\pi nG^2m^2r_c\ln\Lambda} \end{aligned}$$

$$
\begin{aligned}
&= t_{cross}\frac{16\pi^2 G^2 n^2 m^2 r_c^4}{9\times 16\pi n G^2 m^2 r_c \ln\Lambda} \\
&= t_{cross}\frac{N}{12\ln\Lambda}
\end{aligned} \tag{121}
$$

Now a reasonable value of b_{min} is given by setting $\Delta v = 2v$, so

$$
b_{min} = \frac{Gm}{v^2} \tag{122}
$$

and a reasonable value of b_{max} is the average particle spacing

$$
b_{max} = n^{-1/3} \tag{123}
$$

since for larger b's the potential goes over to $\overline{\phi}$. These give

$$
\Lambda = \frac{v^2}{Gmn^{1/3}} = \frac{4\pi Gnmr_c^2}{9Gmn^{1/3}} = \frac{1}{3}\left(\frac{4\pi}{3}\right)^{1/3} N^{2/3} = 0.54N^{2/3} \tag{124}
$$

and the final result that

$$
t_{relax} \approx t_{cross}\frac{N}{8\ln N}. \tag{125}
$$

Note that Binney & Tremaine, in §4 of "Galactic Dynamics", derive $t_{relax}/t_{cross} \approx 0.1N/\ln N$ using different conversions from n and v to N.

Thus for a globular cluster with $\sigma = 7$ km/sec and $r_c = 1.5$ pc, $t_{cross} = 10^{5.4}$ years, and with $N = 10^6$, $t_{relax} = 10^{9.4}$ years. Globular clusters will be relaxed. For a galaxy with $v = 200$ km/sec and $r_c = 3$ kpc, $t_{cross} = 10^{7.5}$ years, and with $N = 10^{11}$, $t_{relax} = 10^{16.2}$ years which is much more than the age of the Universe. Thus galaxies will not be relaxed.

What is the effect of relaxation? It will make the velocity distribution isotropic and Maxwellian. In combination with dynamical friction which gives a greater slowing for massive stars, it will make the velocity dispersion depend on the mass, so heavy stars will have smaller dispersions.

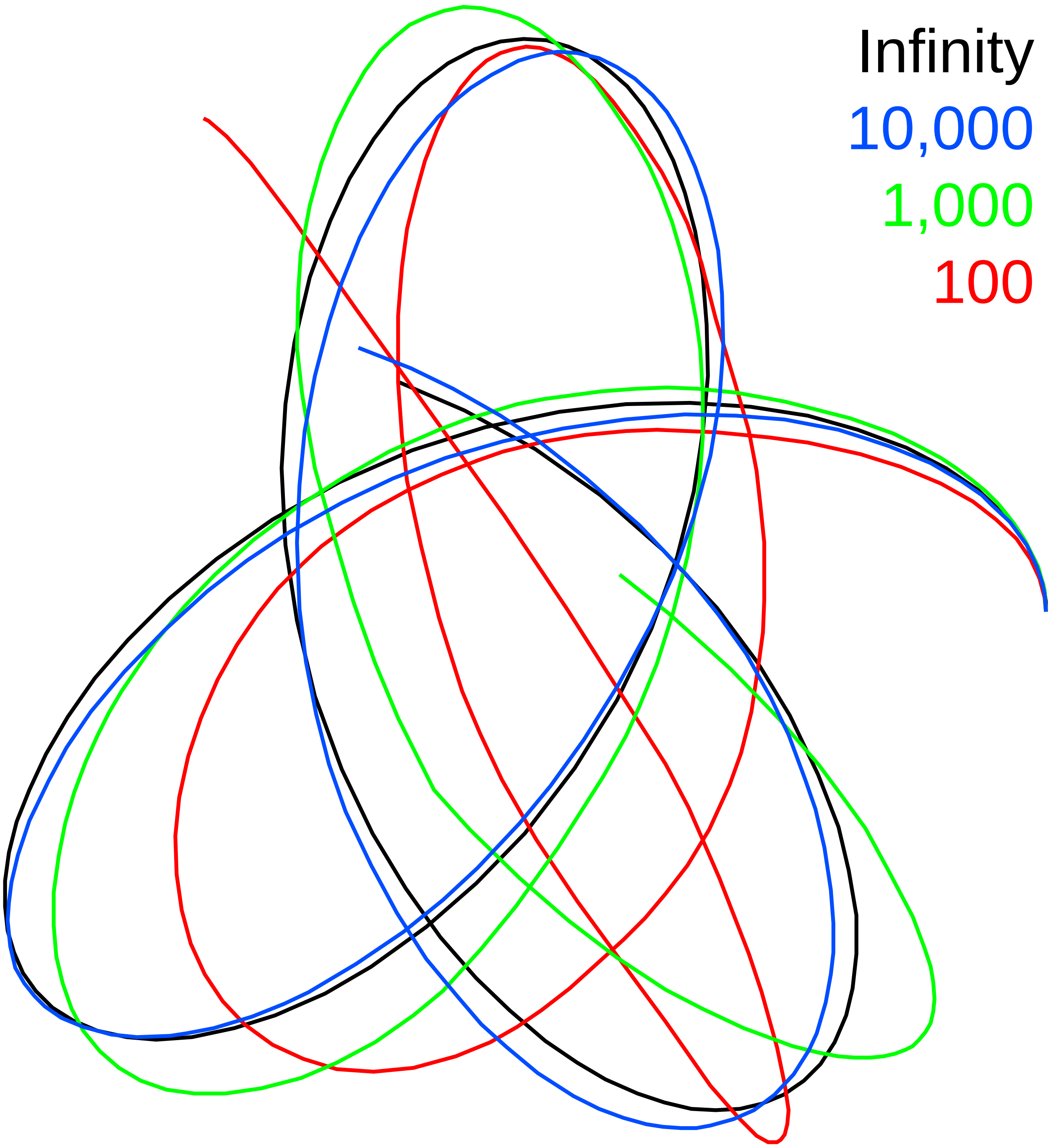

Fig. 16.— The path of a particle moving in the potential from N mass points distributed following the Plummer law for $N = \infty,\ 10^4,\ 10^4$ and 100. The irregularities in the potential for small N scatter particles off of their smooth orbits leading to relaxation.

13. Isothermal models

The most obvious way to make the distribution function depend only on the conserved actions is to have an f that depends only on the energy E. Since velocity distributions in globular cluster are observed to be consistent with Gaussian's, we will use a negative exponential dependence of f on E:

$$f(\vec{x},\vec{v}) = f_\circ \exp\left(\frac{-(v^2/2+\overline{\phi})}{\sigma^2}\right) \tag{126}$$

This makes the one dimensional velocity distribution a Gaussian with standard deviation σ. This gives

$$\rho = \rho_\circ \exp(-\overline{\phi}/\sigma^2) \tag{127}$$

and we also need to satisfy the Poisson equation

$$\nabla^2\overline{\phi} = 4\pi G\rho \tag{128}$$

Let us look for spherically symmetric solutions, so

$$r^{-2}\frac{\partial}{\partial r}r^2\frac{\partial\overline{\phi}}{\partial r} = 4\pi G\rho_\circ \exp(-\overline{\phi}/\sigma^2) \tag{129}$$

The simplest possible solution to this equation would be a power law $\rho = Cr^{-b}$. We can write

$$\begin{aligned}
\frac{d}{dr}\left(r^2\frac{d\ln\rho}{dr}\right) &= -\frac{4\pi G}{\sigma^2}r^2\rho \\
\frac{d}{dr}\left(r^2\left(\frac{-b}{r}\right)\right) &= -\frac{4\pi G}{\sigma^2}r^2Cr^{-b} \\
-b &= -\frac{4\pi GC}{\sigma^2}r^{2-b}
\end{aligned} \tag{130}$$

This only works if $b = 2$ and then implies that

$$C = \frac{\sigma^2}{2\pi G}. \tag{131}$$

Thus

$$\rho(r) = \frac{\sigma^2}{2\pi G}r^{-2} \tag{132}$$

is a solution of the collisionless Boltzmann equation for an isotropic Gaussian velocity distribution with standard deviation σ. This solution is the *singular isothermal sphere* or SIS. The mass contained within radius r is

$$M(<r) = \int_0^r 4\pi r^2\rho(r)dr = \frac{2\sigma^2}{G}r. \tag{133}$$

This gives an infinite mass as $r\to\infty$ which is slightly unrealistic but not too bad, since $r\to\infty$ includes the entire Universe. Circular orbits in this potential field give a flat rotation curve with

$$v_c = \sqrt{\frac{GM(<r)}{r}} = \sqrt{2}\sigma = \text{const.} \tag{134}$$

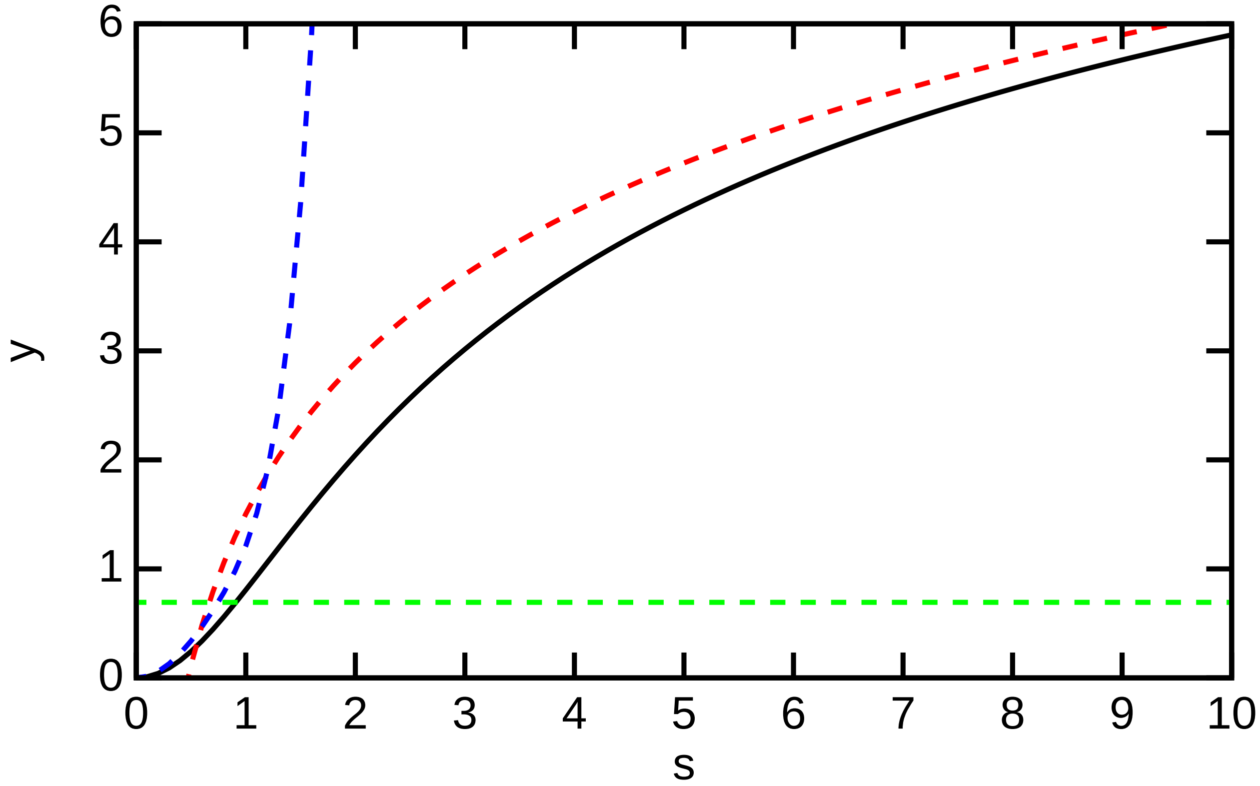

Fig. 17.— Numerical solution of the differential equation for a non-singular isothermal sphere (black solid curve) compared to the singular isothermal sphere with the same velocity dispersion (red dashed curve) and the power law expansion (blue dashed curve). The horizonal green dashed line shows the half-density level: $y = \ln 2$.

If applied to the halo of the Milky Way, this gives $\sigma = 220/\sqrt{2} = 156$ km/sec and a density at the solar circle of

$$\rho(r_\circ) = \frac{v_c^2}{4\pi G} r_\circ^{-2} \approx 10^{-24} \text{ gm/cc} \approx 0.01\ M_\odot/\text{pc}^3 \tag{135}$$

We also want to find non-singular solutions of Eq(129). Now we obviously want to look for a dimensionless solution so define $y = \overline{\phi}/\sigma^2$ giving

$$r^{-2}\frac{\partial}{\partial r} r^2 \frac{\partial y}{\partial r} = \frac{4\pi G\rho_\circ}{\sigma^2}\exp(-y) \tag{136}$$

Finally let $s = r/r_\circ$ with $r_\circ = \sqrt{9\sigma^2/(4\pi G\rho_\circ)}$. Then

$$s^{-2}\frac{\partial}{\partial s} s^2 \frac{\partial y}{\partial s} = 9\exp(-y). \tag{137}$$

Now we let $y = 0$ at $s = 0$ since this just sets the density scale. We try a power series expansion in s:

$$y = c_1 s + c_2 s^2 + c_3 s^3 + c_4 s^4 + \ldots \tag{138}$$

giving

$$2c_1 s^{-1} + 2\cdot 3c_2 s^0 + 3\cdot 4c_3 s^1 + 4\cdot 5c_4 s^2 + \ldots =$$
$$9\left\{1 - \left[c_1 s + c_2 s^2 + c_3 s^3 + c_4 s^4\right] + \frac{1}{2}\left[c_1^2 s^2 + 2c_1c_2 s^3 + (c_2^2 + 2c_1c_3)s^4\right] - \ldots\right\} \tag{139}$$

Reading off successive powers of s gives

$$\begin{aligned} 2c_1 &= 0 \\ 6c_2 &= 9 \\ 12c_3 &= -9c_1 \\ 20c_4 &= -9[c_2 - c_1^2/2] \end{aligned} \tag{140}$$

Clearly all the odd coefficients vanish, so let's rewrite the equation with just the even terms:

$$6c_2 s^0 + 20c_4 s^2 + 42c_6 s^4 + 72c_8 s^6 + \ldots =$$
$$9\left\{1 - \left[c_2 s^2 + c_4 s^4 + c_6 s^6 + \ldots\right] + \frac{1}{2}\left[c_2^2 s^4 + 2c_2c_4 s^6 + \ldots\right] - \ldots\right\} \tag{141}$$

and find $c_2 = 3/2$, $c_4 = -27/40$, $c_6 = 27/70$, so

$$y \approx \frac{3}{2}s^2 - \frac{27}{40}s^4 + \frac{27}{70}s^6 - \ldots. \tag{142}$$

When s is 1, at $r = r_\circ$, the density of this *isothermal* sphere has fallen to about one-half (0.5013...) of the central density, so this is known as the core radius of the isothermal sphere. Note that $r_\circ$ and σ are measureable for globular clusters, so the value of $\rho_\circ$ can be found from the definition of $r_\circ$. As $s \to \infty$, $y \approx 2\ln s + \text{const}$ so the isothermal sphere goes like r^{-2} at large radii, just as the singular isothermal sphere does at all radii. This gives an infinite total mass.

Another class of solutions with a distribution function that depends only on the energy uses

$$f(E) = \begin{cases} F\,(-E)^{n-3/2}, & E < 0; \\ 0, & E > 0. \end{cases} \tag{143}$$

Clearly the density is only non-zero where the potential is negative, so

$$\rho = 4\pi F \int_0^{\sqrt{-2\phi}} (-\phi - \frac{1}{2}v^2)^{n-3/2} v^2 dv = c_n(-\phi)^n \tag{144}$$

Poisson's equation gives us

$$r^{-2}\frac{\partial}{\partial r}r^2\frac{\partial\overline{\phi}}{\partial r} = 4\pi G c_n(-\phi)^n \tag{145}$$

If we let $\psi = \phi/\phi(0)$ and $s = r/b$ with $b = (4\pi G c_n(-\phi(0))^{n-1})^{-1/2}$, we get

$$s^{-2}\frac{\partial}{\partial s}s^2\frac{\partial\psi}{\partial s} = -\psi^n \tag{146}$$

which is the Lane-Emden equation which is also seen for gravitating gas spheres following a *polytropic* equation of state $P = K\rho^{1+1/n}$. The boundary conditions are $\psi(0) = 1$ by definition and

$d\psi/ds = 0$ at $s = 0$ by symmetry. This equation doesn't have elementary solutions for most values of n but for $n = 5$ it does:

$$\psi_5 = \left(1 + \left[\frac{s^2}{3}\right]\right)^{-1/2}. \tag{147}$$

Therefore the density follows the law

$$\rho \propto \psi^5 \propto \left(1 + \left[\frac{s^2}{3}\right]\right)^{-5/2}. \tag{148}$$

This is known as Plummer's law and the potential

$$\phi = -\frac{GM}{\sqrt{r^2 + b^2}} \tag{149}$$

is known as a Plummer potential.

Other analytic solutions for the Lane-Emden equation occur for $n = 0$, which describes an incompressible fluid, with $\psi_0 = 1 - s^2/6$, and $n = 1$, with $\psi_1 = \sin(x)/x$. These all can be described as a power series in s^2, $\psi_n \approx 1 - s^2/6 + ns^4/120 - \ldots$.

In the limit $n \to \infty$ the polytropic equation of state becomes isothermal. In this case the Lane-Emden equation becomes the previous equation for the isothermal sphere if we set $\psi = (1 - y/n)$ giving

$$s^{-2}\frac{\partial}{\partial s}s^2\frac{\partial y}{\partial s} = n(1 - y/n)^n \tag{150}$$

and then change the radial variable to $u = s\sqrt{n/9}$ which gives

$$u^{-2}\frac{\partial}{\partial u}u^2\frac{\partial \psi}{\partial u} = 9(1 - y/n)^n. \tag{151}$$

Finally in the limit $n \to \infty$ we get $(1 - y/n)^n \to \exp(-y)$.

Note we can find an isothermal plane solution by solving

$$\frac{\partial^2\overline{\phi}}{\partial z^2} = 4\pi\rho_\circ \exp(-\overline{\phi}) \tag{152}$$

so an isotropic velocity distribution can give rise to either a spherical or a non-spherical configuration. The next example will show how an anisotropic velocity distribution can give a spherical configuration.

Consider a distribution function that depends on the total angular momentum $L = v_\perp r$ (per unit mass) as well as the energy per unit mass $(v^2/2 + \phi)$:

$$f = f_\circ \exp(-(v^2/2 + \phi)/\sigma^2 - 0.5(v_\perp r/\sigma r_a)^2) \tag{153}$$

Then the integral over the velocity distribution is suppressed by a factor of $1 + (r/r_a)^2$ giving

$$\rho = \rho_\circ \frac{\exp(-\overline{\phi}/\sigma^2)}{1 + (r/r_a)^2} \tag{154}$$

Proceeding to a dimensionless equation as we did for the isothermal sphere gives

$$s^{-2}\frac{\partial}{\partial s}s^2\frac{\partial y}{\partial s} = \frac{9\exp(-y)}{1+(s/s_a)^2}. \tag{155}$$

If this has a "singular" or power-law solution then $y = -b\ln s + C$ and $-b/s^2 = 9e^C s_a^2/s^{b+2}$. This only works for $b = 0$ but then the density has to be zero. Another way to find the asymptotic form of ρ is to assume that the total mass is finite. Then ϕ approachs a constant at large r and the density decline is only due to the $1/(1+(r/r_a)^2)$ term. But this gives $\rho \propto r^{-2}$ at large r and thus an infinite total mass. Thus while this anisotropic velocity distribution gives perfectly reasonable spherical models for globular clusters, they still end up having infinite total mass.

Model globular clusters with finite mass have been found by Ivan King. These King models use a "lowered" Boltzmann distribution:

$$f_K(E) = \begin{cases} f_\circ\left(\exp(-E/\sigma^2) - \exp(-E_t/\sigma^2)\right), & E < E_t; \\ 0, & E > E_t. \end{cases} \tag{156}$$

If we define $\Psi = -(E_t - \phi)$, then the density obtained by integrating f_K over velocities is

$$\rho_K = \rho_1\left[\exp\left(\Psi/\sigma^2\right)\,\mathrm{erf}\left(\frac{\sqrt{\Psi}}{\sigma}\right) - \sqrt{\frac{4\Psi}{\pi\sigma^2}}\left(1+\frac{2\Psi}{3\sigma^2}\right)\right] \tag{157}$$

This will vanish for radii greater than the *tidal radius* r_t at which $\phi = E_t$, so the total mass is finite. There is a family of solutions with different ratios of r_t/r_c which can be found by setting the value of Ψ/σ^2 at $r = 0$ to different initial conditions.

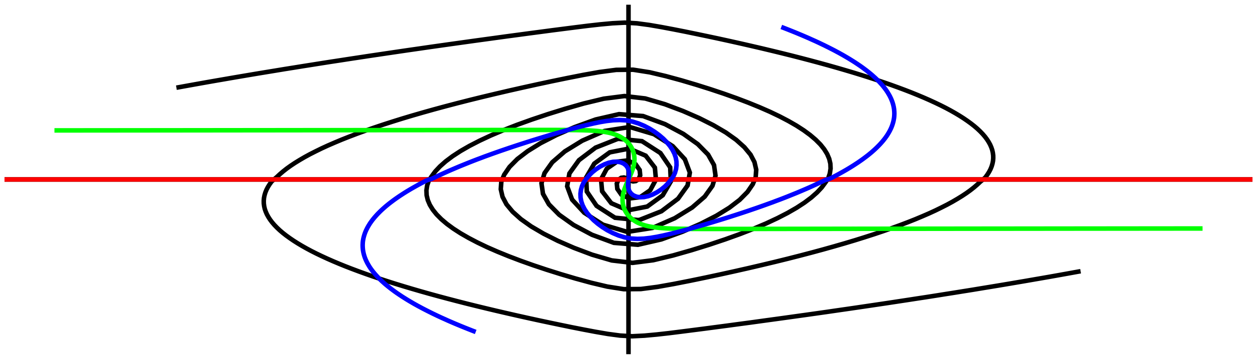

Fig. 18.— Phase diagram for points released from rest at $t = 0$ starting with z from -25 to 25 in the isothermal plane potential $\phi = \ln\cosh(z)$ at times $t = 0$ (red), $t = 2$ (green), $t = 8(blue)$ and $t = 32$ (black). Note how the phase wrapping causes the phase space distribution to approach a smooth function of z and v.

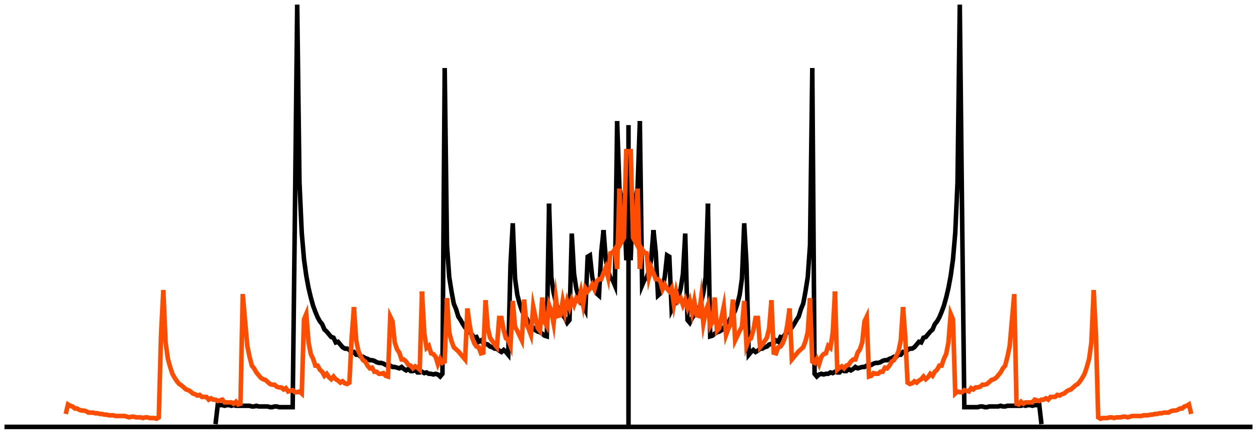

Fig. 19.— Velocity distribution at $z = 0$ for the phase space distributions shown in Figure 18. The distribution $p(v)$ is shown at $t = 128$ (orange) and $t = 32$ (black). Note how the phase wrapping causes the velocity distribution to approach a smooth function of v.

14. Violent Relaxation

This material is covered in Chapter 4 of Binney & Tremaine, "Galactic Dynamics".

Even though galaxies are not relaxed, the velocity distribution of stars tends to be Gaussian. However, unlike a true Maxwellian relaxed case, the heavier stars have the same velocity dispersion as the light stars. This suggests that large-scale gravitational potential gradients have created acceleration fields that accelerate stars of all masses by the same amount.

Unlike two-body relaxation, these large-scale gradients do not make a non-zero Df/dt. But a Gaussian distribution has a large range of values of f, so some changes in f are needed. The processes of *phase wrapping* and *coarse graining* are the mechanism which changes the apparent distribution function. As an example of these processes, consider stars initially all at rest in the potential of the isothermal plane, $\phi = 4\pi G H^2 \rho_\circ \ln(\cosh(z/H))$. Initially this forms a line along the z axis in the $z - v_z$ phase diagram. After $\frac{1}{4}$ of the period for small oscillations, the middle part of the distribution is along the v_z axis, but the outer parts have longer periods, so they are still close to the z axis. After many periods, the distribution is wound into a tight spiral in the phase diagram, as shown in Figure 18. If we look at the phase diagram with coarse bins in space and velocity, this tightly wound spiral will look like a two-dimensional distribution that has a finite, smooth f, while the original f was either infinity or zero because all the stars were concentrated into a one-dimensional track through the phase diagram.

In addition to phase wrapping and coarse graining, many non-Gaussian velocity distributions lead to unstable collisionless systems. A velocity distribution with two peaks suffers from the *two-stream* instability. Normally a density enhancement would create a wake or density enhancement in the down-stream direction. However, this would be the upstream direction for the other stream. Thus the wake in one stream forces a wake in the other stream that is superimposed on the original perturbation. This causes the perturbation to grow, leading to instability.

15. "Non-rotating" disks

One example of a distribution function that would suffer from the two-stream instability is the "non-rotating" disk. Imagine a very flat, rotating disk like the typical spiral galaxy. Now take ever other star and reverse its tangential velocity. This leaves the potential the same and the orbits all remain the same, so this distribution is still a solution of the collisionless Boltzmann equation. But now the net rotational velocity is zero.

In reality, this distribution is unstable: a small perturbation leads to a $\partial f/\partial t$ that increases the perturbation. So a galaxy formed by merging two spirals undergoes violent relaxation giving a more Gaussian velocity distribution, and is thought to become an elliptical galaxy.

16. Quadratic programming to construct models

One technique to find model distribution functions is to pick a potential ϕ that has the desired shape, such as a prolate ellipsoid for an E6 or E7 galaxy. Then follow thousands of orbits in this potential, and for each orbit compute the time-averaged space density $\rho_i(\vec{x})$ for the i^{th} orbit. Let the potential due to the is density be $\phi_i(\vec{x})$. Now pick the number of stars n_i on each orbit to minimize the error between the initial potential and the resultant potential: find n_i to minimize

$$\int w(\vec{x}) \left(\phi(\vec{x}) - \sum_i n_i \phi_i(\vec{x}) \right)^2 d^3\vec{x} \tag{158}$$

subject to the constraint that $n_i \geq 0$. $w(\vec{x})$ is a non-negative weight function used to force the fit to be better in critical regions.

This constrained minimization problem is known as *quadratic programming*, and this technique for finding self-consistent distribution functions was pioneered by Martin Schwarzschild.

17. Windup problem for spiral arms

Spiral arms are not material arms, but rather density waves in the disk. The galactic rotation period at the solar circle is about 0.2 Gyr, while the age of the disk is about 10 Gyr. Thus there have been 50 complete rotations in the age of the disk. At a radius 10% closer to the galactic center, the rotation period is about 10% shorter because $v_c(R)$ is almost flat. Thus there have been 55 complete rotations at the position 10% closer to the GC. Thus an originally straight stripe "painted" on the disk would make one complete turn around the disk in only 2% of the radius, leading to a pitch angle of about 1°. This is nothing like the observed spiral arms.

Instead of physical stripes, spiral arms are waves in the disk. Just as in the case of ocean waves, the material in the disk just moves in small loops as the wave passes. The phases of the loops are synchronized so there is an enhanced density in the spiral arms.

18. Epicyclic frequency in disks

Small loop motions in the disk can be studied using epicycles. The best conserved quantity for identifying the radius of an orbit is the angular momentum $L = v_\perp R$, where as usual we consider a unit mass. The circular orbit of a given L is the lowest energy orbit of all those with the same L. The excess energy of a non-circular orbit is associated with with the radial oscillation between radii R_- (perigalacticon) and R_+ (apogalacticon). The total energy can be written as

$$\begin{aligned} E &= \frac{1}{2}v_r^2 + \frac{1}{2}v_t^2 + \phi(R) \\ &= \frac{1}{2}v_r^2 + \frac{L^2}{2R^2} + \phi(R) \end{aligned} \tag{159}$$

We can separate the radial motion giving an effective potential

$$\phi_e(R) = \frac{L^2}{2R^2} + \phi(R) \tag{160}$$

This gives a radial acceleration of

$$\ddot{R} = -\frac{\partial \phi_e}{\partial R} = -\frac{\partial \phi}{\partial R} + \frac{L^2}{R^3} \tag{161}$$

This acceleration is zero at the radius of the circular orbit with angular momentum L, $R_\circ$:

$$\frac{L^2}{R^3} = \frac{v_\perp^2}{R_c} = \frac{\partial \phi}{\partial R} \tag{162}$$

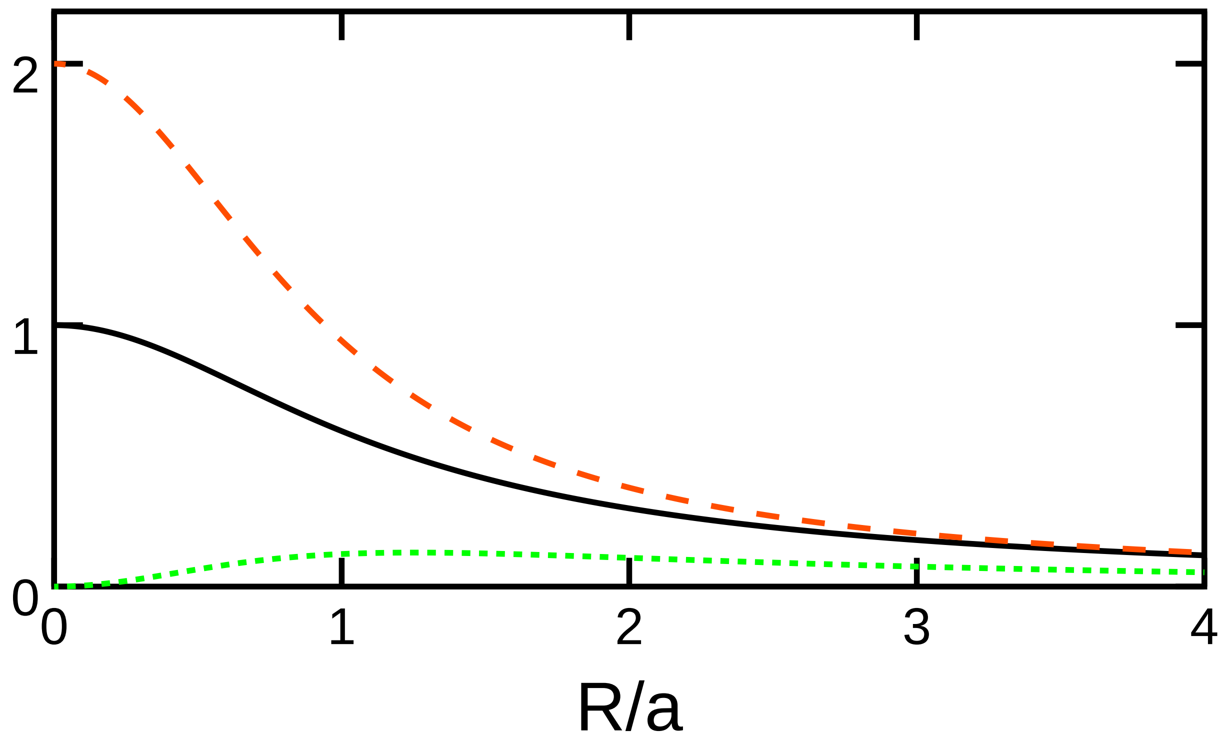

Fig. 20.— Orbital frequency ω (black solid curve), epicyclic frequency κ (orange long dashed curve), and the pattern speed for the inner Lindblad resonance, $\omega - \kappa/2$, in units of $\sqrt{GM/a^3}$ for the Kuzmin disk model.

But the gradient of the acceleration with radius determines the frequency $\omega_{rad} = \kappa$ of the radial oscillation:

$$\begin{aligned}
\kappa &= \sqrt{\frac{\partial^2 \phi_e}{\partial R^2}} \\
&= \sqrt{\frac{\partial^2 \phi}{\partial R^2} + \frac{3L^2}{R^4}} \\
&= \sqrt{\frac{\partial}{\partial R}\left(\frac{v_c^2}{R}\right) + \frac{3v_c^2}{R^2}} \\
&= \sqrt{2\frac{v_c^2}{R^2} + 2\frac{v_c}{R}\frac{\partial v_c}{\partial R}} \\
&= \sqrt{2\frac{v_c}{R}\left(\frac{v_c}{R} + \frac{\partial v_c}{\partial R}\right)}
\end{aligned} \tag{163}$$

This epicyclic frequency can be given in terms of the measured Oort constants as

$$\kappa = \sqrt{4B(B-A)} \tag{164}$$

The ratio of the epicyclic frequency to the orbital frequency $\omega = v_c/R$ is

$$\frac{\kappa}{\omega} = \sqrt{2\left(1 + \frac{\partial v_c/\partial R}{v_c/R}\right)} \tag{165}$$

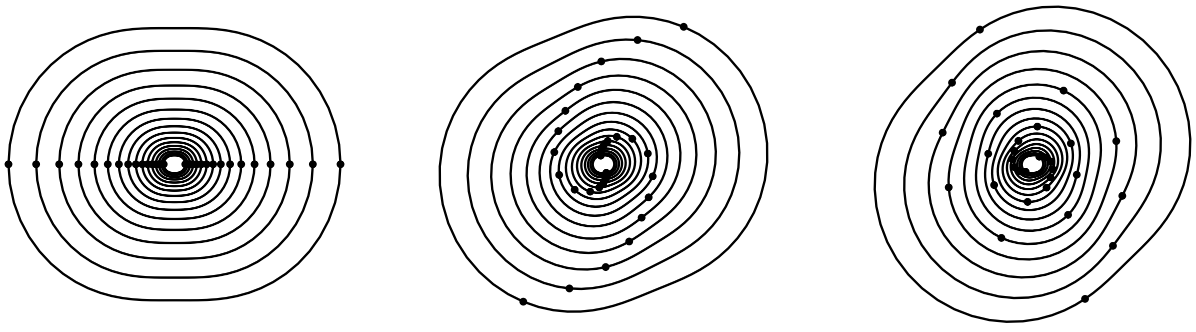

Fig. 21.— A Kuzmin disk model with an $m = 2$ perturbation. Each oval is a set of stars on epicyclic orbits, and the black dots highlight two "material" arms of stars. The three panels shows 3 equally spaced times. Notice how the pattern of ovals rotates and winds up much more slowly than the material arms. If $\omega - \kappa/2$ were constant then the pattern of the ovals would rotate like a solid body with no wind up.

If $v_c(R)$ is flat this ratio is $\sqrt{2}$, while for Keplerian motion $v_c \propto R^{-1/2}$ and the ratio is unity leading to closed orbits. Solid body rotation with $v_c \propto R$ gives $\kappa/\omega = 2$ so the orbits are once again closed ellipses but now are centered instead of having one focus at the center of the system.

The epicyclic motion of stars leads to an observable consequence: the ellipsoid describing the peculiar velocities of stars in the disk is anisotropic, with a larger velocity dispersion in the radial direction than in the tangential direction. When a star reaches its maximum distance from the GC, $R_+ = R_\circ + \Delta v_r/\kappa$, where Δv_r is its peculiar radial velocity when it crosses $R_\circ$, its tangential velocity is given by $v_t = L/R_+$ as always. But this gives a tangential peculiar velocity of $\Delta v_t = v_t - v_c(R_+)$. Thus

$$\begin{aligned}
\Delta v_t &= \frac{v_c(R_\circ)}{1 + \frac{\Delta v_r}{\kappa R_\circ}} - v_c(R_\circ + \frac{\Delta v_r}{\kappa}) \\
&\approx v_c(R_\circ) - \frac{\Delta v_r}{\kappa R_\circ} v_c(R_\circ) - v_c(R_\circ) - \frac{\Delta v_r}{\kappa}\frac{\partial v_c}{\partial R} \\
&= -\frac{\Delta v_r}{\kappa}\left(\frac{v_c}{R} + \frac{\partial v_c}{\partial R}\right)
\end{aligned} \tag{166}$$

The final result is

$$\frac{\Delta v_t}{\Delta v_r} = -\sqrt{\frac{1}{2}\left(1 + \frac{R\partial v_c}{v_c \partial R}\right)} = -\frac{\kappa}{2\omega} =\approx -0.7 \tag{167}$$

for flat rotation curves. Thus the ratio of velocity dispersions will be $\sigma(v_t)/\sigma(v_r) = 0.7$

19. Lindblad Resonances

If a spiral arm pattern persist without wrapping, then the pattern must rotate as a solid body, even though the disk is differentially rotating. Let the angular speed of this pattern be Ω_p. If the pattern has m arms, then the period of perturbations seen by the stars is $m(\omega - \Omega_p)$. If this period is $\pm\kappa$, then the perturbation from the spiral pattern will strongly excite the epicyclic motion of the stars. This condition gives

$$\Omega_p = \omega \pm \frac{\kappa}{m} \tag{168}$$

Since both ω and κ fall with radius, the point at which $\Omega_p = \omega - \kappa/m$ occurs at a smaller radius than the point where $\Omega_p = \omega + \kappa/m$. Thus $\Omega_p = \omega - \kappa/m$ is known as the inner Lindblad resonance while $\Omega_p = \omega + \kappa/m$ is the outer Lindblad resonance. Note that for solid body rotation, $v_c \propto R$, which typically occurs within the core radius of a galaxy, $\kappa = 2\omega$, so a fixed two-armed pattern with zero pattern speed is at the inner Lindblad resonance. For $v_c \approx$ const, $\kappa = \sqrt{2}\omega$, so a two-armed pattern at the inner Lindblad resonance has a pattern speed of 0.3ω.

In general spiral galaxies have velocity curves that give a fairly constant value of $\omega - \kappa/2$, so a two-armed spiral pattern can be close to the inner Lindblad resonance over a wide range of radii. Figure 20 shows the near constancy of $\omega - \kappa/2$ over a wide range of R for the Kuzmin disk model.

20. Toy model of disk instability

Imagine a disk is cut open and rolled out into a set of parallel rods. this will be a reasonable approximation for small-scale perturbations with sizes much less than the radius. Each rod represent the stars at a given angular momentum. We want to see whether purely radial oscillations are stable. Let the surface density of the disk be μ. Then if the rods have spacing Δx, their linear mass density is $\lambda = \mu\Delta x$. Each rod is connected to its "home" position by a spring which makes a harmonic oscillator with frequency κ. In addition, the rods interact gravitationally. The force between two rods separated by Δ is $2G\lambda^2/\Delta$ per unit length. If the rods are displaced from their home position by $\epsilon(x)$, then the net force per unit length on the rod at x is

$$\begin{aligned} F &= 2G\mu^2\Delta x \int_0^\infty \left(\frac{1}{\Delta + \epsilon(x+\Delta) - \epsilon(x)} - \frac{1}{\Delta + \epsilon(x) - \epsilon(x-\Delta)} \right) d\Delta \\ &\approx 2G\mu^2\Delta x \int_0^\infty \frac{2\epsilon(x) - \epsilon(x-\Delta) - \epsilon(x+\Delta)}{\Delta^2} d\Delta \end{aligned} \tag{169}$$

For $\epsilon(x) = \epsilon_\circ \cos kx$, this force is

$$F = 4G\mu^2\Delta x\epsilon_\circ \cos kx \int_0^\infty \frac{1-\cos k\Delta}{\Delta^2} d\Delta = 2\pi G\mu^2\Delta x\epsilon_\circ |k| \cos kx \tag{170}$$

Note that this force is positive in the same places that the displacement is positive: it is a destabilizing force. Now add in a time dependence to the displacement: let $\epsilon = \epsilon_\circ \cos(\kappa x - \omega t)$. Then the total force is

$$F_{total} = -(\mu\Delta x)\kappa^2\epsilon_\circ \cos(\kappa x - \omega t) + 2\pi|k|G\mu^2\Delta x\epsilon_\circ \cos(\kappa x - \omega t) \tag{171}$$

while ma is

$$ma = \mu \Delta x (-\omega^2) \epsilon_\circ \cos(\kappa x - \omega t) \tag{172}$$

These must be equal, so

$$\omega^2 = \kappa^2 - 2\pi |k| G\mu \tag{173}$$

Thus for long wavelength perturbations with $k \to 0$, the oscillation frequency is the epicyclic frequency, but for small-scale perturbations with $k \to \infty$, ω^2 is negative so ω is imaginary and the perturbations grow exponentially instead of oscillating.

To stabilize the disk against these small-scale radial perturbations, we need to have a velocity dispersion. Then the stars with a given angular momentum will be spread out is a Gaussian cloud with profile $\propto \exp(-0.5x^2/\sigma_R^2)$ where the radial standard deviation $\sigma_R = \sigma(v_r)/\kappa$.

The potential generated by a rod will be the convolution of this Gaussian cloud with the previous force, while the force received by a given rod will be the convolution of the Gaussian cloud with the potential gradient. Each convolution introduces a factor of $\exp(-0.5k^2\sigma_R^2)$. Thus the dispersion relation in Eqn(173) becomes

$$\omega^2 = \kappa^2 - 2\pi |k| G\mu \exp(-k^2\sigma_R^2) \tag{174}$$

The maximum value of $k\exp(-k^2\sigma_R^2)$ occurs when $k\sigma_R = \sqrt{0.5}$. At this point

$$\omega^2 = \kappa^2 - \sqrt{2}\pi G\mu \exp(-0.5)/\sigma_R = \kappa\left(\kappa - \frac{\sqrt{2/e}\pi G\mu}{\sigma(v_r)}\right) \tag{175}$$

and this is greater than zero (giving a stabilized disk) only if

$$\frac{\sigma(v_r)\kappa}{\sqrt{2/e}\pi G\mu} = \frac{\sigma(v_r)\kappa}{2.7G\mu} > 1 \tag{176}$$

Note that the least stable k has a wavelength

$$2\pi/k = \sqrt{2}\pi\sigma(v_r)/\kappa \approx \pi\sigma(v_r)/\omega = (\pi\sigma(v_r)/v_c)R_\circ \tag{177}$$

which is about 30% of the radius, so the assumption that the perturbations are very small compared to the radius is not completely accurate.

Toomre has done a much more careful analysis and the Toomre stability criterion is

$$Q = \frac{\sigma(v_r)\kappa}{3.36G\mu} > 1 \tag{178}$$

21. Global Stability

Ostriker and Peebles derived a stability criterion for rotating systems relating the kinetic energy of rotation T_{rot} to the potential energy W:

$$T_{rot} < (0.12 - 0.14)|W| \tag{179}$$

If this limit is violated the system will form a bar, or become triaxial.

Now by the virial theorem the total kinetic energy is $T_{tot} = 0.5|W|$, so the Ostriker-Peebles stability criterion is that $T_{rot} < 0.25T_{tot}$. For the Milky Way disk with $v_{rot} = 220$ km/sec and $\sigma = 20$ km/sec, we have $T_{rot} = 0.98T_{tot}$ and the disk should be grossly unstable to bar formation. But if there is a halo with a mass 2 times larger than the disk mass and velocity dispersion $\sigma = 220/\sqrt{2}$ as in the singular isothermal sphere, then $T_{rot} = 0.25T_{tot}$ and the system is at least marginally stable.

22. Jeans Mass

Let us consider the gravitational effect on a gas with a sound speed of c_s. Let the mean density be $\rho_\circ$, and consider small perturbations $\rho_1, \vec{v}_1, P_1, \phi_1$ and $\vec{g}_1$ in the density, velocity, pressure, gravitational potential and gravitational acceleration. The force per unit volume is $-\rho_\circ \vec{\nabla}\phi_1 - \vec{\nabla}P_1$ so

$$\rho_\circ \frac{\partial v_1}{\partial t} = -\rho_\circ \vec{\nabla}\phi_1 - \vec{\nabla}P_1. \tag{180}$$

Mass conservation says that

$$\frac{\partial \rho}{\partial t} = -\vec{\nabla}\cdot(\rho\vec{v}) \tag{181}$$

and the sound speed gives $P_1 = c_s^2\rho_1$. Poisson's equation give $\nabla^2\phi_1 = 4\pi G\rho_1$.

Assuming a plane wave solution $\rho_1 = \rho_1 \exp(i(\vec{k}\cdot\vec{x} - \omega t))$ allows us to write (noting that $\vec{v}_1$ and $\vec{g}_1$ are in the same direction as $\vec{k}$, so only the magnitudes are used):

$$\begin{aligned} -i\omega\rho_1 &= -i\rho_\circ v_1 k \\ -k^2\phi_1 &= 4\pi G\rho_1 \\ i\frac{4\pi G\rho_\circ\rho_1}{k} - ic_s^2 k\rho_1 &= -i\omega\rho_\circ v_1 \\ &= -i\omega\rho_1\frac{\omega}{k} \end{aligned} \tag{182}$$

Multiplying the last line above by ik/ρ_1 gives

$$\omega^2 = c_s^2 k^2 - 4\pi G\rho_\circ. \tag{183}$$

This gives $\omega = 0$ at the Jeans wavenumber

$$k_J = \frac{\sqrt{4\pi G\rho_\circ}}{c_s}. \tag{184}$$

The Jeans wavelength is $\lambda_J = 2\pi/k_J$. The Jeans mass is the mass contained within a sphere of diameter λ_J. This value is

$$M_J = \frac{4\pi}{3}\rho_\circ(\frac{1}{2}\lambda_J)^3 = \frac{\pi^{5/2}c_s^3}{6G^{3/2}\rho_\circ^{1/2}}. \tag{185}$$

23. Virial Theorem

One of the most useful methods for obtaining masses of systems is the *Virial* Theorem. This states that the total kinetic energy of a bound system is minus one-half of the total potential energy:

$$\mathrm{KE} = -\frac{1}{2}\mathrm{PE} \tag{186}$$

Remember that the potential energy of a bound system is negative so the minus sign gives a positive kinetic energy. To prove this consider the moment of inertia

$$I = \sum m_i \vec{r}_i^{\,2} \tag{187}$$

If the system has settled into a steady state, then the time-averaged value of the second time derivative of I will be zero. So

$$\dot{I} = \sum 2 m_i \vec{r}_i \cdot \dot{\vec{r}}_i \tag{188}$$

and

$$\ddot{I} = \sum 2 m_i \left(\dot{\vec{r}}_i^{\,2} + \vec{r}_i \cdot \ddot{\vec{r}}_i \right). \tag{189}$$

Thus

$$\ddot{I} = 4 \sum \frac{1}{2} m_i \vec{v}_i^{\,2} + 2 \sum_i m_i \vec{r}_i \cdot \left(\sum_{j \neq i} \frac{G m_j (\vec{r}_j - \vec{r}_i)}{|\vec{r}_j - \vec{r}_i|^3} \right). \tag{190}$$

Now

$$\sum_i \sum_{j \neq i} \frac{G m_i m_j \vec{r}_i \cdot (\vec{r}_j - \vec{r}_i)}{|\vec{r}_j - \vec{r}_i|^3} = \sum_j \sum_{i \neq j} \frac{G m_i m_j \vec{r}_j \cdot (\vec{r}_i - \vec{r}_j)}{|\vec{r}_j - \vec{r}_i|^3} \tag{191}$$

because the RHS is obtained by just interchanging the indices i and j, and the names of the indices don't matter when they are summed over. But if these two quantities are equal, we can add them together and divide by two and get the same value again. This gives

$$\begin{aligned} \sum_i \sum_{j \neq i} \frac{G m_i m_j \vec{r}_i \cdot (\vec{r}_j - \vec{r}_i)}{|\vec{r}_j - \vec{r}_i|^3} &= \frac{1}{2} \sum_i \sum_{j \neq i} \frac{G m_i m_j (\vec{r}_i - \vec{r}_j) \cdot (\vec{r}_j - \vec{r}_i)}{|\vec{r}_j - \vec{r}_i|^3} \\ &= -\frac{1}{2} \sum_{i, j \neq i} \frac{G m_i m_j}{|\vec{r}_j - \vec{r}_i|} = \mathrm{PE} \end{aligned} \tag{192}$$

Thus

$$\ddot{I} = 4(\mathrm{KE}) + 2(\mathrm{PE}) = 0 \tag{193}$$

in a steady state. This needs to be taken as a time average, since in many bound systems such as a elliptical binary the ratio of the instantaneous kinetic energy to the instantaneous potential energy varies with phase. But the time-averaged values satisfy

$$\langle \mathrm{KE} \rangle = -\frac{1}{2} \langle \mathrm{PE} \rangle \tag{194}$$

Usually we apply the virial theorem to a distant cluster where only radial velocities can be measured. If the radial velocity dispersion is $\sigma(v_r)$ then the kinetic energy is

$$\mathrm{KE} = \frac{3}{2} M \sigma(v_r)^2 \tag{195}$$

where the "3" comes from assuming isotropy in the velocity distribution. The potential energy is

$$\mathrm{PE} = -\frac{GM^2}{R_e} \tag{196}$$

where R_e, the effective radius, is found from

$$-\frac{1}{2}\sum_{i,j\neq i}\frac{Gm_i m_j}{|\vec{r}_j - \vec{r}_i|} = \mathrm{PE} = -\frac{GM^2}{R_e} \tag{197}$$

so

$$\frac{1}{R_e} = \frac{1}{2}\left(\sum_{i,j\neq i}\frac{(m_i/M)(m_j/M)}{|\vec{r}_j - \vec{r}_i|}\right). \tag{198}$$

Evaluating the 3-dimensional $\langle R^{-1}\rangle$ in terms of the 2-dimensional projected separation on the sky, as done in Eqn(10-19) of "Galactic Dynamics" by Binney & Tremaine, gives a formula that is not suitable for practical work because its variance is infinite. B&T's approach uses

$$\left\langle\frac{1}{|\mathbf{R}_i - \mathbf{R}_j|}\right\rangle_\Omega = \frac{\pi}{2|\vec{r}_i - \vec{r}_j|} \tag{199}$$

But the probability of getting a small projected separation and hence a very large $|\mathbf{R}_i - \mathbf{R}_j|^{-1}$ is

$$P(|\mathbf{R}_i - \mathbf{R}_j|^{-1} > x) = 1 - \sqrt{1 - (|\vec{r}_i - \vec{r}_j|x)^{-2}} \propto x^{-2} \tag{200}$$

for large x so $p(x)dx \propto dx/x^3$. To find the variance one needs the second moment $\int x^2 p(x)dx$ but this is clearly infinite. When the variance is infinite then averaging values together will not improve your estimate for $1/R_e$.

A more stable evaluation of the potential energy can be found if a model for the density law of the cluster, $\rho(r)$, is found. Then

$$\mathrm{PE} = -16\pi^2 G\int_0^\infty \rho(r)r\left(\int_0^r \rho(r')r'^2 dr'\right)dr \tag{201}$$

Peebles in "Physical Cosmology" (the 1971 book, not the later "Principles of Physical Cosmology") gives an interesting way to determine R_e from strip counts. Let $S(\delta)$ be the count of objects in a strip displaced by δ from the center of the cluster on the sky. This strip is really a plane in 3-D. Let x and y be coordinates in that plane. Then

$$S(\delta) = \int\int n(\sqrt{\delta^2 + x^2 + y^2})dxdy = 2\pi\int_0^\infty \eta n(\sqrt{\delta^2 + \eta^2})d\eta \tag{202}$$

where $\eta = \sqrt{x^2 + y^2}$. Now $r = \sqrt{\delta^2 + \eta^2}$ so $rdr = \eta d\eta$. Thus

$$S(\delta) = 2\pi\int_\delta^\infty rn(r)dr \tag{203}$$

and

$$\frac{dS}{d\delta} = -2\pi\delta n(\delta). \tag{204}$$

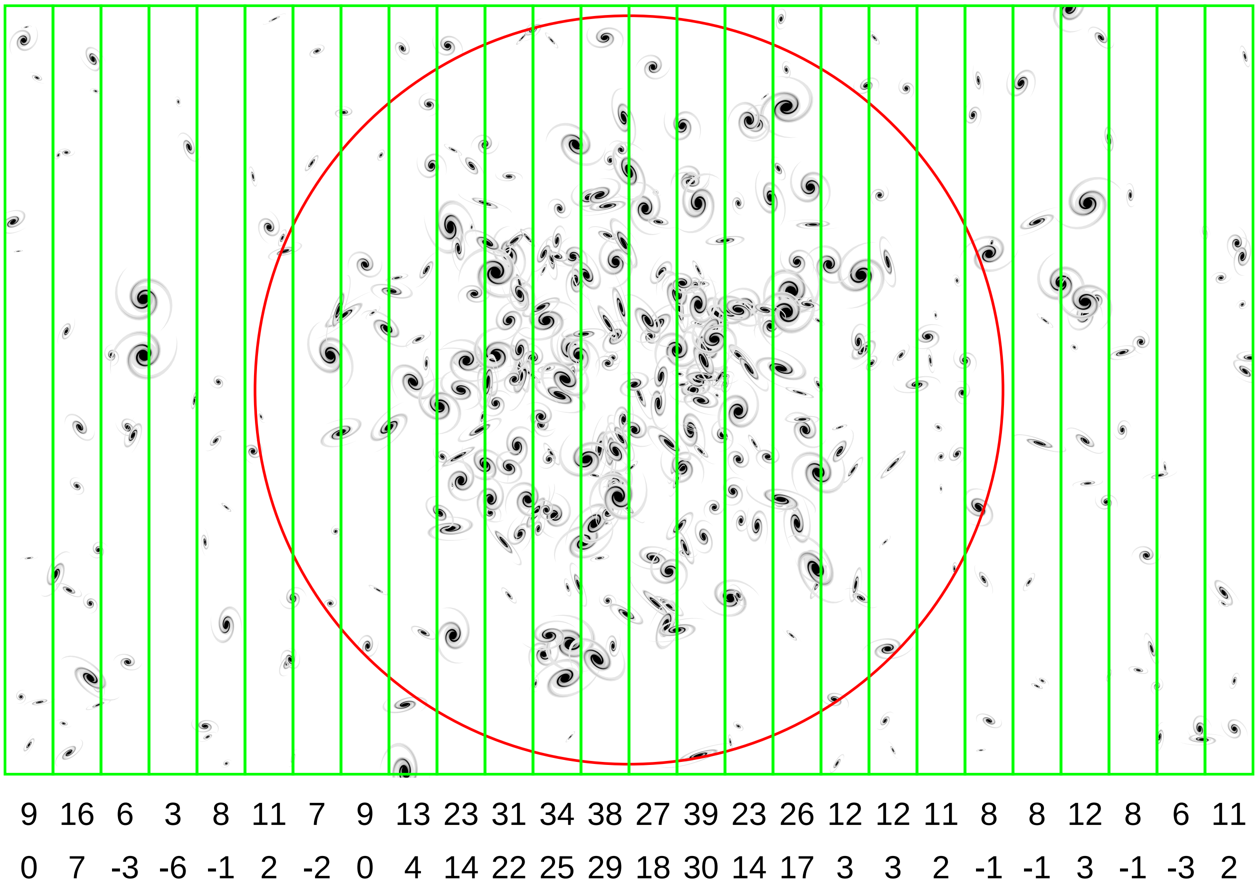

Fig. 22.— A schematic cluster with strip counts, background subtracted strip counts, and R_e computed using Eqn(209) shown by the circle.

Now let m be the mass of an object and M be the mass of the cluster:

$$M = 4\pi m \int_0^\infty n(r) r^2 dr = 2m \int_0^\infty S(\delta) d\delta \tag{205}$$

The potential energy is

$$\begin{aligned} \mathrm{PE} &= -16\pi^2 G m^2 \int_0^\infty n(r) r \left(\int_0^r n(r') r'^2 dr' \right) dr \\ &= -4Gm^2 \int_0^\infty \frac{dS}{dr} \left(\int_0^r r' \frac{dS}{dr'} dr' \right) dr \end{aligned} \tag{206}$$

This can be integrated by parts with $q = S$ and $p = \int_0^r r'(dS/dr')dr'$ giving

$$\mathrm{PE} = 4Gm^2 \int_0^\infty S r \frac{dS}{dr} dr. \tag{207}$$

Once again integrate by parts with $q = S^2/2$ and $p = r$, giving

$$\mathrm{PE} = -2Gm^2 \int_0^\infty S^2 dr \tag{208}$$

Now

$$R_e = -\frac{GM^2}{\mathrm{PE}} = \frac{4Gm^2\left[\int_0^\infty S dr\right]^2}{2Gm^2\int_0^\infty S^2 dr} = \frac{2\left[\int_0^\infty S dr\right]^2}{\int_0^\infty S^2 dr} = \frac{\left[\int_{-\infty}^{+\infty} S dr\right]^2}{\int_{-\infty}^{+\infty} S^2 dr} \tag{209}$$

Applying the virial theorem gives

$$M = \frac{3\sigma(v_r)^2}{G}\frac{2\left[\int_0^\infty S dr\right]^2}{\int_0^\infty S^2 dr} \tag{210}$$

Note that R_e is usually larger than one would naively expect. For a uniform density sphere of radius R, the strip counts are $S(\delta) \propto [1-(\delta/R)^2]$. Then $\int S(\delta)d\delta = 4R/3$ and $\int S(\delta)^2 d\delta = 16R/15$. Then $R_e = (5/3)R$. For the Plummer model with $\rho \propto [1+(r/b)^2]^{-5/2}$ the effective radius is $R_e = (32/3\pi)b$. Fig. 22 shows how R_e is considerably bigger than the half-density radius of a cluster.

If the strip counts follow

$$S \propto \frac{1}{\delta^2 + \delta_\circ^2} \tag{211}$$

then

$$\begin{aligned} \int S(\delta)d\delta &= \frac{\pi}{2}\frac{1}{\delta_\circ} \\ \int S(\delta)^2 d\delta &= \frac{\pi}{4}\frac{1}{\delta_\circ^3} \\ R_e &= \frac{2(\pi/(2\delta_\circ))^2}{\pi/(4\delta_\circ^3)} = 2\pi\delta_\circ \end{aligned} \tag{212}$$

If the strip counts follow $S \propto \exp(-\delta^2/\delta_\circ^2)$ then $R_e = \sqrt{2\pi}\delta_\circ$.

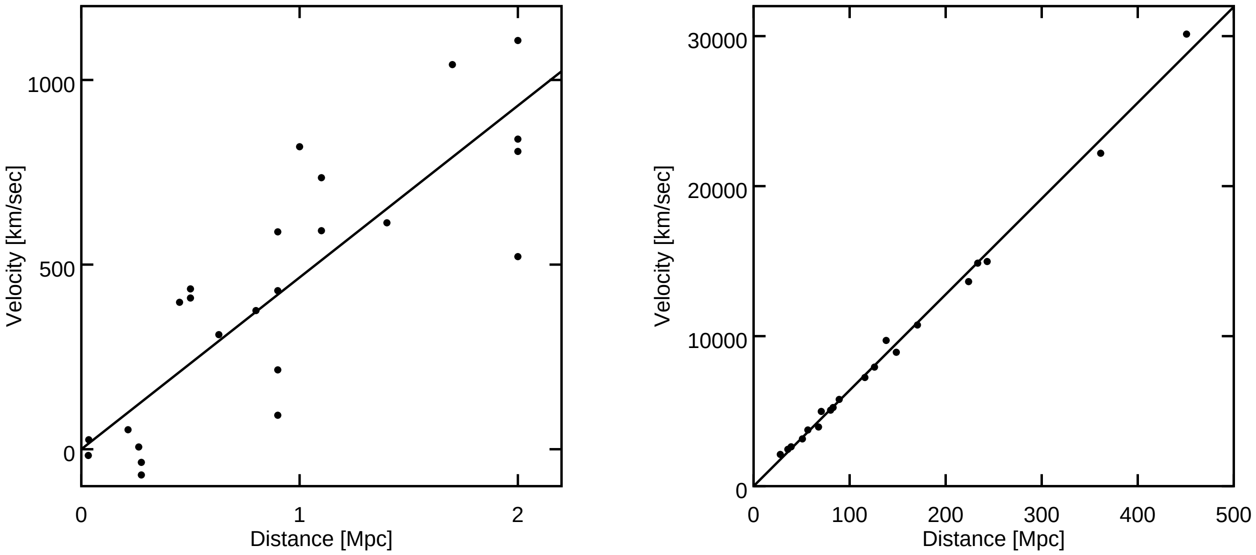

Fig. 23.— Left: Hubble's velocity-distance data in 1929. Right: velocity-distance data from Riess, Press & Kirshner (1996, ApJ, 473, 88).

24. Cosmological Observations

24.1. Recession velocities

Modern cosmology has been driven by observations made in the 20^{th} century. While there were many speculations about the nature of the Universe, little progress was made until data were obtained on distant objects. The first of these observations was the discovery of the expansion of the Universe. In the paper "THE LARGE RADIAL VELOCITY OF NGC 7619" by Milton L. Humason (1929) we read that

"About a year ago Mr. Hubble suggested that a selected list of fainter and more distant extra-galactic nebulae, especially those occurring in groups, be observed to determine, if possible, whether the absorption lines in these objects show large displacements toward longer wave-lengths, as might be expected on de Sitter's theory of curved space-time.

During the past year two spectrograms of NGC 7619 were obtained with Cassegrain spectrograph VI attached to the 100-inch telescope. This spectrograph has a 24-inch collimating lens, two prisms, and a 3-inch camera, and gives a dispersion of 183 Angstroms per millimeter at 4500. The exposure times for the spectrograms were 33 hours and 45 hours, respectively. The radial velocity from these plates has been measured by Miss MacCormack, of the computing division, and by myself, the weighted mean value being +3779 km./sec."

Note that NGC 7619 is a 12^{th} magnitude galaxy and the observational limit now is $B < 24^{th}$ magnitude, and especially note the total exposure time of 78 hours!

"A RELATION BETWEEN DISTANCE AND RADIAL VELOCITY AMONG EXTRA-GALACTIC NEBULAE" by Edwin Hubble (1929) takes the radial velocities for 24 galaxies with "known" distances and fits them to the form

$$v = Kr + X \cos\alpha \cos\delta + Y \sin\alpha \cos\delta + Z \sin\delta \tag{213}$$

where K is the coefficient relating velocity to distance in a linear velocity distance law, while (X, Y, Z) are the contribution of the solar motion to the radial velocity. Hubble found a solution corresponding to solar motion of 280 km/sec toward galactic coordinates $l = 65$, $b = 18$. Modern determinations give 308 km/sec toward $l = 105$, $b = 7$ (Yahil, Tammann & Sandage, 1977), so this part of the fit has remained quite stable. But the value of $K = 500$ km/sec/Mpc derived by Hubble in 1929 is much too large, because his distances were much too small. Nonetheless, this discovery that distant galaxies have recession velocities proportional to their distances is the cornerstone of modern cosmology.

In modern terminology, Hubble's K is denoted $H_\circ$, and called the *Hubble constant.* Since it is not really a constant, but decreases as the Universe gets older, some people call it the Hubble parameter.

This velocity field, $\vec{v} = H_\circ \vec{r}$, has the very important property that its form is unchanged by a either a translation or a rotation of the coordinate system. To have a relation unchanged in form during a coordinate transformation is called an *isomorphism.* The isomorphism under rotations around the origin is obvious, but to see the effect of translations consider the observations made by astronomers on a galaxy A with position $\vec{r}_A$ relative to us and a velocity $\vec{v}_A = H_\circ \vec{r}_A$ relative to us. Astronomers on A would measure positions relative to themselves $\vec{r}' = \vec{r} - \vec{r}_A$ and velocities relative to themselves, $\vec{v}' = \vec{v} - \vec{v}_A$. But

$$\vec{v}' = H_\circ \vec{r} - H_\circ \vec{r}_A = H_\circ (\vec{r} - \vec{r}_A) = H_\circ \vec{r}' \tag{214}$$

so astronomers on galaxy A would see the same Hubble law that we do.

Thus even though we see all galaxies receding from us, this does not mean that we are in the center of the expansion. Observers on any other galaxy would see exactly the same thing. Thus the Hubble law does not define a center for the Universe. Other forms for the distance-velocity law do define a unique center for the expansion pattern: for example neither a constant velocity $\vec{v} = v_\circ \hat{r}$ nor a quadratic law $\vec{v} = M r^2 \hat{r}$ are isomorphic under translations.

In actuality one finds that galaxies have peculiar velocities in addition to the Hubble velocity, with a magnitude of ± 500 km/sec. In order to accurately test the Hubble law, one needs objects of constant or calibratable luminosities that can be observed at distances large enough so the Hubble velocity is $>> 500$ km/sec. Type Ia supernovae are very bright, and after a calibration based on their decay speed, they have very small dispersion in absolute magnitude. Riess, Press & Kirshner (1996) find that slope in a fit of velocities to distance moduli, $\log v = a(m - M) + b$, is $a = 0.2010 \pm 0.0035$ while the Hubble law predicts $a = 1/5$. Thus the Hubble law has a good theoretical basis and is well-tested observationally.

The actual value of the Hubble constant $H_\circ$ is less well determined since it requires the measurement of absolute distances instead of distance ratios. The best value now is $H_\circ = 71 \pm 3.5$ km/sec/Mpc, but this was determined rather indirectly using the anisotropy of the cosmic microwave background. The best direct measurement comes from the Hubble Space Telescope Key Project on the Distance Scale: $H_\circ = 72 \pm 8$ km/sec/Mpc (Freedman *et al.*, 2001, ApJ, 553, 47). We will discuss many ways of measuring $H_\circ$ later in the course. In many cosmological results the uncertain value of $H_\circ$ is factored out using the notation $h = H_\circ/100$ in the standard units of km/sec/Mpc. Thus if a galaxy has a Hubble velocity of 1500 km/sec, its distance is $15/h$ Mpc.

While units of km/sec/Mpc are commonly used for $H_\circ$, the metric units would be second^{-1}. The conversion is simple:

$$H_\circ = 100 \text{ km/sec/Mpc} = \frac{10^7 \text{ cm/sec}}{3.085678 \times 10^{24} \text{ cm}} = 3.2 \times 10^{-18} \text{ s}^{-1} = \frac{1}{9.78 \text{ Gyr}} \tag{215}$$

24.2. Age

Another observable quantity in the Universe is the age of the oldest things in it. There are basically three ways to find ages for very old things. The first and best known example is the use of the HR diagram to determine the age of a star cluster. The luminosity of the stars just leaving the main sequence (main sequence turnoff = MSTO) varies like $L \propto M^4$, so the main sequence lifetime varies like $t \propto M^{-3} \propto L^{-3/4}$. This means that a 10% distance error to globular cluster gives a 15-20% error in the derived age. Distances to globular clusters are determined from the magnitudes of RR Lyrae stars, and there are two different ways to estimate their absolute magnitudes. Using the statistical parallax of nearby RR Lyrae stars gives an age for the oldest globular cluster of 18 ± 2 Gyr, while using the RR Lyrae stars in the LMC (Large Magellanic Cloud) as standards gives an age of 14 ± 2 Gyr. New data from Hipparcos suggests the globular clusters are more distant so the age could be as low as 12 ± 2 Gyr.

The second technique for determining ages of stellar populations is to look for the oldest white dwarf. White dwarves are formed from stars with initial masses less than about 8 $M_\odot$ so the first WDs form after about 20 million years. Once formed, WDs just get cooler and fainter. Thus the oldest WDs will be the least luminous and coolest WDs. These are the hardest to find, of course. However, there appears to be a sharp edge in the luminosity function of white dwarves, corresponding to an age of about 11 Gyr. Since these are disk stars, and the stars in the disk formed after the halo stars and globular clusters, this means that the age of the Universe is at least 12 Gyr. One pitfall in this method is the phenomenon of crystalization in white dwarf nuclei. When the central temperature get low enough, the nuclei arrange themselves into a regular lattice. As this happens, the WDs remain for a long time at a fixed temperature and luminosity. After the heat of solidification is radiated away, the WD cools rapidly. Thus there will naturally be an edge in the luminosity function of WDs caused by crystalization. While the best evidence is that the oldest WDs haven't yet started to crystalize, the expected luminosity of a solidifying WD is only slightly below the observed edge. But Hansen *et al.*(2002, ApJL, 574, 155) managed to see the faintest WDs in the globular cluster M4 and derived an age of 12.7 Gyr for this cluster.

The third way to measure the age of the Universe is to measure the age of the chemical elements. This method relies on radioactive isotopes with long half-lives. It is very easy to make a precise measurement of the time since a rock solidified, and by applying this technique to rocks on the Earth an oldest age of 3.8 Gyr is found. But rocks that fall out of the sky, meteorites, are older. The Allende meteorite is well studied and has an age of 4.554 Gyr. It is much more difficult to get an age for the Universe as a whole, since one has to assume a model for the star formation history and for stellar nucleosynthesis yields. For example, the ratio of ^{187}Re to ^{187}Os is less than that predicted by nucleosynthesis calculations, and ^{187}Re is radioactive. The derived average age of the elements is 9.3 ± 1.5 Gyr. Assuming that the elements in the Solar System (since the ^{187}Re:^{187}Os ratio can only be measured in the Solar System) were made uniformly between the age of the Universe $t_\circ$ and the formation of the Solar System, then $t_\circ = 14 \pm 3$ Gyr.

The dimensionless product $H_\circ t_\circ$ can be used to discriminate among cosmological models. Taking $t_\circ = 14 \pm 2$ Gyr and $H_\circ = 72 \pm 8$ km/sec/Mpc, this product is

$$H_\circ t_\circ = \frac{h t_\circ}{9.78\ \mathrm{Gyr}} = 1.03 \pm 0.19 \tag{216}$$

The WMAP model based on CMB data that gives $H_\circ = 71 \pm 3.5$ km/sec/Mpc also gives $t_\circ = 13.7 \pm 0.2$ Gyr, or

$$H_\circ t_\circ = \frac{h t_\circ}{9.78\ \mathrm{Gyr}} = 0.99 \pm 0.05 \tag{217}$$

24.3. High Redshift Time Dilation

When high redshift events are observed, we should observe a longer time between events than the time in the high redshift objects rest frame. So far the only events with known lifetimes that are bright enough to see at high redshift are supernovae. Since there is an intrinsic spread in decay times, one must either use a collection of objects or else calibrate the decay time by the use of color data, since color and decay rate are correlated for Type Ia SNe. Figure 24 shows the data that have been published to date. Tired light models, which attribute the redshift to an intrinsic redshift caused by an as yet undetermined cause, can be ruled out by these data.

24.4. Number counts

While it took weeks of exposure time to measure the redshifts of galaxies in the 1920's, it was much easier to photograph them and measure their magnitudes and positions. An important observable is the number of sources brighter than a given flux per unit solid angle. This is normally denoted $N(S)$ or $N(>S)$, where S is the limiting flux, and N is the number of objects per steradian brighter than S. In principle $N(S)$ is a function of direction as well as flux. In practice,

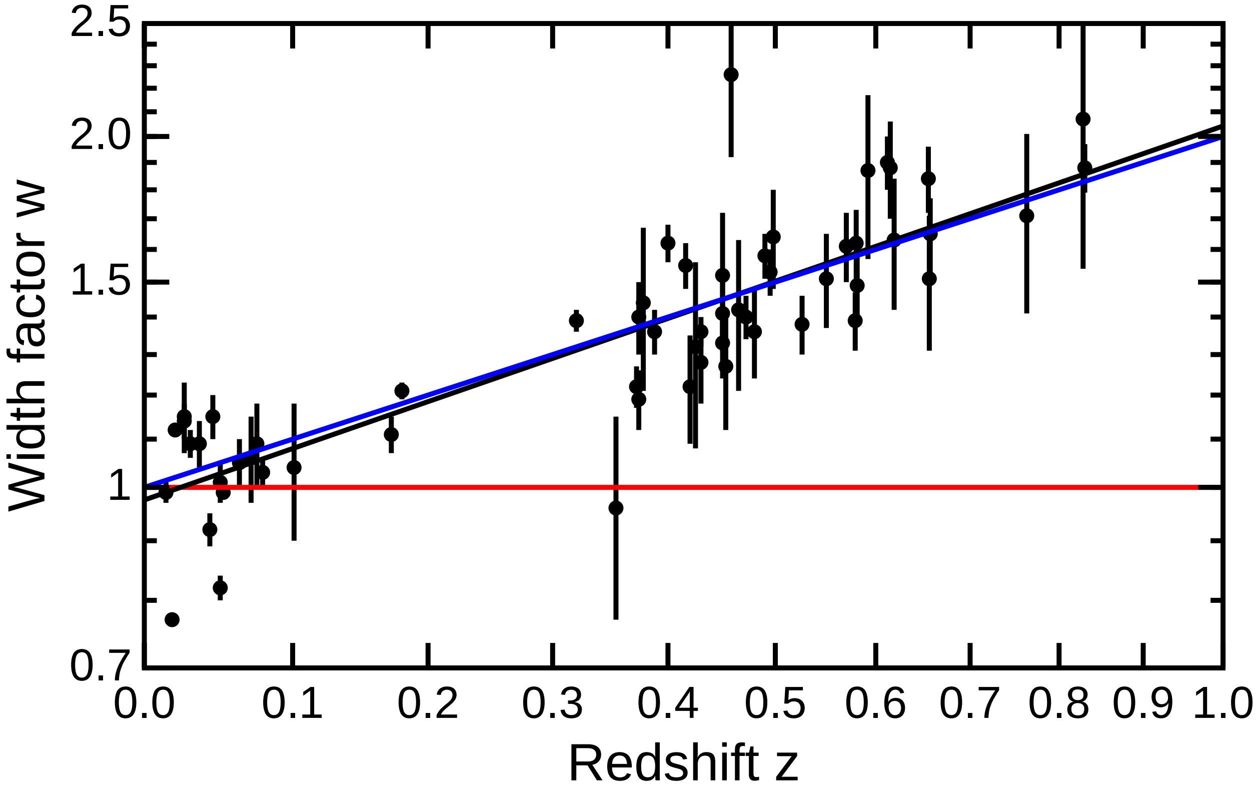

Fig. 24.— High redshift supernovae exhibit slower decays in their lightcurves. The ratio of observed decay time to the decay time of similar low redshift supernovae is plotted *vs.* redshift, and agrees with the expected decay time proportional to $(1+z)$ relation shown in blue. The red curve shows what would happen if some effect other than the Doppler shift produced a gradual decay of photon energy (tired light). The black line is the best fit to these data.

for brighter galaxies, there is a very prominent concentration in the constellation Virgo, known as the Virgo cluster, and toward a plane is the sky known as the supergalactic plane. This larger scale concentration is known as the Local Supercluster.

However, as one looks at fainter and fainter galaxies, the number of galaxies per steradian gets larger and larger, and also much more uniform across the sky. For optical observations the dust in the Milky Way creates a "zone of avoidance" where only a few galaxies are seen, but this effect is not seen in the infrared observations from the IRAS experiment. Thus it is reasonable to postulate that the Universe is *isotropic* on large scales, since the number counts of faint galaxies are approximately the same in all directions outside the zone of avoidance. Isotropic means the same in all directions. Mathematically isotropic means isomorphic under rotations.

The slope of the number counts, $d \ln N / d \ln S$, is another observable quantity. If the sources being counted are uniformly distributed throughout space, than observing to a flux limit 4 times lower will allow one to see object twice as far away. But this volume is 8 times larger, so the slope of the source counts is $-\ln 8/\ln 4 = -3/2$. Hubble observed that the source counts followed this law

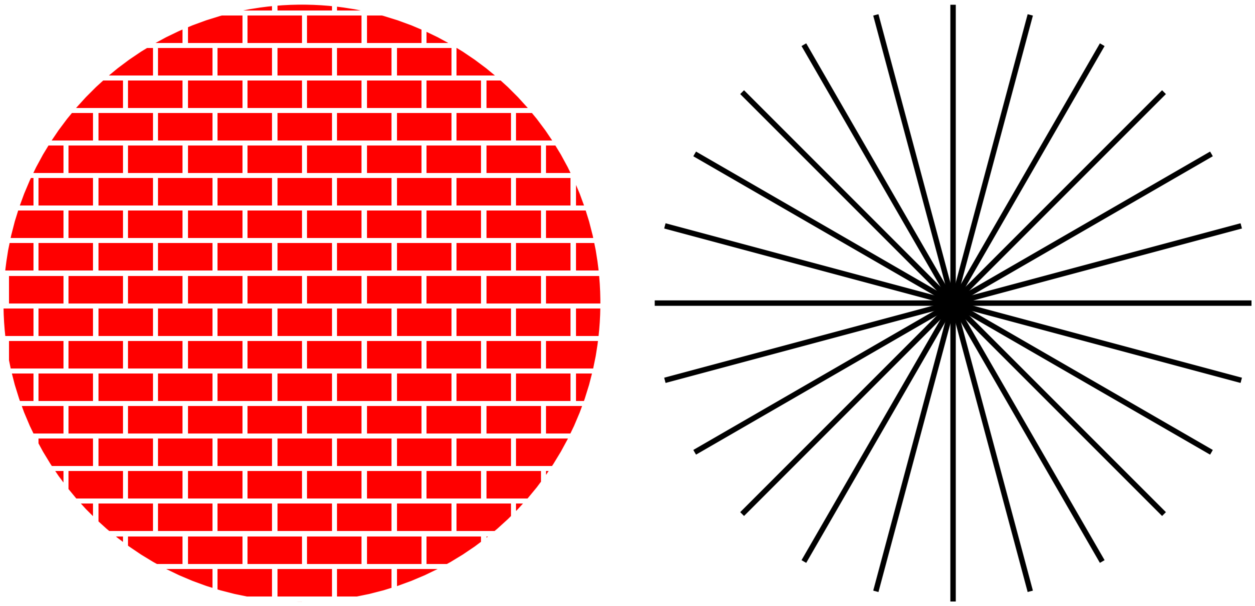

Fig. 25.— The brick wall on the left is homogeneous but not isotropic, while the radial pattern on the right is isotropic around its center but not homogeneous.

rather well, indicating that the galaxies beyond the Local Supercluster but within reach of the 100 inch telescope and old photographic plates were uniformly distributed in space. This implies that the Universe is *homogeneous* on large scales. Just as homogenized milk is not separated into cream and skim milk, a homogeneous Universe is not separated into regions with different properties. Mathematically homogeneous means isomorphic under translations.

It is possible be isotropic without being homogeneous, but the isotropy will only hold at one or two points. Thus a sheet of polar coordinate graph paper is in principle isotropic around its center, but it is not homogeneous. The meridians on a globe form an isotropic pattern around the North and South poles, but not elsewhere.

It is also possible to be homogeneous but not isotropic. Standard square grid graph paper is in principle homogeneous but it is not isotropic since there are four preferred directions. A pattern like a brick wall is homogeneous but not isotropic. Note that a pattern that is isotropic around three or more points is necessarily homogeneous.

24.5. CMBR

In 1964 Penzias & Wilson found an excess of radio noise in the big horn antenna at Bell Labs. This excess noise was equivalent to 3.5 ± 1 K. This means that a blackbody with $T = 3.5$ K would produce the same amount of noise. A blackbody is an object that absorbs all the radiation that hits it, and has a constant temperature. Penzias & Wilson were observing 4 GHz ($\lambda = 7.5$ cm), and if the radiation were truly a blackbody, then it would follow the Planck radiation law

$$I_\nu = B_\nu(T) = 2h\nu \left(\frac{\nu}{c}\right)^2 \frac{1}{\exp(h\nu/kT) - 1} \tag{218}$$

for all frequencies with a constant T. Here I_ν is the intensity of the sky in units of erg/cm^2/sec/sr/Hz or W/m^2/sr/Hz or Jy/sr. Actually this blackbody radiation was first seen at the same 100 inch telescope used to find the expansion of the Universe in the form of excitation of the interstellar cyanogen radical CN into its first excited state. This was seen in 1939 but the resulting excitation temperature at $\lambda = 2.6$ mm of 2.3 ± 1 K was not considered significant. After 1964, many groups measured the brightness of the sky at many different wavelengths, culminating in Matehr *et al.* (1999, ApJ, 512, 511), which finds $T = 2.725 \pm 0.002$ K for 0.5 mm to 5 mm wavelength as shown in Figure 26.

This blackbody radiation was predicted by Gamow and his students Alpher and Herman from their theory for the formation of all the chemical elements during a dense hot phase early in the history of the Universe. Alpher & Herman (1948) predict a temperature of 5 K. But this theory of the formation of elements from 1 to 92 failed to make much of anything heavier than helium because there are no stable nuclei with atomic weights of 5 or 8. Thus the successive addition of protons or neutrons is stopped at the $A = 5$ gap. Because of this failure, the prediction of a temperature of the Universe was not taken seriously until after Penzias & Wilson had found the blackbody radiation. A group at Princeton led by Dicke was getting set to try to measure the radiation when they were scooped by Penzias & Wilson. When Dicke started this project he asked a student to find previous references and the only prior measurement of the temperature of the sky that had been published was $T < 20$ K by Dicke himself. And this paper was published in the same volume of the *Physical Review* that had Gamow's first work.

The CMBR (Cosmic Microwave Background Radiation) is incredibly isotropic (Figure 27). Except for a dipole term due to the Sun's motion relative to the cosmos (like the (X, Y, Z) terms in Hubble's fit), the temperature is constant to 11 parts per million across the sky.

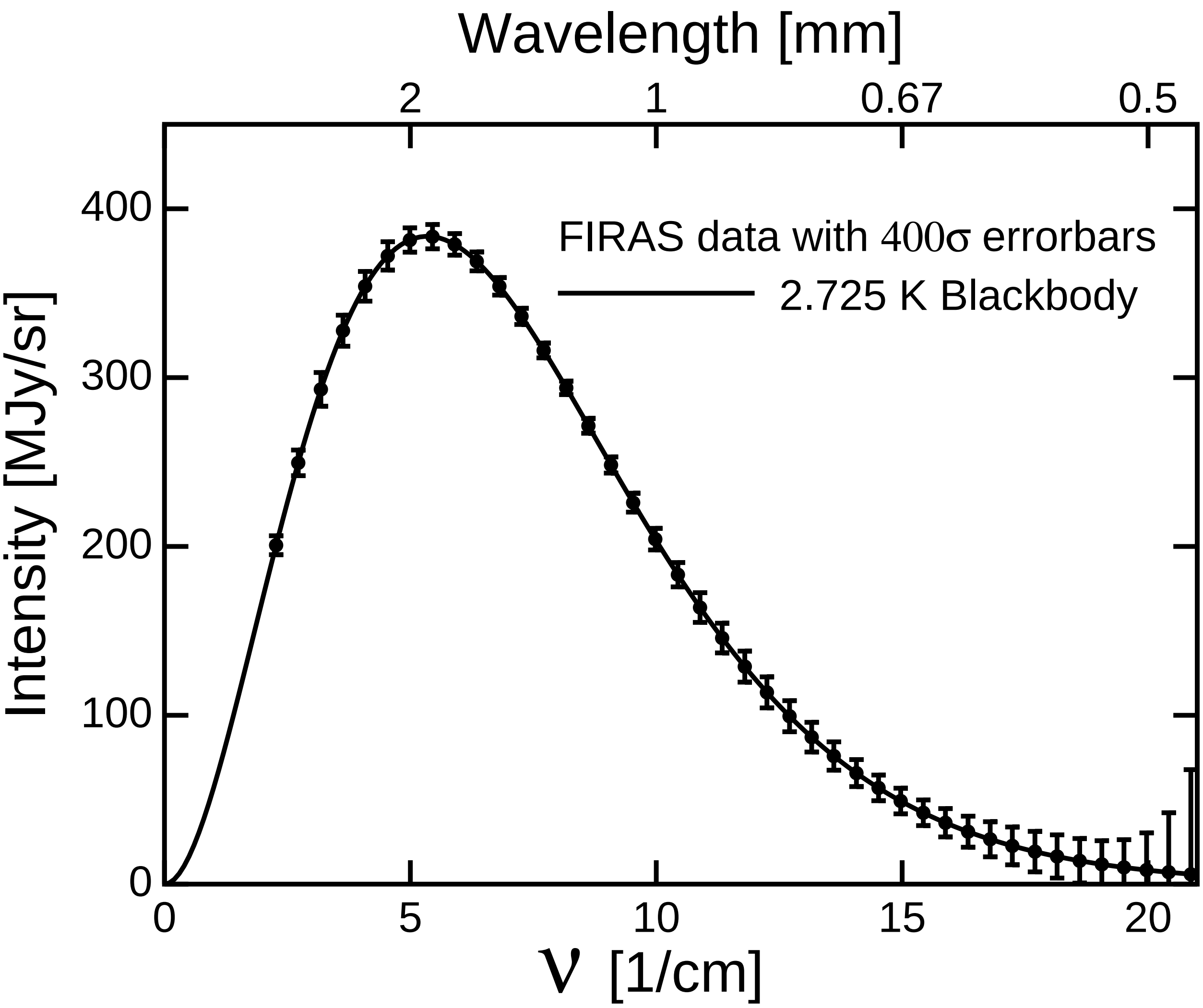

Fig. 26.— The spectrum of the Cosmic Microwave Background: measured data (points with errorbars) compared to a 2.725 K Planck function. The data agrees with the model with an RMS accuracy of 1 part in 20,000.

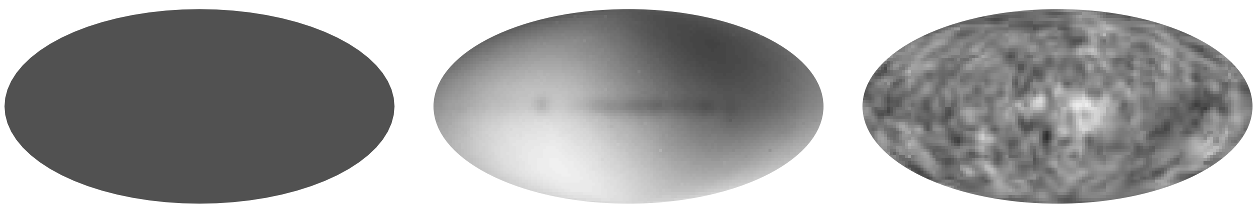

Fig. 27.— Left: true contrast CMB sky. 0 K = white, 4 K = black. Middle: contrast enhanced by 400X, monopole removed, showing dipole and Milky Way. Right: contrast enhanced by 6,667X, with monopole, dipole and Milky Way removed.

25. Cosmological Principle

Because the CMBR is so isotropic, and since isotropic at more than two points implies homogeneity, and taking the Copernican view that the Earth is not in a special place in the cosmos, we come to promulgate the cosmological principle:

$$\boxed{\text{The Universe is Homogeneous and Isotropic}}$$

Since galaxies are receding from each other, the average density of the Universe will be decreasing with time, unless something like the Steady State model were correct [but it's not]. This means that we have to be careful about how we define homogeneity. We have to specify a cosmic time and state that the Universe is homogeneous only on slices through space-time with constant cosmic time. This sounds like it contradicts one of the tenets of special relativity, which states that different observers moving a different velocities will disagree about whether events are simultaneous. However, an observer traveling at 0.1 c relative to us would disagree even more about the Hubble law, since she would see blueshifts of up to 30,000 km/sec on one side of the sky, and redshifts greater than 30,000 km/sec on the other side of the sky. Thus we can define a special class of observers, known as comoving observers, who all see the Hubble law for galaxy redshifts in its simple form. When we ask each comoving observer to determine the local density ρ at a time when the measured age of the Universe is t, then homogeneity means that this ρ is a function only of t and does not depend on the location of the observer.

26. Geometry

The implications of the Cosmological Principle and the Hubble law are substantial. Let the distance between two galaxies A and B at time t be $D_{AB}(t)$. This distance has to be measured by comoving observers all at time t. Once distances are large enough so the light travel time becomes important, this distance must be measured by several comoving observers strung out along the way between A and B. For example, the path from A to B might be $A \rightarrow 1 \rightarrow 2 \ldots \rightarrow B$. An observer on galaxy A can measure the distance $D_{A1}(t)$ by sending out a radar pulse at time $t_S = t - D_{A1}(t)/c$, and receiving the echo at time $t_R = t + D_{A1}(t)/c$. The distance is found using $D = c(t_R - t_S)/2$ as in all radars. Of course since the observer on A is trying to measure the distance, she would have to either guess the correct time, or else send out pulses continually with each pulse coded so the echo can be identified with the correct transmit time. In a time interval t to $t + \Delta t$, each of these small subintervals grows to $(1 + H\Delta t)$ times its initial value. Thus for any pair of galaxies the distance grows by a factor $(1 + H\Delta t)$ even if the distance D_{AB} is quite large. Thus the equation

$$v = \frac{dD_{AB}}{dt} = HD_{AB} \tag{219}$$

which is the Hubble law is exactly true even when the distances are larger than c/H and the implied velocities are larger than the speed of light. This is a consequence of the way distances and times are defined in homogeneous cosmologies, which are consistent with the locally inertial coordinates of a comoving observer only for small distances.

A useful consequence of the Hubble law is that $D_{AB}(t)$, which depends on three variables, can be factored into a time variable part $a(t)$ and a fixed part X_{AB} which depends on the pair of objects but not on the time:

$$D_{AB}(t) = a(t)X_{AB} \tag{220}$$

where $a(t)$ is the cosmic scale factor and applies to the whole Universe, while X_{AB} is the *comoving distance* between A and B. Obviously one can multiply all the X's by 10 and divide $a(t)$ by 10 and get the same D's, so a convention to fix the scale of the scale factor is needed. We will usually use $a(t_\circ) = 1$ where $t_\circ$ is the current age of the Universe. Of course that means that our calculations done today will be off by one part in 4 trillion tomorrow, but this error is so small we ignore it.

The common growth factor $(1 + H\Delta t)$ discussed earlier can be written as $a(t + \Delta t)/a(t)$. Therefore the Hubble constant can be written as

$$H = \frac{1}{a}\frac{da}{dt}. \tag{221}$$

The second derivative of a with respect to time enters into the dimensionaless *deceleration parameter* $q = -a\ddot{a}/\dot{a}^2$. Unless $a \propto e^{Ht}$, the Steady State model, the value of the Hubble constant will change with time. Thus some people call it the Hubble parameter.

26.1. Relation between z and $a(t)$

If the Hubble velocities $v_{AB} = dD_{AB}/dt$ can be larger than c, we probably should not use the special relativistic Doppler formula for redshift,

$$\frac{\lambda_{obs}}{\lambda_{em}} = 1 + z = \sqrt{\frac{1 + v/c}{1 - v/c}}. \tag{222}$$

What technique can we use instead? Let's go back to our series of observers $A \to 1 \to 2 \ldots \to n \to B$. Galaxy B emits light at time $t_B = t_{em}$ and wavelength λ_{em}. This reaches observer n at time $t_n = t_{em} + D_{nB}(t_{em})/c + \ldots$. This distance is small so we can use the first-order approximation

$$\frac{\lambda_n}{\lambda_B} = 1 + \frac{v}{c} = 1 + \frac{HD_{nB}(t_{em})}{c} = 1 + H(t_n - t_{em}) \tag{223}$$

But the rightmost side is just $a(t_n)/a(t_{em})$. The same argument can be applied to show that

$$\frac{\lambda_{n-1}}{\lambda_n} = \frac{a(t_{n-1})}{a(t_n)} \tag{224}$$

Finally we get to galaxy A at time $t_{obs} = t_A$, and can write

$$1 + z = \frac{a(t_A)}{a(t_1)}\frac{a(t_1)}{a(t_2)} \cdots \frac{a(t_n)}{a(t_B)} = \frac{a(t_A)}{a(t_B)} = \frac{a(t_{obs})}{a(t_{em})} \tag{225}$$

This formula only applies to the redshifts of comoving observers as seen by comoving observers. It is always possible to have a spaceship traveling at $v/c = 0.8$ one light-day away from the Solar System emit light yesterday which we see today with a redshift of $1+z = 3$ even though $a(t_{obs})/a(t_{em}) \approx 1$.

We can also derive the law in Eqn(225) by considering a photon bouncing in a mirrored box. A room with mirrors on all walls gives the appearance of a homogeneous space because of the infinite number of images. Let the size of the box be $D(t) = D_\circ a(t)$. Assume that just one of the mirrors is moving to expand the box. Then the time for a roundtrip in the box is $dt = 2D(t)/c$ and the Doppler shift at the moving mirror is $d\lambda/\lambda = 2v/c$ with $v = D_\circ da/dt$. Thus we get

$$\frac{d\lambda}{\lambda} = \frac{2v}{c} = \frac{2D_\circ da}{cdt} = \frac{2D_\circ da}{2D(t)} = \frac{da}{a} \tag{226}$$

This equation has the solution $\lambda \propto a$ so $1 + z = a(t_{obs})/a(t_{em})$.

A third derivation uses the fact that $\oint pdq$ is an *adiabatic invariant*, and since the momentum p of a photon is proportional to ν, we get $\nu a(t) = \text{const}$, so $1 + z = a(t_{obs})/a(t_{em})$.

Because of the relationship between redshift z and $a(t)$ and hence t, we often speak of things happening at a given redshift instead of at a given time. This is convenient because the redshift is observable and usually has a great effect on the rates of physical processes.

27. Dynamics: $a(t)$

We can now find the differential equation that governs the time evolution of the scale factor $a(t)$. Knowing this will tell us a lot about our Universe. In this section we take a strictly Newtonian point of view, so all velocities we consider will be small compared to c. This mean that the pressure satisfies $P << \rho c^2$ and is thus insignificant when compared to the rest mass density. The way to find $a(t)$ is to consider a sphere of radius $R_\circ$ now. A comoving test particle on the surface of this sphere will have a velocity of $v = H_\circ R_\circ$. The acceleration of the test particle due to the gravity of the material inside the sphere is

$$-\frac{dv}{dt} = g = \frac{GM}{R^2} = \frac{4\pi}{3} G\rho R \tag{227}$$

which is the same as the g due to a point mass at the center of the sphere with the same mass as the total mass of the sphere. The gravitational effect of the concentric spherical shells with radii greater than $R_\circ$ is *zero*. Note that even a large pressure would not contribute to the acceleration since only pressure gradients cause forces, but we shall see later that in General Relativity, *pressure has weight* and must be included in the gravitational source term. We have a differential equation for the radius of the sphere $R(t)$ but in order to solve it we need to know how ρ varies with R.

The matter in this problem is all part of the Hubble flow, so the matter inside the sphere with $r < R$ stays inside the sphere since its radial velocity is less than the velocity of the surface of the sphere. The material outside the sphere has larger velocities than the surface of the sphere so it stays outside. This simplifies the problem to the problem of radial orbits in the gravitational field of a point mass.

The velocity at any distance can easily be found from the energy equation:

$$\frac{v^2}{2} = E_{tot} + \frac{GM}{R} \tag{228}$$

If the total energy E_{tot} is positive, the Universe will expand forever. But if the E_{tot} is negative, the Universe will stop expanding at some maximum size, and then recollapse. We can find the total energy by plugging in the velocity $v_\circ = H_\circ R_\circ$ and the density $\rho_\circ$ in the Universe now. This gives

$$E_{tot} = \frac{(H_\circ R_\circ)^2}{2} - \frac{4\pi G\rho_\circ R_\circ^2}{3} = \frac{(H_\circ R_\circ)^2}{2}\left(1 - \frac{\rho_\circ}{\rho_{crit}}\right) \tag{229}$$

with the critical density at time $t_\circ$ being $\rho_{crit} = 3H_\circ^2/(8\pi G)$. Thus if $\rho > \rho_{crit}$ the Universe will recollapse, but if $\rho \le \rho_{crit}$ the Universe will expand forever. We define the ratio of density to critical density as $\Omega = \rho/\rho_{crit}$. Thus $\Omega > 1$ means a recollapse, while $\Omega \le 1$ gives perpetual expansion. Since Ω is not a constant, we use a subscript "naught" to denote its current value, just as we do for the Hubble constant.

The value of the critical density is both large and small. In CGS units,

$$\rho_{crit} = \frac{3H_\circ^2}{8\pi G} = 1.879h^2 \times 10^{-29}\ \mathrm{gm/cc} = 10,539h^2\ \mathrm{eV/c^2/cc} \tag{230}$$

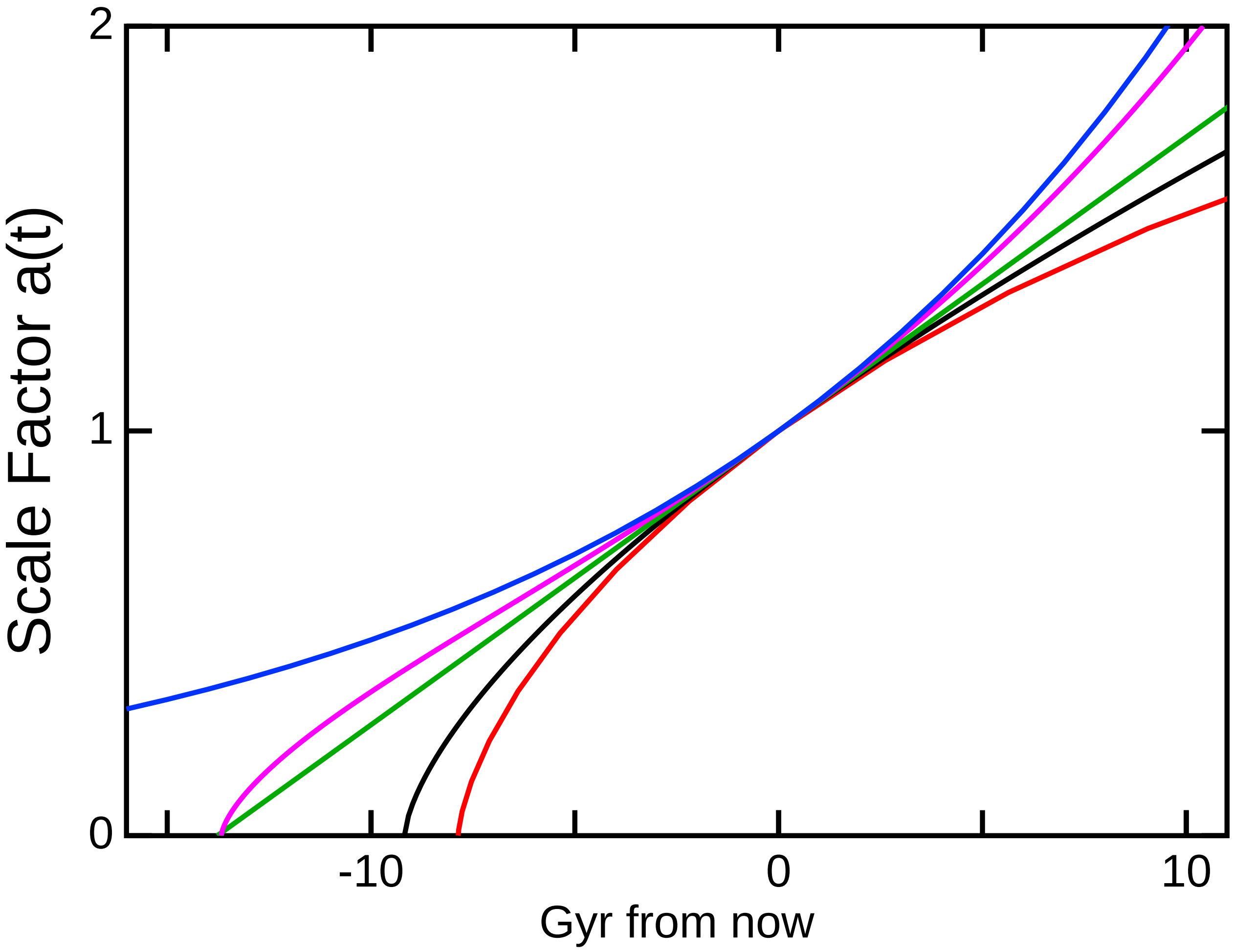

Fig. 28.— Scale factor *vs.* time for five models, all with $H_\circ = 71$. From top to bottom on the right, these are the Steady State, amodel with $\Omega_{m\circ} = 0.27$ and $\Omega_{v\circ} = 0.73$, the empty Universe, a critical density matter dominated Universe, and a closed $\Omega_\circ = 2$ model.

which is $11.2h^2$ protons/m^3. While this is certainly a small density, it appears to be much larger than the density of observed galaxies. Blanton *et al.* (2003, ApJ, 592, 819) gives a local luminosity density of $(1.6 \pm 0.2)h \times 10^8\ L_\odot/\mathrm{Mpc}^3$ in the V band. The critical density is $2.8h^2 \times 10^{11}\ M_\odot/\mathrm{Mpc}^3$ so

$$\Omega_{lum} = \frac{(M/L)/(M_\odot/L_\odot)}{1750h} \tag{231}$$

Thus the density of luminous matter seems to be much less than the critical density.

At the critical density we have the simple equation

$$\left(\frac{dR}{dt}\right)^2 = \frac{2GM}{R} \tag{232}$$

with the solution $R \propto t^{2/3}$ so the normalized scale factor is $a(t) = (t/t_\circ)^{2/3}$. On the other hand, if the density is zero, then $v = \mathrm{const}$ so $a = (t/t_\circ)$.

With this we can rewrite the energy equation Eq(228):

$$v^2 = H^2 R^2 = H_\circ^2 R_\circ^2 (1 - \Omega_\circ) + \frac{8\pi G \rho_\circ R_\circ^3}{3R} \tag{233}$$

which if we divide through by $R^2 H_\circ^2$ gives

$$\frac{H^2}{H_\circ^2} = \frac{R_\circ^2}{R^2}(1 - \Omega_\circ) + \frac{R_\circ^3}{R^3}\Omega_\circ \tag{234}$$

But remember that $R_\circ / R = a(t_\circ)/a(t) = (1+z)$ so this becomes

$$\frac{H^2}{H_\circ^2} = (1+z)^2(1-\Omega_\circ) + (1+z)^3\Omega_\circ = (1+z)^2(1+\Omega_\circ z) \tag{235}$$

One thing we can compute from this equation is the age of the Universe given $H_\circ$ and $\Omega_\circ$. H is given by $H = d\ln a/dt = -d\ln(1+z)/dt$ so

$$\frac{dt}{dz} = -\frac{1}{(1+z)H} = -\frac{1}{H_\circ (1+z)^2 \sqrt{1+\Omega_\circ z}} \tag{236}$$

The age of the Universe $t_\circ$ is obtained by integrating this from $z = \infty$ to $z = 0$ giving

$$H_\circ t_\circ = \int_0^\infty \frac{dz}{(1+z)^2 \sqrt{1+\Omega_\circ z}} \tag{237}$$

If the current density is negligible compared to the critical density, then the Universe is almost empty, and $\Omega_\circ \approx 0$. In this limit $H_\circ t_\circ = 1$. If the Universe has the critical density, $\Omega_\circ = 1$ and $H_\circ t_\circ = 2/3$. The current best observed value for the product $H_\circ t_\circ$ is only 0-1 σ higher than the $\Omega = 1$ model's prediction.

A second thing we can compute is the time variation of Ω. From Eqn(228) we have

$$2E_{tot} = v^2 - \frac{2GM}{R} = H^2 R^2 - \frac{8\pi G \rho R^2}{3} = \text{const} \tag{238}$$

If we divide this equation by $8\pi G \rho R^2/3$ we get

$$\frac{3H^2}{8\pi G \rho} - 1 = \frac{\text{const}'}{\rho R^2} = \Omega^{-1} - 1 \tag{239}$$

Let's calculate what value of Ω at $z = 10^4$ is needed to give $\Omega_\circ = 0.1$ to 2 now. The density scales like $(1+z)^3$ while the radius R scales like $(1+z)^{-1}$ so $\text{const}' = (-0.5\ldots 9)\rho_\circ R_\circ^2$ and $\Omega = 0.9991$ to 1.00005 at $z = 10^4$. This is a first clue that there must be an extraordinarily effective mechanism for setting the initial value of Ω to a value very close to unity. Unity and zero are the only fixed points for Ω, but unity is an unstable fixed point. Thus the fact that $\Omega_\circ$ is close to unity either means that a), it's just a coincidence, or b), there is some reason for Ω to be 1 exactly.

The fact that the dynamics of $a(t)$ are the same as the dynamics of a particle moving radially in the gravitational field of a point mass means that we can use Kepler's equation from orbit calculations:

$$M = E - e\sin E \tag{240}$$

where M is the *mean anomaly* which is just proportional to the time, e is the *eccentricity*, and E is the *eccentric anomaly*. The x and y coordinates are given by

$$\begin{aligned} x &= a_{SM}(e - \cos E) \\ y &= a_{SM}\sqrt{1-e^2}\sin E \end{aligned} \tag{241}$$

with semi-major axis a_{SM}. Since we want radial motion with $y = 0$, clearly we need $e = 1$. Thus we get a parametric equation for $a(t)$:

$$\begin{aligned} t &= A(E - \sin E) \\ a &= B(1 - \cos E) \end{aligned} \tag{242}$$

Clearly these equations apply to a closed Universe since a reaches a maximum of $2B$ at $E = \pi$ and then recollapses. To set the constants A and B, we need to use

$$\begin{aligned} \dot{a} &= \frac{da/dE}{dt/dE} = \left(\frac{B}{A}\right)\frac{\sin E}{1-\cos E} \\ \ddot{a} &= \left(\frac{B}{A^2}\right)\frac{\cos E(1-\cos E) - \sin^2 E}{(1-\cos E)^3} = -\left(\frac{B}{A^2}\right)\frac{1}{(1-\cos E)^2} \\ q &= \frac{-\ddot{a}a}{\dot{a}^2} = \frac{1-\cos E}{\sin^2 E} = \frac{1}{1+\cos E} \end{aligned} \tag{243}$$

Thus $E_\circ = \cos^{-1}(q_\circ^{-1} - 1)$, $B = (2 - q_\circ^{-1})^{-1}$, and $A = t_\circ/(E_\circ - \sin E_\circ)$. Note that

$$H_\circ t_\circ = \frac{\dot{a}}{a}t_\circ = \frac{(E_\circ - \sin E_\circ)\sin E_\circ}{(1-\cos E_\circ)^2} \tag{244}$$

For example, with $q_\circ = 1$ or $\Omega = 2$, we get $E_\circ = \pi/2$. Then $H_\circ t_\circ = \pi/2 - 1 = 0.5708$. The ratio of the time at the Big Crunch ($E = 2\pi$) to the current time is then

$$\frac{t_{BC}}{t_\circ} = \frac{2\pi}{\pi/2 - 1} = 11.008 \tag{245}$$

for $\Omega_\circ = 2$.

For an open Universe we change the parametric equation to

$$\begin{aligned} t &= A(\sinh E - E) \\ a &= B(\cosh E - 1) \end{aligned} \tag{246}$$

We get $E_\circ$ from $q_\circ$ using

$$q_\circ = \frac{1}{1+\cosh E_\circ} \tag{247}$$

27.1. with Pressure

General relativity says that pressure has weight, because it is a form of energy density, and $E = mc^2$. Thus

$$\ddot{R} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right)R \tag{248}$$

This basically replaces the density by the trace of the stress-energy tensor, but we will use this GR result without proof.

This equation actually leads to a very simple form for the energy equation. Consider

$$\frac{\dot{R}^2}{2} = \frac{GM}{R} + E_{tot} \tag{249}$$

If we take the time derivative of this, and remember that if the pressure is not zero the work done during expansion causes the mass to change, we get

$$\dot{R}\ddot{R} = -\frac{GM}{R^2}\dot{R} + \frac{G}{R}\frac{dM}{dR}\dot{R} \tag{250}$$

Now the work done by the expansion is $dW = PdV = P(4\pi R^2)dR$ and this causes the enclosed mass to go down by $dM = -dW/c^2$, so

$$\begin{aligned} \dot{R}\ddot{R} &= -\frac{4\pi G\rho}{3}R\dot{R} - 4\pi GPR\dot{R}/c^2 \\ &= -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right)R\dot{R} \end{aligned} \tag{251}$$

which agrees with the acceleration equation from GR. Thus the GR "pressure has weight" correction leaves the energy equation the same, so the critical density is unchanged, and the relation $(\Omega^{-1} - 1)\rho a^2 = \text{const}$ is also unchanged.

The two characteristic cases where pressure is significant are for radiation density and vacuum energy density. A gas of randomly directed photons (or any relativistic particles) has a pressure given by

$$P = \frac{\rho c^2}{3} \tag{252}$$

This has the effect of doubling the effective gravitational force. But the pressure also changes the way that density varies with redshift. The pressure does work against the expansion of the Universe, and this loss of energy reduces the density. We have $W = PdV = PV3dR/R$. This must be subtracted from the total energy $\rho c^2 V$ giving $d(\rho c^2 V) = -\rho c^2 V dR/R$. Finally we find that $\rho \propto R^{-4} \propto (1+z)^4$ for radiation. Putting this into the force equation Eq(248) gives

$$\ddot{R} = -\frac{8\pi G}{3}\frac{\rho_\circ R_\circ^4}{R^3} \tag{253}$$

which becomes an energy equation

$$v^2 = 2E_{tot} + \frac{8\pi G}{3}\frac{\rho_\circ R_\circ^4}{R^2} \tag{254}$$

Note that the "2" from doubling the effective density through the "weight" of the pressure was just the factor needed to integrate $1/R^3$, and the resulting critical density for a radiation dominated case is still $\rho_{crit} = 3H^2/(8\pi G)$. When the density is critical, $E_{tot} = 0$, and the solution has the form $R \propto t^{1/2}$.

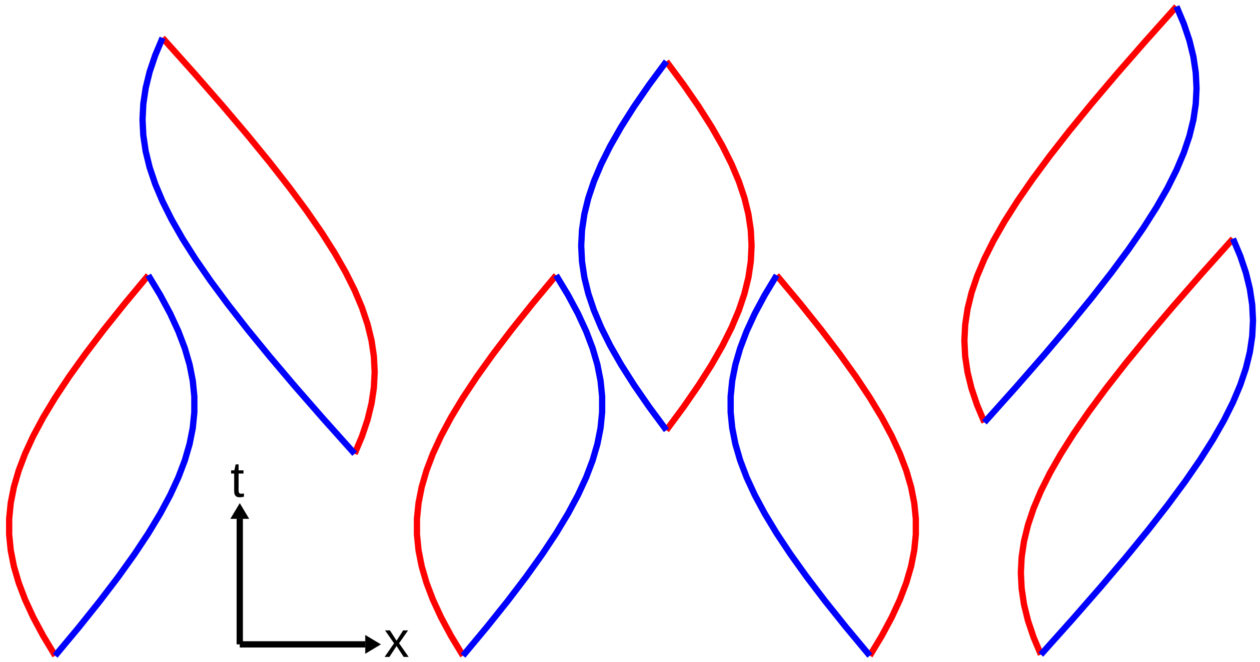

Fig. 29.— A space-time diagram of the vacuum: virtual particle-antiparticle pairs being created and annihilating throughout space. The reason why this quantum process gives a zero energy density is not clear and there may be a small residual energy density, yielding a cosmological constant.

For vacuum energy density the pressure is $P = -\rho c^2$. Naively one thinks that the vacuum has zero density but in principle it can have a density induced by quantum fluctuations creating and annihilating virtual particle pairs as shown in Figure 29. With $P = -\rho c^2$, the stress-energy tensor is a multiple of the metric, and is thus Lorentz invariant. Certainly we expect that the stress-energy tensor of the vacuum has to be Lorentz invariant, or else it would define a preferred frame. Of course we expect the stress-energy tensor of the vacuum to be zero, and the zero tensor is Lorentz invariant, but so is the metric.

Explicitly the stress energy tensor for a fluid in its rest frame is

$$T_{\mu\nu} = \begin{pmatrix} \rho c^2 & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & 0 & 0 & P \end{pmatrix} \tag{255}$$

After a Lorentz boost in the x-direction at velocity $v = \beta c$ we get

$$T'_{\mu\nu} = \begin{pmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \rho c^2 & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & 0 & 0 & P \end{pmatrix} \begin{pmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \gamma^2\rho c^2 + \gamma^2\beta^2 P & \gamma^2\beta(\rho c^2 + P) & 0 & 0 \\ \gamma^2\beta(\rho c^2 + P) & \gamma^2\beta^2\rho c^2 + \gamma^2 P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & 0 & 0 & P \end{pmatrix} \tag{256}$$

While it is definitely funny to have $\rho_{vac} \neq 0$, it would be even funnier if the stress-energy tensor of the vacuum was different in different inertial frames. So we require that $T'_{\mu\nu} = T_{\mu\nu}$. The tx component gives an equation

$$\gamma^2\beta(\rho c^2 + P) = 0 \tag{257}$$

which requires that $P = -\rho c^2$. The tt and xx components are also invariant because $\gamma^2(1-\beta^2) = 1$.

Because the pressure is negative, the work done on the expansion is negative, and the overall energy content of the vacuum grows as the Universe expands. In fact, $W = PdV = -\rho c^2 V(3dR/R)$ which changes the energy content by $d(\rho c^2 V) = 3\rho c^2 V dR/R$ so $\rho =$ const during the expansion. This is reasonable, because if the density is due to quantum fluctuation, they shouldn't care about what the Universe is doing. The pressure term in the force equation makes the force -2 times what it would have been, giving

$$\ddot{R} = \frac{8\pi G\rho}{3} R \tag{258}$$

The solutions of this equation are

$$a \propto \exp\left(\pm t\sqrt{\frac{8\pi G\rho}{3}}\right) \tag{259}$$

After a few e-foldings only the positive exponent contributes and

$$H = \sqrt{\frac{8\pi G\rho}{3}} \tag{260}$$

We see once again that the critical density is

$$\rho_{crit} = \frac{3H^2}{8\pi G}. \tag{261}$$

For this vacuum-dominated situation, $\Omega = 1$ is a stable fixed point, and this exponential growth phase offers a mechanism to set $\Omega = 1$ to great precision.

27.2. General case

It is quite easy to find $a(t)$ with a combination of the different kinds of matter. The potential energies all add linearly, so

$$v^2 = 2E_{tot} + \frac{8\pi G R^2}{3}\left(\rho_{v\circ} + \rho_{m\circ}\frac{R_\circ^3}{R^3} + \rho_{r\circ}\frac{R_\circ^4}{R^4}\right) \tag{262}$$

where $\rho_{m\circ}$ is the density of ordinary zero-pressure matter at $t_\circ$, etc. Dividing by R^2 gives

$$H^2 = H_\circ^2 \left([1 - \Omega_{v\circ} - \Omega_{m\circ} - \Omega_{r\circ}] (1+z)^2 + \Omega_{v\circ} + \Omega_{m\circ}(1+z)^3 + \Omega_{r\circ}(1+z)^4 \right) \tag{263}$$

This gives

$$H_\circ(1+z)\frac{dt}{dz} = \left([1 - \Omega_{v\circ} - \Omega_{m\circ} - \Omega_{r\circ}] (1+z)^2 + \Omega_{v\circ} + \Omega_{m\circ}(1+z)^3 + \Omega_{r\circ}(1+z)^4 \right)^{-1/2} \tag{264}$$

For *flat, vacuum dominated* models with $\Omega_{v\circ} + \Omega_{m\circ} = 1,\ \Omega_{r\circ} = 0$, the $H_\circ t_\circ$ product is

$$H_\circ t_\circ = \int_0^\infty \frac{dz}{(1+z)\sqrt{1 + \Omega_{m\circ}(3z + 3z^2 + z^3)}} \tag{265}$$

which is larger than 1 for $\Omega_{m\circ} < 0.27$. Some people feel that this is useful way to fit the high observed value of $H_\circ t_\circ$, and the current data suggesting the Universe is accelerating now ($q_\circ < 0$) also favors a non-zero vacuum energy density.

Even in this general case, the energy equation is still

$$2E_{tot} = v^2 - \frac{8\pi G \rho R^2}{3} = H^2 R^2 - \frac{8\pi G \rho R^2}{3} = \text{const} \tag{266}$$

Thus we still get

$$\Omega^{-1} - 1 = \frac{\text{const}'}{\rho a^2} \tag{267}$$

Currently $\Omega_{r\circ} \approx 10^{-4}$ but for $z > 10^4$ the radiation will dominate the density. At $z = 10^{31}$ the density will be $10^{-4}(10^{31})^4 = 10^{120}$ times higher than the current density. The scale factor is of course 10^{-31} times the current scale factor, so

$$\Omega^{-1} - 1 = 10^{-58} \left(\Omega_\circ^{-1} - 1 \right) \tag{268}$$

Thus to get $\Omega_\circ$ between 0.1 and 2 requires

$$\begin{aligned} &0.999999999999999999999999999999999999999999999999999999999991 \\ &\qquad < \Omega < \\ &1.00000000000000000000000000000000000000000000000000000000005 \end{aligned} \tag{269}$$

which is really stretching the bounds of coincidence. This is known as the *flatness-oldness* problem in cosmology. “Flatness” is used because the critical density Universe has flat spatial sections, while “oldness” is used because if the Universe had $\Omega = 1.1$ at a time 10^{-43} seconds after the Big Bang, it would recollapse in 10^{-42} seconds. Thus in order to have a Universe as old as ours with $\Omega_\circ$ close to 1 requires Ω very very close to 1 at early times.

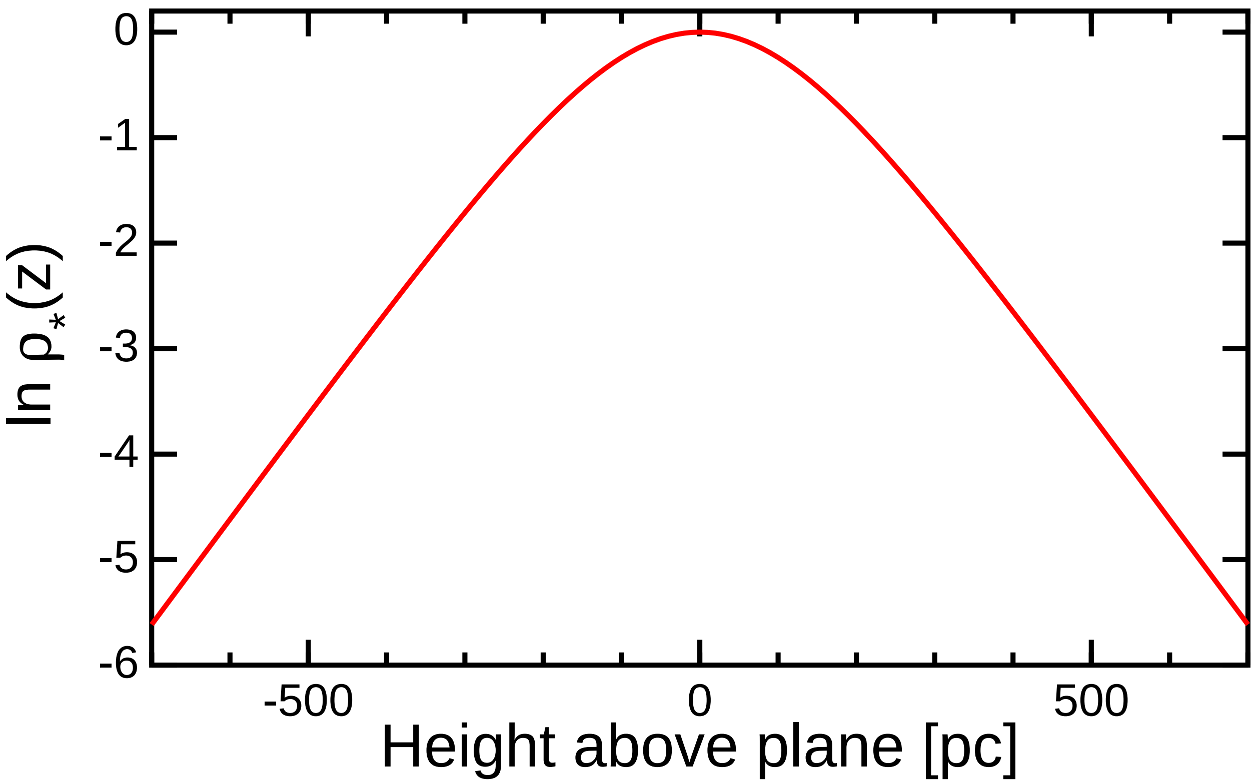

Fig. 30.— Density of tracer stars as a function of distance from the galactic plane. The curvature near the plane gives the total mass density at midplane, $\rho_T = \langle v_z^2 \rangle d^2[\ln(\rho_*)]/dz^2/(4\pi G)$, while the slope well off the plane gives the total surface mass density, $\Sigma_T = \langle v_z^2 \rangle d[\ln(\rho_*)]/dz/(2\pi G)$.

28. Density of the Universe

This material is covered in Chapter 6 of Rowan-Robinson, and also in Chapter 10 of Binney & Tremaine, "Galactic Dynamics".

The "missing mass" problem is really the missing light problem, since the mass is there.

The observed stars in the Solar neighborhood add up to $\rho = 0.11\ M_\odot/\mathrm{pc}^3$. The observed luminosity density is $4\pi j_V = 0.067\ L_\odot/\mathrm{pc}^3$ so the mass to light ratio is $M/L = 1.7$ solar units in the V band. However, the hot young and luminous stars are concentrated near the plane, so the column densities obtained by integrating through the disk are $\Sigma = 50\ M_\odot/\mathrm{pc}^2$ and $4\pi I_V = 15\ L_\odot/\mathrm{pc}^2$ so $(M/L) = 3.3$ solar units for the Milky Way disk at $R_\circ = 8.5$ kpc. These values are all obtained by counting stars and multiplying by the mass and luminosity for each type of star.

We can estimate the total mass using gravitational dynamics as well. Let the distribution function be $f(z, v_z) = f(E_z)$. This will be true if the distribution is in a steady state so $\partial f/\partial t = 0$. The motions parallel to the plane don't matter so we ignore them for this discussion. Thus

$$E_z = \frac{v_z^2}{2} + \phi(z) \tag{270}$$

If we assume a Gaussian velocity distribution then we have an isothermal distribution, so

$$f(E_z) = \exp\left[-\frac{E_z}{\langle v_z^2\rangle}\right] = \exp\left[-\frac{1}{2}\frac{v_z^2}{\langle v_z^2\rangle}\right]\exp\left[-\frac{\phi(z)}{\langle v_z^2\rangle}\right] \tag{271}$$

Near the plane, the density is approximately constant since the density is a maximum in the plane. Thus

$$\nabla^2\phi = 4\pi G\rho = \frac{\partial^2\phi}{\partial z^2} \tag{272}$$

so

$$\phi \approx 2\pi G\rho_T(0)z^2 \tag{273}$$

where ρ_T is the total mass density. We then get

$$\rho_*(z) \simeq \rho_*(0)\exp\left[-\frac{2\pi G\rho_T(0)z^2}{\langle v_z^2\rangle}\right] \tag{274}$$

where ρ_* is the density of stellar tracers with vertical velocity dispersion $\langle v_z^2\rangle$.

It is not necessary that the tracers used include all of the density. It is only necessary that the tracers have a steady-state distribution function with a Gaussian velocity dispersion. To be useful tracers stars must have known distances and be easy to find.

Thus the variation of the tracer density near the plane gives the value of $\rho_T(0)$. If we measure the density of the tracers high above the plane, then we expect that $\rho_T \approx 0$ so $\partial^2\phi/\partial z^2 = 0$ Thus ϕ will be a linear function of height above the plane:

$$\begin{aligned}
\phi &= z\frac{\partial\phi}{\partial z} + \text{const} \\
\frac{\partial\phi}{\partial z} &= \int_0^\infty 4\pi G\rho_T dz \\
&= 4\pi G\frac{1}{2}\Sigma
\end{aligned} \tag{275}$$

With the potential increasing linearly with height, the tracer density falls exponentially with height

$$\rho_*(z) \propto \exp\left[-\frac{2\pi G\Sigma_T z}{\langle v_z^2\rangle}\right] \tag{276}$$

From observations we find that

$$\begin{aligned}
\rho_T(0) &= (0.18 \pm 0.03)\ M_\odot/\text{pc}^3 \\
\Sigma_T(|z| < 700\text{ pc}) &= 75\ M_\odot/\text{pc}^2
\end{aligned} \tag{277}$$

Comparing the density determined dynamically to the density obtained by counting stars, we find a small amount of missing light in the disk:

$$\frac{\text{DARK MATTER}}{\text{KNOWN STARS}} = (50 \pm 21)\% \tag{278}$$

Now the rotation speed of the Milky Way is 220 km/sec at the solar circle and approximately flat. A Mestel disk with $\Sigma = \Sigma_\circ R_\circ / R$ has a flat rotation curve and

$$v_c^2 = 2\pi G \Sigma_\circ R_\circ \tag{279}$$

so $\Sigma_\circ = 210\ M_\odot/\mathrm{pc}^2$ for $R_\circ = 8.5$ kpc. Thus the total mass needed to produce the rotation of the Milky Way is 3 times higher than the mass of the disk deduced from the vertical distribution of tracers and 4 times higher than the observed star counts would give. Combining the excess surface density (135 $M_\odot/\mathrm{pc}^2$) and the excess local density ($< 0.07\ M_\odot/\mathrm{pc}^3$) implies that this extra mass must be more extended above and below the disk than the observed stars. It could be a spherical halo with 2/3 the mass of the Milky Way.

A spherical halo would have $\rho \propto R^{-2}$ for $v_c =$ const, with $\rho(R_\circ) = 0.0123\ M_\odot/\mathrm{pc}^3$. But 25% of this is known stars, so

$$\begin{array}{ccc} 0.01 & < \rho_{DM} < & 0.07 \pm 0.03\ M_\odot/\mathrm{pc}^2 \\ \text{SPHERICAL HALO} & & \text{FLATTENED DISK} \end{array} \tag{280}$$

in the solar neighborhood.

How far out does the $v_c =$const go? The estimated total mass of the Milky Way can come from the fastest stars in the solar neighborhood. Since these are bound, they give a minimum depth of the potential well, and hence the range out to which $M \propto R$ for $v_c =$const. This gives

$$M_{MW} = 3.8 \times 10^{11}\ M_\odot \pm 40\% \tag{281}$$

and the luminosity of the Milky Way is about $L = 1.4 \times 10^{10}\ L_\odot$. The ratio of these gives

$$\left(\frac{M}{L}\right)_{MW} = 27 \frac{M_\odot}{L_\odot} \tag{282}$$

28.1. Local Group Timing

Another way of estimating the mass of the Milky Way is through Local Group timing. The radial velocity of M31 is $v_r = -119$ km/sec relative to the galactic center. The distance of M31 is 0.73 Mpc. Let us assume that M31 and the Milky Way are on a radial orbit and that they started at the periapsis. It doesn't look like M31 and the Milky Way have collided so assume that the orbital period is about 1.5 times the age of the Universe so the next periapsis after the Big Bang is still a few Gyr in the future. Now Kepler's Third Law gives us

$$\frac{G(M+m)}{4\pi^2} = \frac{a^3}{P^2} \tag{283}$$

Also, the Kepler equation gives us a parametric solution for the orbit

$$\begin{aligned} R &= a(1 - e\cos\eta) \\ t &= \sqrt{\frac{a^3}{G(M+m)}}(\eta - e\sin\eta) \end{aligned} \tag{284}$$

where η is the eccentric anomaly. Let $e = 1$ because the initial separation of M31 and the Milky Way was much smaller than the current distance. Then

$$\frac{dR}{d\eta} = a \sin\eta \tag{285}$$

and

$$\frac{dt}{d\eta} = \sqrt{\frac{a^3}{G(M+m)}}(1-\cos\eta) = \sqrt{\frac{a}{G(M+m)}}R \tag{286}$$

Thus

$$v_r = \frac{dR}{dt} = \frac{a\sin\eta}{R\sqrt{a/(G(M+m))}} \tag{287}$$

Now t is probably between 2/3 and 1 times $1/H_\circ$. Thus $R/t = (1.25 \pm 0.25)RH_\circ = 77$ km/sec for $H_\circ = 70$ and $\Omega_\circ = 1$. Combining our equations gives

$$\frac{R/t}{dR/dt} = \frac{(1-\cos\eta)^2}{\sin\eta(\eta - \sin\eta)} = -0.64 \tag{288}$$

which gives $\eta = 4.08$ radians. Then $a = 0.73/(1-\cos\eta) = 0.458$ Mpc and the period is 14.66 Gyr. Finally

$$M + m = 4 \times 10^{12}\ M_\odot \tag{289}$$

The M/L ratio for M31 and the Milky Way combined is 93.

Note that as $H_\circ$ goes down, the period goes up, and since the size of the orbit is not derived from $H_\circ$, the mass goes down. Thus $M_{M31} + M_{MW} \propto H_\circ^{0.97}$ and thus is essentially $M/L \propto h$ since the luminosity is determined without using Hubble distances. Thus we get

$$\frac{M_{M31} + M_{MW}}{L_{M31} + L_{MW}} = 93 \left(\frac{H_\circ}{70 \text{ km/sec/Mpc}}\right)^{0.97} \tag{290}$$

28.2. External Galaxies

Consider a spiral galaxy with redshift z so its mean radial velocity is cz, but with a variation due to rotation of $\pm v_{rot}$. The mass inside radius R is

$$M(<R) = \frac{Rv_{rot}^2}{G} \tag{291}$$

but the radius is determined from an observed angular radius θ and a distance $cz/H_\circ$ computed from using the Hubble constant. Thus the derived mass is $M \propto H^{-1}$. The luminosity is the observed flux F times the distance squared, $L = 4\pi(cz/H_\circ)^2 F$, so $L \propto h^{-2}$. Thus $M/L \propto h$.

28.3. Virial Mass of Cluster

The mass determined from the virial theorem is

$$M = \frac{3\sigma(v_r)^2 R_e}{G} \tag{292}$$

so the mass goes like $M \propto h^{-1}$. The luminosity goes like $L \propto h^{-2}$, so

$$\frac{M}{L} \approx 500h\frac{M_\odot}{L_\odot} \tag{293}$$

28.4. Closing the Universe

The mass-to-light ratio required to close the Universe is given by the critical density divided by the luminosity density. The luminosity density is based on counting n galaxies with a given flux F, computing a distance limit D and then finding

$$4\pi j_\nu \propto \frac{n4\pi D^2 F}{D^3} \propto D^{-1} \propto h \tag{294}$$

The luminosity density determined by Davis & Huchra (1982) and Kirshner *et al.* (1983) is of $(1.7 \pm 0.6)h \times 10^8\ L_\odot/\mathrm{Mpc}^3$ in the V band. Since the critical density is $2.8h^2 \times 10^{11}\ M_\odot/\mathrm{Mpc}^3$, the critical mass-to-light ratio is $(1600^{+900}_{-400})h$ solar units. This is a few times greater than the cluster mass-to-light ratio, suggesting that the matter density in the Universe is less than critical. On the other hand this requires an incredible coincidence (the flatness-oldness problem), and also we see that M/L increases with scale size from the solar neighborhood to galaxies to the Local Group to clusters. Perhaps another factor of 3 can be found giving $\Omega = 1$.

One possible cause of this trend would be that the dark matter is more spread out than the stars (stars are made from baryons). The dark matter would have to be smoothed out to about $10/h$ Mpc scale. If the dark matter particles have some initial momentum they will drift away from the dense regions of the primordial perturbations, producing this smoothing effect. We can calculate the amount of the this drift if we know the way that peculiar velocities evolve with time. Consider a particle with momentum p bouncing back and forth between comoving mirrored walls of a small piece of the Universe. The action $J = \oint p dq$ is an adiabatic invariant so it will be conserved during the expansion of the Universe. Thus the peculiar momentum p varies like a^{-1} since dq in the cell grows with the Universe. When applied to photons this rule gives us the redshift law again, since for photons the frequency is proportional to the momentum. Now the velocity of a particle with momentum p and mass m is

$$v = \frac{pc^2}{E} = \frac{p/m}{\sqrt{1 + (p/mc)^2}} \tag{295}$$

The distance traveled is vdt but the *comoving distance* is $(1+z)vdt$. Thus the total comoving distance traveled by a particle with $p = (1+z)p_\circ$ is

$$X = \int \frac{(1+z)^2(p_\circ/mc)cdt}{\sqrt{1 + ((1+z)p_\circ/mc)^2}} = \int \frac{(1+z)^2(p_\circ/mc)cdz}{H_\circ(1+z)^2\sqrt{(1+\Omega_{m\circ}z)(1+((1+z)p_\circ/mc)^2)}} \tag{296}$$

for a matter dominated model. If $p_\circ >> mc$ then the particle is always traveling at the speed of light and we get the distance to the *horizon*:

$$X_h = \int (1+z)cdt \tag{297}$$

which is $2c/H_\circ$ for a matter dominated model with $\Omega_{m\circ} = 1$. If we are looking for $X = 10/h$ Mpc then we need $p_\circ << mc$ giving

$$X \approx \frac{2c}{H_\circ\sqrt{\Omega_{m\circ}}}\sqrt{\frac{p_\circ}{mc}} \tag{298}$$

This will be $10/h$ Mpc if $p_\circ/mc = 3\times10^{-6}$. If $p_\circ$ is thermally generated then it should be $\mathcal{O}(3kT_\circ/c)$ so

$$mc^2 \approx 10^6 kT_\circ \approx 200 \text{ eV} \tag{299}$$

This is one motivation for the *Hot Dark Matter* model which has dark matter particles which were relativistic during the early history of the Universe. The leading candidate for HDM particles is a neutrino species with a mass of several eV.

Particles which are much more massive than 1 keV are known as *Cold Dark Matter* particles. Their thermal velocities are always negligible. Candidates for CDM particles are WIMPs (Weakly Interacting Massive Particles), axions, and MACHOs (MAssive Compact Halo Objects). Axions are hypothetical particles suggested by Peccei & Quinn, and could be CDM if they have a mass of about 10^{-5} eV. This very small mass would make axions HDM if they had a thermal distribution in equilibrium with the 2.73 K photons, but they are assumed to form with zero thermal velocities. The lightest neutral supersymmetric particle could be a WIMP. In supersymmetry, particles have supersymmetric partners: photon & photino, electron & selectron, Z boson & zino, and of course the W boson & the wino. The lightest neutral supersymmetric particle or neutralino should be stable and weakly interacting, with a mass of $\mathcal{O}(10^2)$ Gev and is a candidate for a WIMP. MACHOs are seen by microlensing experiments, with masses of a fraction of a solar mass. They could be white dwarf or brown dwarf stars, or they could be black holes. If made in the first few seconds after the Big Bang, then primordial black holes could be CDM.

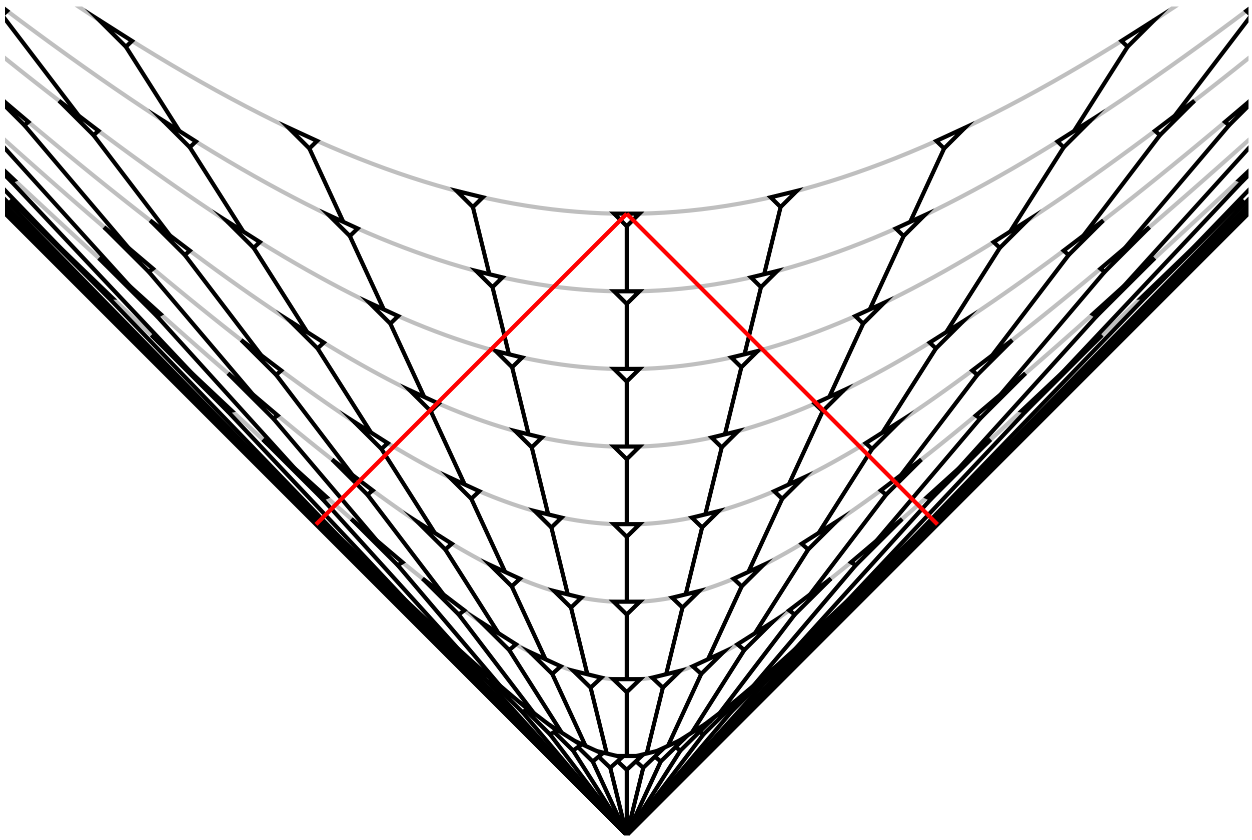

Fig. 31.— A space time diagram in the coordinates of special relativity showing the comoving galaxies of an $\Omega = 0$ cosmological model. World lines have been decorated with little light cones, and the entire past light cone of an event on the central world line is shown. The light gray hyperbolae show curves of constant proper time since the Big Bang for comoving observers. This is the cosmic time variable.

29. Distant Objects

The appearance of distant objects (flux and angular size) can not be calculated in a mirrored box. In general we need to use a metric calculated using GR, but for the simple case of $\Omega = 0$, the Universe is empty and there are no gravitational forces, so special relativity can be used. In special relativity the metric is

$$ds^2 = c^2dt^2 - (dx^2 + dy^2 + dz^2) = c^2dt^2 - r^2(d\delta^2 + \cos^2\delta d\alpha^2) \tag{300}$$

The worldlines of comoving galaxies all have to intersect at some event which we identify as the Big Bang. Let's choose this event as the zero point for our coordinates. Without gravity all the comoving galaxies move on straight lines so for any particular galaxy B we have

$$\begin{aligned} x_B(t) &= a(t)X_B \\ y_B(t) &= a(t)Y_B \\ z_B(t) &= a(t)Z_B \end{aligned} \tag{301}$$

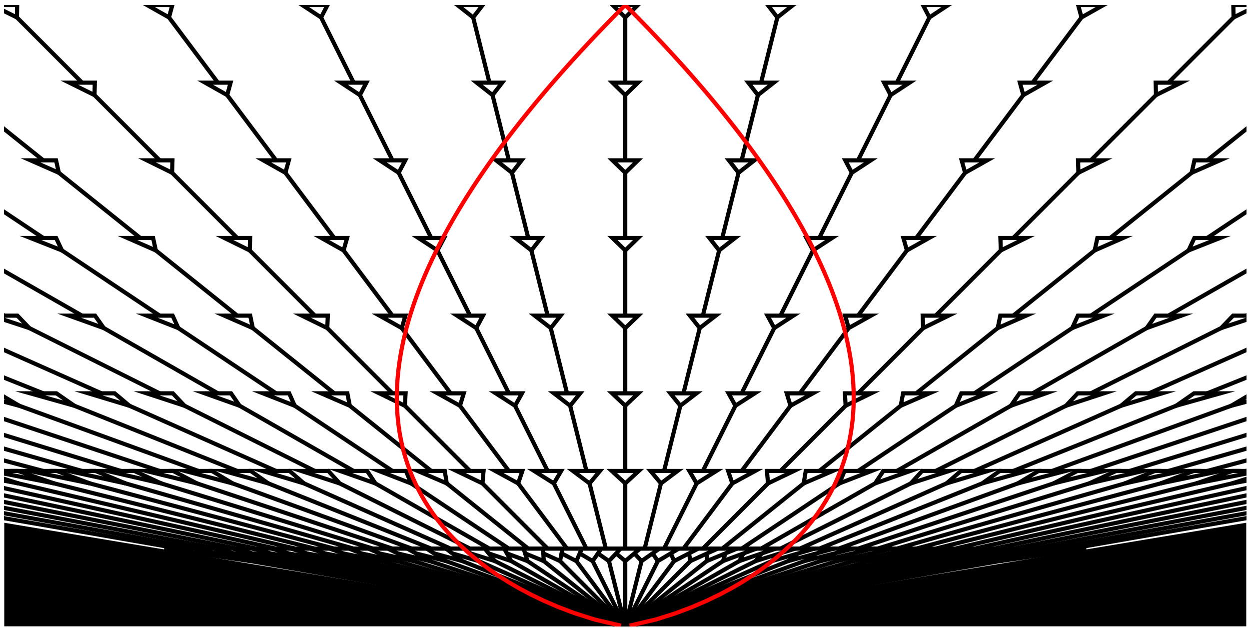

Fig. 32.— The $\Omega = 0$ cosmological model plotted using the cosmic time τ and radial proper distance $c\tau\psi$ as variables.

with $a(t) = t/t_\circ$. However, the special relativistic time variable t can not be used as the cosmic time variable, because objects at the same t have different proper times since the Big Bang for comoving obervers. The events that do have the same proper time τ since the Big Bang for comoving observers lie on a hyperbola defined by $\tau^2 = t^2 - r^2/c^2$.

Thus a constant τ hyperbola has to be flattened into a plane. This immediately gives expansion velocities greater than c in distant regions of the Universe. This reinforces the point made earlier that the Hubble law velocities $v = HD$ can be larger than the speed of light. The scale factor becomes $a(\tau) = \tau/t_\circ$. Thus the Hubble constant is given by $H_\circ = a^{-1}da/dt = t_\circ^{-1}$ which agrees with our earlier calculation.

29.1. Angular size distance

Now let us consider an observation we make of an object at special relativistic coordinates $x = d_A$ and $t = t_\circ - d_A/c$. This object is clearly on our past light cone, since us-now is the event at $x = 0$ and $t = t_\circ$. If the object has a dimension R perpendicular to the line-of-sight, then we know that it will subtend an angle $\Delta\theta = R/d_A$ because the SR coordinates describe a simple geometry. This distance defined by $d_A = R/\Delta\theta$ is known as the *angular size distance*. The redshift of the object at $x = d_A$ can be found several different ways, but $cz = H_\circ d_A$ *is not one of them*. The first way uses the rule that $1+z = a(\tau_{em})^{-1}$. The cosmic time $\tau = \sqrt{(t_\circ - d_A/c)^2 - d_A^2/c^2}$ so

$$1 + z = \frac{t_\circ}{\sqrt{(t_\circ - d_A/c)^2 - d_A^2/c^2}} = \frac{1}{\sqrt{1 - 2d_A/ct_\circ}} \tag{302}$$

Solving this equation gives

$$d_A = ct_\circ \frac{z(1+z/2)}{(1+z)^2} \tag{303}$$

The second way to find z at d_A is to look at the SR velocity $v = d_A/(t_\circ - d_A/c)$ and compute the SR Doppler shift

$$1 + z = \sqrt{\frac{1+v/c}{1-v/c}} = \sqrt{\frac{ct_\circ}{ct_\circ - 2d_A}} \tag{304}$$

which clearly gives the same result.

29.2. Luminosity distance

The flux from an object subtending an angle $\Delta\theta$ can be found using the fact that the number of photons per mode is not changed during the expansion of the Universe. For a blackbody the number of photons per mode is $(\exp(h\nu/kT) - 1)^{-1}$. For the an object at redshift z, the photons emitted at ν_{em} arrive with frequency $\nu_{obs} = \nu_{em}/(1+z)$. Since the number of photons per mode stays the same, a blackbody emitting at a temperature T_{em} will appear to be a blackbody of temperature $T_{obs} = T_{em}/(1+z)$. Thus an object with luminosity $L = 4\pi R^2 \sigma_{SB} T_{em}^4$ has a flux $F = (\Delta\theta)^2 \sigma_{SB} T_{obs}^4$. The *luminosity distance* d_L is defined by

$$F = \frac{L}{4\pi d_L^2} \tag{305}$$

so

$$d_L = \sqrt{\frac{L}{4\pi F}} = \sqrt{\left(\frac{R}{\Delta\theta}\right)^2 \left(\frac{T_{em}}{T_{obs}}\right)^4} = d_A(1+z)^2 \tag{306}$$

29.3. Radial Distance

The actual distance that should go into the Hubble Law can be measured by comoving observers using radar pulses sent just before and received just after the cosmic time τ. In order to compute this distance, let's use the hyperbolic sine and cosine since the slice of constant proper time since the Big Bang is a hyperbola in special relativistic coordinates. So let

$$\begin{aligned} t &= \tau \cosh\psi \\ x &= c\tau \sinh\psi \end{aligned} \tag{307}$$

where ψ is the hyperbolic "angle". The distance at constant τ between ψ and $\psi + d\psi$ is given by $-ds^2 = dx^2 - c^2dt^2$ with $dx = c\tau\cosh\psi d\psi$ and $dt = \tau\sinh\psi d\psi$ so $-ds^2 = (c\tau)^2(\cosh^2\psi - \sinh^2\psi)d\psi^2$. Hence the distance is $c\tau d\psi$ and the total radial distance is $c\tau\psi$. But the circumference of a circle is given by $2\pi x = 2\pi c\tau\sinh\psi$. Since $\sinh\psi > \psi$, the spatial sections of the zero density Universe are negatively curved. The space-time diagram in Figure 32 uses $c\tau$ for time and $c\tau\psi$ for space, so uniformly spaced lines on the diagram are all separated by the same distance in actuality.

29.4. Friedmann-Robertson-Walker metric

We can write the entire metric in cosmological variables now:

$$ds^2 = c^2d\tau^2 - (c\tau)^2\left(d\psi^2 + \sinh^2\psi(d\delta^2 + \cos^2\delta d\alpha^2)\right) \tag{308}$$

This is often rewritten using $r = \sinh\psi$ as the radial variable. Since $dr = \cosh\psi d\psi$ and $\cosh\phi = \sqrt{1+r^2}$, this gives

$$ds^2 = c^2d\tau^2 - (c\tau)^2\left(\frac{dr^2}{1+r^2} + r^2(d\delta^2 + \cos^2\delta d\alpha^2)\right) \tag{309}$$

This is a Friedmann-Robertson-Walker (FRW) metric with negative curvature. It describes one of the three 3D spaces which are isotropic and homogeneous. The other 2 are Euclidean space and the hyperspherical 3D surface of a 4D ball. Usually t is used for the cosmic time variable instead of τ.

A general form for the metric of an isotropic and homogeneous cosmology is

$$ds^2 = c^2dt^2 - a(t)^2R_\circ^2\left(\frac{dr^2}{1-kr^2} + r^2(d\delta^2 + \cos^2\delta d\alpha^2)\right) \tag{310}$$

where $k = -1, 0$ or 1 for the negatively curved (hyperboloidal), flat or positively curved (hyperspherical) cases, $a(t)$ is computed from the $H(t)$ that we have found, and finally the radius of curvature $R_\circ$ of the Universe is given by

$$R_\circ = \frac{c/H_\circ}{\sqrt{|1 - \Omega_{v\circ} - \Omega_{m\circ} - \Omega_{r\circ}|}} \tag{311}$$

If $1 - \Omega_{v\circ} - \Omega_{m\circ} - \Omega_{r\circ} > 0$ then $k = -1$, while if $1 - \Omega_{v\circ} - \Omega_{m\circ} - \Omega_{r\circ} < 0$ then $k = +1$. We take this General Relativity result without proof.

Often the combination $a(t)R_\circ$ is changed to $a(t)$. This puts the dimensions of distance onto $a(t)$. In this class, I will stay with a dimensionless a and $a(t_\circ) = 1$.

It is possible for a closed Universe with $k = +1$ to expand forever if the cosmological constant $\Omega_{v\circ}$ is large enough. The usual association of closed Universes with recollapse works when the vacuum energy density is zero.

30. General formula for angular sizes

We have worked out the angular size distance versus redshift for the empty $\Omega = 0$ Universe. We have also worked the general FRW metric which we can use to find the general answer for angular size versus redshift. The angular size distance is obviously given by $\sqrt{-ds^2/d\delta^2}$ with $dt, d\alpha, dr = 0$. This must be evaluated at the emission time t_{em}, so $D_A = a(t_{em})R_\circ r$. But we have to find r by solving a differential equation to follow the past light cone:

$$\frac{a(t)R_\circ dr}{\sqrt{1-kr^2}} = -cdt \tag{312}$$

so

$$R_\circ \int \frac{dr}{\sqrt{1-kr^2}} = \int_{t_{em}}^{t_\circ} (1+z)cdt \tag{313}$$

This can be viewed as follows: light always travels at c, so the distance covered in dt is cdt. But this distance expands by a factor $(1+z)$ between time t and now, and since the comoving distance is measured now, this $(1+z)$ factor is needed. The integral on the LHS of Eqn(313) is either $\sin^{-1} r$, r, or $\sinh^{-1} r$ depending on whether $k = +1$, 0 or -1.

30.1. Critical Density Universe

For $\Omega_{m\circ} = 1$, $\Omega_{v\circ} = \Omega_{r\circ} = 0$, the integral of

$$\int_{t_{em}}^{t_\circ} (1+z)cdt = \frac{c}{H_\circ}\int (1+z)^{-3/2}dz = 2\frac{c}{H_\circ}\left(1-(1+z)^{-1/2}\right) \tag{314}$$

Also, the $R_\circ$'s cancel out and $k = 0$. Thus the angular size distance for this model is

$$D_A = 2\frac{c}{H_\circ}\left((1+z)^{-1} - (1+z)^{-3/2}\right) \tag{315}$$

Therefore the luminosity distance for the $\Omega = 1$ model is

$$D_L = 2\frac{c}{H_\circ}\left(1+z-\sqrt{1+z}\right) = \frac{cz}{H_\circ}\left(1+\frac{z}{4}+\ldots\right) \tag{316}$$

30.2. Steady State Universe

Another easy special case is the Steady State Universe which is a critical density vacuum-dominated model. Since H is a constant, $a(t) = \exp(H(t-t_\circ))$. Then

$$\int_{t_{em}}^{t_\circ} (1+z)cdt = \int_{t_{em}}^{t_\circ} \exp(H(t_\circ - t))cdt = \frac{cz}{H} \tag{317}$$

The Steady State model has $k = 0$ since if $k = \pm 1$, then $R_\circ$ is measurable, but the radius of curvature grows with the expansion of the Universe, and hence one doesn't have a Steady State. Thus $k = 0$ and $R_\circ$ cancels out. This gives

$$D_A = \frac{c}{H}\frac{z}{1+z} \tag{318}$$

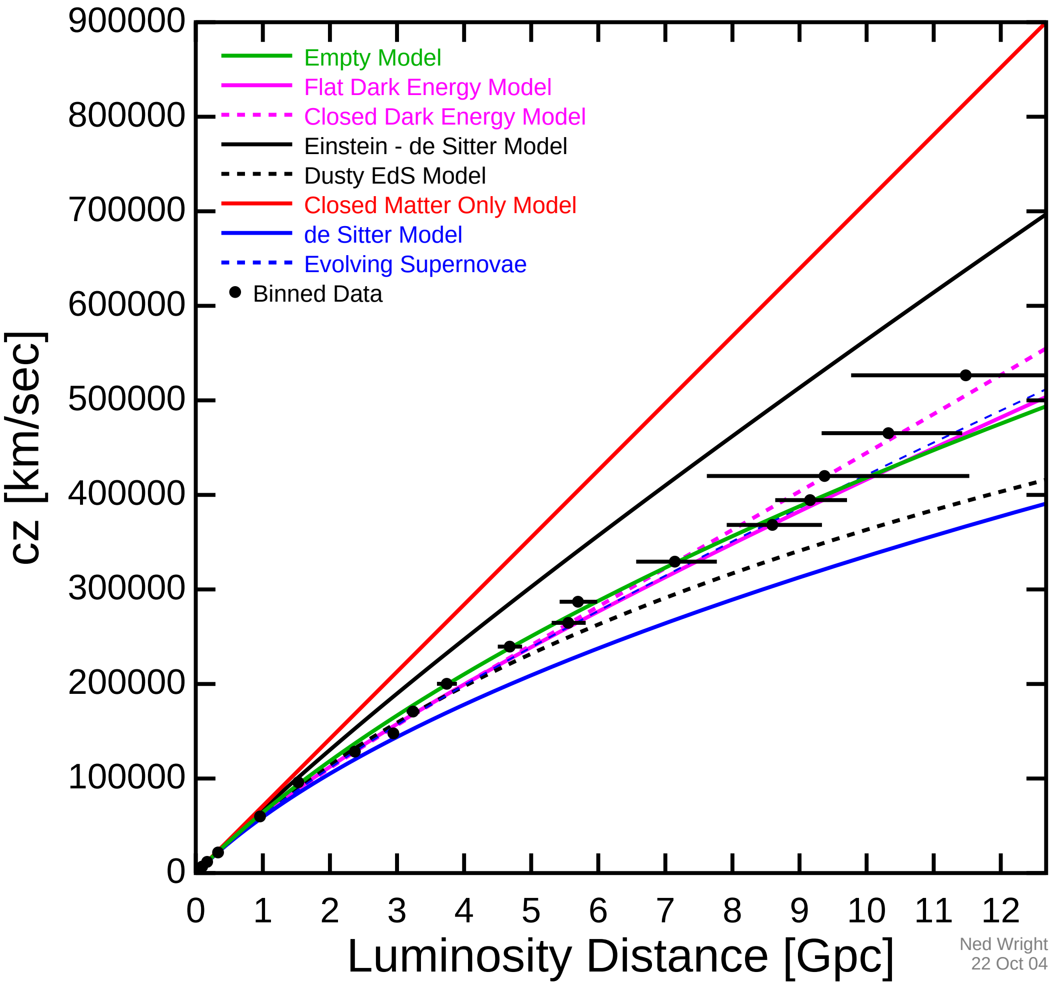

Fig. 33.— Luminosity distance *vs.* redshift for high redshift Type Ia supernovae.

and the luminosity distance is

$$D_L = \frac{c}{H} z(1+z) \tag{319}$$

As a final special case consider the $\Omega_{r\circ} = 1$, $\Omega_{m\circ} = \Omega_{v\circ} = 0$ critical density radiation dominated Universe. Since $\Omega = 1$, $R_\circ \rightarrow \infty$ but it cancels out in determining D_A. Since $a(t) \propto t^{1/2}$, $(1+z) \propto t^{-1/2}$ and

$$c\frac{dt}{dz} = -\frac{c}{H_\circ}\frac{1}{(1+z)^3} \tag{320}$$

Thus

$$\int (1+z)c\,dt = \frac{c}{H_\circ}\left(1 - \frac{1}{1+z}\right) \tag{321}$$

and

$$D_A = a(t_{em}) \int (1+z)c\,dt = \frac{c}{H_\circ}\frac{z}{(1+z)^2}. \tag{322}$$

Finally $D_L = cz/H_\circ$ exactly.

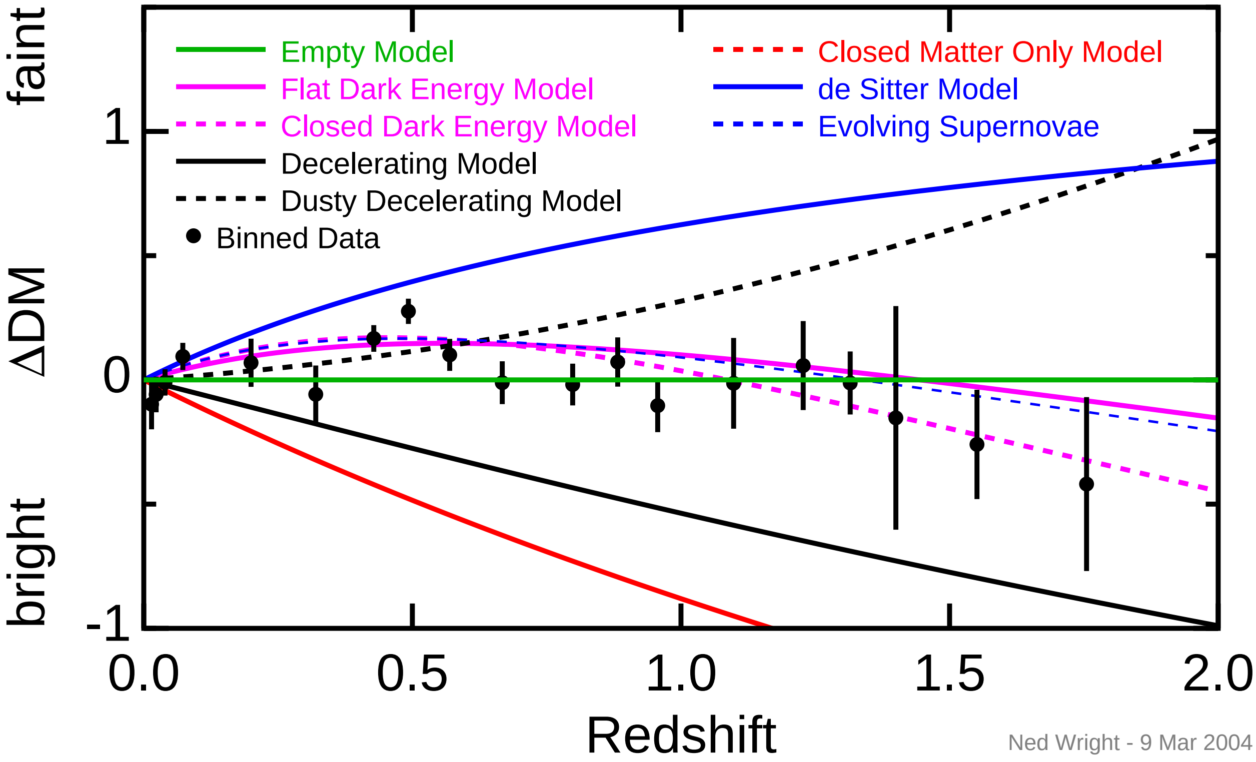

Fig. 34.— Distance modulus relative to an $\Omega = 0$ model *vs.* redshift for high redshift Type Ia supernovae.

For the more general case we note that $\sin r$ and $\sinh r$ differ from r only in the cubic term. However, the integral on the RHS of Eqn(313) depends on $a(t)$ and differs from a linear approximation $cz/H_\circ R_\circ$ in the second order. The second order deviation of the angular size distance away from the linear approximation $D_A = cz/H_\circ$ thus depends only on the time history of the scale factor $a(t)$. We can write

$$a(t_\circ + \Delta t) = a(t_\circ)\left(1 + H_\circ \Delta t - \frac{1}{2} q_\circ (H_\circ \Delta t)^2 + \ldots\right) \tag{323}$$

which defines the *deceleration parameter*

$$q_\circ = -\frac{a\ddot{a}}{\dot{a}^2} \tag{324}$$

The force equation $\ddot{a} = -(4\pi G/3)(\rho + 3P/c^2)a$ from our previous analysis gives us

$$q_\circ = -\frac{a\ddot{a}}{\dot{a}^2} = \frac{4\pi G}{3H_\circ^2}\left(\rho + \frac{3P}{c^2}\right) = \frac{\Omega_{m\circ}}{2} + \Omega_{r\circ} - \Omega_{v\circ} \tag{325}$$

Thus the $\Omega = 0$ empty Universe has $q_\circ = 0$, the critical density Universe has $q_\circ = 0.5$, and the Steady State model has $q_\circ = -1$.

Given the expansion for $a(t)$ we find

$$d\left(\frac{a(t)}{a(t_\circ)}\right) = d\left(\frac{1}{1+z}\right) = \frac{-dz}{(1+z)^2} = H_\circ(1 - q_\circ(H_\circ \Delta t) + \ldots)dt \tag{326}$$

Since $H_\circ \Delta t = -z + \mathcal{O}(z^2)$ we get

$$\frac{dt}{dz} = \frac{-H_\circ^{-1}}{(1+z)^2(1+q_\circ z+\ldots)} \tag{327}$$

The integral on the RHS of Eqn(313) is then given by

$$\int_{t_{em}}^{t_\circ} (1+z)cdt = \frac{c}{H_\circ}\int_0^z \frac{dz}{(1+z)(1+q_\circ z)} = \frac{cz}{H_\circ}\left(1+\frac{z}{2}[-1-q_\circ]+\ldots\right) \tag{328}$$

Eqn(313) then gives

$$D_A = R_\circ(r+\mathcal{O}(r^3))/(1+z) = \frac{cz}{H_\circ}\left(1+\frac{z}{2}[-3-q_\circ]+\ldots\right) \tag{329}$$

Finally

$$D_L = D_A(1+z)^2 = \frac{cz}{H_\circ}\left(1+\frac{z}{2}[1-q_\circ]+\ldots\right) \tag{330}$$

This is consistent with our four special cases:
$D_L = (cz/H_\circ)(1+z)$ for $q_\circ = -1$,
$D_L = (cz/H_\circ)(1+z/2)$ for $q_\circ = 0$,
$D_L = (cz/H_\circ)(1+z/4+\ldots)$ for $q_\circ = 0.5$, and
$D_L = (cz/H_\circ)$ for $q_\circ = 1$.

Recent work on distant Type Ia SNe by Perlmutter *et al.* (1998) and Garnavich *et al.* (1998) (Figure 33) shows that $q_\circ < 0$, which favors a Universe dominated by a cosmological constant.

Finally, a useful formula found by Mattig for matter-dominated models with $\Omega_{r\circ} = \Omega_{v\circ} = 0$ is

$$D_L = \frac{cz}{H_\circ}\left[\frac{4-2\Omega+2z}{(1+\sqrt{1+\Omega z})(1-\Omega+\sqrt{1+\Omega z})}\right] \tag{331}$$

Note that for $\Omega = 2$, $q_\circ = 1$, we have

$$D_L = \frac{cz}{H_\circ} \tag{332}$$

exactly. This particular case simplifies because for $\Omega = 2$, the radius of curvature of the Universe is $R_\circ = c/H_\circ$.

30.3. K correction, Evolution

The formula $F = L/(4\pi D_L^2)$ applies to *bolometric* or total fluxes and luminosities. When converting it to band fluxes such as V magnitudes or F_ν, we need to do two things. The first is to properly transform the frequency, so we compute the flux F_ν from the luminosity $L_{\nu(1+z)}$. The second is to properly transform the bandwidth of the observation into the bandwidth of the emission. This is trivial if we use the flux per octave and luminosity per octave, since the fractional bandwidth or number of octaves doesn't change with redshift. Thus

$$\nu F_\nu = \frac{\nu(1+z)L_{\nu(1+z)}}{4\pi D_L^2} \tag{333}$$

Thus the flux vs redshift law for the flux per octave is the same as the one for bolometric flux. From this we easily get

$$
\begin{aligned}
F_\nu &= \frac{(1+z)L_{\nu(1+z)}}{4\pi D_L^2} \\
F_\lambda &= \frac{L_{\lambda/(1+z)}}{4\pi D_L^2(1+z)}
\end{aligned}
\tag{334}
$$

The difference between $\nu(1+z)L_{\nu(1+z)}$ and νL_ν leads to a correction known as the K-correction. Expressed as magnitudes to be added to the apparent magnitude, the K-correction is

$$
K(\nu, z) = -2.5\log\left(\frac{\nu(1+z)L_{\nu(1+z)}}{\nu L_\nu}\right) \tag{335}
$$

If working with only V band data, we can write

$$
V = M_V + 5\log\left(\frac{D_L(z)}{10\ \mathrm{pc}}\right) + K(\nu_V, z) \tag{336}
$$

Obviously observations or models are needed to predict how the luminosity depends on frequency away from the V band.

When all observations were made in the photographic blue, the K-corrections could be quite large for galaxies since the flux drops precipitously at the 400 nm edge due to the H and K lines of ionized calcium plus the Balmer edge in hydrogen. But with modern multiband data, we can usually use the R or I band to observe galaxies with $z \approx 0.5$, and compare these fluxes to B or V band data on nearby galaxies. This reduces the magnitude and uncertainty in the K-correction.

A more serious difficulty is the possibility of evolution. A galaxy at $z = 0.5$ is approximately 5 Gyr younger than the nearby galaxies we use for calibration. If new stars are not being formed, the brighter more massive stars will reach the end of their main sequence life, become red giants and then fade away. As a result, galaxies get fainter as they get older, and this leads to a correction to the flux-redshift law that has a large uncertainty. Evolution introduces a correction that is proportional to z just like the $q_\circ$ term in D_L. This has prevented the use of galaxies to determine $q_\circ$, and increases the utility of the distant Type Ia SNe work. Type Ia SNe are thought to be due to white dwarfs in binaries slowly accreting material until they pass the Chandrasekhar limit and explode. Since the Chandrasekhar limit doesn't evolve with time, the properties of Type Ia SNe should not depend on z. However, the peak brightness of a Type Ia SNe is not a constant, but depends on the decay rate after the peak. Faster decaying Type Ia SNe are fainter, while slower decaying Type Ia SNe are brighter. The cause of this correlation is not understood, and it thus might depend on redshift. The typical mass of a white dwarf does evolve with time, and was higher in the past, so there is still the possibility of a systematic error in the Type Ia SNe work.

31. General formula for D_{ltt}, D_{now}, D_A and D_L

The following formulae are used in my cosmology calculator on the World Wide Web. The metric is given by

$$ds^2 = c^2dt^2 - a(t)^2R_\circ^2(d\psi^2 + S^2(\psi)[d\theta^2 + \sin^2\theta d\phi^2]) \tag{337}$$

where $S(x)$ is $\sinh(x)$ if $\Omega_{tot} < 1$ and $\sin(x)$ for $\Omega_{tot} > 1$. $R_\circ = (c/H_\circ)/\sqrt{|1-\Omega_{tot}|}$. The past light cone is given by $cdt = a(t)R_\circ d\psi$ so

$$D_{now} = R_\circ\psi = \int \frac{cdt}{a} = \int_{1/(1+z)}^{1} \frac{cda}{a\dot{a}} \tag{338}$$

and of course the light travel time distance is given by

$$D_{ltt} = \int cdt = \int_{1/(1+z)}^{1} \frac{cda}{\dot{a}} \tag{339}$$

We can write $\dot{a}$ as $H_\circ\sqrt{X}$ with

$$X(a) = \Omega_{m\circ}/a + \Omega_{r\circ}/a^2 + \Omega_{v\circ}a^2 + (1-\Omega_{tot}) \tag{340}$$

Let us define

$$Z = \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{X}} \tag{341}$$

so $D_{now} = (cZ/H_\circ)$ and $D_{ltt} = (c/H_\circ)\int_{1/(1+z)}^{1} da/\sqrt{X}$. Then

$$\begin{aligned} D_A &= \frac{c}{H_\circ}\frac{S(\sqrt{|1-\Omega_{tot}|}Z)}{(1+z)\sqrt{|1-\Omega_{tot}|}} \\ &= \frac{D_{now}}{(1+z)}\left(1+\frac{1}{6}(1-\Omega_{tot})Z^2 + \frac{1}{120}(1-\Omega_{tot})^2Z^4 + \ldots\right) \end{aligned} \tag{342}$$

We can define a function $J(x)$ given by

$$J(x) = \begin{cases} \frac{\sin\sqrt{-x}}{\sqrt{-x}}, & x < 0; \\ \frac{\sinh\sqrt{x}}{\sqrt{x}}, & x > 0; \\ 1 + x/6 + x^2/120 + \ldots + x^n/(2n+1)! + \ldots, & x \approx 0. \end{cases} \tag{343}$$

Then

$$\begin{aligned} D_A &= \frac{cZ(z)}{H_\circ}\frac{J([1-\Omega_{tot}]Z^2)}{1+z} \\ D_L &= (1+z)^2 D_A \end{aligned} \tag{344}$$

Applying these formulae to four simples cases and one hard case gives results consistent with our earlier calculations:

- $\Omega = 0$

$$
\begin{aligned}
Z &= \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{X}} = \int_{1/(1+z)}^{1} \frac{da}{a} = \ln(1+z) \\
D_{now} &= \frac{cZ}{H_\circ} = \frac{c}{H_\circ}\ln(1+z) \\
D_{ltt} &= \frac{c}{H_\circ}\frac{z}{1+z} \\
D_A &= \frac{cZ(z)}{H_\circ}\frac{J([1-\Omega_{tot}]Z^2)}{1+z} = \frac{c}{H_\circ}\frac{\sinh((\ln(1+z))}{1+z} \\
&= \frac{c}{H_\circ}\frac{(1+z)-1/(1+z)}{2(1+z)} = \frac{c}{H_\circ}\frac{z(1+z/2)}{(1+z)^2} \\
D_L &= (1+z)^2 D_A = \frac{cz}{H_\circ}(1+0.5\times z)
\end{aligned}
$$

- $\Omega_m = 1$

$$
\begin{aligned}
Z &= \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{X}} = \int_{1/(1+z)}^{1} a^{-1/2}da = 2(1-\frac{1}{\sqrt{1+z}}) \\
D_{now} &= \frac{cZ}{H_\circ} = \frac{2c}{H_\circ}\left(1-\frac{1}{\sqrt{1+z}}\right) \\
D_{ltt} &= \frac{2c}{3H_\circ}\left(1-(1+z)^{-3/2}\right) \\
D_A &= \frac{cZ(z)}{H_\circ}\frac{J([1-\Omega_{tot}]Z^2)}{1+z} = \frac{2c}{H_\circ}\left(\frac{1}{1+z}-\frac{1}{(1+z)^{3/2}}\right) \\
D_L &= (1+z)^2 D_A = \frac{2c}{H_\circ}\left(1+z-\sqrt{1+z}\right) \\
&= \frac{cz}{H_\circ}(1+0.25\times z+\ldots)
\end{aligned}
$$

- $\Omega_r = 1$

$$
\begin{aligned}
Z &= \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{X}} = \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{1/a^2}} = \frac{z}{1+z} \\
D_{now} &= \frac{cZ}{H_\circ} = \frac{c}{H_\circ}\left(\frac{z}{1+z}\right) \\
D_{ltt} &= \frac{c}{2H_\circ}\left(1-(1+z)^{-2}\right) \\
D_A &= \frac{cZ(z)}{H_\circ}\frac{J([1-\Omega_{tot}]Z^2)}{1+z} = \frac{cz}{H_\circ(1+z)^2} \\
D_L &= (1+z)^2 D_A = \frac{cz}{H_\circ} \\
&= \frac{cz}{H_\circ}(1+0.0\times z)
\end{aligned}
$$

- $\Omega_m = 2$

$$
\begin{aligned}
Z &= \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{X}} = \int_{1/(1+z)}^{1} \frac{da}{\sqrt{2a-a^2}} = 2\sin^{-1}\sqrt{\frac{a}{2}}\Bigg|_{1/(1+z)}^{1} \\
&= \frac{\pi}{2} - 2\sin^{-1}(1/\sqrt{2(1+z)}) \\
\sin\psi &= \sin Z = \cos\left[2\sin^{-1}\left(\frac{1}{\sqrt{2(1+z)}}\right)\right] \\
&= \cos^2\left[\sin^{-1}\left(\frac{1}{\sqrt{2(1+z)}}\right)\right] - \sin^2\left[\sin^{-1}\left(\frac{1}{\sqrt{2(1+z)}}\right)\right] \\
&= 1 - \frac{1}{2(1+z)} - \frac{1}{2(1+z)} = \frac{z}{1+z} \\
D_A &= \frac{c}{H_\circ}\frac{\sin\psi}{1+z} = \frac{cz}{H_\circ(1+z)^2} \\
D_L &= (1+z)^2 D_A = \frac{cz}{H_\circ} \\
&= \frac{cz}{H_\circ}\,(1 + 0.0 \times z)
\end{aligned}
$$

- $\Omega_v = 1$

$$
\begin{aligned}
Z &= \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{X}} = \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{a^2}} = z \\
D_{now} &= \frac{cZ}{H_\circ} = \frac{cz}{H_\circ} \\
D_{ltt} &= \frac{c}{H_\circ}\ln(1+z) \\
D_A &= \frac{cZ(z)}{H_\circ}\frac{J([1-\Omega_{tot}]Z^2)}{1+z} = \frac{cz}{H_\circ(1+z)} \\
D_L &= (1+z)^2 D_A = \frac{cz}{H_\circ}(1+z) \\
&= \frac{cz}{H_\circ}\,(1 + 1.0 \times z)
\end{aligned}
$$

32. Number Counts

One kind of cosmological observation is the number versus flux law, $N(S)$. We can compute the expected $N(S)$ law for various cosmological models using the distances D_A and D_L. The physical volume of the shell between redshift z and $z+dz$ is given by the surface area of the sphere which is $4\pi D_A(z)^2$ times the thickness of the shell which is $(cdt/dz)dz$. Thus if we have conserved objects, so their number density varies like $n(z) = n_\circ(1+z)^3$, then the number we expect to see in the redshift range is

$$\frac{dN}{dz} = n_\circ(1+z)^3 D_A(z)^2 \frac{cdt}{dz} \tag{345}$$

where N is the number of sources with redshift less than z per steradian.

However, we generally don't have a complete survey of all the objects closer than a given redshift. It is much more common to have a survey complete to a given flux or magnitude. Let S be the flux and L be the luminosity of the objects. We will consider a single class of objects, all with the same luminosity. The case with a range of luminosities is easily constructed from a superposition of several standard candle cases. With these assumptions, the luminosity distance is

$$D_L = \sqrt{\frac{L}{4\pi S}} \tag{346}$$

and the counts versus flux are given by

$$\frac{dN}{dS} = n_\circ(1+z)^3 \frac{D_L^2}{(1+z)^4} \frac{d(D_L)}{dS} \frac{dz}{d(D_L)} \frac{cdt}{dz} \tag{347}$$

Now $d(D_L)/dS = -0.5S^{-3/2}\sqrt{L/4\pi}$ and $D_L^2 = L/(4\pi S)$ so

$$\frac{dN}{dS} = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}} \left[(1+z)^{-1} \frac{dz}{d(D_L)} \frac{cdt}{dz}\right] \tag{348}$$

The first factor on the RHS is the "Euclidean" dN/dS which one would get for uniformly distributed sources in a non-expanding Euclidean Universe. The term in brackets contains the corrections due to cosmology.

The total intensity from all sources is given by

$$J = \int SdN = \int S \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}} \left[(1+z)^{-1} \frac{dz}{d(D_L)} \frac{cdt}{dz}\right] dS \tag{349}$$

Without the cosmological correction term this is $\int S^{-3/2}dS$ which diverges as $S \to 0$. This divergence is another statement of Olber's paradox.

If we use the expansions to second order in z

$$\begin{aligned}
\frac{dt}{dz} &= \frac{-H_\circ^{-1}}{(1+z)^2(1+q_\circ z+\ldots)} \\
D_L &= \frac{cz}{H_\circ}\left(1+\frac{z}{2}[1-q_\circ]+\ldots\right) \\
\frac{d(D_L)}{dz} &= \frac{c}{H_\circ}\left(1+z[1-q_\circ]+\ldots\right)
\end{aligned} \tag{350}$$

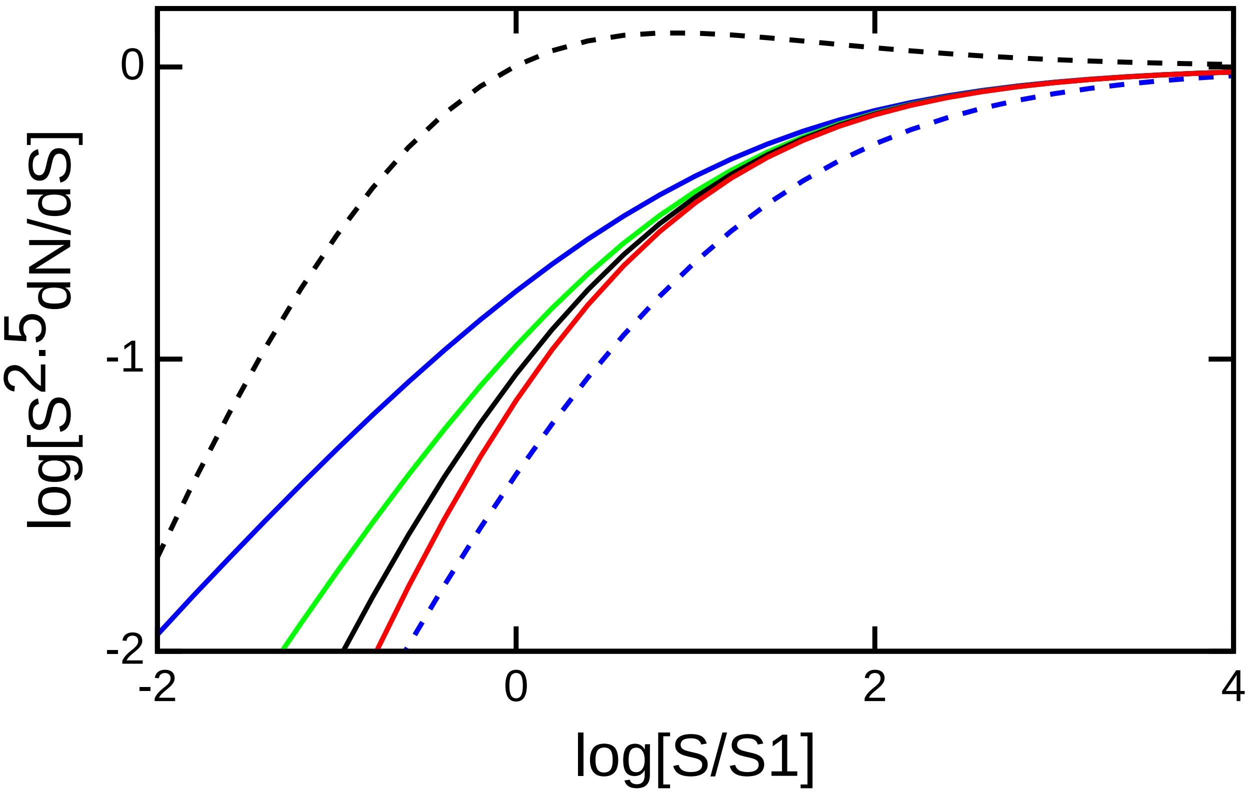

Fig. 35.— The predicted source counts for standard candles in various cosmological models, normalized to the Euclidean counts. This is for bolometric fluxes or sources with $L_\nu \propto \nu^{-1}$. S_1 is the "Euclidean" flux at a distance of $c/H_\circ$. The models are (from right to left: Steady State, blue dashed; $\Omega = 2$, red solid; $\Omega = 1$, black solid; empty, green solid; vacuum-dominated, blue solid; and "reality" for radio sources, black dashed.

we get

$$\frac{dN}{dS} = \frac{n_\circ (L/4\pi)^{3/2}}{2S^{5/2}} \left[\frac{1}{(1+z)^3(1+q_\circ z+\ldots)(1+z\,[1-q_\circ]+\ldots)} \right] \tag{351}$$

The $q_\circ$ dependence cancels out in the first two terms so

$$\frac{dN}{dS} = \frac{n_\circ (L/4\pi)^{3/2}}{2S^{5/2}} \left[\frac{1+\mathcal{O}(z^2)}{(1+z)^4} \right] \tag{352}$$

We see that the correction term decreases the source counts below the Euclidean expectation. This flattening of dN/dS avoids the divergence implied by Olber's paradox. The redshift where the count reduction is substantial is rather small since the correction term is $\approx (1+z)^{-4}$, so by $z = 0.25$ the correction is already a factor of 0.4.

We can easily work out the exact form of the relativistic correction for a few simple cases. For the empty Universe we get

$$D_L = \frac{cz}{H_\circ}(1+z/2)$$

$$\begin{aligned}
\frac{d(D_L)}{dz} &= \frac{c}{H_\circ}(1+z) \\
\frac{cdt}{dz} &= \frac{c}{H_\circ}\frac{1}{(1+z)^2} \\
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(1+z)^4}\right]
\end{aligned} \tag{353}$$

For large z we have $D_L \propto z^2$ in this case so $S \propto z^{-4}$. Hence $dN/dS \propto S^{-3/2}$ as $S \to 0$. This flattening is enough to make the total intensity finite, solving Olber's paradox, but the total number of observable sources is infinite.

We can express the redshift in terms of the flux by defining $\zeta = \sqrt{S_1/S}$ where $S_1 = L/(4\pi(c/H_\circ)^2)$ is the flux the source would have in a Euclidean Universe with distance $d = c/H_\circ$. Think of ζ ("zeta") as a Euclidean distance in redshift units. Solving $\zeta = z(1+z/2)$ gives $z = -1+\sqrt{1+2\zeta}$ so for $\Omega = 0$ the $N(S)$ law is

$$\frac{dN}{dS} = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(1+2\zeta)^2}\right] = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{\left(1+2\sqrt{S_1/S}\right)^2}\right] \tag{354}$$

For an $\Omega = 2$ matter dominated Universe we get

$$\begin{aligned}
D_L &= \frac{cz}{H_\circ} \\
\frac{d(D_L)}{dz} &= \frac{c}{H_\circ} \\
\frac{cdt}{dz} &= \frac{c}{H_\circ}\frac{1}{(1+z)^2\sqrt{1+2z}} \\
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(1+z)^3\sqrt{1+2z}}\right]
\end{aligned} \tag{355}$$

For large z we have $S \propto z^{-2}$ in this model, so the source counts flatten to $dN/dS \propto S^{-3/4}$. This not only solves Olber's paradox but also gives a finite total number of observable sources. For any $\Omega > 0$ we have $D_L \propto z$ for large z, so this finite total source count applies to all models with $\Omega > 0$, even though the models with $\Omega < 1$ are open models with infinite volumes. The total source count gives the number of sources in the *observable* Universe, and this is less than the total Universe unless $\Omega = 0$. For this model $z = \zeta$ so

$$\frac{dN}{dS} = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{\left(1+\sqrt{S_1/S}\right)^3\sqrt{1+2\sqrt{S_1/S}}}\right] \tag{356}$$

For a critical density matter-dominated Universe we get

$$D_L = \frac{2c}{H_\circ}\left(1+z-\sqrt{1+z}\right)$$

$$\begin{aligned}
\frac{d(D_L)}{dz} &= \frac{c}{H_\circ}\left(2-\frac{1}{\sqrt{1+z}}\right) \\
\frac{cdt}{dz} &= \frac{c}{H_\circ}\frac{1}{(1+z)^{2.5}} \\
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(2\sqrt{1+z}-1)(1+z)^3}\right]
\end{aligned} \tag{357}$$

If we let $u=\sqrt{1+z}$, then $2u(u-1)=\zeta$ so

$$u=\sqrt{1+z}=\frac{1+\sqrt{1+2\zeta}}{2} \tag{358}$$

and the source counts $N(S)$ are given by

$$\begin{aligned}
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{8}{\sqrt{1+2\zeta}\left(1+\zeta+\sqrt{1+2\zeta}\right)^3}\right] \\
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{8}{\sqrt{1+2\sqrt{S_1/S}}\left(1+\sqrt{S_1/S}+\sqrt{1+2\sqrt{S_1/S}}\right)^3}\right]
\end{aligned} \tag{359}$$

For a vacuum-dominated model with $q_\circ=-1$, the cosmological functions are

$$\begin{aligned}
D_L &= \frac{c}{H_\circ}z(1+z) \\
\frac{d(D_L)}{dz} &= \frac{c}{H_\circ}(1+2z) \\
\frac{cdt}{dz} &= \frac{c}{H_\circ}\frac{1}{1+z}
\end{aligned} \tag{360}$$

Thus $\zeta=z(1+z)$, or

$$z=\frac{-1+\sqrt{1+4\zeta}}{2} \tag{361}$$

The number counts will be given by

$$\begin{aligned}
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(1+z)^2(1+2z)}\right] \\
&= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{4}{\left(1+\sqrt{1+4\sqrt{S_1/S}}\right)^2\sqrt{1+4\sqrt{S_1/S}}}\right]
\end{aligned} \tag{362}$$

Since $q_\circ$ has canceled out in the leading terms, the source count test is an insensitive method to find the geometry of the Universe. Even so, source counts were used to rule out the Steady State model even before the microwave background was discovered. The reason this was possible is that the observed source counts of radio sources and quasars are not consistent with Eqn(352)

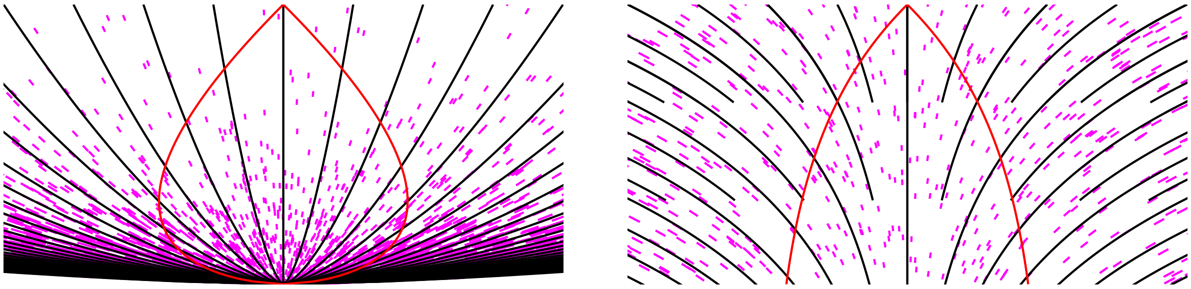

Fig. 36.— The excess of faint quasars and radio sources can be explained by an evolving population in a homogeneous Universe as shown at left above. The quasars are shown as little magenta line segments because they have a short lifetime compared to the age of the Universe. There are now only a few quasars, so we see only a few bright quasars. But when we look far away we are also looking into the past, and see many more quasars. In the Steady State model on at right above, this evolution with time is not allowed, so there is no explanation for the excess of faint quasars and radio sources.

for any value of $q_\circ$. Instead of being flatter than the $dN/dS \propto S^{-5/2}$ Euclidean prediction, the observed source counts are steeper for medium fluxes, and only flatten for very faint flux levels. These steeper than Euclidean source counts can only be obtained by changing our assumptions: either the density or luminosity of the sources was greater in the past. To fit the quasar counts with an increase of density, we need to increase our assumed n by a factor of $\approx (1+z)^6$. This increase is faster than the $(1+z)^{-4}$ correction term due to cosmology, and matches the steep observed number counts. Thus the total density of quasars must vary like $n(z) = n_\circ(1+z)^3(1+z)^6 \propto (1+z)^9$ back to redshifts of $z \approx 2.5$. How does this result rule out the Steady State model? The geometry of the Steady State model is just that of a $q_\circ = -1$ vacuum-dominated model, and $q_\circ$ has canceled out in dN/dS. But the Steady State model also asserts that $n(z)$ is *constant* so the $(1+z)^3$ term in Eqn(345) is dropped. This gives

$$\frac{dN}{dS} = \frac{n_\circ (L/4\pi)^{3/2}}{2S^{5/2}} \left[\frac{32}{\left(1+\sqrt{1+4\sqrt{S_1/S}}\right)^5 \sqrt{1+4\sqrt{S_1/S}}} \right] \tag{363}$$

in the Steady-State model, and there is no freedom to adjust the source density evolution with z. Since the apparent source density evolves like $(1+z)^9$ which is clearly not constant, the Steady State model is ruled out.

32.1. Gamma Ray Bursts

γ-ray bursts are short bursts (≈ 10 seconds) of 100-300 keV γ-rays. The total energy fluence is 10^{-5} erg/cm^2 for fairly bright bursts that occur about once a month. Thus the flux from these bursts is about 10^{-6} erg/cm^2/sec, which is the same as the bolometric flux of a 1^{st} magnitude star. But since the bursts occur without warning, very little is known about their sources. I participated in a search for optical counterparts in the 1970's using the Prairie Network of meteor cameras. These cameras photographed the entire visible sky every dark night to look for meteors. Multiple cameras separated by 10's of kilometers were used to provide stereo views of the meteor trails. Grindlay, Wright & McCrosky (1974, ApJL, 192,L113) searched these films for "dots" in the error boxes of γ-ray bursts. Stars left long circular trails, but a "dot" would be a source that flashed on then went out quickly. If two cameras recorded a dot in the same place, we would have a hit. Unfortunately there were lots of dots (mainly dirt and "plate" flaws), but no coincidences. But we were able to prove that the optical power was less than the γ-ray power.

GRB's are isotropic on the sky, with no preference for the galactic plane or the galactic center. Furthermore, the bright bursts follow a $N(>S) \propto S^{-1.5}$ source count law appropriate for uniformly distributed sources. This indicates that they originate either from sources less than one disk scale height from the Sun, or from sources further away than the Virgo cluster of galaxies. So the distance is either < 100 pc or > 100 Mpc. When GRO was launched, many people expected the isotropic pattern to break down for the fainter bursts with fluences down to 10^{-7} erg/cm^2 that the BATSE experiment can detect. Instead, the faint burst distribution is still isotropic, but the source counts flatten for fainter bursts. Thus the GRB distribution has an "edge", but is spherically symmetric around the Solar System.

One of leading models for this is that the "edge" is the edge of the observable Universe. We have seen that the source counts will naturally break away from the Euclidean $N(>S) \propto S^{-1.5}$ law when $z \approx 0.25$. When dealing with GRB's there is an additional factor of $1/(1+z)$ in the source count correction because the rate of bursts from high redshift regions is reduced by the time dilation factor: all rates transform like the apparent rate of oscillation of atomic clocks, and are thus slower by a factor of $1/(1+z)$. If we take $z = 0.2$ for a burst with a fluence of 10^{-6} erg/cm^2, we find a source energy release of

$$\Delta E = 4\pi D_L(z)^2 \times (\text{fluence})/(1+z) \approx (5/h^2) \times 10^{49} \text{ ergs} \tag{364}$$

The $1/(1+z)$ factor is due to the stretching of the burst by the redshift, so the emitted burst is shorter than the observed burst. This total energy is considerably larger than the optical light emitted by a supernova. Producing this large quantity of γ-rays is the principal difficulty with the cosmological model for GRB's.

But new data obtained using the Beppo-SAX satellite have proved that GRB's are at cosmological distances. That satellite has a wide-field of view hard X-ray camera, that can locate the position of a GRB to $4'$ accuracy. Ground controllers slew the satellite to point a higher resolution X-ray telescope at the burst position, and often a fading X-ray transient source is seen for several hours after the burst. The X-ray telescope provides $1'$ positions. Ground-based telescopes see a fading optical transient at the same position, and in a few cases a large redshift has been determined for the optical transient because absorption lines due to an intervening cloud of gas at redshift z_{abs} are seen. The redshift of the GRB source must be larger that z_{abs}. GRB 970508 showed $z_{abs} = 0.825$ (Metzger *et al.*, 1998, Nature, 387, 878). GRB 990123 was observed optically within 22 seconds of the BATSE trigger, and was seen to rise to a peak brightness of 9^{th} magnitude, even though the redshift is $z_{abs} = 1.6$ (Akerlof *et al.*, astro-ph/9903271). The γ-ray fluence for $E > 20$ keV of this burst was 3×10^{-4} erg cm^{-2} (Band *et al.*, 1999, ApJ, TBD, TBD, astro-ph/9903247). After a few months the optical transient has faded away, and there is usually a faint galaxy at the former position of the fading optical transient. These galaxies have large redshifts: for example, the host galaxy of GRB 971214 has $z = 3.42$ (Kulkarni *et al.*, 1998, Nature, 393, 35). This burst had a γ-ray fluence of 1.1×10^{-5} erg cm^{-2}. While these data confirm the cosmological nature of the GRB's predicted by their number counts and isotropy, the high redshifts seen for bursts clearly on the $S^{-1.5}$ part of the $N(S)$ curve requires a very wide distribution of intrinsic GRB brightness.

33. Horizon Problem

The *horizon* is the greatest distance we can see – the distance at which $z = \infty$. If we write the FRW metric so the radial spatial part is simple, we get

$$ds^2 = c^2dt^2 - a(t)^2\left[d\psi^2 + R_\circ^2\left\{\begin{array}{c}\sinh^2\left(\psi R_\circ^{-1}\right)\\ \left(\psi R_\circ^{-1}\right)^2\\ \sin^2\left(\psi R_\circ^{-1}\right)\end{array}\right\}\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right] \tag{365}$$

and the proper radial distance to an object at $z = \infty$ is just

$$D_H = \psi = \int \frac{cdt}{a(t)} = \int_0^\infty (1+z)c\frac{dt}{dz}dz \tag{366}$$

This distance is known as the horizon distance. For an $\Omega = 1$ matter-dominated Universe D_H is $2c/H_\circ$.

If we evaluate the horizon for an observer at redshift z, then in comoving units (lengths scaled to the current time), we have

$$D_H(z) = \int_z^\infty (1+z)c\frac{dt}{dz}dz = \frac{2c}{H_\circ\sqrt{1+z}} \tag{367}$$

If we look at two blobs of gas at $Z_{rec} \approx 1100$ when the Universe became transparent, separated by an angle θ on the sky, then the comoving distance between them is $2\sin(\theta/2)[D_H(0) - D_H(z)]$, and if this is greater than $2D_H(z)$ the two blobs of gas have disjoint domains of influence. This means that there is no event in spacetime that is in or on the past light cones of both of the two blobs. This occurs whenever $\sin(\theta/2) > 1/(\sqrt{1+z} - 1)$ or $\theta > 3.6°$. But the whole sky has a uniform CMBR temperature to within 1 part in 10^5. This appears to require a very special initial condition for the Universe.

This can be seen clearly in a *conformal* spacetime diagram. This is a diagram that plots ψ *vs.* the conformal time η, defined as $d\eta = a(t)^{-1}cdt = (1+z)cdt$. In terms of this time variable, the metric becomes

$$ds^2 = a(t)^2\left(d\eta^2 - \left[d\psi^2 + R_\circ^2\left\{\begin{array}{c}\sinh^2\left(\psi R_\circ^{-1}\right)\\ \left(\psi R_\circ^{-1}\right)^2\\ \sin^2\left(\psi R_\circ^{-1}\right)\end{array}\right\}\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right]\right) \tag{368}$$

and the paths of light rays are obviously the $\pm 45°$ lines $\psi = \psi_\circ \pm (\eta - \eta_\circ)$. In order to make a conformal spacetime diagram from an ordinary spacetime diagram we first divide the spatial coordinate by $a(t)$. This makes the worldlines of comoving galaxies run straight up and down. We then stretch the time axis near the Big Bang to keep the slope of null rays at $\pm 45°$.

In order to solve the horizon problem, one needs to have more conformal time before recombination than after it. This will happen when the integral for the conformal time diverges as $a \to 0$:

$$\eta = \int \frac{cdt}{a} = \int \frac{cda}{a\dot{a}} \tag{369}$$

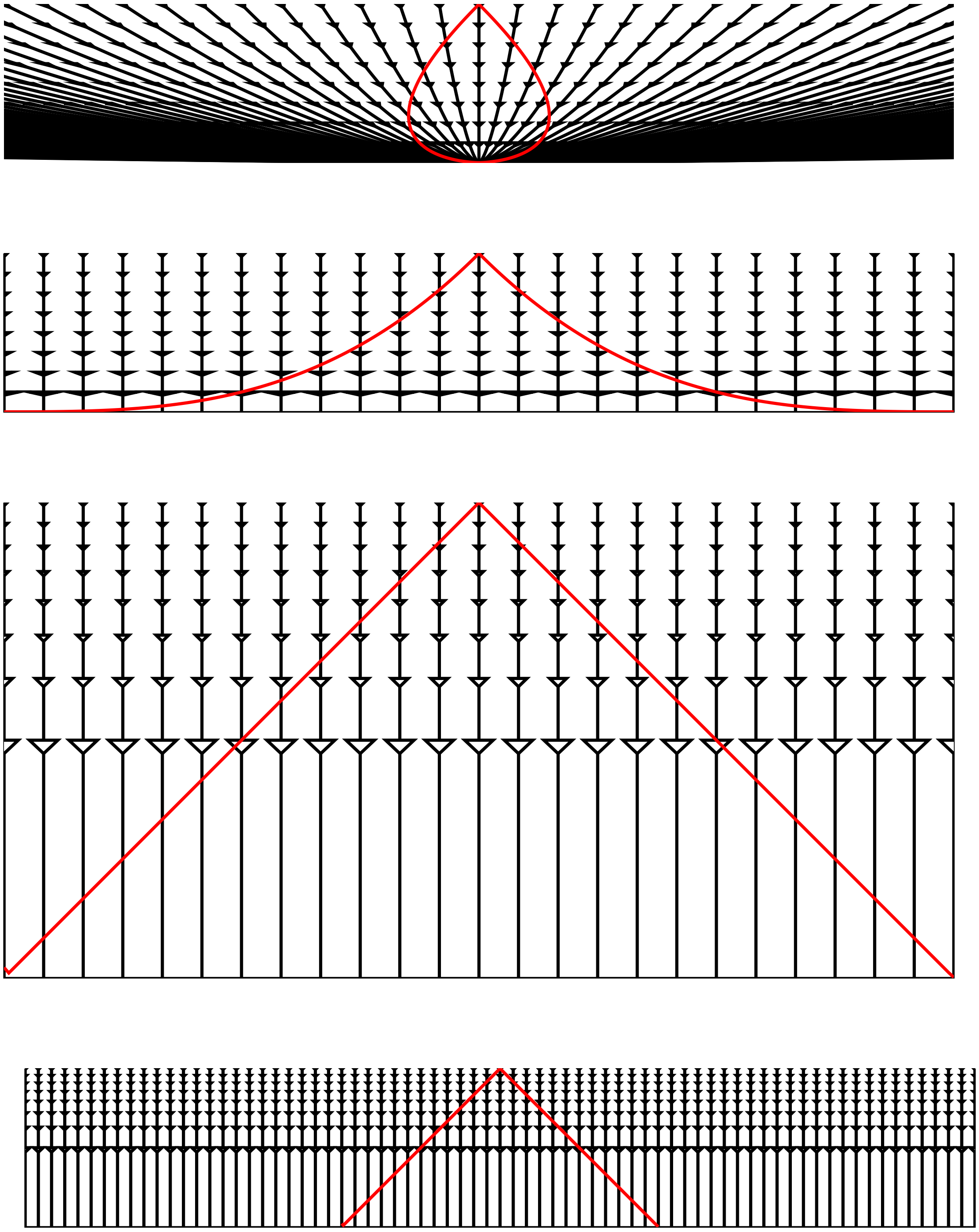

Fig. 37.— From top to bottom: a) $\Omega = 1$ S-T in standard form; b) with distances divided by $a(t)$; c) time axis "stretched" into conformal time; d) a wider view showing the Universe is much bigger than the observable Universe.

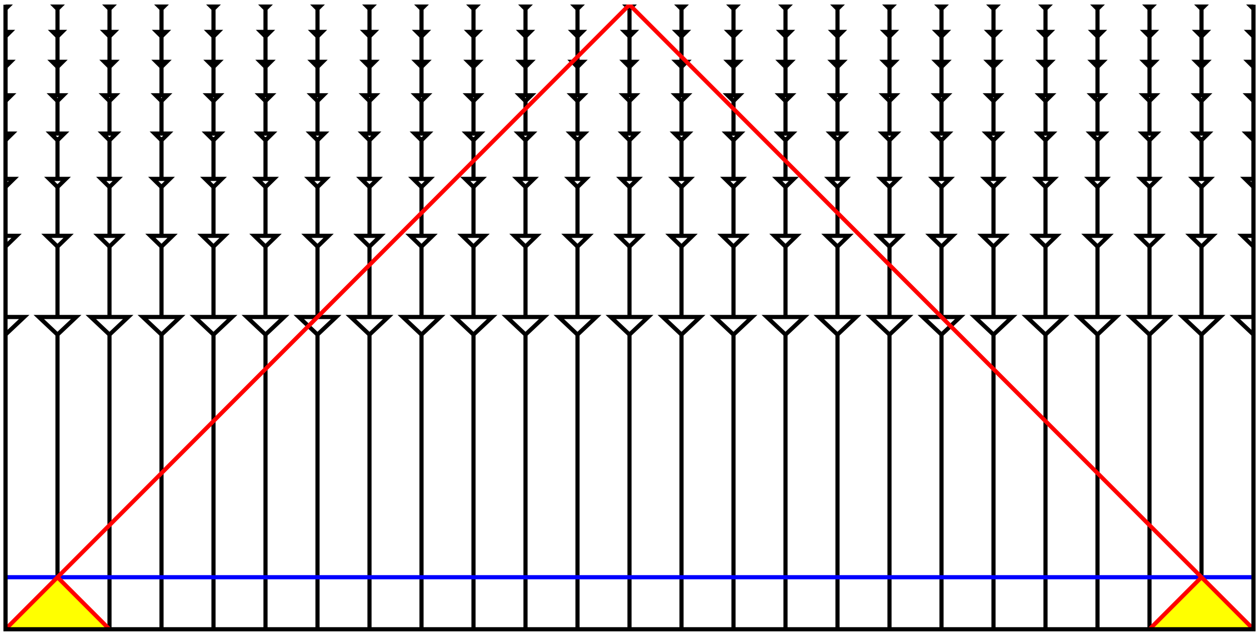

Fig. 38.— A conformal space-time diagram for the $\Omega = 1$ model. Recombination is indicated by the blue line. The gas blob seen on the left and the gas blob seen on the right have disjoint past light cones. So why is the CMB temperature so uniform?

This integral will diverge as $a \to 0$ if $a(t) \propto t^n$ with $n \geq 1$, but for the critical density model $n = 2/3$ and the integral converges. The integral also diverges for a vacuum-dominated model with $a \propto \exp(Ht)$. In this case $\dot{a} = Ha$ and the integral $\int da/a^2$ diverges for small a.

34. Inflationary Scenario

34.1. Spontaneous Symmetry Breaking

Modern theories of particle physics invoke spontaneous symmetry breaking to explain the multiple manifestations of a presumably unified interaction. For example, the electroweak theory at ordinary energies shows two very different behaviours: the long-range electromagnetic force mediated by massless photons, and the very short-range weak nuclear force carried by the massive W and Z bosons. The way that spontaneous symmetry breaking produces these effects is to have a symmetric model whose lowest energy states are not symmetric. An an example of this is a vacuum energy density which is a function of two fields ϕ_1 and ϕ_2 given by

$$V(\phi_1, \phi_2) = \lambda(\sigma^2 - (\phi_1^2 + \phi_2^2))^2 \tag{370}$$

This potential has a ring-shaped minimum that is reminiscent of a sombrero, so it is often called the Mexican-hat potential. It is obvious that the potential is symmetric under rotations in the two dimensional ϕ space, and that ϕ_1 and ϕ_2 are treated identically in the theory. But once the system settles into one of the states with lowest energy, then there will be two very different modes of oscillation. We can assume that the system settles into the state with $\phi_1 = \sigma$ and $\phi_2 = 0$. This means that there is a non-zero *vacuum expectation value* since $\langle 0|\phi_1|0\rangle >= \sigma$. Let $\psi = \phi_1 - \sigma$ and expand the potential energy for small values of ψ and ϕ_2, giving

$$V(\psi, \phi_2) \approx 4\lambda\sigma^2\psi^2 + \ldots \tag{371}$$

The Lagrangian density is then

$$\begin{aligned} \mathcal{L} &= \partial_\mu\psi\partial^\mu\psi - 4\lambda\sigma^2\psi^2 \\ &+ \partial_\mu\phi_2\partial^\mu\phi_2 \end{aligned} \tag{372}$$

This Lagrangian describes a massless boson ϕ_2 and a massive boson ψ with mass $2\sqrt{\lambda}\sigma$. These equations are written using $\hbar = 1$ and $c = 1$, so the units for an energy density are M^4, and the units of ∂_μ are M^1, so the units of ϕ are also M^1. Since σ and ϕ both have the units of M, the coefficient λ is a dimensionless number in the theory. This mechanism for spontaneous symmetry breaking is known as the Higgs mechanism, and the particles predicted are called Higgs bosons.

34.2. Topological Defects

Obviously this model is not elaborate enough to produce the electroweak force because that has 4 different bosons. But this simple model with two scalar fields does demonstrate a fundamental property of spontaneous symmetry breaking: *topological defects.* Consider a field configuration with spatially varying ϕ_1 and ϕ_2, given by

$$\begin{aligned} \phi_1 &= \sigma f(r)\frac{x}{\sqrt{x^2+y^2}} \\ \phi_2 &= \sigma f(r)\frac{y}{\sqrt{x^2+y^2}} \end{aligned} \tag{373}$$

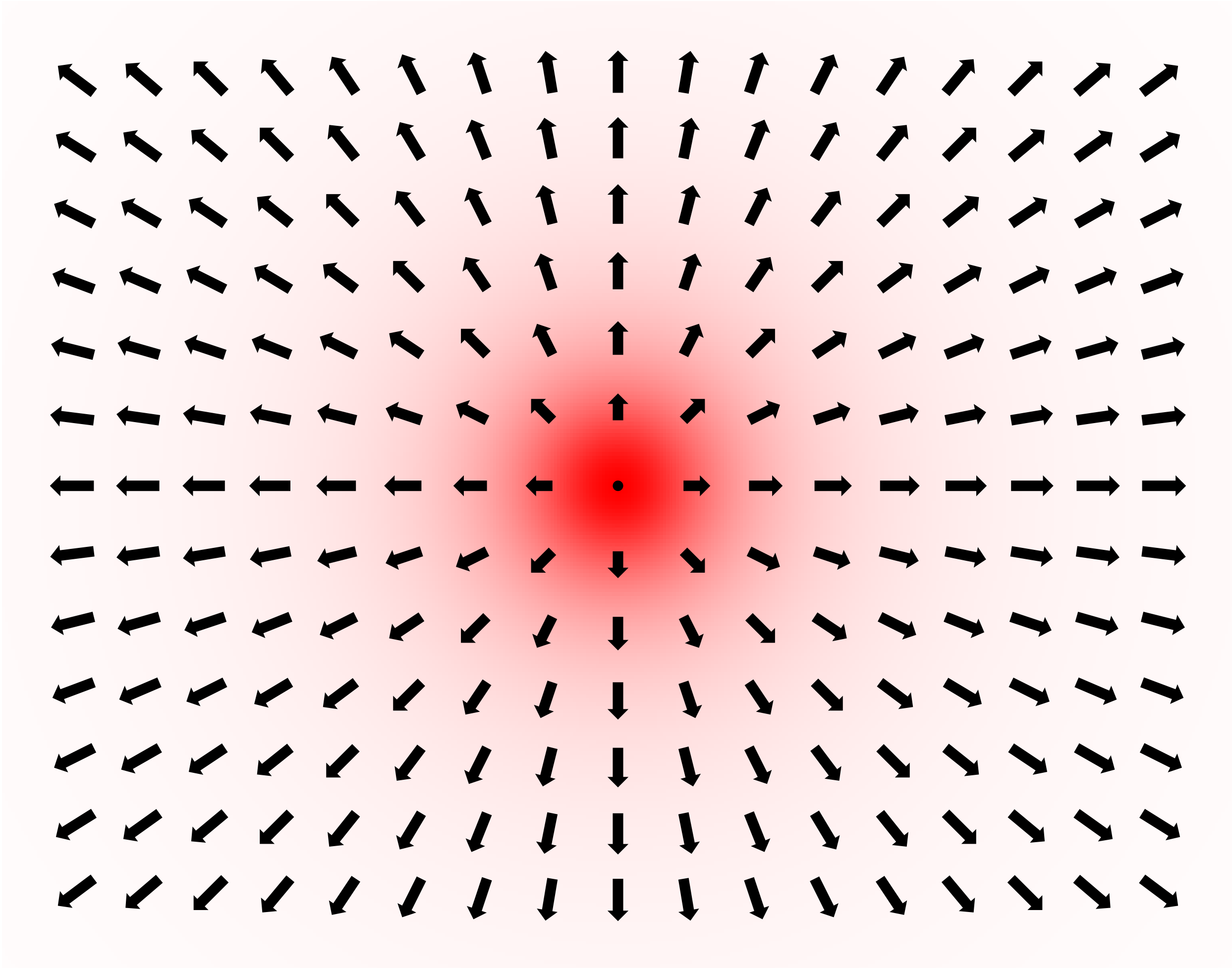

Fig. 39.— Figure showing the fields ϕ_1 as the x-component of a vector and ϕ_2 as the y-component as a function of (x, y) position around a topological defect known as a cosmic string. The red shading indicates the energy density due to both the gradient term and the vacuum energy term.

This field configuration has the ϕ field making a full rotation in ϕ space as the spatial coordinates make one loop around the z-axis. The energy per unit dz is given by

$$T = 4\pi \int \left(\sigma^2 \left[\left(\frac{df}{dr}\right)^2 + \left(\frac{f}{r}\right)^2 \right] + \lambda\sigma^4 (1 - f(r)^2)^2 \right) r dr \tag{374}$$

To minimize the last term we want $f(r) = 1$, but to make the $(f/r)^2$ term less than infinity we need $f(0) = 0$, and the $(df/dr)^2$ term limits the rate at which the field magnitude can approach the minimum at σ. Thus there is a minimum possible energy per unit length for this configuration, which defines the *tension* of the *cosmic string* to be this minimum $T = \mathcal{O}(\sigma^2)$. For $\sigma \approx 100$ GeV, which would be appropriate for the breaking the electroweak symmetry, the string tension is about 6×10^{17} GeV/cm or an equivalent linear mass density of 1 μg/cm. For the $\sigma \approx 10^{16}$ GeV needed to break a grand unified theory, the mass per unit length is 10^{22} gm/cm.

Another kind of topological defect is possible with three scalar fields and a vacuum energy density $V = \lambda(\sigma^2 - \sum \phi_i^2)^2$. Now the fields can be arranged in a radial pattern around a point, leading to a pointlike topological defect with a mass $M \approx \sigma/\sqrt{\lambda} \approx 10^{16}$ GeV for a typical GUTs σ. If the spontaneous symmetry breaking due to these fields leads to the standard model, then this pointlike defect has a magnetic charge, and is thus an ultramassive magnetic monopole.

34.3. Monopole Problem

If the spontaneous symmetry breaking occurs when $kT \approx 10^{15}$ GeV, then the time is about $t = 10^{-36}$ sec, and the Higgs fields will probably only be uniform on patches of size $ct = 3\times10^{-26}$ cm. Thus a density of about 10^{78} monopoles per cc could easily be generated. The expansion of the Universe since this time would reduce the density by a factor of 10^{81}, and only a small fraction of the monopoles would have avoided annihilating with an oppositely charged monopoles, but the expected current density of monopoles is still about one per cubic meter. Given their high mass, this leads to $\Omega = 10^{15}$! This is the *monopole problem* is the standard hot Big Bang model.

34.4. Inflation to the rescue

The hot Big Bang model has three problems:

1. the special initial conditions needed to explain the flatness and oldness of the current Universe,
2. the special initial conditions needed to explain the near isotropy of the CMBR, and
3. a surfeit of monopoles.

Fortunately a slight modification of the potential $V(|\phi|)$ can remove these problems. The modification is to make the central hump in the potential extremely flat. Thus for $0 < |\phi| < \sigma$, we suppose that $V \approx \mathcal{O}(\sigma^4)$ but the slope $|dV/d\phi| << \mathcal{O}(\sigma^3)$. [The slope is negative of course.] Now when the Universe cools to the point where the spontaneous symmetry breaking will occur, the ϕ fields will probably have some value between 0 and σ. Because ϕ is already non-zero, the monopoles will have already been generated. Now, because the slope of the potential is so small, it will take a long time for the ϕ fields to reach the global minimum at $|\phi| = \sigma$. During this time, the Universe has a large vacuum energy density $V \approx \sigma^4$. Because of this vacuum energy density, the Universe undergoes exponential expansion, with $a(t) \propto \exp(Ht)$. Now it is only necessary to have the number of e-foldings during the exponential expansion be larger than about 70 so the scale factor grows by a factor of 10^{30} or more. Figure 40 shows $1 + z = 1/a$ *vs.* time with and without inflation.

This expansion will solve the three problems of the Big Bang:

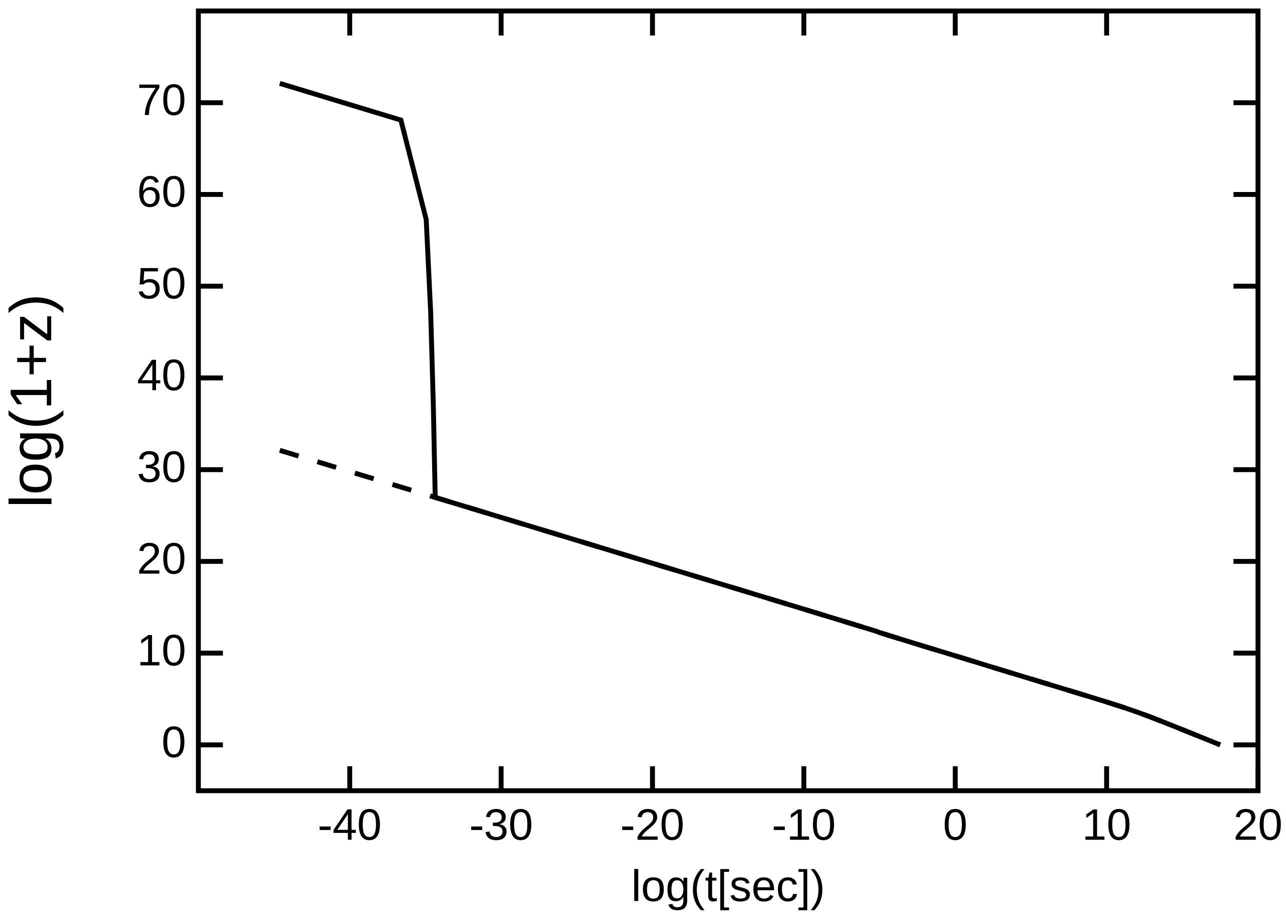

Fig. 40.— The redshift *vs.* time for the standard hot Big Bang model (dashed) and for a model with inflation.

1. Because the scale factor grows by 10^{30} during inflation, but we normalize $a(t_\circ) = 1$ now, the value of a at the Planck time becomes 10^{-30} times smaller. But the density ρ does not change during the vacuum dominated inflationary epoch. Using the equation

$$\left(\frac{1}{\Omega(t)} - 1\right) = \left(\frac{1}{\Omega_\circ} - 1\right)\left(\frac{\rho_\circ a(t_\circ)^2}{\rho(t)a(t)^2}\right) \tag{375}$$

we see that the initial conditions on Ω are relaxed by a factor of 10^{60} so that just about any starting value will work.

2. The horizon problem arose because the conformal time before recombination was much less than the conformal time after recombination, and the conformal time measures how far light can travel on a scale with all lengths scaled up to their current size. But the conformal time is given by

$$\eta = \int (1+z)c\,dt = \int ct(1+z)\,d\ln t \tag{376}$$

so the contribution per logarithmic time interval is given by $t \times (1+z)$, shown in Figure 41. This peaks after recombination in the standard model but before recombination in the

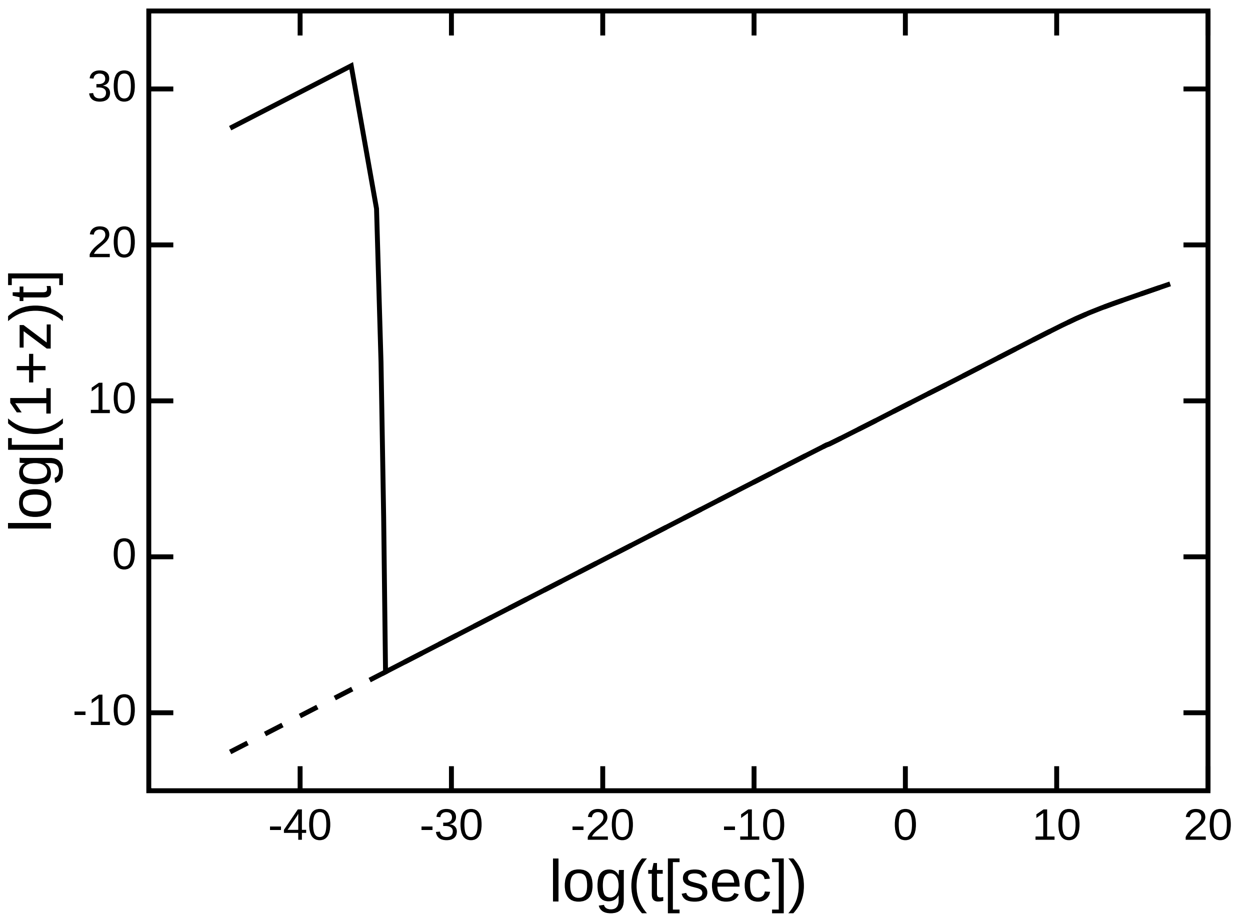

Fig. 41.— The contribution to the conformal time per octave in time, $(1+z)t$, for the standard hot Big Bang model (dashed) and for a model with inflation.

inflationary model. To say it in words, a region 3×10^{-26} cm across becomes homogeneous before inflation starts, and then grows to be 3×10^{3} cm at the end of inflation. This region then grows another factor of 3×10^{27} in the normal hot Big Bang phase following inflation, leading to a homogeneous patch 10^{31} cm across, which is larger than the observable Universe.

3. The monopole density is reduced by a factor of 10^{90} because the inflation occurs *after* the spontanteous symmetry breaking. So instead of one monopole per cubic meter, there is at most one monopole per observable Universe.

While inflation will get rid of monopoles, it will also get rid of baryons by driving the net baryon density down by a factor of 10^{90}. GUTs allow the possibility of proton decay and also of *baryogenesis*, where a combination of a violation of baryon number conservation and a violation of time reversal invariance and the expansion of the Universe leads to the creation of more baryons than antibaryons. Obviously this process must occur *after* inflation. The latest allowable time for baryogenesis is the electroweak $\rightarrow$ EM+weak transition, which occurs about 1 picosecond after the Big Bang. The earliest time that inflation could occur is the Planck time, and this would require a separate solution to the monopole problem. Beyond this limits, very little can be said for certain about inflation. So most papers about inflationary models are more like historical novels than real history, and they describe possible interactions that would be interesting instead of interactions that have to occur. As a result, inflation is usually described as the *inflationary scenario* instead of a theory or a hypothesis. However, it seems quite likely the inflation did occur, even though we don't know when or what the potential was. If inflation occured, then there is a fairly definite prediction made about the primordial density fluctuations that are the seeds for galaxy formation and also produce the small anisotropy of the CMBR seen by COBE. There is also a fairly definitite prediction about the value of Ω: since 71 e-foldings during inflation are just as likely as 70 or 72 e-folding, the probability of a given range of Ω has to be proportional to dN where N is the number of e-foldings. Thus the probability of $0.9 < \Omega < 0.99$ is the same as the probability of $0.99 < \Omega < 0.999$ which is the same as the probability of $0.999 < \Omega < 0.9999$ etc. There is an infinite accumulation of probability at $\Omega = 1$. An value of Ω that is definitely not equal to 1 is evidence against inflation. But this is Ω_{tot}, so a flat vacuum-dominated model with $\Omega_{m\circ} + \Omega_{v\circ} = 1$ is consistent with inflation.

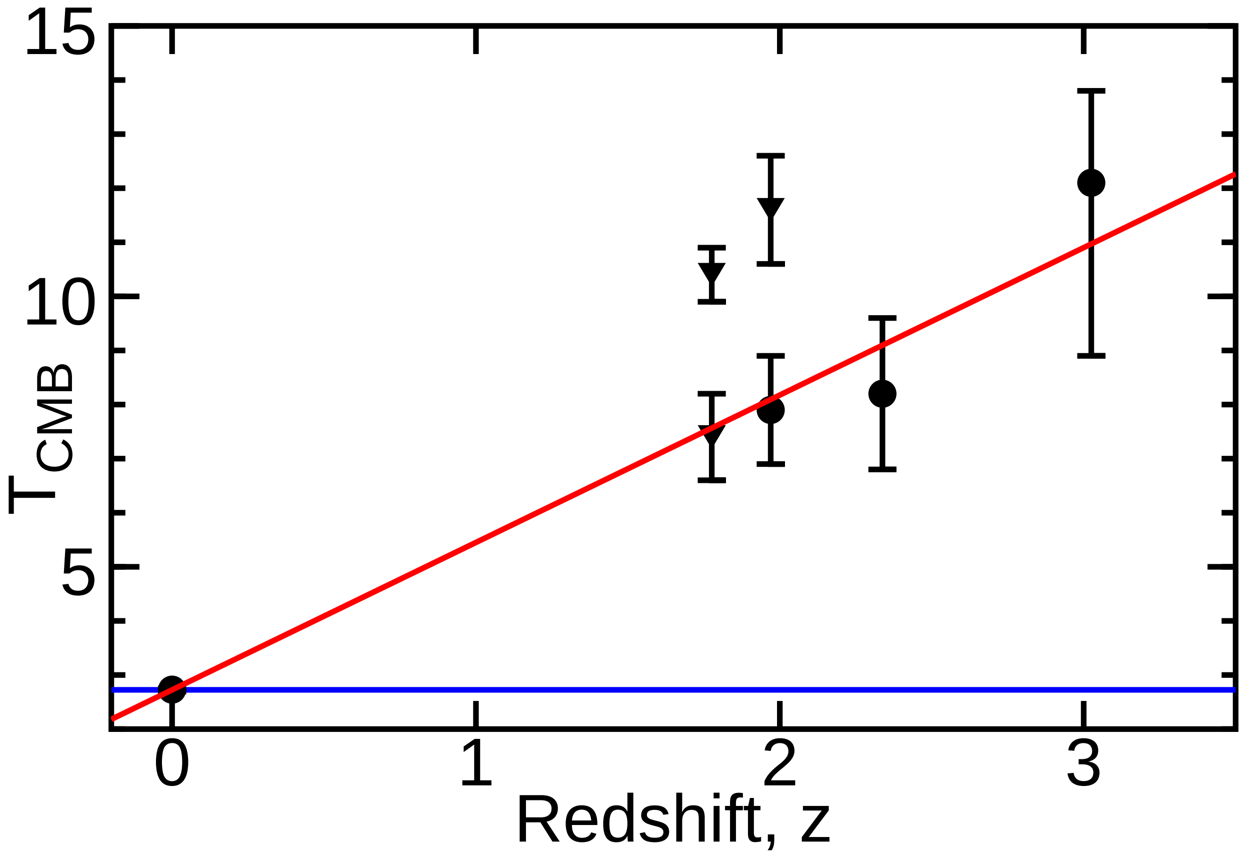

Fig. 42.— Measured temperature of the cosmic microwave background as a function redshift.

35. Temperature vs. Time

As the Universe expands, certain quantities are conserved. The energy, however, is not conserved. If it were, the fact that the energy density of a blackbody scales like T^4 would imply that a^3T^4 was conserved, giving $T \propto a^{-3/4} = (1+z)^{0.75}$. But the redshift is always changing the energy per photon, while the number of photons is conserved because the photon creation and destruction rates are very slow. Since the number of photons in a blackbody scales like T^3, conservation of photons gives a constant a^3T^3, or $T \propto (1+z)$. This has actually been observed by looking at excited levels in interstellar ions and molecules in high redshift clouds, as shown in Figure 42.

But what would happen if the photon creation and destruction rates were large compared to the Hubble rate? This happens during the first few months after the Big Bang. What is conserved then? We can determine the answer by considering a comoving cube in the Universe, one that expands along with the Universe. Since the cube faces are moving with the general expansion of the Universe, and the Universe is homogeneous, there is no net flux of energy into or out of the cube. The cube is thus an adiabatic enclosure and will have a conserved entropy. Thus in the presence of photon creation or destruction, we still know that the entropy density s will scale like $(1+z)^3$ so that the total entropy in the cube, which scales like a^3s, will remain constant. Thus we

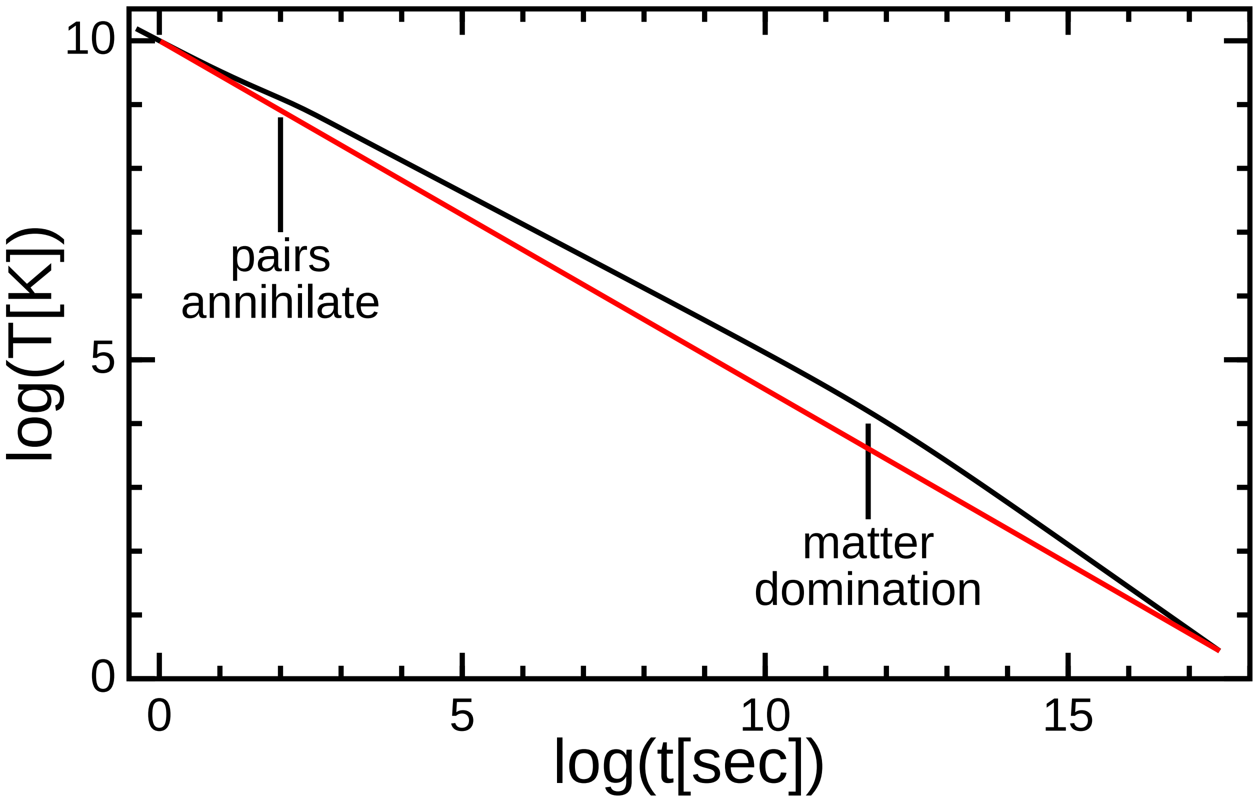

Fig. 43.— Photon temperature as a function of time.

have this equation for $T(z)$:

$$s(T(z)) = s(t_\circ)(1+z)^3 \tag{377}$$

How can we evalute the current entropy density $s(t_\circ)$? It is dominated by the most numerous particles, the photons. The next most numerous are the neutrinos, but they have been decoupled since one second after the Big Bang so we will treat them as a separate system with independently conserved entropy. We can evaluate the entropy density of the photons at temperature T using

$$V\,ds = dS = \frac{dQ}{T} = 4aT^2 V\,dT \tag{378}$$

so $s = (4/3)aT^3$ for photons. This formula is correct for all temperatures $\ll m_e c^2/k$ where the density of thermally generated electron-positron pairs is negligible. For higher temperatures we need to add the energy of the pair plasma to Q in Eqn(378). This gives

$$u = aT^4 + 2 \times 2 \times 4\pi h^{-3} \int \frac{\sqrt{(m_e c^2)^2 + p^2c^2}\, p^2 dp}{\exp(\sqrt{(m_e c^2)^2 + p^2c^2}/kT) + 1} \tag{379}$$

The two factors of 2 in front of the integral are for the two spin states and for the two types of particles (electrons and positrons). The $e^x + 1$ in the denominator is correct for fermions as opposed to the $e^x - 1$ form for bosons. For $kT \gg m_e c^2$ the pair plasma energy density simplifies to

$$u_{e^+e^-} = 16\pi kT(kT/hc)^3 \int \frac{x^3\,dx}{e^x + 1} \tag{380}$$

Now

$$\begin{aligned}
\int \frac{x^n\,dx}{e^x+1} &= \int x^n(e^{-x} - e^{-2x} + e^{-3x} - e^{-4x} + \ldots)dx \\
&= \Gamma(n+1)(1 - 2^{-(n+1)} + 3^{-(n+1)} - 4^{-(n+1)} + \ldots) \\
&= \Gamma(n+1)(1 + 2^{-(n+1)} + 3^{-(n+1)} + 4^{-(n+1)} + \ldots \\
&\quad -2 \times 2^{-(n+1)}(1 + 2^{-(n+1)} + \ldots)) \\
&= \Gamma(n+1)\zeta(n+1)(1 - 2^{-n}) \qquad (381)
\end{aligned}$$

That electrons are fermions instead of bosons reduces the energy density by a factor $1 - 2^{-3} = 7/8$ while the two types of particles factor doubles the energy density. Thus for temperatures higher than the electron rest mass, the pair plasma contributes $7/4$ times more to the energy than the photons. Since the entropy density is computed from the energy density, it will be a total of $11/4$ times higher when the pair plasma is fully developed. This makes the temperature a factor of $(4/11)^{1/3} = 0.714$ times lower that a straight $T \propto (1+z)$ extrapolation. The temperature of $(4/11)^{1/3}T_\circ = 1.945$ K is the current neutrino background temperature. After the pair plasma annihilates the energy density of the photons plus neutrinos is $u = aT_\gamma^4 + 3 \times (7/8) \times aT_\nu^4 = (1 + (21/8)(4/11)^{4/3})aT_\gamma^4 = 1.68aT_\gamma^4$. There are three species of neutrinos, and they are fermions. But there is only one spin state allowed for neutrinos and one for anti-neutrinos. Thus the ratio to photons is $3 \times (7/8)$. Before the pair plasma annihilates the energy density is $u = (1 + (21/8) + (7/4))aT^4 = 5.375aT^4$ and there is only one temperature. When the Universe is radiation dominated and $\Omega_r = 1$, $H = 0.5/t$ and $u = \rho_{crit}c^2 = 3H^2c^2/(8\pi G) = 3c^2/(32\pi G t^2)$. Thus

$$t = \frac{1.78 \times 10^{20}\ \mathrm{K}^2\,\mathrm{sec}}{T_\gamma^2} \qquad (382)$$

after the pair plasma annihilates, and

$$t = \frac{0.995 \times 10^{20}\ \mathrm{K}^2\,\mathrm{sec}}{T^2} \qquad (383)$$

before the pair plasma annihilates but after the muons, pions and other exotics have disappeared. Figure 43 shows the time history of the temperature of the Universe, with a kink at 100 seconds when the pairs annihilate and a break at $10^{4.5}$ years when the Universe became matter dominated.

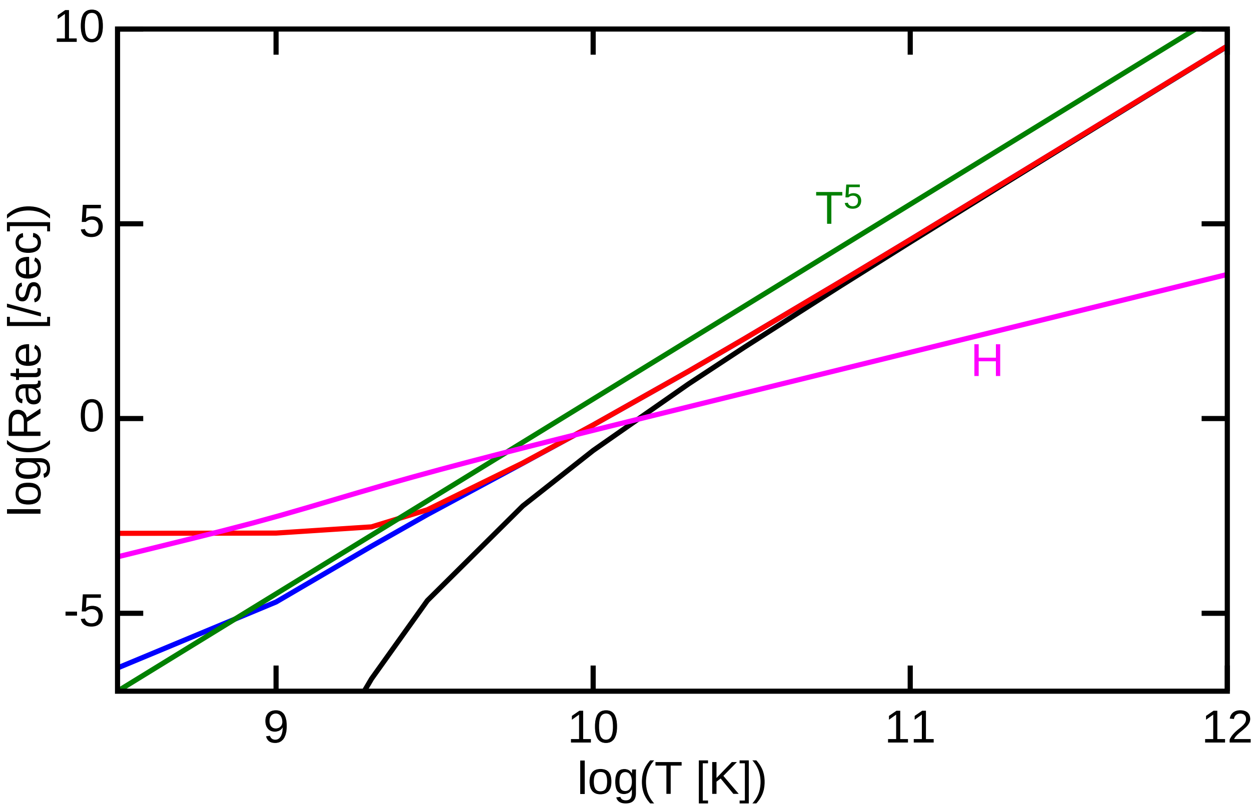

Fig. 44.— Rates for the expansion of the Universe, H, in magenta; and nuclear reaction rates $n \rightarrow p$, in blue without the free decay and red with the free neutron decay; and for $p \rightarrow n$ in black. A T^5 line is shown in green.

36. Big Bang Nucleosynthesis

The formation of the light elements (H, D, He and Li) is one of the three confirmed predictions of the hot Big Bang model in cosmology. This model was originally proposed as an explanation for the formation of all the elements by Gamow, Herman and Alpher. But the absence of any stable nuclei with atomic mass number A = 5 makes it impossible to proceed past Li in the brief time and relatively low baryon density available during the Big Bang.

As a result, the formation of carbon and other heavy elements proceeds through the triple-α reaction in the centers of red giants. The conditions where this occurs have densities of 10^7 gm/cc and temperatures of 10^8 K, and time scales of 10^{14} sec.

When the Universe makes He the temperature is about 10^9 K but the density of baryons is only $2 \times 10^{-29} \Omega_B h^2 (T/T_\circ)^3$ gm/cc $= 2 \times 10^{-5}$ gm/cc and the timescale is only 3 minutes, so the probability of a three body collision occurring is negligible even though the temperature is high.

Because of the low matter density, we need to concentrate on reactions that involve mainly abundant particles, which are the photons, neutrinos and the electron-positron plasma (for $T >$

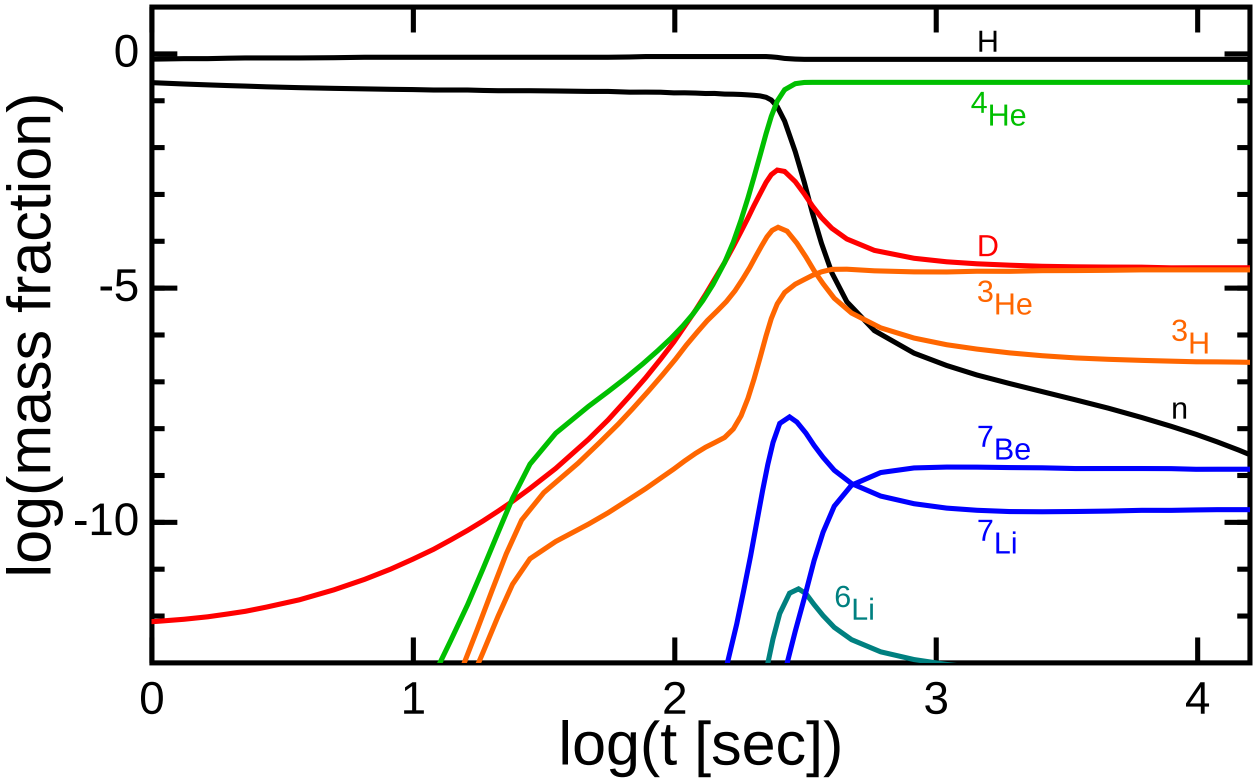

Fig. 45.— Mass fraction of various isotopes *vs.* time during the first few minutes after the Big Bang. This run is for a baryon density of $\Omega_b h^2 \approx 0.03$ which is higher than the best fit.

10^9).

The thermal equilibrium between neutrons and protons at $t \approx 1$ sec is maintained by weak interactions. The rate for the reaction $\nu + n_e \rightarrow p + e^-$ is given by

$$\langle n\sigma v\rangle \propto T^3 E^2 c \propto T^5 \tag{384}$$

The Hubble constant is the rate of expansion of the Universe and it is $\propto T^2$. Thus the weak interactions freeze out when when the weak interation rate $\langle n\sigma v\rangle$ equals H at about $T_f \approx 10^{10}$ K, and this leaves the $n/(p+n) \approx 0.14$. These neutrons then undergo the standard decay of free neutrons with a mean lifetime of 887 ± 2 seconds (Copi *et al.*, 1995, Science, 267, 192) until the temperature falls enough to allow deuterium to form.

The binding energy of deuterium is 2.2 MeV, and the temperature at the freeze out of the weak interactions is only 1 MeV, so one might expect that deuterium would form quite readily. But the reaction $p + n \leftrightarrow d + \gamma$ has two rare particles on the left hand side and only one rare particle on the right. Since the photon to baryon ratio is about 3×10^9, deuterium will not be favored until $\exp(-\Delta E/kT) = 10^{-9.5}$ which occurs when $T = 10^{9.1}$ K. This happens when $t = 10^{2.1}$ seconds. As a result about 14% of the neutrons decay into protons before they form deuterons, leaving a net neutron fraction of $0.14 \times (1 - 0.14) = 0.12$. Essentially all of these deuterons get incorporated into He^4, so the final helium abundance by weight is $Y_{pri} \approx 0.24$. This number depends only

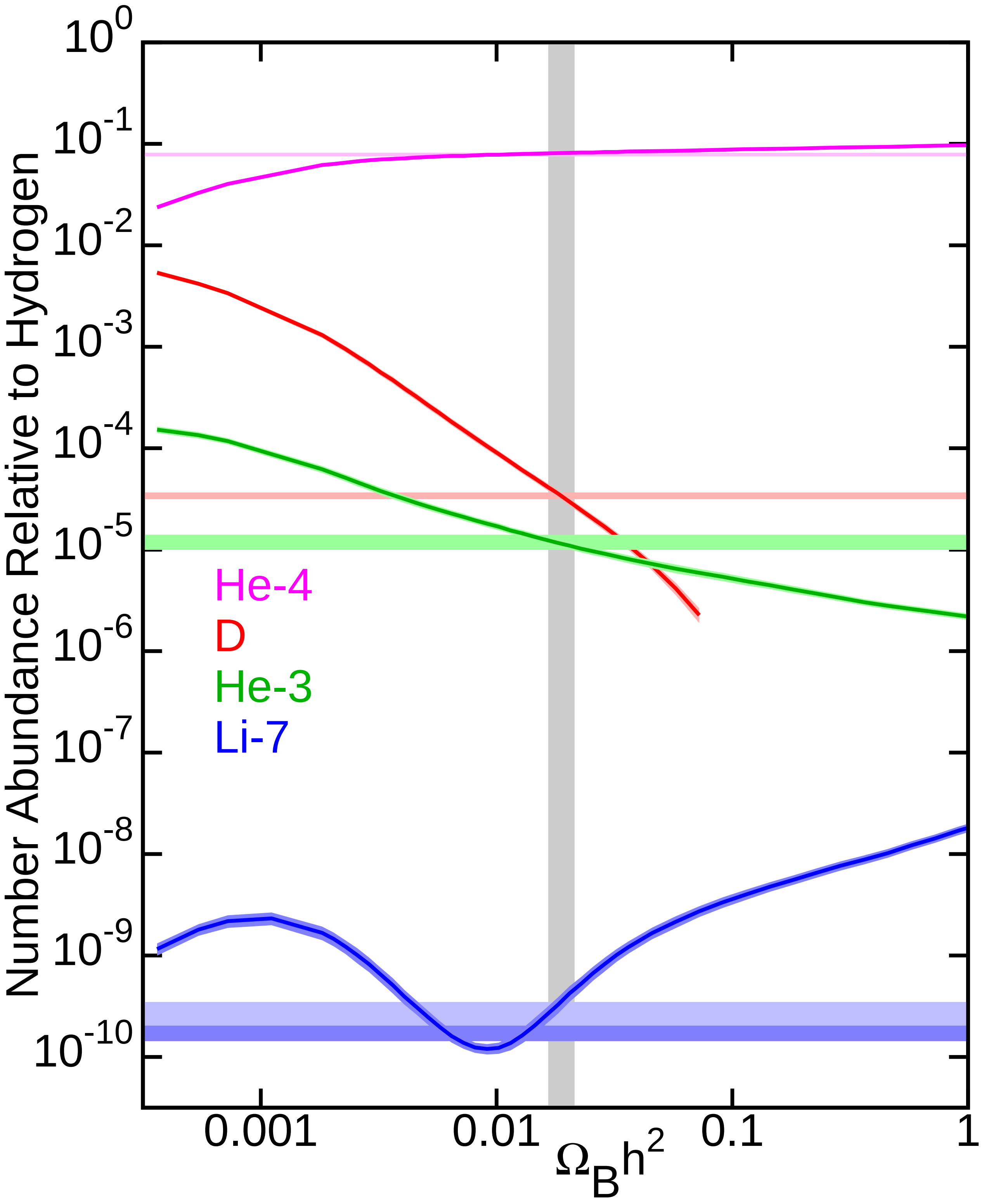

Fig. 46.— Final abundances of the light elements as a function of the baryon density. The horizontal bands show the observed values, and the vertical band shows the best fit value of the baryon density.

weakly on the photon to baryon ratio, and is determined primarily by the strength of the weak interactions, the neutron-proton mass difference, and the number of particle types with masses less than 1 MeV that contribute to the expansion rate of the Universe during the weak freeze out. The close agreement between the predicted 24% and the observed value is a very important confirmation of the hot Big Bang model. A 20% change in the weak interaction rates or the expansion rate of the Universe during the first 3 minutes after the Big Bang would destroy this agreement. Figure 45 shows the time evolution of the abundances of various isotopes during the first few minutes after the Big Bang.

The deuterium abundance can be used to determine the baryon density of the Universe. We can simplify the reaction network that makes helium to $d + d \rightarrow \mathrm{He} + \gamma$. The binding energy is 24 Mev, so the reverse reaction will not occur since deuterium doesn't form until $kT < 100$ keV. For this one reaction, we get the following equation for the deuteron fraction X_d:

$$\frac{dX_d}{dt} = -2\alpha(T) n_B X_d^2 = -2\alpha(T_1(t/t_1)^{1/2}) n_B(t_1)(t/t_1)^{-3/2} X_d^2 \tag{385}$$

where $\alpha(T)$ is the "recombination coefficent" for deuterons which will be small for high T, peak at intermediate T, and then be exponentially suppressed at low T by the Coulomb barrier. This equation has the solution

$$X_d^{-1} = 2n_B(t_1)\int_{t_1}^{\infty} \alpha(T_1(t/t_1)^{-1/2})(t/t_1)^{-3/2} dt + X_d(t_1)^{-1} \tag{386}$$

The time versus temperature is given by

$$\begin{aligned} \frac{1.68aT^4}{c^2} &= \frac{3}{32\pi G t^2} \\ t &= \frac{1.78 \times 10^{20}\ \mathrm{K}^2}{T^2} \end{aligned} \tag{387}$$

since the deuterium forms after the annihilation of the thermal e^+e^- plasma. Almost all of the deuterium will be swept up into helium so the final deuterium abundance is only slightly dependent on $X_d(t_1)$. Changing variables to $T = T_1\sqrt{t_1/t}$ gives

$$X_d = \frac{1}{n_B(t_1)} \frac{T_1}{4t_1 \int_0^{T_1} \alpha(T) dT} = \frac{T_1^3}{n_B(t_1)} \frac{1}{7.1 \times 10^{20}\ \mathrm{K}^2 \int_0^{T_1} \alpha(T) dT} \tag{388}$$

and is thus inversely proportional to the baryon to photon ratio since $n_\gamma(t_1) \propto T_1^3$. This ratio is usually quoted in terms of $\eta = n_B/n_\gamma$, or in terms of $\eta_{10} = 10^{10}\eta$. Since the photon density is known to be $n_\gamma = 411$ cm^{-3} for $T_\circ = 2.725$ K, we find $n_B(t_\circ) = 0.411\eta_{10} \times 10^{-7}$/cc. The critical density for $H_\circ = 100$ km/sec/Mpc corresponds to $n_B = 1.12 \times 10^{-5}$/cc, so

$$\Omega_B h^2 = \frac{0.411\eta_{10} \times 10^{-7}}{1.12 \times 10^{-5}} = 0.00367\eta_{10} \tag{389}$$

For $\eta_{10} > 7$ some D is converted into He^3, but the sum of D+He^3 continues to follow an inverse baryon density law.

A small amount of Li^7 is also produced in the Big Bang, and the predicted abundance agrees with the observed abundance in stars with very low metallicity which should have close to primordial abundances, as long as the stars have radiative envelopes. Convective envelopes carry the lithium down to hot regions of the star, and lithium is destroyed at high temperatures.

Comparison of observed abundances with predicted abundances gives an allowed range of baryon abundances of $\eta_{10} = 5.9 \pm 0.5$ (Kirkman *et al.*, astro-ph/0302006), which corresponds to $\Omega_B h^2 = 0.0214 \pm 9.5\%$. Schramm & Turner (1997, astro-ph/9706069) gave $\eta_{10} = 6 \pm 1$ or $\Omega_B h^2 = 0.022 \pm 0.004$. Burles, Nollett, Truran & Turner (1999, PRL, 82, 4176) gave $\eta_{10} = 5.1 \pm 0.5$ or $\Omega_B h^2 = 0.019 \pm 0.024$ so the baryon density is well constrained. Figure 46 shows the abundances of various isotopes as a function of the baryon density. For $H_\circ = 71$ this gives $\Omega_B = 0.044$ which is much less than the measured Ω and very much less than 1, so the Universe is primarily made of matter which did not take part in the reactions leading to light elements. Thus most of the Universe must be *non-baryonic* dark matter.

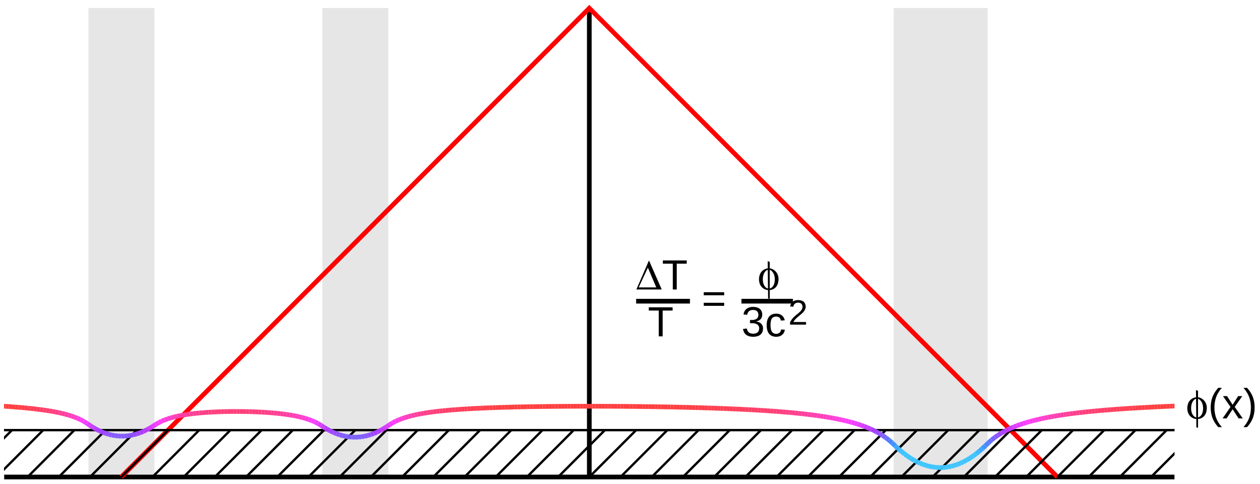

Fig. 47.— A conformal space-time diagram illustrating the Sachs-Wolfe effect. Dense regions have negative gravitational potentials leading to cold spots in the CMB sky.

37. Inhomogeneities & Anisotropies

The Universe is *almost* homogeneous and isotropic, but obviously not perfectly homogeneous since we are here with a density 10^{29} times higher than the average density of the Universe. So what is the nature of the deviations from homogeneity and isotropy?

We will see that

- When $\Omega = 1$ and pressure gradients are not important, the density contrast grows in a way that gives a constant perturbed gravitational potential, $\Delta\phi = \text{const}$, during the expansion of the Universe.

- During the inflationary epoch, the Steady State like nature of the Universe during inflation gives a perturbation amplitude $\Delta\phi$ that is independent of the the physical scale.

- For large angular scales, the anisotropy of the CMB is given by $\Delta T/T = (1/3)\Delta\phi/c^2$, the Sachs-Wolfe effect.

- For smaller angular scales that were less than the Hubble radius at last-scattering, the effects of pressure gradients cause the density contrast to oscillate instead of grow.

- These oscillations have been seen in the small angular scale anisotropy of the CMB.

37.1. $\Delta\phi$ Constant During Expansion

The perturbed gravitational potential is given by

$$\Delta\phi = -\frac{G\Delta M}{R} = -\frac{4\pi G\Delta\rho R^3}{3R} \tag{390}$$

This can be compared to the energy equation for a homogeneous Universe:

$$\begin{aligned} 2E &= \dot{R}^2 - \frac{8\pi G\rho R^3}{3R} = \text{const} \\ \text{const} &= \frac{8\pi G\rho R^2}{3}\left(\frac{3(\dot{R}/R)^2}{8\pi G\rho} - 1\right) \\ \text{const'} &= \rho a^2\left(\Omega^{-1} - 1\right) \end{aligned} \tag{391}$$

As long as $\Omega \approx 1$, we can write $\Omega^{-1} - 1 \approx -\Delta\rho/\rho$. We can treat large regions of the Universe with sizes bigger than ct as independent homogeneous Universes. Thus we get

$$\Delta\phi = -(4\pi G/3)\Delta\rho R^2 = (4\pi G/3)(\Omega^{-1} - 1)\rho R^2 = \text{const} \tag{392}$$

37.2. $\Delta\phi$ Constant *vs.* Scale

During inflation quantum fluctuations create a constant power spectrum of potential perturbations. The Universe is in a Steady State like phase, so the amplitude of potential fluctuations at scale c/H is independent of time. But the exponential expansion of the Universe pulls the scale c/H to $2.71828 \times c/H$ in a time interval of $1/H$. Of course the amplitude at scale c/H is still the same. These scales are larger than the horizon and evolve without any effects of pressure gradients. Ω is very close to one, so the amplitude does not change due to expansion. Another $\Delta t = 1/H$ later, these scales are at $(e^2 = 7.4) \times c/H$ and $(e^1 = 2.72) \times c/H$. After 70 e-foldings all scales from c/H to $10^{30} \times c/H$ have the same amplitude of potential fluctuations.

37.3. $\Delta T/T = (1/3)\Delta\phi/c^2$

The easiest way to look at the effect of density perturbations on the CMB anisotropy is to use a *Newtonian* gauge, which specifies the metric to be in the form:

$$ds^2 = (1 + 2\psi/c^2)c^2dt^2 - a(t)^2(1 - 2\phi/c^2)(dx^2 + dy^2 + dz^2) \tag{393}$$

The potentials ψ and ϕ will actually be equal for scales larger than c/H. Both are then given by $\nabla^2\phi = 4\pi G\Delta\rho$. If we assume a scale size R, this gives $\phi \approx 4\pi G\Delta\rho R^2 = 1.5(8\pi G/3)(\Delta\rho/\rho)$ If $R = c/H$, then we get $\phi/c^2 = 1.5\Omega(\Delta\rho/\rho)$. Thus for scales larger than c/H, $\Delta\rho/\rho \ll \phi/c^2$. We can ignore the real density contrast for the scales observed by COBE.

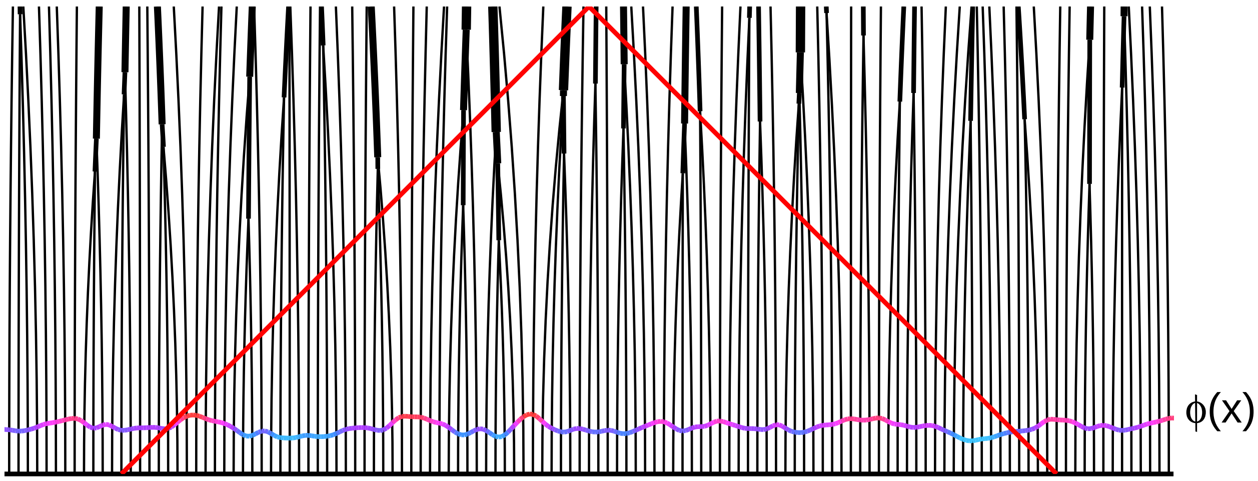

Fig. 48.— A conformal space-time diagram showing galaxies moving in the gravitational field produced by the gradient of the potential which can be measured using $\Delta T/T$.

However, for the Newtonian gauge metric, the time variable is not the proper time for comoving observers. A perturbation ϕ/c^2 gives a proper time τ that is $1+\phi/c^2$ times larger than the coordinate time t. This greater proper time makes the scale factor $a \propto t^{2/3}$ larger by a factor $1+(2/3)\phi/c^2$ and the local temperature cooler by a factor of $1-(2/3)\phi/c^2$. When combined with the gravitational redshift factor $1+\phi/c^2$ the net result is $\Delta T/T = (1/3)\phi/c^2$. This result is called the Sachs-Wolfe (1967, ApJ, 147, 73) effect.

37.4. Equal Power on All Scales

Because the potential is the same on all scales, and $\Delta T/T$ is proportional to the potential, one expects $\Delta T/T$ to be the same on all scales. One normally expresses the anisotropy map as

$$\frac{\Delta T(l,b)}{T} = \sum a_{\ell m} Y_{\ell m}(l,b) \tag{394}$$

where l, b are angular coordinates on the sky [specifically galactic coordinates in this case]. Since there are $\Delta\ell \approx \ell$ spherical harmonic orders corresponding to a given scale $\theta = 180°/\ell$, and there are $2\ell+1$ values of m for each ℓ, the variance on the sky is roughly $\ell(2\ell+1) < a_{\ell m}^2 > /4\pi$ at a given scale. [The mean value of $Y_{\ell m}^2$ is $1/4\pi$.] The mean value $< a_{\ell m}^2 >$ is called C_ℓ, the angular power spectrum. For equal power on all scales, $C_\ell \propto 1/[\ell(\ell+1)]$. It is usual to plot $\ell(\ell+1)C_\ell/2\pi$ is make an equal power on all scales spectrum appear as a horizontal line. Another way to plot the angular power spectrum is as Q_{flat} or Q_{rms-ps}. Here Q is the RMS quadrupole on the sky, which is given by $Q^2 = 5[\ell(\ell+1)/6]C_\ell/4\pi$.

The "equal power on all scales" spectrum of perturbations is known as the Harrison-Zeldovich spectrum. The argument used to derive the H-Z spectrum was the $\Delta\phi$ should be a power law function of the scale, since there is no preferred scale in the problem. If the power in the power

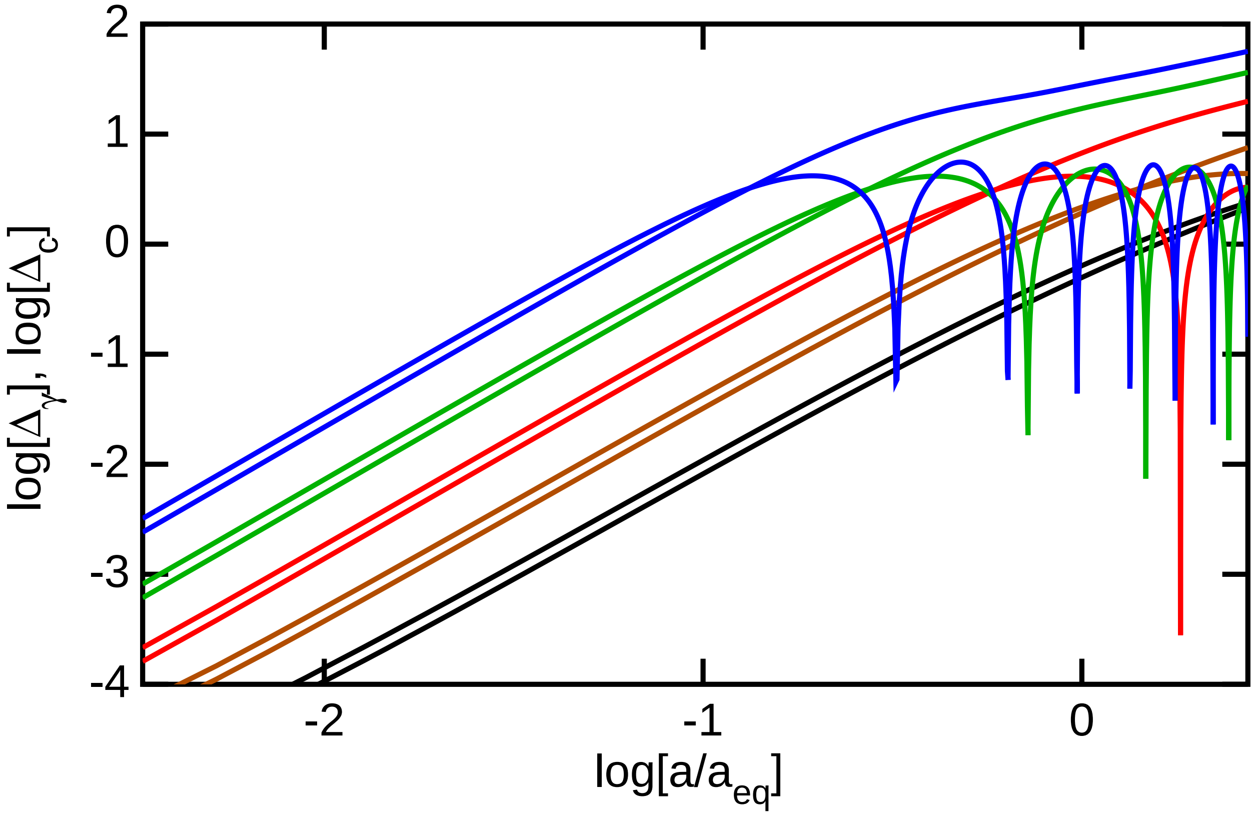

Fig. 49.— The density contrasts for cold dark matter and for the photon-baryon fluid prior to recombination for a model with $H_\circ = 65$, $\Omega_b = 0.05$ and $\Omega_m = 0.3$. Fives different scales $\kappa = 5, 10, 20, 40$ & 80 are shown from bottom to top at left where κ is the wavenumber in units of the horizon at recombination. The photon-baryon density contrast oscillates instead of growing.

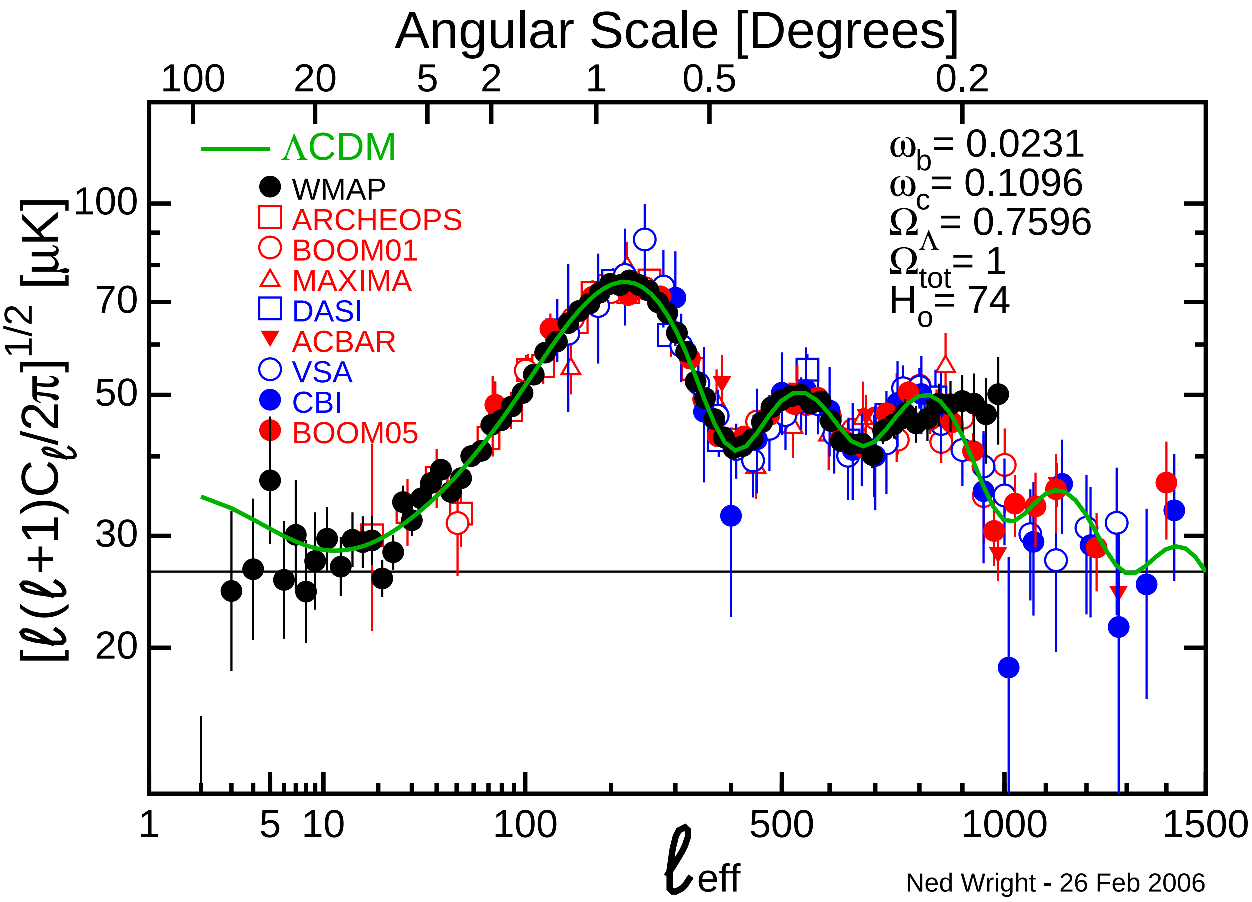

Fig. 50.— The angular power spectrum as of 2006. The data points show the results from several ground-based and balloon-borne experiments that have mapped small parts of the sky, and from the WMAP satellite that mapped the entire sky.

law is positive, then for large scales $\Delta\phi/c^2$ is large and the Universe is grossly inhomogeneous. If the power in negative then $\Delta\phi/c^2$ is large for small scales and many black holes will be formed. Neither is observed to happen, so the power has to be zero which is neither positive nor negative.

37.5. Smaller Scales

On scales smaller than c/H at the last scattering surface, $z \approx 1100$, pressure gradients cause the density contrast to oscillate instead of grow as the Universe expands. Since the density contrast is out of phase with the potential effect at the largest scales, there will be a particular scale which undergoes one-half cycle of this oscillation before recombination, and this scale will have a much larger amplitude of temperature fluctuation than the larger scales. There will be a second peak for a wavelength that makes one full oscillation by recombination, but this will be weaker than both the first and third peaks because the baryon density is once again out of phase with the effect of the dark matter gravitational potential. The third peak will occur for an even smaller scale which undergoes 1.5 cycles of oscillation prior to recombination. Thus there will be series of acoustic peaks in the angular power spectrum.

Because the oscillations are acoustic, and the speed of sound is $c/\sqrt{3}$, the wavelength of the first peak is roughly $2/\sqrt{3}$ times the radius of the horizon on the last scattering surface. This gives a spherical harmonic order of $\ell_{pk} \approx 200$. The actual location of the first peak was measured by WMAP to be $\ell_{pk} = 220.1 \pm 0.8$.