Astronomy 275 Lecture Notes, Spring 2005

1. Cosmological Observations

1.1. 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

$$v = Kr + X\cos\alpha\cos\delta + Y\sin\alpha\cos\delta + Z\sin\delta \tag{1}$$

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. Figure 1 shows the velocities corrected for the motion of the galaxy, $v - (X\cos\alpha\cos\delta + Y\sin\alpha\cos\delta + Z\sin\delta)$, plotted *vs.* distance. 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

$$\vec{v}' = H_\circ \vec{r} - H_\circ \vec{r}_A = H_\circ(\vec{r} - \vec{r}_A) = H_\circ \vec{r}' \tag{2}$$

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 $\vec{v} = Mr^2\hat{r}$ 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 most likely range now is $H_\circ = 72 \pm 8$ km/sec/Mpc from the HST Key Project on the distance scale (Freedman *et al.*, 2001, ApJ, 553, 47) and 71 ± 3.5 km/sec/Mpc from CMB anisotropy studies. 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:

$$H_\circ = 100 \text{ km/sec/Mpc} = \frac{10^7 \text{ cm/sec}}{3.085678 \times 10^{24} \text{ cm}} = 3.2 \times 10^{-18} \text{ s}^{-1} = \frac{1}{9.78 \text{ Gyr}} \tag{3}$$

1.2. 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. Using HIPPARCOS observations of nearby subdwarfs to calibrate the main sequence of globular clusters gives 11.7 ± 1.4 Gyr (Chaboyer, Demarque, Kernan & Krauss 1998, Ap,J 494, 96). The globular clusters needed some time to form, so I will take $t_\circ = 12.2 \pm 1.5$ Gyr for this method.

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, and I will take $t_\circ = 12.5 \pm 1.5$ Gyr. Brad Hansen (of UCLA) *et al.*, 2002, astro-ph/0205087 give 12.7 ± 0.7 Gyr for the age of the white dwarf population in the globular cluster M4. 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.

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. Models consistent with the CMB anisotropy have a very narrow range of ages: $t_\circ = 13.7 \pm 0.2$ Gyr, but this is quite model-dependent.

The dimensionless product $H_\circ t_\circ$ can be used to discriminate among cosmological models. Taking $t_\circ = 13.7 \pm 0.2$ Gyr (the CMB value) and $H_\circ = 71 \pm 3.5$ km/sec/Mpc, this product is

$$H_\circ t_\circ = \frac{h t_\circ}{9.78 \text{ Gyr}} = 0.99 \pm 0.05. \tag{4}$$

1.3. 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, 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 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 hold 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.

1.4. 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

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

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 Mather *et al.* (1999, 512, 511), which finds $T = 2.72528 \pm 0.00065$ K for 0.5 mm to 5 mm wavelength.

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. 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.

1.5. Time Dilation

One consequence of the standard interpretation of the redshift is that the durations of lightcurves should be increased when looking at high redshift objects. This has been confirmed in the sample of 61 high redshift supernovae from Goldhaber and the Supernova Cosmology Project (2001, astro-ph/0104382) as shown in Figure 6. The width factor w is the ratio of the duration of the observed decay to a nominal decay duration. The stretch factor used to calibrate the Type Ia brightness is given by $s = w/(1+z)$. If the population of fast *vs.* slow decayers does not change with redshift one expects $w = (1+z)$. An L_1 norm (least sum of absolute values of errors) fit to the Goldhaber *et al.* data is shown and has the coefficients

$$w = 0.985(1+z)^{1.045 \pm 0.089}. \tag{6}$$

An L_1 fit is robust against outliers in the data, and the uncertainty in the exponent was evaluated using the half-sample bootstrap method. Clearly this fit is perfectly consistent with the Big Bang prediction of $w = (1+z)$ and 11 standard deviations away from the tired light prediction of $w = (1+z)^0$.

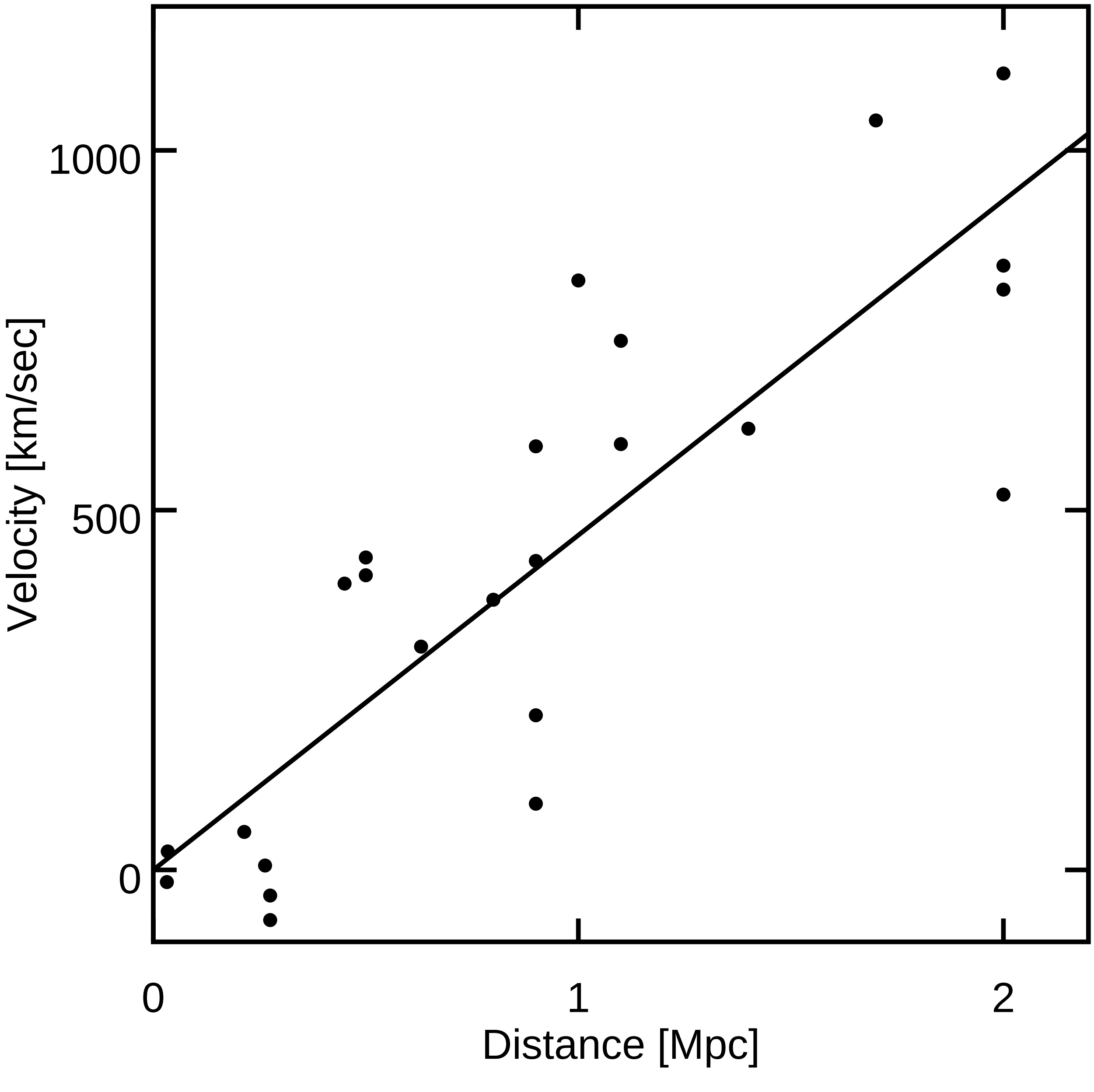

Fig. 1.— Hubble's data in 1929.

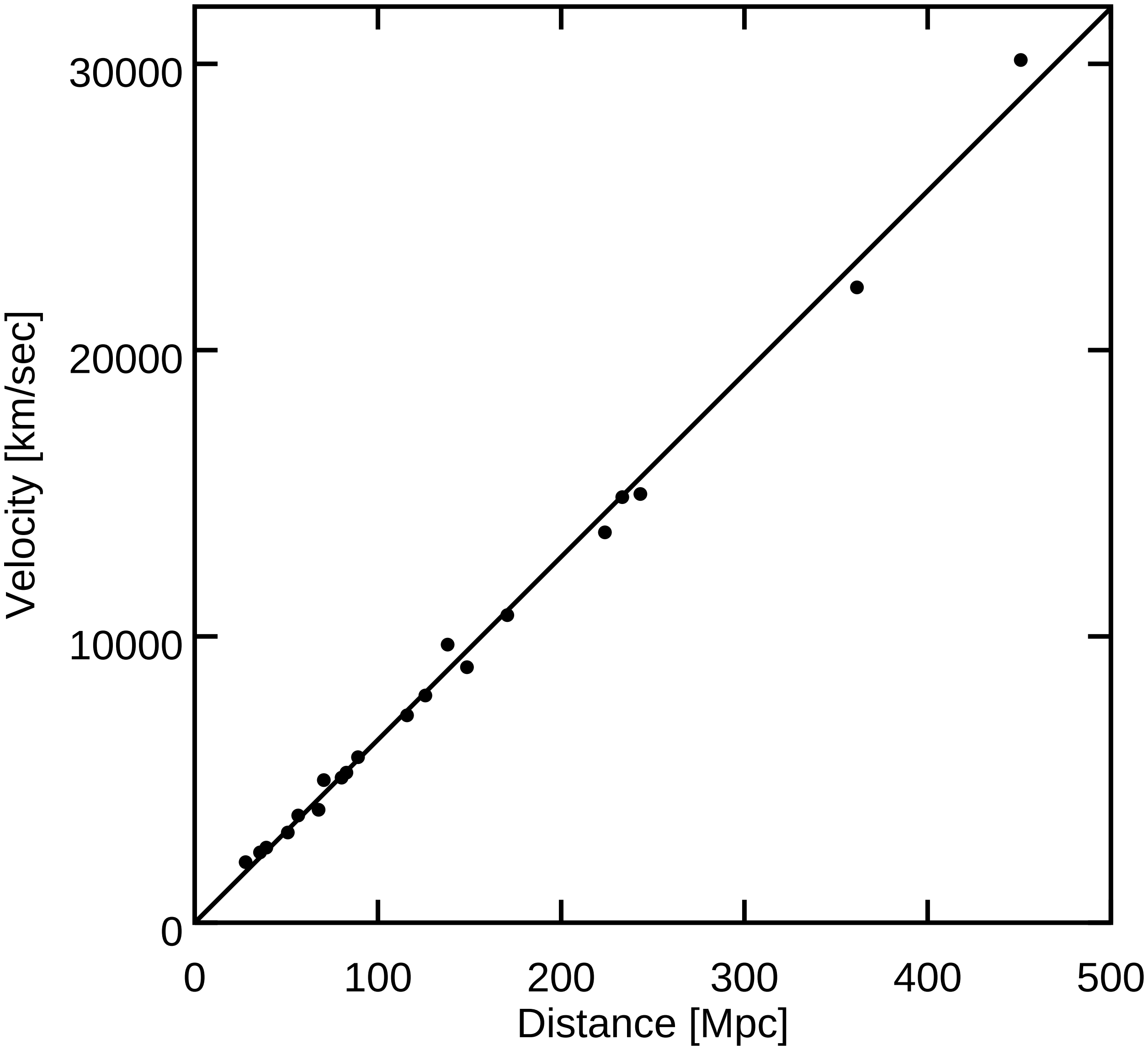

Fig. 2.— Distance *vs.* redshift for Type Ia Sne

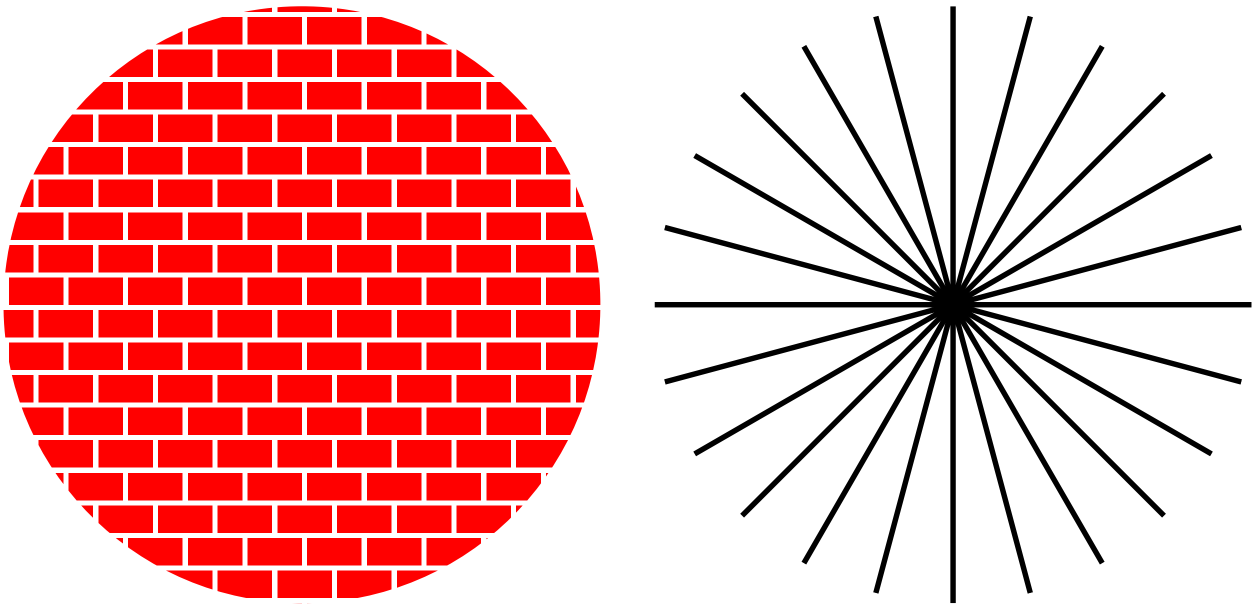

Fig. 3.— The brick wall on the left is a homogeneous but not isotropic pattern, while the pattern on the right is isotropic (about the center) but no homogeneous.

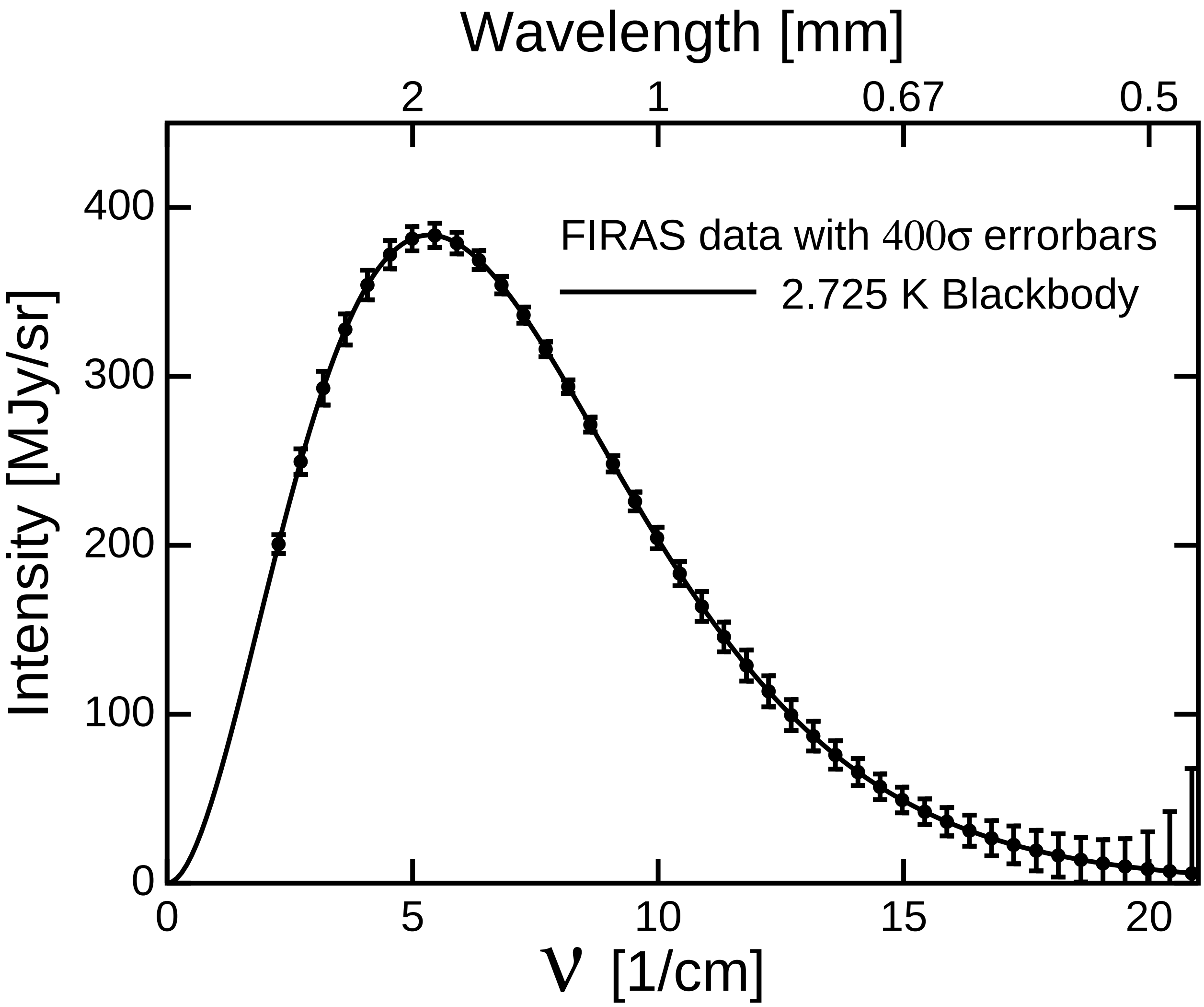

Fig. 4.— CMB Spectrum Measured by FIRAS on COBE.

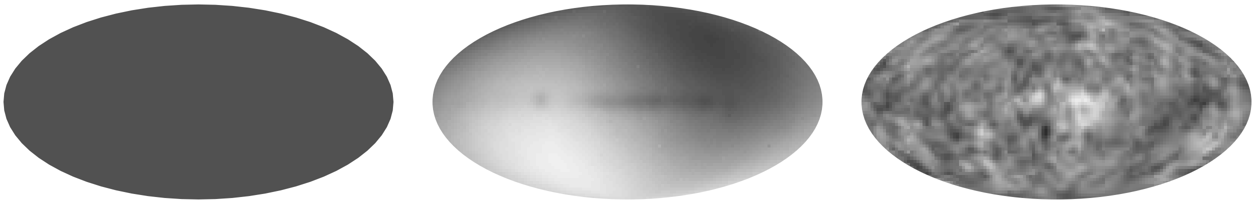

Fig. 5.— Left: true contrast CMB sky. 0 K = white, 4 K = black. Middle: contrast enhanced by 400X, monopole removed, showing dipole and Milky Way. Right: contrast enhanced by 6,667X, with monopole, dipole and Milky Way removed.

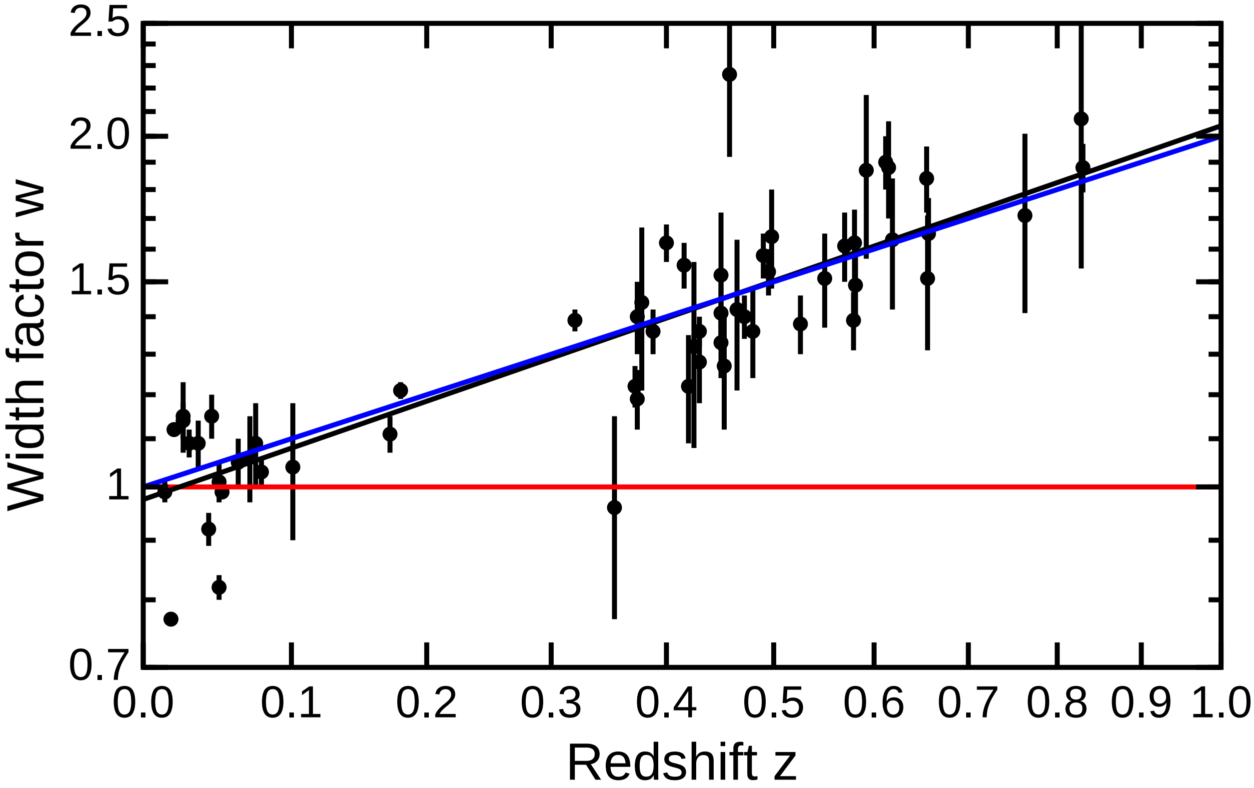

Fig. 6.— The time dilation (ratio of observed duration to nominal duration at $z = 0$) *vs.* redshift for high redshift supernovae.

2. 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.

3. 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

$$v = \frac{dD_{AB}}{dt} = HD_{AB} \tag{7}$$

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:

$$D_{AB}(t) = a(t)X_{AB} \tag{8}$$

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

$$H = \frac{1}{a}\frac{da}{dt}. \tag{9}$$

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.

3.1. 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,

$$\frac{\lambda_{obs}}{\lambda_{em}} = 1 + z = \sqrt{\frac{1+v/c}{1-v/c}}. \tag{10}$$

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

$$\frac{\lambda_n}{\lambda_B} = 1 + \frac{v}{c} = 1 + \frac{HD_{nB}(t_{em})}{c} = 1 + H(t_n - t_{em}) \tag{11}$$

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

$$\frac{\lambda_{n-1}}{\lambda_n} = \frac{a(t_{n-1})}{a(t_n)} \tag{12}$$

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

$$1 + z = \frac{a(t_A)}{a(t_1)}\frac{a(t_1)}{a(t_2)} \cdots \frac{a(t_n)}{a(t_B)} = \frac{a(t_A)}{a(t_B)} = \frac{a(t_{obs})}{a(t_{em})} \tag{13}$$

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 $a(t_{obs})/a(t_{em}) \approx 1$.

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.

3.2. Metrics

In General Relativity it is important to realize that coordinates are just used as names for events. An *event* is the analog of a point in space-time. We can name an event S, or R, or Z, or we can give it a name using 4 real variables, typically x, y, z and t. But we can change between different systems of coordinates, and GR gives us the rules for necessary transformations. In particular, GR defines a *metric* that is used to determine measurable distances and time intervals from coordinate differences. For example, in plane Euclidean geometry, the metric is

$$ds^2 = dx^2 + dy^2 \tag{14}$$

We can write this as

$$ds^2 = g_{xx}dx^2 + g_{xy}dxdy + g_{yx}dydx + g_{yy}dy^2 \tag{15}$$

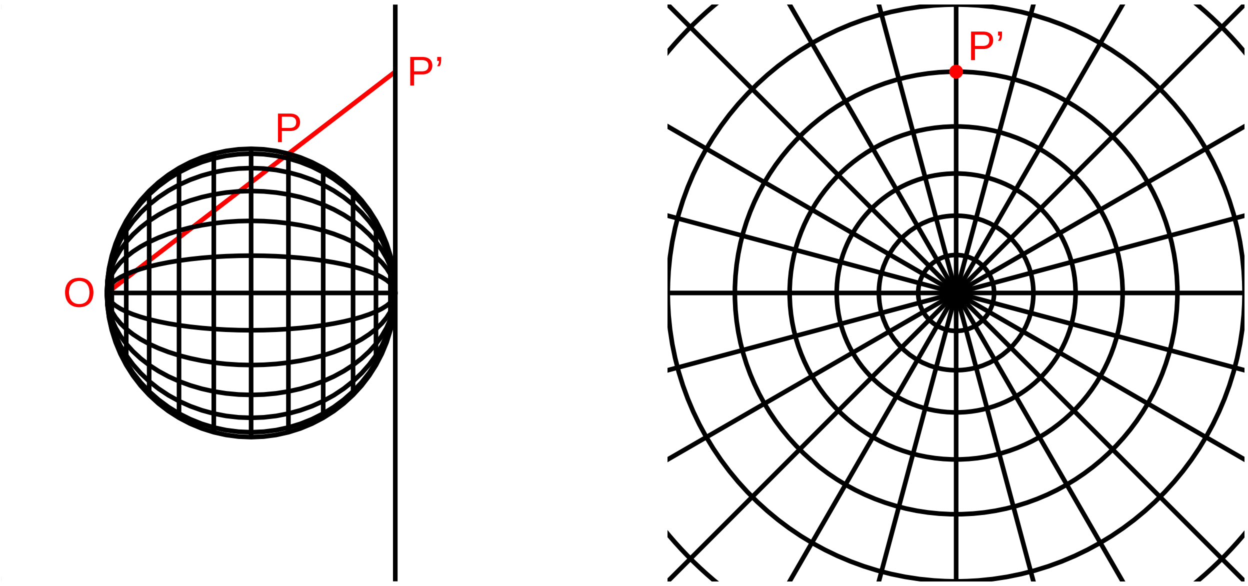

Fig. 7.— Projection used to make a stereographic map from a sphere onto an infinite plane.

with $g_{xx} = g_{yy} = 1$ and $g_{xy} = g_{yx} = 0$. This two index *tensor* $g_{\mu\nu}$ is called the metric. If you transform to rotated and shifted coordinates

$$\begin{aligned} x' &= a + x\cos\theta - y\sin\theta \\ y' &= b + x\sin\theta + y\cos\theta \end{aligned} \quad (16)$$

then $ds^2 = dx'^2 + dy'^2$ which has the same form as Eqn (14). Thus translations and rotations are *isomorphisms* of Euclidean geometry. But other coordinate transformations do change the form of the metric. For example, in polar coordinates

$$ds^2 = dr^2 + r^2 d\theta^2 \quad (17)$$

but this metric still describes Euclidean plane geometry. The metric

$$ds^2 = d\theta^2 + \sin^2\theta d\phi^2 \quad (18)$$

is differs only slightly from Eqn (17) but it describes a *spherical* geometry which is non-Euclidean. The tools of tensor calculus show how to compute the curvature of manifolds described by a metric, and determine whether the manifold is really curved or not.

The metrics needed in cosmology have to be homogeneous and isotropic. Homogeneity means that translations in all three spatial directions have to be isomorphisms. Isotropy means that rotations about all three axes have to isomorphisms. These powerful restrictions mean that there are only three possible spatial manifolds that satisfy the cosmological principle. These can all be described by the metric

$$ds^2 = R_\circ^2 \left(\frac{dr^2}{1 - kr^2} + r^2 \left[d\theta^2 + \sin^2\theta d\phi^2 \right] \right) \quad (19)$$

Here $R_\circ$ is the radius of curvature of the space, which has to be the same everywhere by homogeneity, and $k = +1, 0$ or -1. For $k = 0$ this metric describes Euclidean space, while $k = +1$ describes a 3-dimensional spherical space (the surface of a 4-dimensional ball – but the crutch of thinking about curved spaces embedded in a higher dimensional space is not very useful.) For $k = -1$ the space has negative curvature. Note that there are many different coordinate systems that can be used to describe the same geometry. By setting $\psi = \int dr/\sqrt{1-kr^2}$ we get

$$ds^2 = R_\circ^2 \left(d\psi^2 + \left\{ \begin{array}{c} \sin^2\psi \\ \psi^2 \\ \sinh^2\psi \end{array} \right\} \left[d\theta^2 + \sin^2\theta d\phi^2 \right] \right) \tag{20}$$

where the cases from top to bottom are for $k = +1, 0$ and -1. Yet another form for the homogeneous and isotropic spatial metrics can be found by analogy to the *stereographic* map projection. This is a projection from one end of a diameter of a sphere to a plane tangent to the other end of the diameter, as shown in Figure 7. If θ is the angle from the tangent point to a point on the sphere, then the angle at the projection point is $\theta/2$ and the radius on the projection plane is $r = 2\tan(\theta/2)$. Thus $dr = \sec^2(\theta/2)d\theta$ so the radial component of the metric, $ds^2 = d\theta^2$, can be written

$$ds^2 = \frac{dr^2}{\sec^4(\theta/2)} = \frac{dr^2}{(1+r^2/4)^2} \tag{21}$$

The tangential component of the metric on the sphere, $ds^2 = \sin^2\theta d\phi^2$, can be written as

$$ds^2 = \frac{r^2 d\phi^2}{\sec^4(\theta/2)} = \cos^4(\theta/2) 4\tan^2(\theta/2) d\phi^2 = [2\cos(\theta/2)\sin(\theta/2)]^2 d\phi^2 \tag{22}$$

Therefore the metric of a sphere can be written

$$ds^2 = \frac{dx^2 + dy^2}{(1 + (x^2+y^2)/4)^2} \tag{23}$$

where x and y are the coordinates on a stereographic map. The 3-dimensional version of this is

$$ds^2 = R_\circ^2 \frac{dx^2 + dy^2 + dz^2}{(1 + kr^2/4)^2} \tag{24}$$

where $k = +1, 0$ or -1 as before. Note that because the angle at the projection plane is $\theta/2$, the angle between the projected ray and the sphere normal is also $\theta/2$, so the distortion due to passing obliquely through the sphere is exactly canceled when passing obliquely through the projection plane. Thus this map is a *conformal* map: shapes of small regions are preserved. This implies that a small circle is mapped into a small circle, and not into an ellipse. In fact, any circle, no matter how big, is mapped into a circle. This includes great circles, which are the geodesics in this space.

The $k = -1$ version corresponds to projecting through a plane tangent to the vertex of one branch of the hyperbola from the other vertex of the hyperbola. The radius on the projection plane is given by

$$\begin{aligned}
r &= \frac{2\sinh\psi}{1+\cosh\psi} \\
1 - \frac{r^2}{4} &= \frac{1 + 2\cosh\psi + \cosh^2\psi - \sinh^2\psi}{(1+\cosh\psi)^2} = \frac{2}{1+\cosh\psi} \\
dr &= \left(\frac{2\cosh\phi}{1+\cosh\psi} - \frac{2\sinh^2\phi}{(1+\cosh\psi)^2} \right) d\psi
\end{aligned}$$

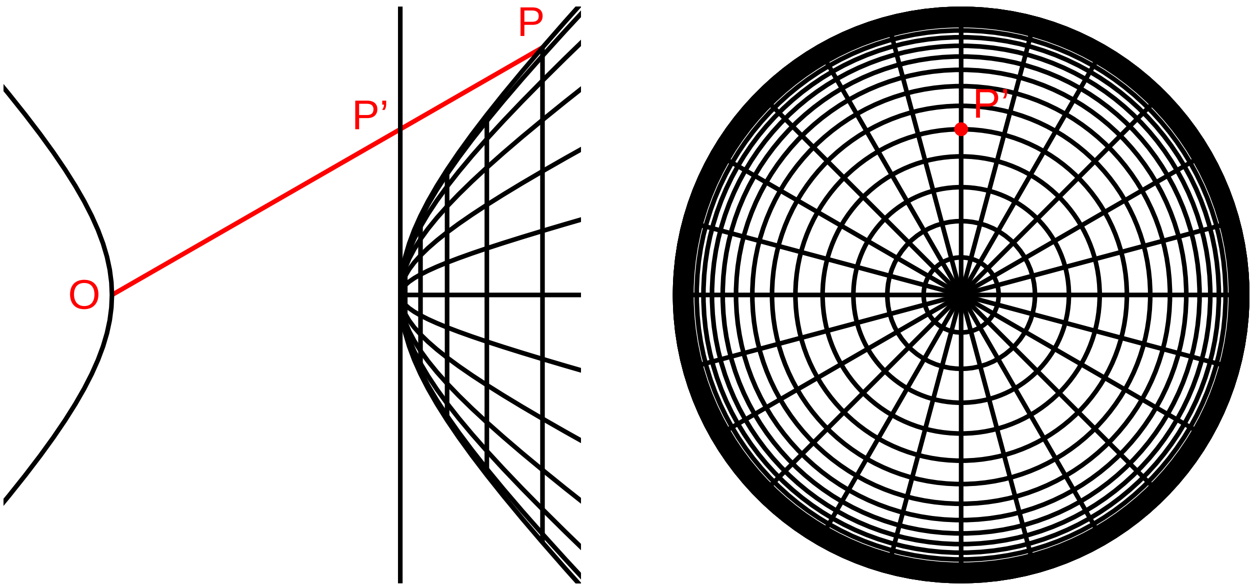

Fig. 8.— Projection used to make a "stereographic" map from a hyperboloid onto a circle.

$$\begin{aligned} &= \frac{2d\psi}{1+\cosh\psi} \\ ds = d\psi &= \frac{dr}{1-r^2/4} \end{aligned} \tag{25}$$

The tangential part of the $k = -1$ metric, $ds^2 = \sinh^2\psi d\theta^2$, becomes

$$ds^2 = \frac{r^2 d\theta^2}{(1-r^2/4)^2} \tag{26}$$

so the total metric is

$$ds^2 = \frac{dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)}{(1-r^2/4)^2} = \frac{dx^2+dy^2+dz^2}{(1-(x^2+y^2+z^2)/4)^2} \tag{27}$$

where the last step just substitutes Cartesian (x, y, z) coordinates for the spherical polar coordinates $r, \theta, \phi)$. This is the same as the $k = -1$ metric given above.

Note that the stereographic map of the finite spherical space is an infinite space, while the stereographic map of the infinite $k = -1$ negatively curved space is confined to a ball of radius 2.

Normally one visualizes the hyperbolic $k = -1$ space as a saddle-shaped thing. But here we have projected from a convex hyperboloid of revolution in Figure 8 onto a plane. The reason is that we have embedded the $k = -1$ space as the surface $x^2 + y^2 + z^2 - w^2 = -1$ in a flat Minkowski space with metric $ds^2 = dx^2 + dy^2 + dz^2 - dw^2$. This embedding of the $k = -1$ curved space into a Minkowski space is exact, but the usual saddle-shaped embedding ino a Euclidean space is only valid over a limited range. The six isometries corresponding to translations and rotations are generated by the Lorentz matrices in the group SL(4) applied to the hyperboloid $x^2 + y^2 + z^2 - w^2 = -1$. For the spherical $k = 1$ case, these six isometries are generated by the 4 dimensional rotation matrices in SO(4).

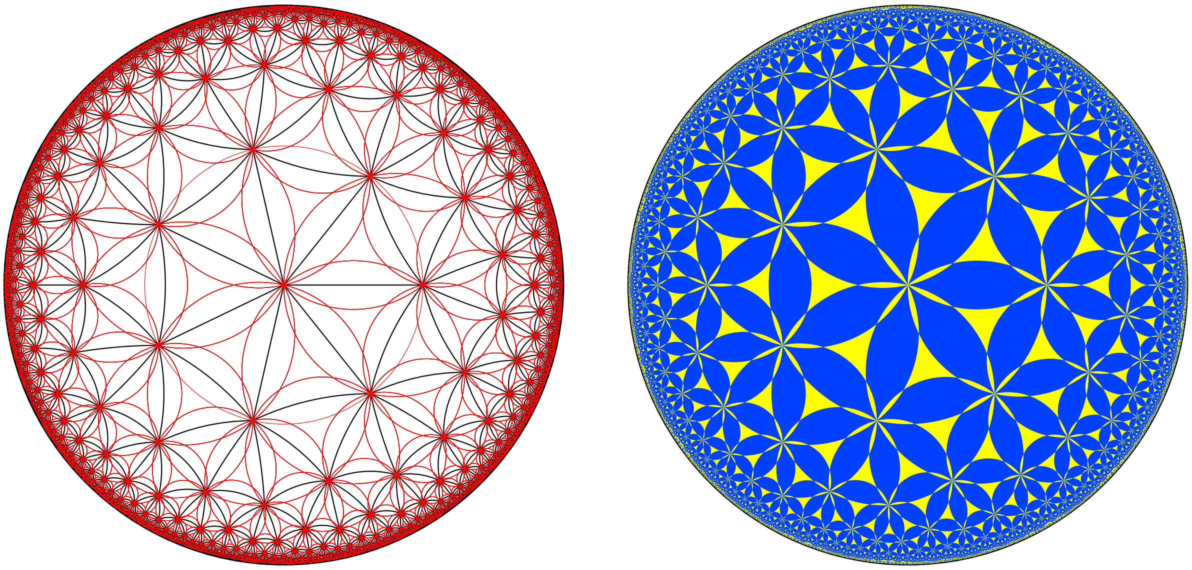

Fig. 9.— Seven fold symmetric pattern in $k = -1$ hyperbolic space, projected onto a circle, with apologies to M. C. Escher. Left: the triangulation of the space with equilateral triangles having angles of $(360/7)^\circ$ in black, with circles drawn around each group of 7 triangles in red. Right: filling the circles with UCLA's colors.

Because this form of the metric has a factor (which depends on position) times the metric for Euclidean space, these mappings are *conformal*. In geography and in cosmology this means that shapes of small regions are correctly portrayed. In particular, angles around a point are correct. The conformal factor in front of the Euclidean metric shows that areas (or volumes) are not preserved in these mappings. Also note that for any of the three cases the map of a circle is a circle. This is a useful property of the stereographic map of a globe.

In a negatively curve $k = -1$ space, an equilateral triangle has angles that are less than 60°. For one particular side length the angles are $(360/7)^\circ$, so 7 equilateral triangles can be placed around a point. Additional equilateral triangles can be added to the outside of the resulting regular heptagon leading to a tiling of the hyperbolic space with seven-fold symmetry. Drawing circles around each triangle vertex with radii all equal to the side length gives a beautiful pattern. In the positively curved $k = +1$ space there is a particular side length for an equilateral triangle such that the angles are $(360/5) = 72^\circ$, and this leads to a tiling of the sphere with five-fold symmetry, the twenty-sided regular polyhedron or icosahedron.

The metric of the expanding space-time that has homogeneous and isotropic spatial sections is

$$\begin{aligned} ds^2 &= c^2 dt^2 - a(t)^2 R_\circ^2 \left(\frac{dr^2}{1 - kr^2} + r^2 \left[d\theta^2 + \sin^2\theta d\phi^2 \right] \right) \\ &= c^2 dt^2 - R(t)^2 \left(\frac{dr^2}{1 - kr^2} + r^2 \left[d\theta^2 + \sin^2\theta d\phi^2 \right] \right) \end{aligned} \tag{28}$$

This form of the metric for an expanding homogeneous and isotropic Universe is called the Friedmann-Robertson-Walker or FRW metric. Comoving observers have constant (r, θ, ϕ) so these are called comoving coordinates. With $dr = d\theta = d\phi = 0$ for comoving observers, the proper time is just dt so the variable t is the proper time since the Big Bang for comoving observers – the cosmic time. When dealing with light cones it is often convenient to use the *conformal time*, defined by $d\eta = cdt/a(t)$. With this variable the FRW metric becomes

$$ds^2 = a(t)^2 \left[d\eta^2 - R_\circ^2 \left(d\psi^2 + S(\psi)^2 \left[d\theta^2 + \sin^2\theta d\phi^2\right]\right)\right] \tag{29}$$

where $S(x) = \sinh x, x$ or $\sin x$ for $k = -1, 0$ or $+1$. Because this looks like a factor times a Minkowski metric, at least for the variables η and $R_\circ\psi$, this is a "conformal" version of the FRW metric.

With the metric we can rederive the relationship between the scale factor $a(t)$ and the redshift z. The equation for a light ray is $ds = 0$, and with the observer at the origin we have $d\theta = d\phi = 0$, so

$$cdt = a(t) R_\circ dr/\sqrt{1-kr^2} \tag{30}$$

or

$$\int_{t_e}^{t_\circ} \frac{cdt}{a} = R_\circ \int_0^{r_s} \frac{dr}{\sqrt{1-kr^2}} \tag{31}$$

For a source fixed in comoving coordinates, r_s is fixed, so the RHS doesn't depend on time. A light pulse emitted at a slightly later time, $t_e + dt_e$ will be received at a time $t_\circ + dt_\circ$, where

$$\int_{t_e}^{t_\circ} \frac{cdt}{a} = \int_{t_e+dt_e}^{t_\circ+dt_\circ} \frac{cdt}{a} \tag{32}$$

which gives $dt_\circ/a(t_\circ) = dt_e/a(t_e)$. The observed wavelength is proportional to $dt_\circ$, while the emitted wavelength is proportional to dt_e, giving

$$1 + z = \frac{\lambda_{obs}}{\lambda_{em}} = \frac{dt_\circ}{dt_e} = \frac{a(t_\circ)}{a(t_e)} \tag{33}$$

4. 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

$$-\frac{dv}{dt} = g = \frac{GM}{R^2} = \frac{4\pi}{3} G\rho R \tag{34}$$

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:

$$\frac{v^2}{2} = E_{tot} + \frac{GM}{R} \tag{35}$$

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

$$E_{tot} = \frac{(H_\circ R_\circ)^2}{2} - \frac{4\pi G \rho_\circ R_\circ^2}{3} = \frac{(H_\circ R_\circ)^2}{2}\left(1 - \frac{\rho_\circ}{\rho_{crit}}\right) \tag{36}$$

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" do 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,

$$\rho_{crit} = \frac{3H_\circ^2}{8\pi G} = 1.879h^2 \times 10^{-29}\ \text{gm/cc} = 10,539h^2\ \text{eV/c}^2\text{/cc} \tag{37}$$

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

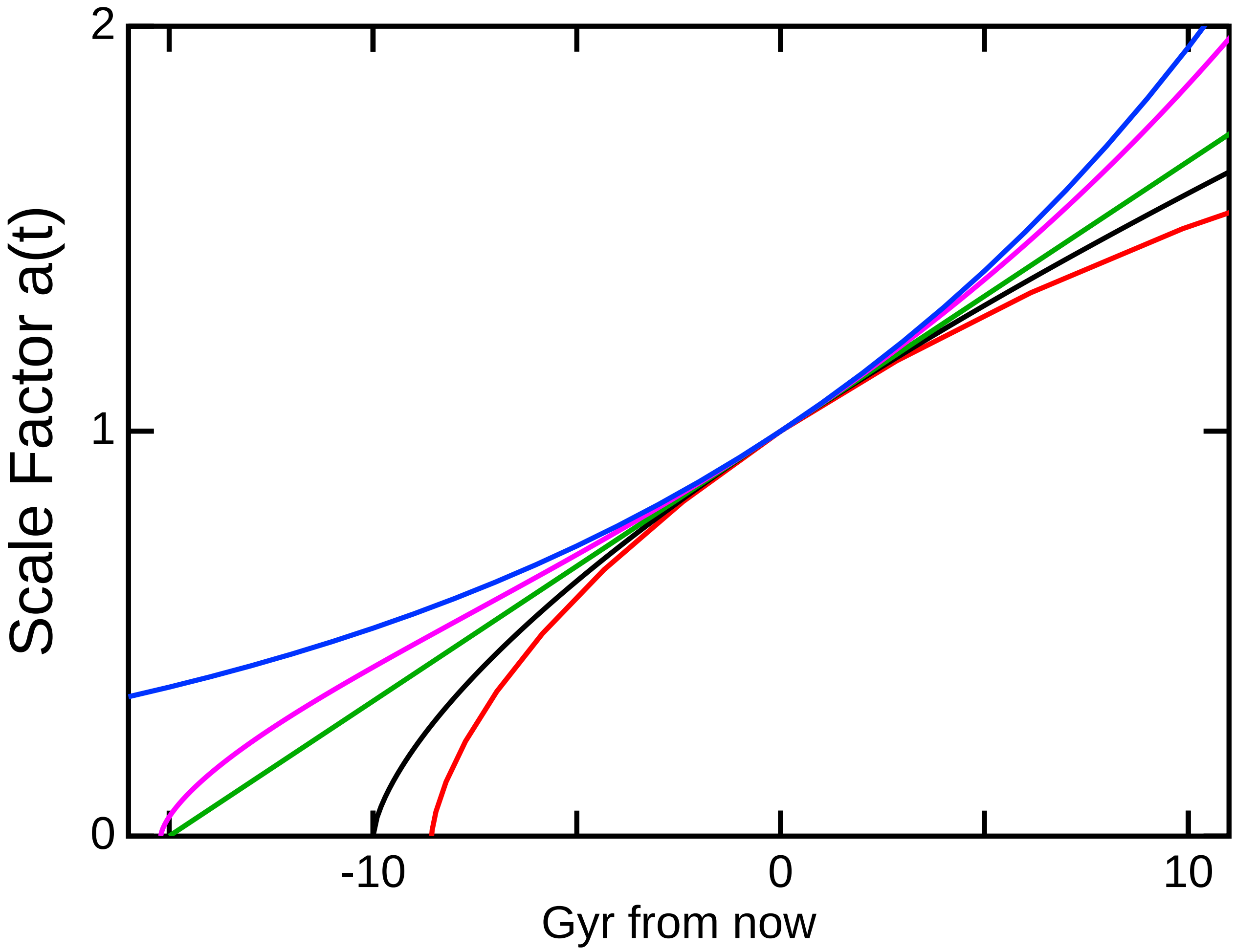

Fig. 10.— Scale factor *vs.* time for 5 different models: from top to bottom having $(\Omega_{m\circ}, \Omega_{v\circ}) = (0, 1)$ in blue, (0.25,0.75) in magenta, (0,0) in green, (1,0) in black and (2,0) in red. All have $H_\circ = 65$.

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

$$\Omega_{lum} = \frac{(M/L)/(M_\odot/L_\odot)}{1750h} \tag{38}$$

The mass-to-light ratio near the Sun is $(M/L)/(M_\odot/L_\odot) = 3.3$ so $\Omega_{lum} \approx 0.003$. Thus the density of luminous matter seems to be much less than the critical density.

At the critical density we have the simple equation

$$\left(\frac{dR}{dt}\right)^2 = \frac{2GM}{R} \tag{39}$$

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(35):

$$v^2 = H^2 R^2 = H_\circ^2 R_\circ^2 (1 - \Omega_\circ) + \frac{8\pi G \rho_\circ R_\circ^3}{3R} \tag{40}$$

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

$$\frac{H^2}{H_\circ^2} = \frac{R_\circ^2}{R^2}(1 - \Omega_\circ) + \frac{R_\circ^3}{R^3}\Omega_\circ \tag{41}$$

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

$$\frac{H^2}{H_\circ^2} = (1+z)^2(1-\Omega_\circ) + (1+z)^3\Omega_\circ = (1+z)^2(1+\Omega_\circ z) \tag{42}$$

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

$$\frac{dt}{dz} = -\frac{1}{(1+z)H} = -\frac{1}{H_\circ(1+z)^2\sqrt{1+\Omega_\circ z}} \tag{43}$$

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

$$H_\circ t_\circ = \int_0^\infty \frac{dz}{(1+z)^2\sqrt{1+\Omega_\circ z}} \tag{44}$$

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 values for the product $H_\circ t_\circ$ are 1-2 σ higher than the $\Omega = 1$ model's prediction.

A second thing we can compute is the time variation of Ω. From Eqn(35) we have

$$2E_{tot} = v^2 - \frac{2GM}{R} = H^2R^2 - \frac{8\pi G\rho R^2}{3} = \text{const} \tag{45}$$

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

$$\frac{3H^2}{8\pi G\rho} - 1 = \frac{\text{const}'}{\rho R^2} = \Omega^{-1} - 1 \tag{46}$$

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 of a particle moving radially in the gravitational field of a point mass means that we can use Kepler's equation from orbit calculations:

$$M = E - e\sin E \tag{47}$$

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

$$\begin{aligned} x &= a_{SM}(e - \cos E) \\ y &= a_{SM}\sqrt{1-e^2}\sin E \end{aligned} \tag{48}$$

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 $a(t)$:

$$\begin{aligned} t &= A(E - \sin E) \\ a &= B(1 - \cos E) \end{aligned} \tag{49}$$

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

$$\begin{aligned} \dot{a} &= \frac{da/dE}{dt/dE} = \left(\frac{B}{A}\right)\frac{\sin E}{1-\cos E} \\ \ddot{a} &= \left(\frac{B}{A^2}\right)\frac{\cos E(1-\cos E) - \sin^2 E}{(1-\cos E)^3} = -\left(\frac{B}{A^2}\right)\frac{1}{(1-\cos E)^2} \\ q &= \frac{-\ddot{a}a}{\dot{a}^2} = \frac{1-\cos E}{\sin^2 E} = \frac{1}{1+\cos E} \end{aligned} \tag{50}$$

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

$$H_\circ t_\circ = \frac{\dot{a}}{a}t_\circ = \frac{(E_\circ - \sin E_\circ)\sin E_\circ}{(1-\cos E_\circ)^2} \tag{51}$$

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

$$\frac{t_{BC}}{t_\circ} = \frac{2\pi}{\pi/2 - 1} = 11.008 \tag{52}$$

for $\Omega_\circ = 2$.

For an open Universe we change the parametric equation to

$$\begin{aligned} t &= A(\sinh E - E) \\ a &= B(\cosh E - 1) \end{aligned} \tag{53}$$

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

$$q_\circ = \frac{1}{1+\cosh E_\circ} \tag{54}$$

This gives

$$\begin{aligned} H_\circ t_\circ = \frac{\dot{a}}{a}t_\circ &= \frac{\sinh E_\circ\,(\sinh E_\circ - E_\circ)}{(\cosh E_\circ - 1)^2} \\ &= \frac{(E_\circ + (1/6)E_\circ^3 + \ldots)((1/6)E_\circ^3 + (1/120)E_\circ^5 + \ldots)}{((1/2)E_\circ^2 + (1/24)E_\circ^4 + \ldots)^2} \\ &= \frac{(1/6)E_\circ^4 + (13/360)E_\circ^6 + \ldots}{(1/4)E_\circ^4 + (1/24)E_\circ^6 + \ldots} = \frac{2}{3}\left(1 + \frac{1}{20}E_\circ^2 - \ldots\right) \end{aligned} \tag{55}$$

For example, if $q_\circ = 0.1$, then $\cosh E_\circ = 9$, $\sinh E_\circ = 8.994$, $E_\circ = 2.887$ and $H_\circ t_\circ = 0.846$.

For both the open and closed Universe cases, the variable E is proportional to the *conformal time* which is usually denoted by η. Conformal time follows the equation $dt = a(t)d\eta$ and we easily see that $dt = (A/B)a(t)dE$. For the closed Universe case the metric in terms of E and ψ is

$$ds^2 = a(E)^2 \left[c^2 \left(\frac{A}{B}\right)^2 dE^2 - R_\circ^2 d\psi^2\right] \tag{56}$$

but

$$R_\circ = \frac{c}{H_\circ\sqrt{|1-\Omega_\circ|}} = c\left(\frac{A}{B}\right)\frac{1-\cos E_\circ}{\sin E_\circ}\sqrt{\frac{1+\cos E_\circ}{1-\cos E_\circ}} = c\left(\frac{A}{B}\right) \tag{57}$$

so the metric is

$$ds^2 = R_\circ^2 a(E)^2 \left[dE^2 - d\psi^2\right] \tag{58}$$

In a conformal space-time diagram, which plots ψ as the spatial coordinate and E as the time coordinate, worldlines of light rays always have slopes of ± 1. Since the range of E is 0 to 2π from the Big Bang to the Big Crunch, and since the range of ψ is 0 to 2π for one trip around the closed Universe, we see that it takes light the entire time from Big Bang to Big Crunch to circumnavigate the Universe.

4.1. with Pressure

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

$$\ddot{R} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) R \tag{59}$$

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

$$\frac{\dot{R}^2}{2} = \frac{GM}{R} + E_{tot} \tag{60}$$

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

$$\dot{R}\ddot{R} = -\frac{GM}{R^2}\dot{R} + \frac{G}{R}\frac{dM}{dR}\dot{R} \tag{61}$$

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

$$\begin{aligned} \dot{R}\ddot{R} &= -\frac{4\pi G\rho}{3}R\dot{R} - 4\pi GPR\dot{R}/c^2 \\ &= -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right)\dot{R} \end{aligned} \tag{62}$$

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

$$P = \frac{\rho c^2}{3} \tag{63}$$

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(59) gives

$$\ddot{R} = -\frac{8\pi G}{3}\frac{\rho_\circ R_\circ^4}{R^3} \tag{64}$$

which becomes an energy equation

$$v^2 = 2E_{tot} + \frac{8\pi G}{3}\frac{\rho_\circ R_\circ^4}{R^2} \tag{65}$$

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 $R \propto t^{1/2}$.

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. 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.

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 $\rho = \text{const}$ during the expansion. This is reasonable, because if the density is due to quantum fluctuations, 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

$$\ddot{R} = \frac{8\pi G\rho}{3}R \tag{66}$$

The solutions of this equation are

$$a \propto \exp\left(\pm t\sqrt{\frac{8\pi G\rho}{3}}\right) \tag{67}$$

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

$$H = \sqrt{\frac{8\pi G\rho}{3}} \tag{68}$$

We see once again that the critical density is

$$\rho_{crit} = \frac{3H^2}{8\pi G}. \tag{69}$$

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.

4.2. 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

$$v^2 = 2E_{tot} + \frac{8\pi G R^2}{3}\left(\rho_{v\circ} + \rho_{m\circ}\frac{R_\circ^3}{R^3} + \rho_{r\circ}\frac{R_\circ^4}{R^4}\right) \tag{70}$$

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

$$H^2 = H_\circ^2\left([1-\Omega_{v\circ}-\Omega_{m\circ}-\Omega_{r\circ}](1+z)^2 + \Omega_{v\circ} + \Omega_{m\circ}(1+z)^3 + \Omega_{r\circ}(1+z)^4\right) \tag{71}$$

Using $H^{-1} = (1+z)dt/dz$ gives

$$H_\circ(1+z)\frac{dt}{dz} = \left([1-\Omega_{v\circ}-\Omega_{m\circ}-\Omega_{r\circ}](1+z)^2 + \Omega_{v\circ} + \Omega_{m\circ}(1+z)^3 + \Omega_{r\circ}(1+z)^4\right)^{-1/2} \tag{72}$$

while using $H = \dot{a}/a$ gives

$$\dot{a} = H_\circ\left([1-\Omega_{v\circ}-\Omega_{m\circ}-\Omega_{r\circ}] + \Omega_{v\circ}a^2 + \Omega_{m\circ}/a + \Omega_{r\circ}/a^2\right)^{1/2} \tag{73}$$

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

$$H_\circ t_\circ = \int_0^\infty \frac{dz}{(1+z)\sqrt{1+\Omega_{m\circ}(3z+3z^2+z^3)}} \tag{74}$$

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 $H_\circ t_\circ$ but I feel that a systematic error in the distances is a more likely explanation.

Note that the redshift actually depends on three parameters: the distance of the object, the observation time, and the time of emission. These three parameters are constrained by the requirement that light travel at speed c, but that still leaves two free parameters. The formulae for dt/dz given above assume that the time of observation is fixed, and that the distance to the source varies as a function of the emission time to satisfy the light speed constraint. We can also ask a very different question: for a source with fixed comoving distance, how does the redshift vary with observation time? In this case the emission time varies as a function of the observation time to satisfy the light speed constraint. In principle this is a way to directly measure the deceleration parameter of the Universe (Loeb, 1998 astro-ph/9802122).

In order to calculate the rate at which observed redshifts for comoving objects will change, we need to carry Eqn(32) to the next higher order, and in this order we need to use $a(t_e + \Delta t_e) = a(t_e)(1 + H(t_e)\Delta t_e + \ldots)$ and $a(t_o + \Delta t_o) = a(t_o)(1 + H(t_o)\Delta t_o + \ldots)$. This gives

$$\frac{c}{a_e}\left(\Delta t_e - \frac{1}{2}H(t_e)\Delta t_e^2 + \ldots\right) = \frac{c}{a_o}\left(\Delta t_o - \frac{1}{2}H(t_o)\Delta t_o^2 + \ldots\right) \tag{75}$$

Solving for Δt_e gives

$$\Delta t_e = \frac{a_e}{a_o}\left(\Delta t_o - \frac{1}{2}H(t_o)\Delta t_o^2\right) + \frac{1}{2}H(t_e)\Delta t_e^2 \tag{76}$$

We can use the first order solution $\Delta t_e = (a_e/a_o)\Delta t_o$ to eliminate the $(\Delta t_e)^2$ term, giving

$$\Delta t_e = \frac{a_e}{a_o}\left(\Delta t_o - \frac{1}{2}H(t_o)\Delta t_o^2\right) + \frac{1}{2}H(t_e)\left(\frac{a_e}{a_o}\right)^2 \Delta t_o^2 \tag{77}$$

The redshift is

$$1 + z = \frac{d\Delta t_o}{d\Delta t_e} = \frac{(a_o/a_e)}{1 - [H(t_o) + (a_e/a_o)H(t_e)]\Delta t_o} = \frac{a_o}{a_e} + \left[\frac{a_o}{a_e}H(t_o) - H(t_e)\right]\Delta t_o \tag{78}$$

Thus the rate of change of the redshift of a comoving object is

$$\begin{aligned}\frac{d(1+z)}{dt_o} &= (1+z)H_\circ - H(t_e) \\ &= H_\circ(1+z)\left(1 - \sqrt{([1-\Omega_{tot,\circ}] + \Omega_{v\circ}/(1+z)^2 + \Omega_{m\circ}(1+z) + \Omega_{r\circ}(1+z)^2)}\right)\end{aligned} \tag{79}$$

For example, consider a source with $z = 3$ in a Universe with $\Omega_{m\circ} = 1$. We get $dz/dt_o = H_\circ(1 + z)(1 - \sqrt{1+z}) = -4H_\circ$. This is negative because this model is decelerating, so redshifts decrease with time. Unfortunately, the velocity change associated with this redshift change is only $dv/dt = c(dz/dt)/(1+z) = -2$ cm/sec/yr for $H_\circ = 65$, so it will be very difficult to measure.

5. Flatness-Oldness

Even in the general case with radiation, matter and vacuum densities, the energy equation is still

$$2E_{tot} = v^2 - \frac{8\pi G\rho R^2}{3} = H^2R^2 - \frac{8\pi G\rho R^2}{3} = \text{const} \tag{80}$$

Thus we still get

$$\Omega^{-1} - 1 = \frac{\text{const}'}{\rho a^2} \tag{81}$$

Currently $\Omega_{r\circ} \approx 10^{-4}$ but for $z > 10^4$ the radiation will dominate the density.

In order to correctly calculate the density during the radiation dominated epoch at $z > 10^4$ we need to consider the *entropy density*, s. This is given by

$$s = V^{-1}\int \frac{dQ}{T} = \int \frac{d(aT^4)}{T} = \frac{4}{3}aT^3 \tag{82}$$

for a blackbody radiation field of photons. But there are also neutrinos of 3 types, each with associated anti-neutrinos. For fermions the energy density in thermal equilibrium is

$$u = g_s \int \frac{E(p)}{\exp(E(p)/kT)+1}\frac{4\pi p^2 dp}{h^3} \tag{83}$$

where g_s is the number of spin degrees of freedom. For neutrinos with spin 1/2, one would expect $g_s = 2$ but since only one helicity of neutrinos seems to exist we set $g_s = 1$. For massless particles $E = pc$ and we get

$$u = 4\pi g_s \left(\frac{kT}{hc}\right)^3 kT \int \frac{x^3 dx}{e^x+1} = \frac{7g_s}{16}aT^4 \tag{84}$$

Thus the entropy density of massless fermions is given by

$$s = \int \frac{du}{T} = \frac{4}{3}\frac{7g_s}{16}aT^3 \tag{85}$$

Since we have a total g_s of 6 for neutrinos and anti-neutrinos one might expect an additional contribution from neutrinos that is 42/16 times the photon entropy but the actual number is smaller because the photons were "heated" by the annihilation of the $e^\pm$ pairs as the temperature of the Universe fell below an MeV but the neutrinos were already decoupled were not heated. Since no heat is transferred into or out of piece of the Universe because of homogeneity, the entropy of photons plus $e^\pm$ plasma at temperature $T_\nu(1+z)$ is entirely transferred to a photon gas at temperature $T_\gamma(1+z)$. The spin factor for electrons is $g_s = 2$ and it is also 2 for positrons. This gives an equation

$$\frac{4}{3}\left(a + 4a\frac{7}{16}\right)T_\nu(1+z)^3 = \frac{4}{3}aT_\gamma(1+z)^3 \tag{86}$$

or

$$\frac{T_\nu(1+z)}{T_\gamma(1+z)} = \left(\frac{4}{11}\right)^{1/3} \tag{87}$$

or $T_{\nu,\circ} = 1.95$ K. Because of this factor, the additional contribution of neutrinos to the current energy density is $(42/16)(4/11)^{4/3} = 0.68$. Thus the current energy density is given by $u = (g_*/2)aT_\circ^4$ with the "effective statistical weight" $g_* = 3.36$. But the current entropy density is given by

$$s = \frac{4}{3}aT_\circ^3 + 3\frac{4}{3}\frac{7}{8}aT_{\nu\circ}^3 = \frac{4}{3}\frac{g_{*S}}{2}aT_\circ^3 \tag{88}$$

where the "effective statistical weight" for entropy is $g_{*S} = 43/11 = 3.91$ at the current time. Now we want to calculate the redshift and density at the earliest time we can reasonably consider, the Planck time given by

$$t_{Pl} = \frac{\hbar}{m_{Pl}c^2} = \frac{\hbar^{1/2}G^{1/2}}{c^{5/2}} = 5.4 \times 10^{-44} \text{ sec} \tag{89}$$

At this time the Hubble constant is $H = 1/(2t)$ for a radiation-dominated critical density Universe, so

$$\rho = \frac{3H^2}{8\pi G} = \frac{3c^5}{32\pi G^2\hbar} = 1.54 \times 10^{92} \text{ gm/cc} \tag{90}$$

We can calculate the temperature using

$$\rho = \frac{ag}{2c^2}T^4 \tag{91}$$

where g is the sum of all the g_s's for bosons plus 7/8 of the sum of the g_s's for fermions. Only particles with rest masses less than kT/c^2 are included in g. In the Standard Model g rises to 106.75 (Kolb & Turner, "The Early Universe", Figure 3.5) at high T. To find the redshift we use

$$s = \frac{4}{3}\frac{g}{2}aT^3 = (1+z)^3 s_\circ = (1+z)^3\,\frac{4}{3}\frac{43}{22}aT_\circ^3 \tag{92}$$

Thus

$$(1+z) = \left(\frac{g}{3.91}\right)^{1/3}\left(\frac{2\rho c^2}{agT_\circ^4}\right)^{1/4} = 3\times 10^{31}\times\left(\frac{g}{106.75}\right)^{1/12}\left(\frac{\rho}{1.54\times 10^{92}\ \mathrm{gm/cc}}\right)^{1/4} \tag{93}$$

Note that the net dependence on the relatively uncertain g factor is only $g^{1/12}$. The quantity $\rho R^2 = \rho/(1+z)^2$ needed to find $\Omega^{-1}-1$ is given by

$$\frac{\rho}{(1+z)^2} = \left(\frac{3.91}{g}\right)^{2/3}\left(\frac{agT_\circ^4\rho}{c^2}\right)^{1/2} = \sqrt{\rho}\left(\frac{106.75}{g}\right)^{1/6}\times 2\times 10^{-17}\ \mathrm{gm/cc} \tag{94}$$

For the density at the Planck time this is 2.16×10^{29} gm/cc if $g = 106.75$. But the current value of ρR^2 is just $\rho_\circ = \Omega_\circ\rho_{crit} = 1.8788\ \Omega_\circ h^2\times 10^{-29}$ gm/cc. Finally

$$\Omega^{-1}-1 = 8.7\times 10^{-59}\Omega_\circ h^2 (g/106.75)^{1/6}\left(\Omega_\circ^{-1}-1\right) \tag{95}$$

Thus to get $\Omega_\circ$ between 0.1 and 2 requires with $h = 0.65$ and $g = 106.75$ requires

$$\begin{array}{r} 0.99999999999999999999999999999999999999999999999999999999997 \\ <\Omega< \\ 1.00000000000000000000000000000000000000000000000000000000004 \end{array} \tag{96}$$

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.

5.1. The Flatness-Oldness Figure

The calculations involved in the flatness-oldness figure are as follows: The $R(t)$ for the closed, re-collapsing case is found parametrically using:

$$\begin{aligned} R &= 1-\cos E \\ t &= E-\sin E \end{aligned} \tag{97}$$

the plotted range is 0–2π in E and t.

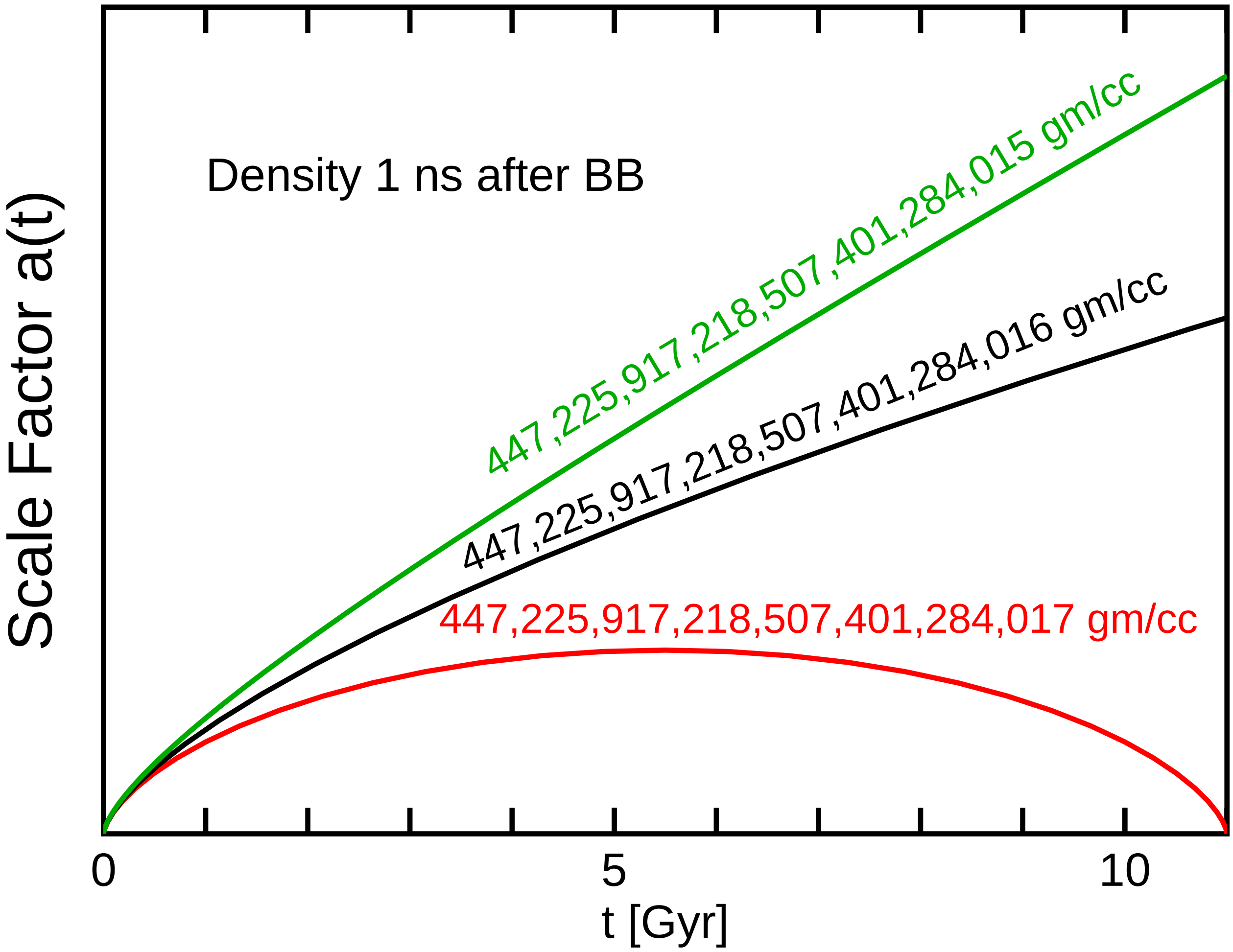

Fig. 11.— Scale factor *vs.* time for open, critical and closed Universes with density at 1 nanosecond after the Big Bang indicated.

The flat model with the same scaling is given by the leading terms of the equations above: $R = (6^{2/3}/2)t^{2/3}$. When $t = 2\pi$, then $R = 5.62156$.

The value of Ω when R is small is determined from $q = 1/(1+\cos E)$. Thus $\Omega = 2/(1+\cos E)$. When $R = 0.00562156$, which is $(1+z) = 1000$ for the critical density case, then $\Omega = 2/(2-R) = 1 + R/2 = 1.0028$ for the closed case. Note that even at this early time, the expansion rate H for the closed case is already reduced relative to the critical density case. We can compute the amount of this reduction using

$$\begin{aligned} Ht &= \frac{\sin E(E - \sin E)}{(1-\cos E)^2} \\ &= \frac{(E - (1/6)E^3 + \ldots)((1/6)E^3 - (1/120)E^5 + \ldots)}{[(1/2)E^2 - (1/24)E^4 + \ldots]^2} \\ &= \frac{2}{3}\left(1 - (1/20)E^2 + \ldots\right) \end{aligned} \quad (98)$$

If we compare this to

$$\Omega = \frac{2}{1+\cos E} = 1 + \frac{E^2}{4} - \ldots \tag{99}$$

we see that only 60% of the increase in Ω is caused by increased density, while 40% is caused by the lower value of H reducing ρ_{crit}. Thus the actual fractional density increase in the closed case is $0.6 \times 0.0028 = 0.0017$. The fractional density increase clearly goes like $E^2 \propto R(t)$ for early times. This growth of $\Delta\rho/\rho$ like $R(t)$ describes the linear (small amplitude) growth of perturbations in a critical density Universe. If we extrapolate the closed case excess density of 0.17% from $(1+z) = 1000$ to $z = 0$, we get a value of $\delta_c = (3/20)(12\pi)^{2/3} = 1.6865$ for the linearly extrapolated density excess necessary to cause collapse.

The open Universe model is

$$\begin{aligned} R &= \cosh\psi - 1 \\ t &= \sinh\psi - \psi \end{aligned} \tag{100}$$

The range is 0 to ψ_m with $\sinh\psi_m - \psi_m = 2\pi$. The $R(\psi_m)$ for this model is the largest and it is used to normalize the curves. The value $\Omega = 2/(1+\cosh\psi_m) = 0.1951$.

Now the general evolution of Ω is $(1/\Omega - 1)\rho R^2 =$const. For Ω very close to 1 this becomes $\Delta\rho R^2 =$const. Therefore, at early times the difference between the critical density and ρ in this Figure is given by $\Delta\rho = \delta_c(3H_\circ^2/8\pi G)(1+z)^2$. We want this to be equal to 1 gm/cc. For an age of $t_\circ = 11$ Gyr, ($H_\circ = 59.3$ km/sec/Mpc for the critical density case), this is $\Delta\rho = 1.11\times 10^{-29}(1+z)^2$. Thus we need $(1+z) = 3 \times 10^{14}$.

In order to calculate the time corresponding to this redshift, we need to use the entropy evolution law,

$$s = (4/3)(g/2)aT^3 = s_\circ(1+z)^3 \tag{101}$$

The entropy density now is given by $s_\circ = (4/3)(43/22)aT_\circ^3$. The factor (43/22) allows for the neutrinos, which are assumed to be massless. The factor $g/2$ allows for the number of effectively massless particle species at a given temperature. It appears to be $g \approx 106.75$ for $kT \approx 100$ GeV. Except during transitions where a large part of the particles are non-relativistic, we have the energy density

$$u = \frac{3}{4}sT = \frac{3.91^{4/3}}{2g^{1/3}} aT_\circ^4(1+z)^4 \tag{102}$$

For $(1+z) = 3\times 10^{14}$, and $g = 106.75$, this is $u = 2.2\times 10^{45}$ erg/cm^3 or an equivalent mass density of $\rho = 2.44 \times 10^{24}$ gm/cm^3. We convert this to a time using

$$\rho = \rho_{crit} = \frac{3H^2}{8\pi G} = \frac{3}{32\pi G t^2} \tag{103}$$

This corresponds to a time of 0.43 ns. I have rounded this off to 1 ns (which is fair enough in the exponent), and adjusted the critical density to match this time. Thus the actual $\Delta\rho$ should be 0.43 gm/cc, so the figure is conservative in the implied accuracy in initial conditions needed to produce a Universe like ours.

Note that $\Delta\rho \propto \rho_{crit}(1+z)^2 \propto H_\circ^2 g^{1/6}/t$.

6. Distant Objects

In general to find the appearance of distant objects (flux and angular size) 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

$$ds^2 = c^2dt^2 - (dx^2 + dy^2 + dz^2) = c^2dt^2 - r^2(d\delta^2 + \cos^2\delta d\alpha^2) \tag{104}$$

The worldlines of comoving galaxies all have to intersect at some event which we identify as the Big Bang. Let's choose the is 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

$$\begin{aligned} x_B(t) &= a(t)X_B \\ y_B(t) &= a(t)Y_B \\ z_B(t) &= a(t)Z_B \end{aligned} \tag{105}$$

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 observers. 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.

6.1. 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

$$1 + z = \frac{t_\circ}{\sqrt{(t_\circ - d_A/c)^2 - d_A^2/c^2}} = \frac{1}{\sqrt{1 - 2d_A/ct_\circ}} \tag{106}$$

Solving this equation gives

$$d_A = ct_\circ \frac{z(1 + z/2)}{(1 + z)^2} \tag{107}$$

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

$$1+z = \sqrt{\frac{1+v/c}{1-v/c}} = \sqrt{\frac{ct_\circ}{ct_\circ - 2d_A}} \tag{108}$$

which clearly gives the same result.

6.2. 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 $T_{obs}/(1+z)$. 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

$$F = \frac{L}{4\pi d_L^2} \tag{109}$$

so

$$d_L = \sqrt{\frac{L}{4\pi F}} = \sqrt{\left(\frac{R}{\Delta\theta}\right)^2 \left(\frac{T_{em}}{T_{obs}}\right)^4} = d_A(1+z)^2 \tag{110}$$

6.3. 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

$$\begin{aligned} t &= \tau \cosh \psi \\ x &= c\tau \sinh \psi \end{aligned} \tag{111}$$

where ψ is the hyperbolic "angle". The distance at constant τ between ψ and $\psi + d\psi$ is given by $-ds^2 = dx^2 - c^2 dt^2$ 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 $2\pi x = 2\pi c\tau \sinh\phi$. Since $\sinh\psi > \psi$, the spatial sections of the zero density Universe are negatively curved.

6.4. Robertson-Walker metric

We can write the entire metric in cosmological variables now:

$$ds^2 = c^2 d\tau^2 - (c\tau)^2 \left(d\psi^2 + \sinh^2\psi(d\delta^2 + \cos^2\delta d\alpha^2)\right) \tag{112}$$

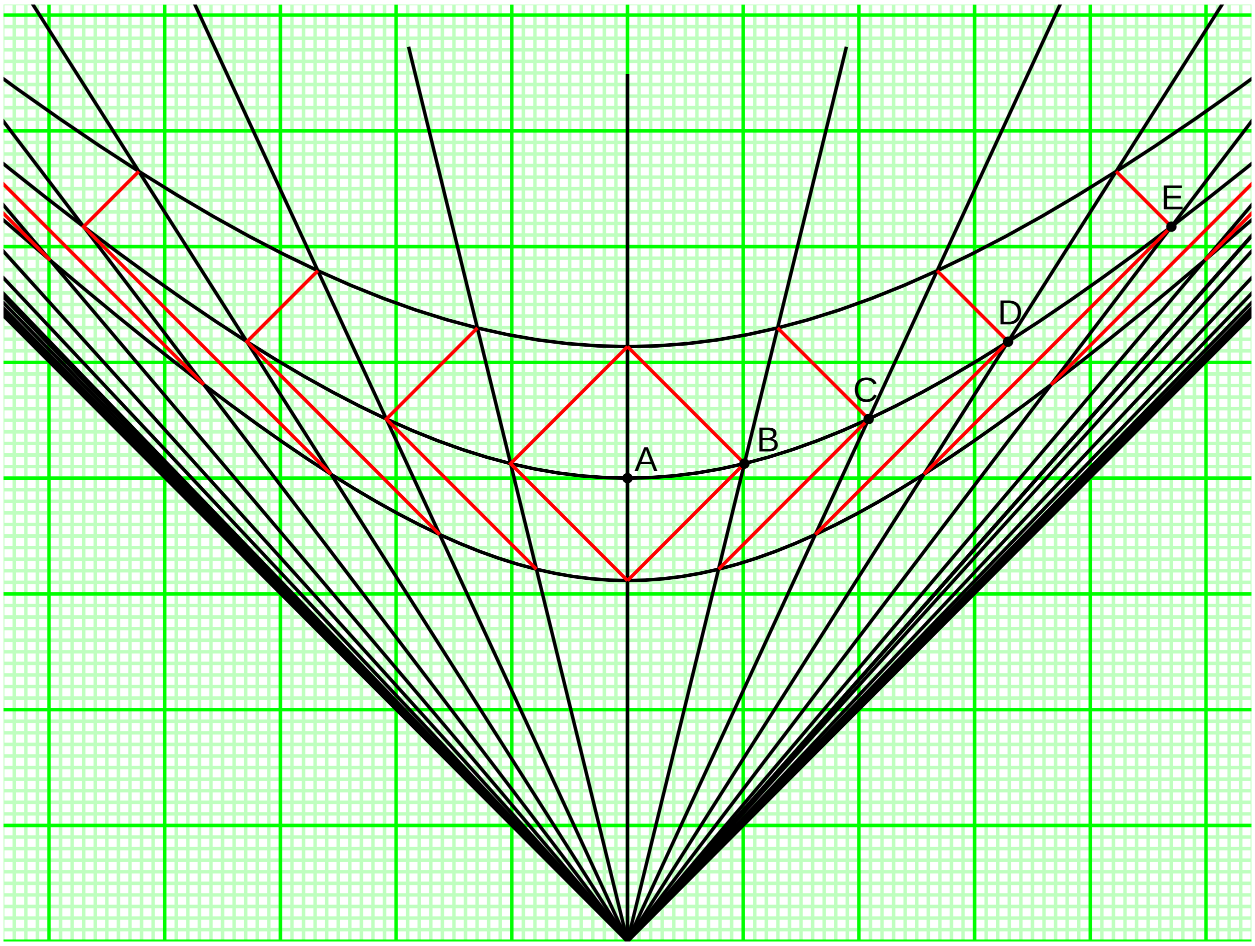

Fig. 12.— Space-time diagram plotted in special relativistic coordinates showing the constant cosmic time τ hyperbolae, and the radar pulses used to measure the radial distance from A to E is $R = 4$ big squares, while the circumference of the circle with center A that contains E is $C = 2\pi$ times 4.7 big squares which is greater than $2\pi R$. Thus this model has hyperbolic geometry.

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

$$ds^2 = c^2 d\tau^2 - (c\tau)^2 \left(\frac{dr^2}{1+r^2} + r^2(d\delta^2 + \cos^2\delta d\alpha^2)\right) \tag{113}$$

This is a Robertson-Walker 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 hypersperical 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

$$ds^2 = c^2 dt^2 - a(t)^2 R_\circ^2 \left(\frac{dr^2}{1-kr^2} + r^2(d\delta^2 + \cos^2\delta d\alpha^2)\right) \tag{114}$$

where $k = -1, 0$ or 1 for the negatively curved (hyperboloidal), flat or positively curved (hyperspherical) cases, $a(t)$ will be computed later using energy arguments, and finally the radius of curvature $R_\circ$ of the Universe is given by

$$R_\circ = \frac{c/H_\circ}{\sqrt{|1 - \Omega_{v\circ} - \Omega_{m\circ} - \Omega_{r\circ}|}} \tag{115}$$

where $\Omega_{v\circ}$ is the current ratio of the vacuum density (cosmological constant) to the critical density $3H^2/8\pi G$, $\Omega_{m\circ}$ is the current ratio of the density of pressureless matter to the critical density, and $\Omega_{r\circ}$ is the current ratio of radiation density ($\rho = u/c^2$) to the critical density. 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.

7. 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:

$$\frac{a(t)R_\circ dr}{\sqrt{1-kr^2}} = -cdt \tag{116}$$

so

$$R_\circ \int \frac{dr}{\sqrt{1-kr^2}} = \int_{t_{em}}^{t_\circ} (1+z)cdt \tag{117}$$

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(117) is either $\sin^{-1} r$, r, or $\sinh^{-1} r$ depending on whether $k = +1$, 0 or -1.

7.1. Critical Density Universe

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

$$\int_{t_{em}}^{t_\circ} (1+z)cdt = \frac{c}{H_\circ} \int (1+z)^{-3/2} dz = 2\frac{c}{H_\circ}\left(1-(1+z)^{-1/2}\right) \tag{118}$$

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

$$D_A = 2\frac{c}{H_\circ}\left((1+z)^{-1} - (1+z)^{-3/2}\right) \tag{119}$$

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

$$D_L = 2\frac{c}{H_\circ}\left(1+z-\sqrt{1+z}\right) = \frac{cz}{H_\circ}\left(1+\frac{z}{4}+\ldots\right) \tag{120}$$

7.2. 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

$$\int_{t_{em}}^{t_\circ} (1+z)cdt = \int_{t_{em}}^{t_\circ} \exp(H(t_\circ - t))cdt = \frac{cz}{H} \tag{121}$$

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

$$D_A = \frac{c}{H}\frac{z}{1+z} \tag{122}$$

and the luminosity distance is

$$D_L = \frac{c}{H} z(1+z) \tag{123}$$

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

$$c\frac{dt}{dz} = -\frac{c}{H_\circ}\frac{1}{(1+z)^3} \tag{124}$$

Thus

$$\int (1+z)cdt = \frac{c}{H_\circ}\left(1 - \frac{1}{1+z}\right) \tag{125}$$

and

$$D_A = a(t_{em})\int (1+z)cdt = \frac{c}{H_\circ}\frac{z}{(1+z)^2}. \tag{126}$$

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

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(117) 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

$$a(t_\circ + \Delta t) = a(t_\circ)\left(1 + H_\circ \Delta t - \frac{1}{2}q_\circ (H_\circ \Delta t)^2 + \ldots\right) \tag{127}$$

which defines the *deceleration parameter*

$$q_\circ = -\frac{a\ddot{a}}{\dot{a}^2} \tag{128}$$

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

$$q_\circ = -\frac{a\ddot{a}}{\dot{a}^2} = \frac{4\pi G}{3H_\circ^2}\left(\rho + \frac{3P}{c^2}\right) = \frac{\Omega_{m\circ}}{2} + \Omega_{r\circ} - \Omega_{v\circ} \tag{129}$$

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

$$d\left(\frac{a(t)}{a(t_\circ)}\right) = d\left(\frac{1}{1+z}\right) = \frac{-dz}{(1+z)^2} = H_\circ(1 - q_\circ(H_\circ \Delta t) + \ldots)dt \tag{130}$$

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

$$\frac{dt}{dz} = \frac{-H_\circ^{-1}}{(1+z)^2(1+q_\circ z + \ldots)} \tag{131}$$

The integral on the RHS of Eqn(117) is then given by

$$\int_{t_{em}}^{t_\circ} (1+z) c dt = \frac{c}{H_\circ} \int_0^z \frac{dz}{(1+z)(1+q_\circ z)} = \frac{cz}{H_\circ} \left(1 + \frac{z}{2}[-1-q_\circ] + \ldots\right) \tag{132}$$

Eqn(117) then gives

$$D_A = R_\circ(r + \mathcal{O}(r^3))/(1+z) = \frac{cz}{H_\circ}\left(1 + \frac{z}{2}[-3-q_\circ] + \ldots\right) \tag{133}$$

Finally

$$D_L = D_A(1+z)^2 = \frac{cz}{H_\circ}\left(1 + \frac{z}{2}[1-q_\circ] + \ldots\right) \tag{134}$$

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 tentative work on distant Type Ia SNe by Perlmutter *et al.* (1998) and Garnavich *et al.* (1998) (Figure 13) suggests that $q_\circ < 0$, which favors an empty Universe or one 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

$$D_L = \frac{cz}{H_\circ}\left[\frac{4 - 2\Omega + 2z}{(1+\sqrt{1+\Omega z})(1-\Omega+\sqrt{1+\Omega z})}\right] \tag{135}$$

Note that for $\Omega = 2$, $q_\circ = 1$, we have

$$D_L = \frac{cz}{H_\circ} \tag{136}$$

exactly. This particular case simplifies because for $\Omega = 2$, the radius of curvature of the Universe is $R_\circ = c/H_\circ$.

7.3. 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

$$\nu F_\nu = \frac{\nu(1+z)L_{\nu(1+z)}}{4\pi D_L^2} \tag{137}$$

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

$$
\begin{aligned}
F_\nu &= \frac{(1+z)L_{\nu(1+z)}}{4\pi D_L^2} \\
F_\lambda &= \frac{L_{\lambda/(1+z)}}{4\pi D_L^2(1+z)}
\end{aligned}
\tag{138}
$$

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

$$K(\nu, z) = -2.5\log\left(\frac{\nu(1+z)L_{\nu(1+z)}}{\nu L_\nu}\right) \tag{139}$$

If working with only V band data, we can write

$$V = M_V + 5\log\left(\frac{D_L(z)}{10 \text{ pc}}\right) + K(\nu_V, z) \tag{140}$$

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.

8. 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

$$ds^2 = c^2dt^2 - a(t)^2R_\circ^2(d\psi^2 + S^2(\psi)[d\theta^2 + \sin^2\theta d\phi^2]) \tag{141}$$

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

$$D_{now} = R_\circ\psi = \int \frac{cdt}{a} = \int_{1/(1+z)}^{1} \frac{cda}{a\dot{a}} \tag{142}$$

and of course the light travel time distance is given by

$$D_{ltt} = \int cdt = \int_{1/(1+z)}^{1} \frac{cda}{\dot{a}} \tag{143}$$

We can write $\dot{a}$ as $H_\circ\sqrt{X}$ with

$$X(a) = \Omega_{m\circ}/a + \Omega_{r\circ}/a^2 + \Omega_{v\circ}a^2 + (1-\Omega_{tot}) \tag{144}$$

Let us define

$$Z = \int_{1/(1+z)}^{1} \frac{da}{a\sqrt{X}} \tag{145}$$

so $D_{now} = (cZ/H_\circ)$ and $D_{ltt} = (c/H_\circ)\int_{1/(1+z)}^{1} da/\sqrt{X}$. Then

$$\begin{aligned} D_A &= \frac{c}{H_\circ}\frac{S(\sqrt{|1-\Omega_{tot}|}Z)}{(1+z)\sqrt{|1-\Omega_{tot}|}} \\ &= \frac{D_{now}}{(1+z)}\left(1+\frac{1}{6}(1-\Omega_{tot})Z^2 + \frac{1}{120}(1-\Omega_{tot})^2Z^4 + \ldots\right) \end{aligned} \tag{146}$$

We can define a function $J(x)$ given by

$$J(x) = \begin{cases} \frac{\sin\sqrt{-x}}{\sqrt{-x}}, & x < 0; \\ \frac{\sinh\sqrt{x}}{\sqrt{x}}, & x > 0; \\ 1 + x/6 + x^2/120 + \ldots + x^n/(2n+1)! + \ldots, & x \approx 0. \end{cases} \tag{147}$$

Then

$$\begin{aligned} D_A &= \frac{cZ(z)}{H_\circ}\frac{J([1-\Omega_{tot}]Z^2)}{1+z} \\ D_L &= (1+z)^2 D_A \end{aligned} \tag{148}$$

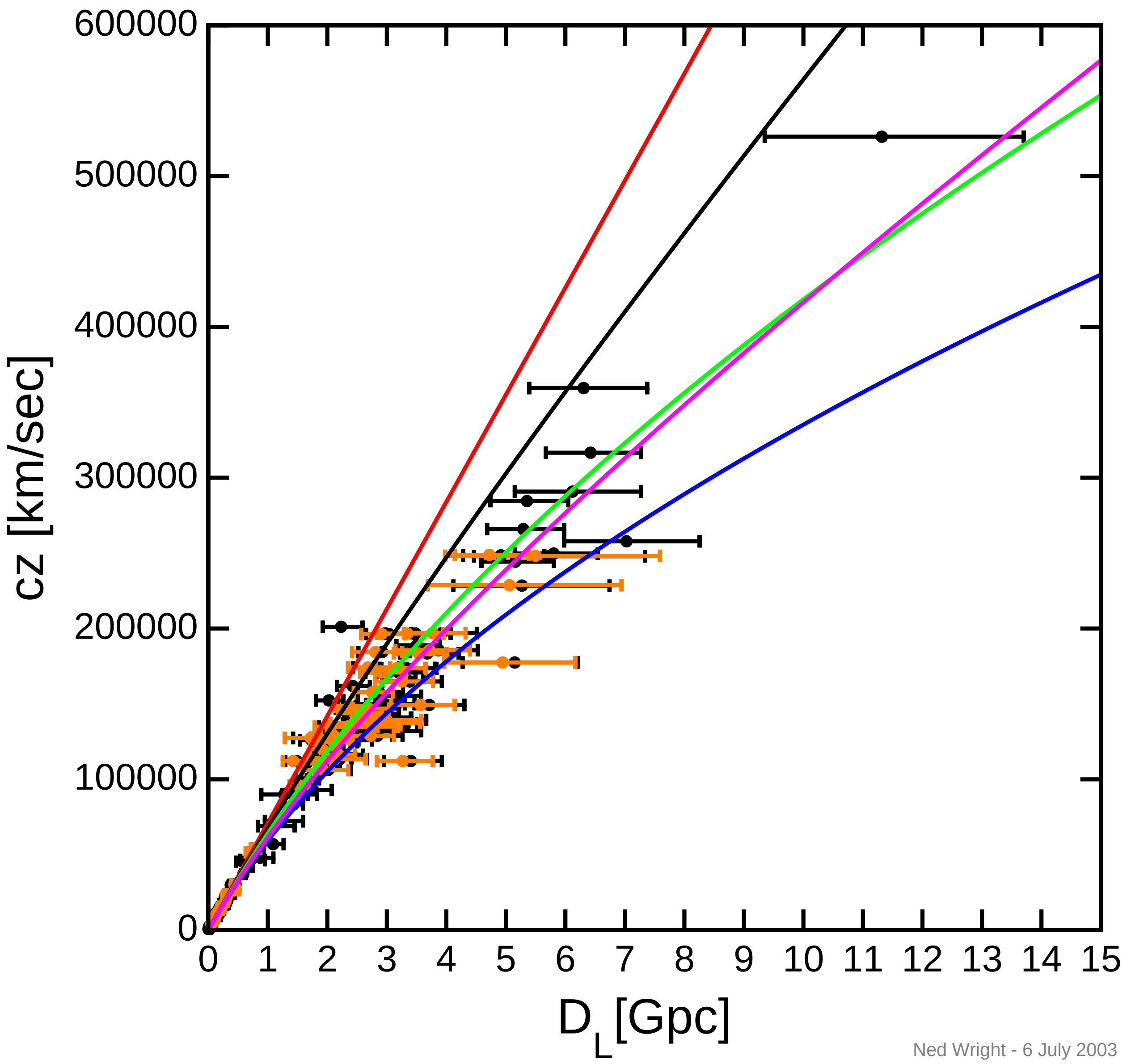

Fig. 13.— Redshift *vs.* luminosity distance for high redshift Type Ia supernovae. The curves (from top to bottom on the right) are for $q_\circ = 1$, $q_\circ = 0.5$, $q_\circ = 0$ and $q_\circ = -1$ (the Steady State).

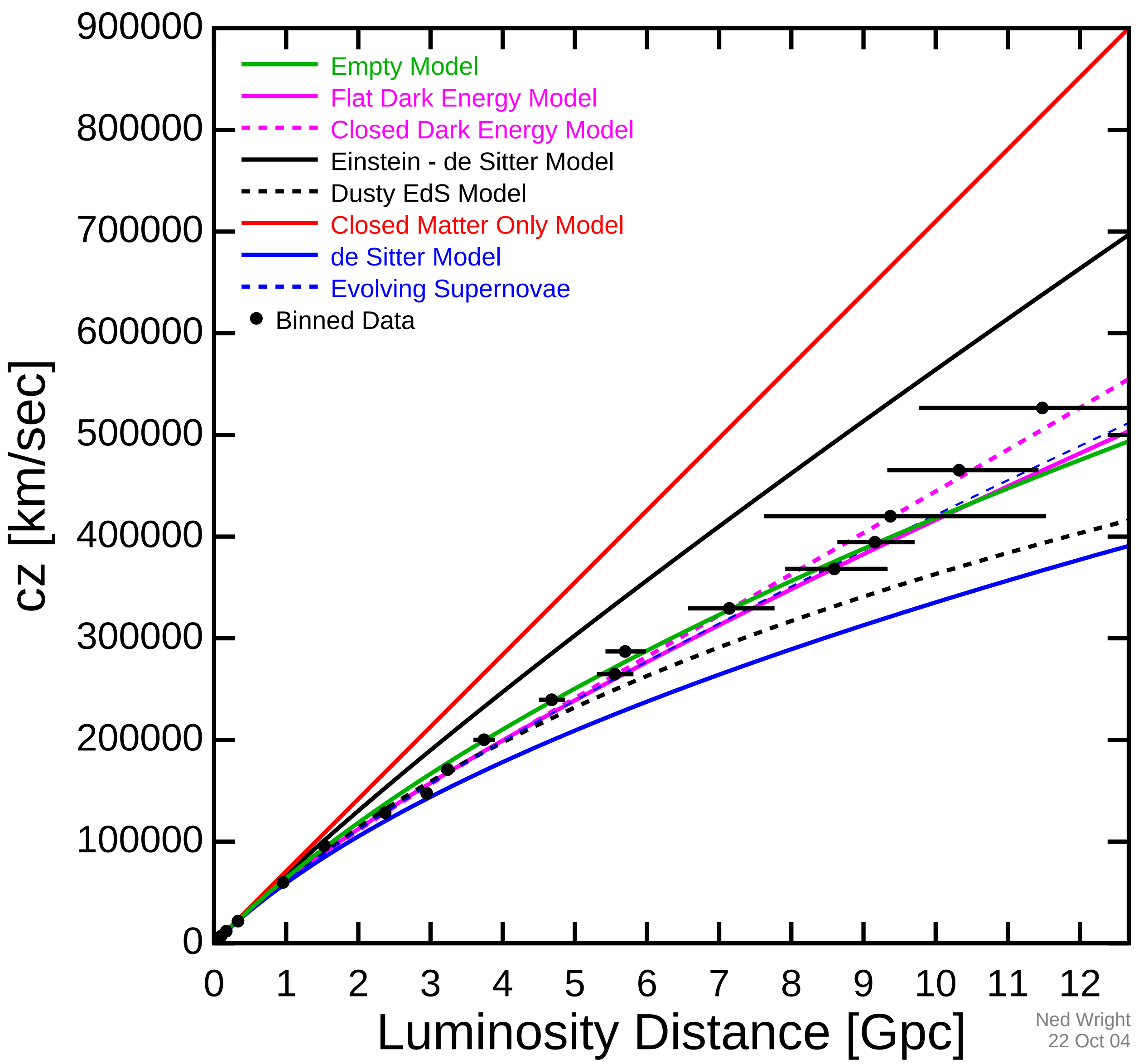

Fig. 14.— Distance modulus *vs.* redshift for high redshift Type Ia supernovae. The distance modulus is $DM = 5\log(D_L(z)/D_\circ)$, normalized to $D_\circ = c/H_\circ$ for this plot.

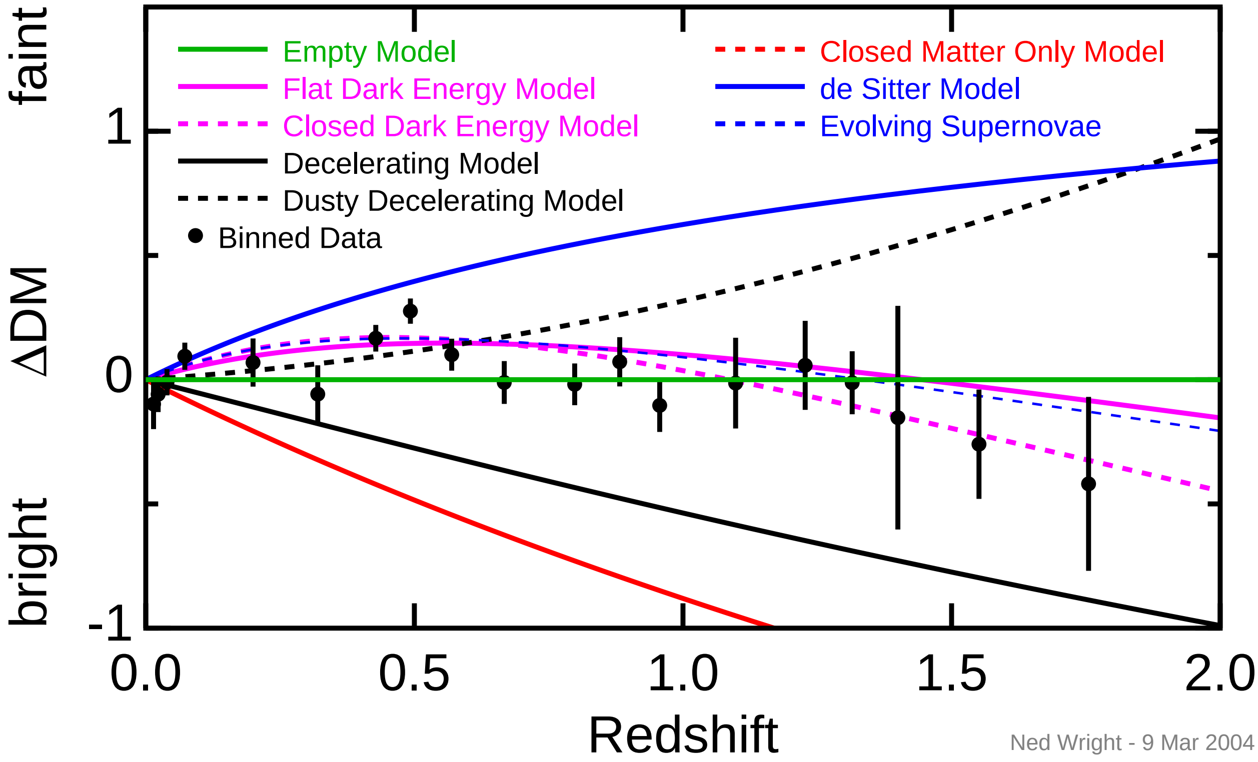

Fig. 15.— Distance modulus relative to an $\Omega = 0$ model *vs.* redshift for high redshift Type Ia supernovae.

9. 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

$$\frac{dN}{dz} = n_\circ(1+z)^3 d_A(z)^2 \frac{cdt}{dz} \tag{149}$$

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

$$d_L = \sqrt{\frac{L}{4\pi S}} \tag{150}$$

and the counts versus flux are given by

$$\frac{dN}{dS} = n_\circ(1+z)^3 \frac{d_L^2}{(1+z)^4} \frac{d(d_L)}{dS} \frac{dz}{d(d_L)} \frac{cdt}{dz} \tag{151}$$

Now $d(d_L)/dS = -0.5S^{-3/2}\sqrt{L/4\pi}$ and $d_L^2 = L/(4\pi S)$ so

$$\frac{dN}{dS} = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}} \left[(1+z)^{-1} \frac{dz}{d(d_L)} \frac{cdt}{dz}\right] \tag{152}$$

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

$$J = \int SdN = \int S \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}} \left[(1+z)^{-1} \frac{dz}{d(d_L)} \frac{cdt}{dz}\right] dS \tag{153}$$

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

$$\begin{aligned} \frac{dt}{dz} &= \frac{-H_\circ^{-1}}{(1+z)^2(1+q_\circ z+\ldots)} \\ d_L &= \frac{cz}{H_\circ}\left(1+\frac{z}{2}[1-q_\circ]+\ldots\right) \\ \frac{d(d_L)}{dz} &= \frac{c}{H_\circ}\left(1+z[1-q_\circ]+\ldots\right) \end{aligned} \tag{154}$$

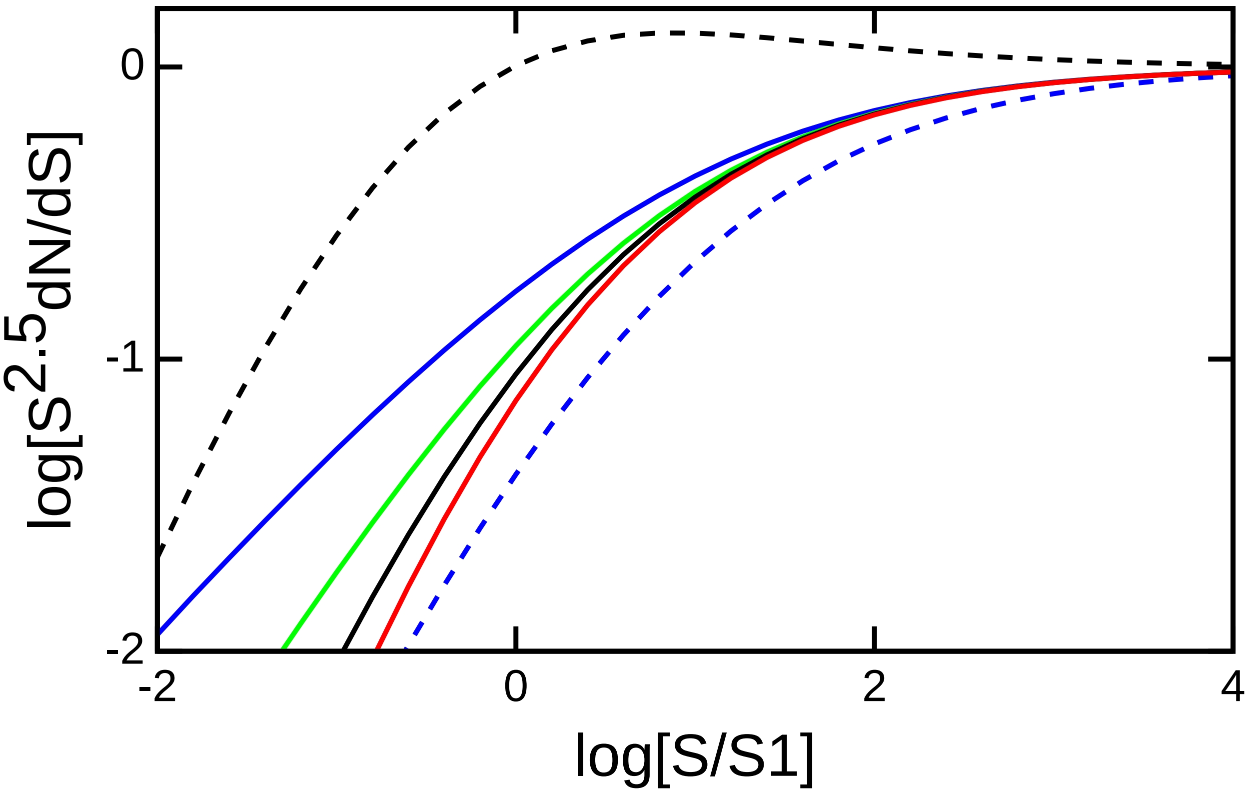

Fig. 16.— The predicted source counts for standard candles in various cosmological models, normalized to the Euclidean counts. This is for bolometric fluxes or sources with $L_\nu \propto \nu^{-1}$. S_1 is the "Euclidean" flux at a distance of $c/H_\circ$. The models are (from right to left: Steady State, 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

$$\frac{dN}{dS} = \frac{n_\circ (L/4\pi)^{3/2}}{2S^{5/2}} \left[\frac{1}{(1+z)^3 (1+q_\circ z + \ldots)(1+z\,[1-q_\circ] + \ldots)} \right] \tag{155}$$

The $q_\circ$ dependence cancels out in the first two terms so

$$\frac{dN}{dS} = \frac{n_\circ (L/4\pi)^{3/2}}{2S^{5/2}} \left[\frac{1+\mathcal{O}(z^2)}{(1+z)^4} \right] \tag{156}$$

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

$$d_L \;\; = \;\; \frac{cz}{H_\circ}(1+z/2)$$

$$\begin{aligned}
\frac{d(d_L)}{dz} &= \frac{c}{H_\circ}(1+z) \\
\frac{cdt}{dz} &= \frac{c}{H_\circ}\frac{1}{(1+z)^2} \\
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(1+z)^4}\right]
\end{aligned} \tag{157}$$

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

$$\frac{dN}{dS} = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(1+2\zeta)^2}\right] = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{\left(1+2\sqrt{S_1/S}\right)^2}\right] \tag{158}$$

For an $\Omega = 2$ matter dominated Universe we get

$$\begin{aligned}
d_L &= \frac{cz}{H_\circ} \\
\frac{d(d_L)}{dz} &= \frac{c}{H_\circ} \\
\frac{cdt}{dz} &= \frac{c}{H_\circ}\frac{1}{(1+z)^2\sqrt{1+2z}} \\
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(1+z)^3\sqrt{1+2z}}\right]
\end{aligned} \tag{159}$$

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

$$\frac{dN}{dS} = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{\left(1+\sqrt{S_1/S}\right)^3\sqrt{1+2\sqrt{S_1/S}}}\right] \tag{160}$$

For a critical density matter-dominated Universe we get

$$d_L = \frac{2c}{H_\circ}\left(1+z-\sqrt{1+z}\right)$$

$$
\begin{aligned}
\frac{d(d_L)}{dz} &= \frac{c}{H_\circ}\left(2-\frac{1}{\sqrt{1+z}}\right) \\
\frac{cdt}{dz} &= \frac{c}{H_\circ}\frac{1}{(1+z)^{2.5}} \\
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(2\sqrt{1+z}-1)(1+z)^3}\right]
\end{aligned}
\tag{161}
$$

If we let $u = \sqrt{1+z}$, then $2u(u-1) = \zeta$ so

$$
u = \sqrt{1+z} = \frac{1+\sqrt{1+2\zeta}}{2} \tag{162}
$$

and the source counts $N(S)$ are given by

$$
\begin{aligned}
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{8}{\sqrt{1+2\zeta}\left(1+\zeta+\sqrt{1+2\zeta}\right)^3}\right] \\
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{8}{\sqrt{1+2\sqrt{S_1/S}}\left(1+\sqrt{S_1/S}+\sqrt{1+2\sqrt{S_1/S}}\right)^3}\right]
\end{aligned}
\tag{163}
$$

For a vacuum-dominated model with $q_\circ = -1$, the cosmological functions are

$$
\begin{aligned}
d_L &= \frac{c}{H_\circ}z(1+z) \\
\frac{d(d_L)}{dz} &= \frac{c}{H_\circ}(1+2z) \\
\frac{cdt}{dz} &= \frac{c}{H_\circ}\frac{1}{1+z}
\end{aligned}
\tag{164}
$$

Thus $\zeta = z(1+z)$, or

$$
z = \frac{-1+\sqrt{1+4\zeta}}{2} \tag{165}
$$

The number counts will be given by

$$
\begin{aligned}
\frac{dN}{dS} &= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{1}{(1+z)^2(1+2z)}\right] \\
&= \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{4}{\left(1+\sqrt{1+4\sqrt{S_1/S}}\right)^2\sqrt{1+4\sqrt{S_1/S}}}\right]
\end{aligned}
\tag{166}
$$

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(156)

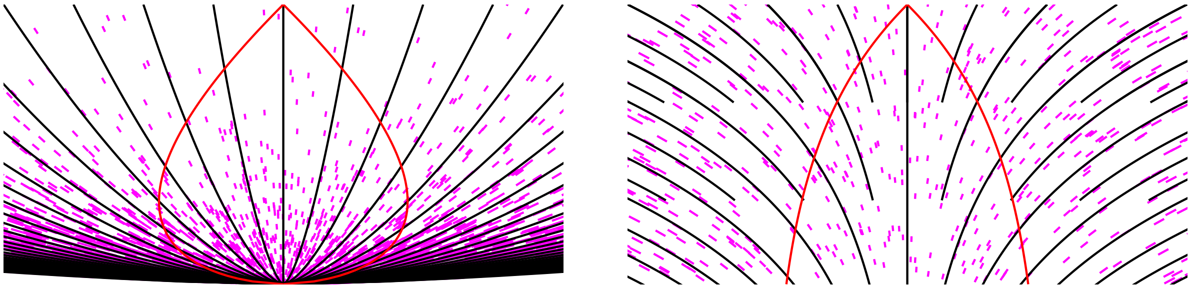

Fig. 17.— The excess of faint quasars and radio sources can be explained by an evolving population in a homogeneous Universe as shown at left above. The quasars are shown as little magenta line segments because they have a short lifetime compared to the age of the Universe. There are now only a few quasars, 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(149) is dropped. This gives

$$\frac{dN}{dS} = \frac{n_\circ(L/4\pi)^{3/2}}{2S^{5/2}}\left[\frac{32}{\left(1+\sqrt{1+4\sqrt{S_1/S}}\right)^5\sqrt{1+4\sqrt{S_1/S}}}\right] \tag{167}$$

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.

9.1. Gamma Ray Bursts

γ-ray bursts are short bursts ($\approx$ 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

$$\Delta E = 4\pi d_L(z)^2 \times (\text{fluence})/(1+z) \approx (5/h^2) \times 10^{49} \text{ ergs} \tag{168}$$

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.

10. Evolution of Diffuse Backgrounds

We can calculate the evolution of diffuse background radiation fields without knowing the brightness of distant galaxies, because if a radiation field is homogeneous and isotropic like the Universe, we only need to consider a small region of space. Imagine a cubical region bounded by comoving mirrors. In a barber shop with mirrors on two opposing walls one sees what looks like an infinite number of images receding off to infinity, and this looks just like a homogeneous and isotropic Universe.

So let us consider the equation of radiative transfer:

$$\frac{dI_\nu}{ds} = j_\nu - \alpha_\nu I_\nu \tag{169}$$

where I_ν is the specific intensity in erg/cm^2/sec/sr/Hz, s is the path length along the ray in cm, j_ν is the emissivity in erg/cm^3/sec/sr/Hz, and α_ν is the absorption coefficient in cm^{-1}. We define the optical depth $\tau_\nu = \int \alpha_\nu ds$ and get the equation

$$\frac{dI_\nu}{d\tau_\nu} = \frac{j_\nu}{\alpha_\nu} - I_\nu = S_\nu - I_\nu \tag{170}$$

where the source function S is the ratio of emissivity to absorption coefficient.

10.1. Olber's Paradox

If we assume that j_ν is due to blackbody stars with number density n, radius R, and temperature T_*, then the luminosity per unit frequency of a single star is $L_\nu = 4\pi^2 R^2 B_\nu(T_*)$. The emissivity is then $j_\nu = nL_\nu/(4\pi) = n\pi R^2 B_\nu(T_*)$. The absorption coefficient of the stars is just one over the mean free path, and is given by $\alpha_\nu = n\pi R^2$. When we solve the radiative transfer equation we get

$$I_\nu = \exp(-\tau_\nu) I_\nu(0) + [1 - \exp(-\tau_\nu)]\, B_\nu(T_*) \tag{171}$$

Thus when $s \to \infty$, the intensity goes to $B_\nu(T_*)$, the surface brightness of a star. Since the sky is much darker than this, something is wrong with Olber's assumptions.

Let's calculate α_ν from the observed luminosity density of $1.6h \times 10^8\ L_\odot/\mathrm{Mpc}^3$:

$$\alpha_\nu = n\pi R^2 = \frac{1.6 \times 10^8}{(3.08 \times 10^{24})^3} \pi (7 \times 10^{10})^2 \approx 10^{-43}\ \mathrm{cm}^{-1} \tag{172}$$

Thus to reach $\tau = 1$ we need to go to $s = 10^{43}$ cm which is 10^{25} light years. Since the lifetime of the Universe is much less than this, we need to consider the effect of the expanding Universe on the equation of radiative transfer.

10.2. Scattering

Often the absorption coefficient includes scattering denoted by σ_ν which generates a contribution to j_ν of $\sigma_\nu J_\nu$, where J_ν is the average over solid angle of the intensity. When averaged over solid angle, scattering cancels out for a diffuse background. This is because scattering only changes the direction of photons, and all directions are equal in an isotropic situation. However, when looking at individual sources, scattering is just as effective in reducing their visibility as true absorption. Therefore we are interested in knowing the optical depth to electron scattering between the event "us-now" and some event at redshift z on our past lightcone. Assuming that all matter is ionized, the optical depth is obtained by using $ds = c(dt/dz)dz$ and $n_e(z) = (1+z)^3 n_e(0)$. For a matter-dominated Universe with $\Omega_{m\circ} >>$ any other density, we have

$$\tau_e = n_e(0)\sigma_T \frac{c}{H_\circ} \int \frac{(1+z)^3 dz}{(1+z)^2\sqrt{1+\Omega_{m\circ} z}} \tag{173}$$

The coefficient in front has the simple interpretation of being the optical depth to a distance equal to the Hubble radius $c/H_\circ$ in a non-expanding Universe. For a typical $n_e(0) = 1/4$ per cubic meter allowed by Big Bang Nucleosynthesis and the CMB acoustic peaks, this is $n_e(0)\sigma_T(c/H_\circ) = 1.7 \times 10^{-3}/h$. Thus this optical depth only becomes large for large z so $\Omega z >> 1$ and the optical depth is

$$\tau \approx 1.7 \times 10^{-3} \frac{2z^{3/2}}{3\sqrt{\Omega_{m\circ} h^2}} \tag{174}$$

This could be significant for $z \approx 10^2$ but only if the gas is ionized, and since this before the existence of quasars the gas is probably neutral until $z = 10^3$. At this point, once the ionized fraction is even a few percent, the optical depth is quite large and no point sources can be seen.

10.3. Cosmological Equation of Transfer

Once we have disposed of scattering, and replaced ds by cdt, we still need to allow for the continual change in frequencies caused by the expansion of the Universe. The easiest way to do this is to replace I_ν by a quantity that does not change during the expansion. The number of photons per mode, $I_\nu/(2h\nu[\nu/c]^2)$, evaluated at $\nu = \nu_\circ(1+z)$ is one such quantity. This gives

$$\frac{\partial}{\partial z}\left(\frac{I_{\nu_\circ(1+z)}}{(1+z)^3}\right) = \frac{cdt/dz}{(1+z)^3}\left(j_{\nu_\circ(1+z)} - \alpha_{\nu_\circ(1+z)} I_{\nu_\circ(1+z)}\right) \tag{175}$$

Consider Olber's paradox again. Assume we have a luminosity density $\mathcal{L} = 4\pi \int j_\nu d\nu$ now and that the only thing that changes with redshift is the density going like $(1+z)^3$, so the *comoving* luminosity density is constant. Thus $j(\nu, z) = j(\nu, z = 0)(1+z)^3$. [Remember that comoving volumes expand and contract as the Universe does.] Since α is so small we will ignore it, giving the equation

$$I_{\nu_\circ} = \int j_{\nu_\circ(1+z)} cdt \tag{176}$$

For the total or bolometric intensity we integrate over frequency giving

$$I = \int I_{\nu_\circ} d\nu_\circ = \int\int j_{\nu_\circ(1+z)} cdt d\nu_\circ = \frac{\mathcal{L}}{4\pi}\int \frac{cdt}{1+z} \tag{177}$$

Thus the bolometric intensity is reduced by factor of $(1+z)^{-1}$ and is also limited by the finite age of the Universe. Both effects have comparable magnitude so it is pointless to argue about which factor is more important, but people do. For $\mathcal{L} \approx 2h \times 10^8\ L_\odot/\mathrm{Mpc}^3$ and $\Omega_m = 1$, we get

$$I = \frac{2 \times 10^8 \times 4 \times 10^{33} \times 3000}{4\pi(3.086 \times 10^{24})^2} \int \frac{dz}{(1+z)^{3.5}} = 8 \times 10^{-6}\ \mathrm{erg/cm^2/sec/sr} = 8\ \mathrm{nW/m^2/sr} \tag{178}$$

This is an energy density of $4\pi J/c = 0.002\ \mathrm{eV/cm^3}$ which is 0.8% of the CMB energy density and 5×10^{-7} of the critical density for $H_\circ = 65$.

For another example, consider the X-ray background produced by a hot intergalactic medium (IGM) with $n_e = n_e(0)(1+z)^3$. The emissivity of a hot plasma is $j_\nu = An_e^2 \exp(-h\nu/kT)/\sqrt{T}$ and the absorption is negligible. This gives

$$\left.\frac{I_{\nu_\circ(1+z)}}{(1+z)^3}\right|_{z=0} = An_e(0)^2 \frac{c}{H_\circ} \int \frac{(1+z)^6 \exp(-h\nu_\circ(1+z)/kT(z))/\sqrt{T}dz}{(1+z)^3(1+z)^2\sqrt{1+\Omega_{m\circ}z}} \tag{179}$$

so

$$I_{\nu_\circ} = An_e(0)^2 \frac{c}{H_\circ} \int \frac{(1+z)\exp(-h\nu_\circ(1+z)/kT(z))dz}{\sqrt{1+\Omega_{m\circ}z}} \tag{180}$$

Unfortunately, for any likely value of $n_e(0)$ this background is extremely small, and the expected distortion of the spectrum of the CMB produced by scattering off the hot electrons is not seen, so this model is incorrect.

As a final example, consider a model for dark matter consisting of heavy neutrinos that decay into a photon and a light neutrino. Conservation of momentum gives $h\nu = 0.5m_\nu c^2$. The emissivity is

$$j_\nu = \frac{\rho_\nu c^2 \delta(\nu - 0.5m_\nu c^2/h)}{8\pi\tau} \tag{181}$$

where τ is the lifetime of the neutrinos which we will assume is much larger than the age of the Universe so only a small fraction will have decayed by now. Then

$$\begin{aligned} I_{\nu_\circ} &= \frac{\rho_{\nu\circ}c^2 c}{8\pi H_\circ \tau} \int \frac{(1+z)^3\delta(\nu_\circ(1+z) - 0.5m_\nu c^2/h)dz}{(1+z)^3(1+z)^2\sqrt{1+\Omega_{m\circ}}} \\ &= \frac{\rho_{\nu\circ}c^2 c}{8\pi H_\circ \tau \nu_\circ} \left.\frac{1}{(1+z)^2\sqrt{1+\Omega_{m\circ}z}}\right|_{1+z=0.5m_\nu c^2/(h\nu_\circ)} \end{aligned} \tag{182}$$

For $\Omega_m = 1$ this gives $I_\nu \propto \nu^{3/2}$ from $\nu = 0$ up to the edge at $\nu = 0.5m_\nu c^2/h$. A reasonable limit on the extragalactic background light in the blue band is 10^{-6} of the critical density or $4\pi\nu I_\nu/c^3 < 10^{-6}\rho_{crit}$ so

$$\frac{\rho_\nu}{\rho_{crit}} \frac{1}{2H_\circ\tau} \left(\frac{h\nu_B}{0.5m_\nu c^2}\right)^{5/2} < 10^{-6} \tag{183}$$

Thus the lifetime of these hypothetical decaying neutrinos has to be larger than 10^{15} years.

11. Effects from Electron Scattering

For early epochs when the Universe was ionized, electron scattering is the dominant mechanism for transferring energy between the radiation field and the matter. As long as the electron temperature is less than about 10^8 K, the effect of electron scattering on the spectrum can be calculated using the Kompaneets (1957, Sov. Phys. JETP, 4, 730) equation:

$$\frac{\partial n}{\partial y} = x^{-2}\frac{\partial}{\partial x}\left[x^4\left(n + n^2 + \frac{\partial n}{\partial x}\right)\right] \tag{184}$$

where n is the number of photons per mode ($n = 1/(e^x - 1)$ for a blackbody), $x = h\nu/kT_e$, and the Kompaneets y is defined by

$$dy = \frac{kT_e}{m_e c^2} n_e \sigma_T c dt. \tag{185}$$

Thus y is the electron scattering optical depth times the electron temperature in units of the electron rest mass. Note that the electrons are assumed to follow a Maxwellian distribution, but that the photon spectrum is completely arbitrary.

Since the Kompaneets equation is describing electron scattering, which preserves the number of photons, one finds that the y derivative of the photon density N vanishes:

$$\begin{aligned}\frac{\partial N}{\partial y} &\propto \int x^2 \frac{\partial n}{\partial y} dx \\ &= \int \frac{\partial}{\partial x}\left[x^4\left(n + n^2 + \frac{\partial n}{\partial x}\right)\right] dx \\ &= 0\end{aligned} \tag{186}$$

The stationary solutions of the equation $\partial n/\partial y = 0$ are the photon distributions in thermal equilibrium with the electrons. Since photons are conserved, the photon number density does not have to agree with the photon number density in a blackbody at the electron temperature. Thus a more general Bose-Einstein thermal distribution is allowed: $n = 1/(\exp(x+\mu)-1)$. This gives $\partial n/\partial y = 0$ for all μ. Since the Bose-Einstein spectrum is a stationary point of the Kompaneets equation, it is the expected form for distortions produced at epochs when

$$(1+z)\frac{\partial y}{\partial z} = \sigma_T n_{e,\circ} \frac{kT_\circ}{m_e c^2}\frac{c}{H}(1+z)^4 > 1 \tag{187}$$

For the $\Omega_B h^2$ given by BBNS and the CMB acoustic peak heights, this redshift z_y where this inequality is crossed is well within the radiation dominated era, with a value $z_y = 10^{5.0}/\sqrt{\Omega_B h^2/0.0224}$.

There is a simple solution to the Kompaneets equation with non-zero $\partial n/\partial y$ which gives the Sunyaev-Zeldovich (1969, Ap. & Sp. Sci., 4, 301) or y spectral distortion. This is normally defined for the case where $h\nu << kT_e << m_e c^2$. Since the x variable is defined as $h\nu/kT_e$, these assumptions make dx very small, and thus the only term that contributes significantly to $\partial n/\partial y$ is the $\partial n/\partial x$ term: the n and n^2 terms are unimportant in this limit. But when the initial photon field is a blackbody with a temperature T_γ which is only slightly below the electron temperature we get a distortion of the same shape. Letting $f = T_e/T_\gamma$, we find that the initial photon field is

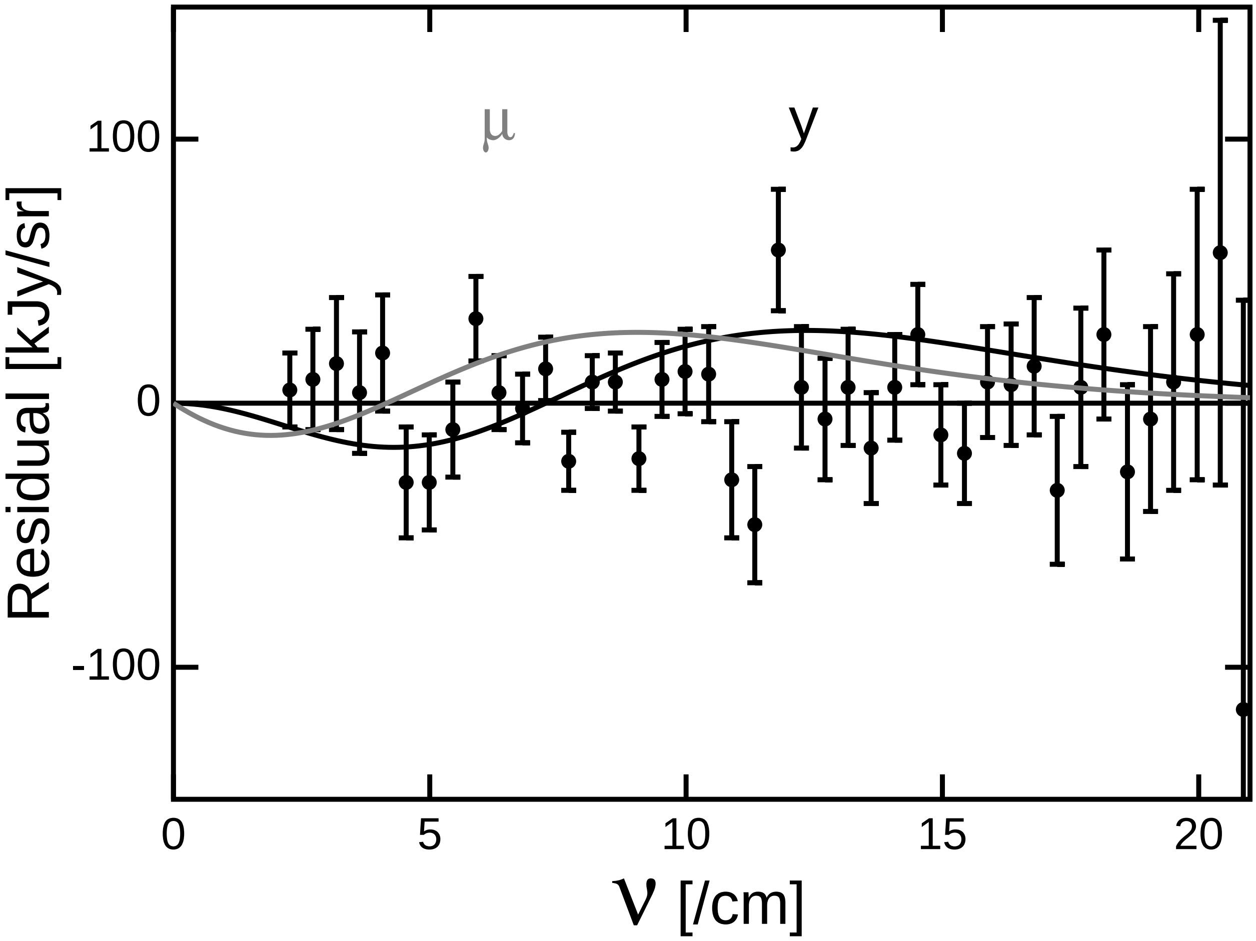

Fig. 18.— FIRAS residuals and the maximum allowed y and μ distortions: $y = 1.5 \times 10^{-5}$ and $\mu = 9 \times 10^{-5}$.

given by $n = 1/(\exp(fx) - 1)$. Therefore

$$\begin{aligned}
\left(n + n^2 + \frac{\partial n}{\partial x}\right) &= \frac{1}{\exp(fx) - 1} + \frac{1}{(\exp(fx) - 1)^2} - \frac{f\exp(fx)}{(\exp(fx) - 1)^2} = \\
\frac{(\exp(fx) - 1 + 1 - f\exp(fx)}{(\exp(fx) - 1)^2} &= \frac{(1-f)\exp(fx)}{(\exp(fx) - 1)^2} \\
&= (1 - f^{-1})\frac{\partial n}{\partial x}. \qquad (188)
\end{aligned}$$

Thus the distortion has a Sunyaev-Zeldovich shape but is reduced in magnitude by a factor $(1 - T_\gamma/T_e)$. Defining the "distorting" y as

$$dy_D = \frac{k(T_e - T_\gamma)}{m_e c^2} n_e \sigma_T c dt \qquad (189)$$

we find that the final spectrum is given by a frequency-dependent temperature given by

$$T_\nu = T_\circ \left[1 + y_D \left(\frac{x(e^x + 1)}{e^x - 1} - 4\right) + \ldots\right]. \qquad (190)$$

where $x = h\nu/kT_\circ$. The FIRAS spectrum in Figure 18 shows that $|y_D| < 1.5 \times 10^{-5}$.

The energy density transferred from the hotter electrons to the cooler photons in the y distortion is easily computed. The energy density is given by $U \propto \int x^3 n dx$ so

$$\frac{\partial U}{\partial y_D} = \int x \frac{\partial}{\partial x}\left(x^4 \frac{\partial n}{\partial x}\right) dx \tag{191}$$

which when integrated by parts twice gives

$$\begin{aligned} \frac{\partial U}{\partial y_D} &= -\int \left(x^4 \frac{\partial n}{\partial x}\right) dx \\ &= 4\int x^3 n dx = 4U. \end{aligned} \tag{192}$$

Thus the limit on y_D gives a corresponding limit on energy transfer: $\Delta U/U < 6 \times 10^{-5}$. Any energy which is transferred into the electrons at redshifts $z > 7$ where the Compton cooling time is less than the Hubble time will be transferred into the photon field and produce a y distortion. Since there are 10^9 times more photons than any other particles except for the neutrinos, the specific heat of the photon gas is overwhelmingly dominant, and the electrons rapidly cool (in a Compton cooling time) back into equilibrium with the photons. The energy gained by the photons is $\Delta U = 4yaT_\gamma^4$ which must be equal to the energy lost by the electrons and ions: $1.5(n_e+n_i)k\Delta T_e$. Since $y = \sigma_T n_e (kT_e/m_e c^2) c\Delta t$ we find the Compton cooling time

$$t_C = \frac{1.5(1+n_i/n_e)m_e c^2}{4\sigma_T c U_{rad}} = \frac{7.4 \times 10^{19}\ \mathrm{sec}}{(1+z)^4} \tag{193}$$

for an ion to electron ratio of 14/15. This becomes equal to the Hubble time, $(3.08568\times 10^{17}\mathrm{sec})/(h(1+z)^{1.5})$ for $\Omega_m = 1$, at $(1+z) = 9h^{0.4}$.

At redshifts $z > z_y = 10^5$, there will be enough electron scattering to force the photons into a thermal distribution with a μ distortion instead of a y distortion. However, the normal form for writing a μ distortion does not preserve the photon number density. Thus we should combine the μ distortion with a temperature change to give an effect that preserves photon number. The photon number density change with μ is given by

$$\begin{aligned} N &\propto \int \frac{x^2 dx}{\exp(x+\mu)-1} \\ &= \sum_{k=1}^{\infty} e^{-k\mu} \int x^2 e^{-kx} dx \\ &= 2\sum_{k=1}^{\infty} \frac{e^{-k\mu}}{k^3} \\ &= 2\left(\zeta(3) - \mu\zeta(2) + \ldots\right). \end{aligned} \tag{194}$$

A similar calculation for the energy density shows that

$$U \propto 6\left(\zeta(4) - \mu\zeta(3) + \ldots\right). \tag{195}$$

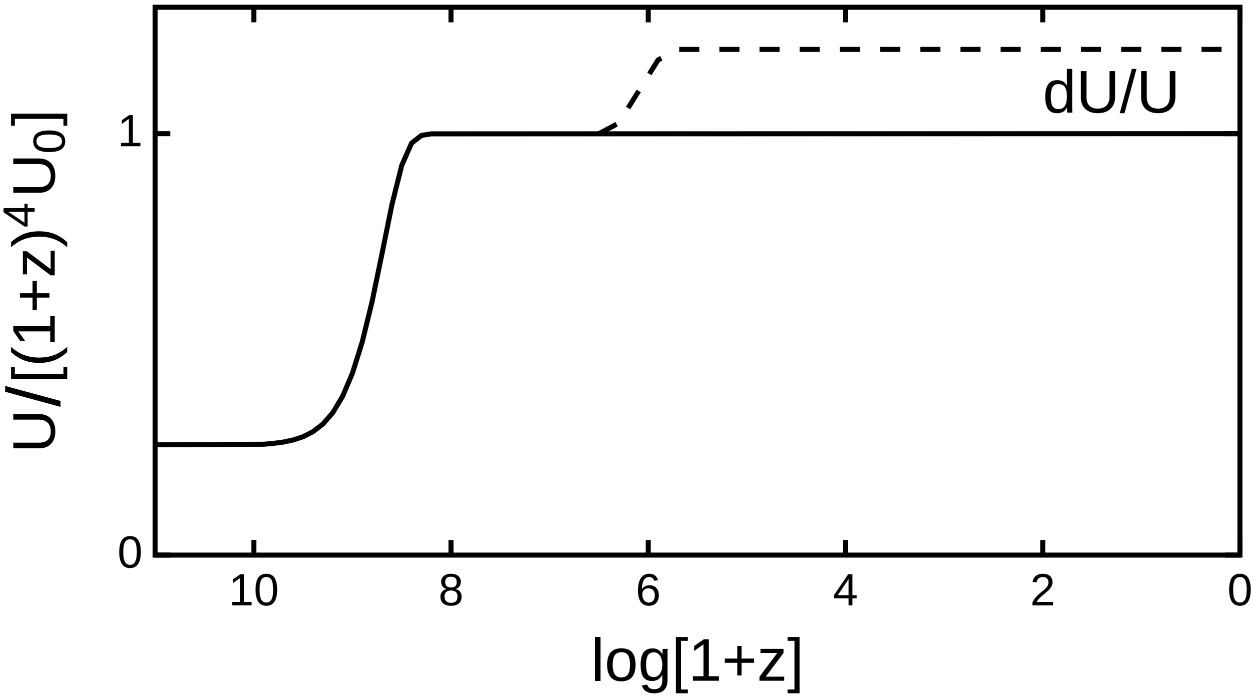

Fig. 19.— The radiation energy density U divided by $U_\circ(1+z)^4$. The large jump at $\log z \approx 8.3$ is when the e^+e^- pair plasma annihilated. Any further input dU occurring after $z_{th} = 10^{6.3}$ is limited to be $< 6 \times 10^{-5}$ instead of the 0.2 shown in the Figure.

In order to maintain $N = const$, the temperature of the photon field changes by an amount $\Delta T/T = \mu\zeta(2)/(3\zeta(3))$. Therefore, the energy density change at constant N is

$$\frac{\Delta U}{U} = \left(\frac{4\zeta(2)}{3\zeta(3)} - \frac{\zeta(3)}{\zeta(4)}\right)\mu = 0.714\mu. \tag{196}$$

Thus the FIRAS limit $|\mu| < 9 \times 10^{-5}$ implies $\Delta U/U < 6 \times 10^{-5}$. The "improved" form of the μ distortion, with a ΔT added to keep N constant, can be given as a frequency dependent brightness temperature:

$$T_\nu = T_\circ \left(1 + \mu\left[\frac{\zeta(2)}{3\zeta(3)} - x^{-1}\right] + \ldots\right). \tag{197}$$

In this form it is clear that a μ distortion has a deficit of low energy photons and a surplus of high energy photons with respect to a blackbody. In this it is like the y distortion, but the crossover frequency is lower. The "improved" form of the μ distortion is plotted in Figure 18.

Finally, at high enough redshift the process of double photon Compton scattering becomes fast enough to produce the extra photons needed to convert a distorted spectrum into a blackbody. Whenever a photon with frequency ν scatters off an electron, there is an impulse $\propto h\nu/c$ transferred to the electron. This corresponds to an acceleration $a \propto h\nu^2/(m_e c)$ for a time interval $\Delta t \propto 1/\nu$. The energy radiated in new photons is thus $\propto e^2h^2\nu^3/(m_e^2c^5)$ which is $\propto \alpha h\nu(h\nu/(m_ec^2))^2$. Since the rate of scatterings per photon is $\propto \Omega_B h^2(1+z)^3$, the overall rate of new photon creation is $\propto \Omega_B h^2(1+z)^5$ while the Hubble time is $\propto (1+z)^{-2}$. Thus the photon creation rate per Hubble time is $\propto (1+z)^3$ and at earlier enough times a blackbody spectrum is produced no matter how much energy is transferred to the photon field. The emissivity for double photon Compton scattering has approximately the same spectral shape as free-free emission, since both are due to impulsive accelerations. Thus the addition of photons to the radiation field can be described by

$$\frac{\partial n}{\partial y} = A\frac{1-e^{-x}}{x^3}\left(\frac{1}{e^x-1} - n\right) \tag{198}$$

where A is the ratio of photon creation via double photon Compton scattering or free-free emission to the increase of y. For double photon Compton scattering A scales like $(1+z)$. Because the photons are created primarily at low frequencies, $n(x)$ approaches a blackbody for $x \to 0$, while the frequency shifts due to Compton scattering are simultaneously transferring photons to larger values of x. A is very small at redshifts where distortions can survive, so the photon creation occurs primarily at $x \ll 1$. For any $n = (\exp(x+\mu(x)) - 1)^{-1}$,

$$\left(n + n^2 + \frac{\partial n}{\partial x}\right) = -n(n+1)\frac{\partial \mu}{\partial x} \tag{199}$$

so when $\mu(x) = \mu_\circ \exp(-x_\circ/x)$ and $x, x_\circ, \mu$ and A are all $\ll 1$, implying $n \approx n+1 \approx x^{-1}$,

$$x^{-2}\frac{\partial}{\partial x}\left(x^4\left(n + n^2 + \frac{\partial n}{\partial x}\right)\right) \approx \frac{-\mu_\circ \exp(-x_\circ/x)x_\circ^2}{x^4} \tag{200}$$

This cancels the $\partial n/\partial y$ due to photon addition, which is

$$\frac{\partial n}{\partial y} = A\frac{1-e^{-x}}{x^3}\left(\frac{1}{e^x-1} - n\right) \approx \frac{A\mu_\circ \exp(-x_\circ/x)}{x^4} \tag{201}$$

giving a quasi-equilibrium solution, if $x_\circ = \sqrt{A}$. With this form for n one finds a net photon addition rate of

$$\begin{aligned}\frac{\partial N}{\partial y} &= \int x^2 \frac{\partial n}{\partial y} dx \\ &= \int x^2 A\frac{1-e^{-x}}{x^3}\left(\frac{1}{e^x-1} - n\right) dx \\ &= A\mu/x_\circ = \mu\sqrt{A}. \end{aligned} \tag{202}$$

Since the deficit of photons associated with μ is $\mu\pi^2/3$, one finds a thermalization rate per unit y of

$$\frac{\partial \ln \mu}{\partial y} = \frac{3\sqrt{A}}{\pi^2} \propto \sqrt{1+z} \tag{203}$$

Since $(1+z)\partial y/\partial z \propto \Omega_B h^2(1+z)^2$, the overall rate for eliminating a μ distortion scales like $\Omega_B h^2(1+z)^{5/2}$ per Hubble time. A proper consideration (Burigana *et al.* 1991, ApJ, 379, 1-5) of this interaction of the photon creation process with the Kompaneets equation shows that the redshift from which $1/e$ of an initial distortion can survive is

$$z_{th} = \frac{4.24 \times 10^5}{[\Omega_B h^2]^{0.4}} \tag{204}$$

which is $z_{th} = 1.9 \times 10^6$ for $\Omega_B h^2 = 0.0224$.

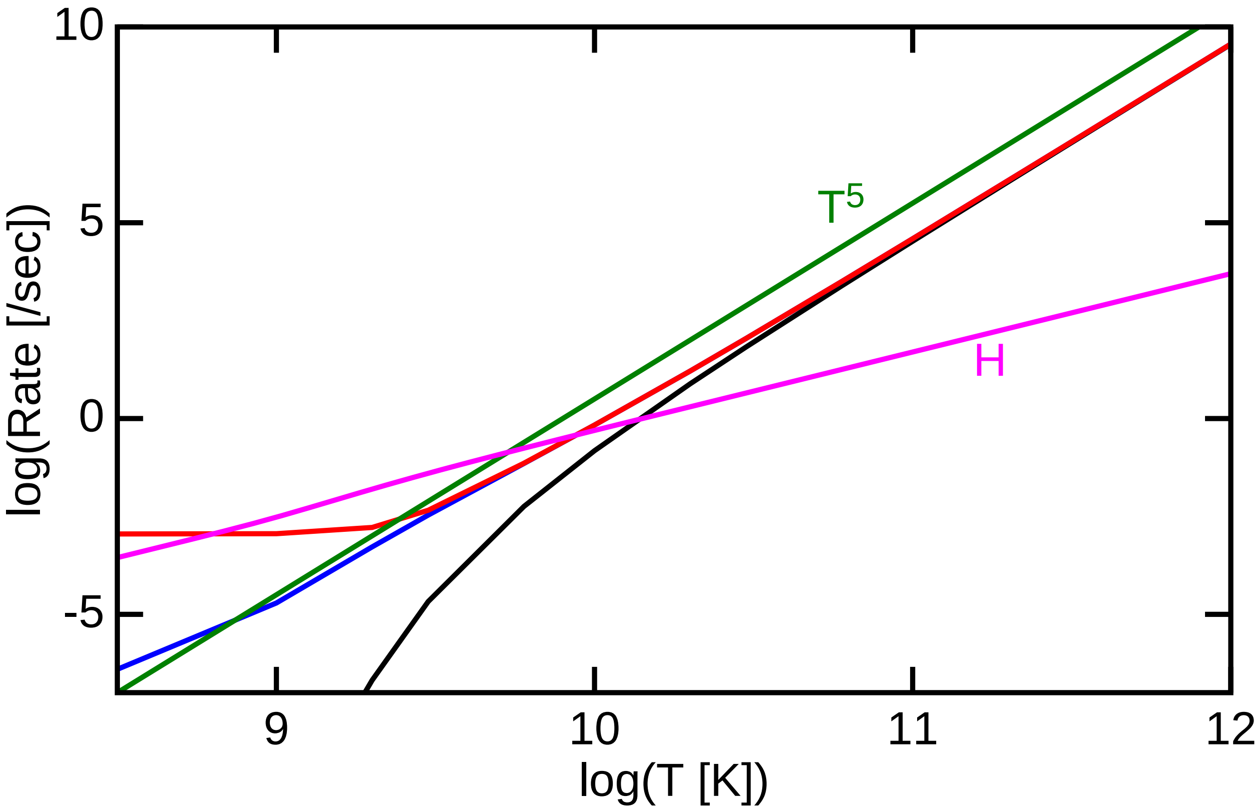

Fig. 20.— Rates for the expansion of the Universe, H, in magenta; and nuclear reaction rates $n \to p$, in blue without the free decay and red with the free neutron decay; and for $p \to n$ in black. A T^5 line is shown in green.

12. 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. The nucleus He^5 is unbound by 0.957 MeV relative to $\mathrm{He}^4 + n$, while Li^5 is unbound by 1.967 Mev relative to $\mathrm{He}^4 + p$. Be^8 is unbound by 94 keV relative to He^4+He^4. Li^8 beta decays in 0.8 seconds, and B^8 decays in 0.8 seconds by electron capture, which does not happen at the high temperatures during the first few minutes. But the route to carbon and heavier elements via B^8 is ineffective due to the low densities and high Coulomb barriers. 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 $T > 10^9$).

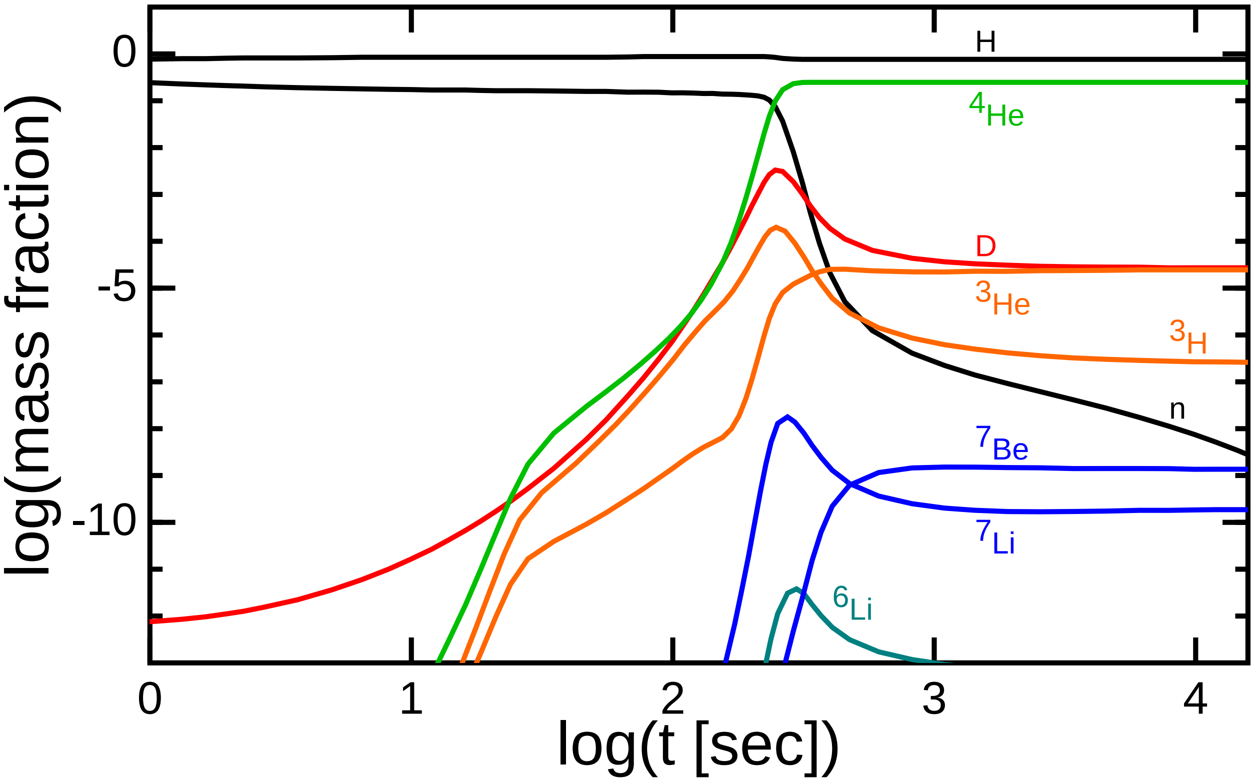

Fig. 21.— Mass fraction of various isotopes *vs.* time during the first few minutes after the Big Bang. This run is for a baryon density of $\Omega_b h^2 \approx 0.03$ which is higher than the best fit.

The thermal equilibrium between neutrons and protons at $t \approx 1$ sec is maintained by weak interactions. The cross-section for the reaction $\nu_e + n \to p + e^-$ is given by

$$\sigma = \frac{2\pi^2 \hbar^3 v_e (E_\nu + Q)^2}{f \tau_n m_e^5 c^9} \tag{205}$$

where $Q = (m_n - m_p)c^2 = 1.293$ MeV, the factor $f = 1.634$ (Peebles 1971) and τ_n is the neutron mean life-time of 878.5 ± 1 sec. For $E_\nu = 1$ MeV this gives $\sigma = 6 \times 10^{-43}$ cm^2, and the cross-section scales roughly like E_ν^2 for high energy. The neutrino density is in thermal equilibrium, so the interaction rate $\langle n\sigma v\rangle$ is given approximately by

$$\begin{aligned}
\langle n\sigma v\rangle &\approx \int \left[\frac{4\pi g_s p^2 dp/h^3}{\exp(pc/kT)+1} \times 10^{-43}\ \mathrm{cm}^2 \left(\frac{pc+Q}{1\ \mathrm{MeV}}\right)^2 c\right] \\
&= 4\pi g_s c \left(\frac{kT}{hc}\right)^3 \times 10^{-43}\ \frac{\frac{15}{16}\Gamma(5)\zeta(5)(kT)^2 + \frac{7}{4}\Gamma(4)\zeta(4)(kTQ) + \frac{3}{4}\Gamma(3)\zeta(3)Q^2}{(1\ \mathrm{MeV})^2} \\
&= 4\pi g_s c \left(\frac{kT}{hc}\right)^3 \times 10^{-43} \left(23.33\left[\frac{kT}{1\ \mathrm{MeV}}\right]^2 + 14.69\left[\frac{kT}{1\ \mathrm{MeV}}\right] + 3.014\right) \\
&= 0.5\ \mathrm{sec}^{-1}\ \text{at}\ T = 10^{10}\ \mathrm{K}
\end{aligned} \tag{206}$$

for the $g_s = 1$ appropriate for neutrinos. The neutrino density is $n_\nu = 7 \times 10^{30}(T/10^{10}\ \mathrm{K})^3$ cm^{-3}. The reaction $e^+ + n \to p + \overline{\nu_e}$ essentially doubles the rate of $n \leftrightarrow p$ interchanges. The expansion

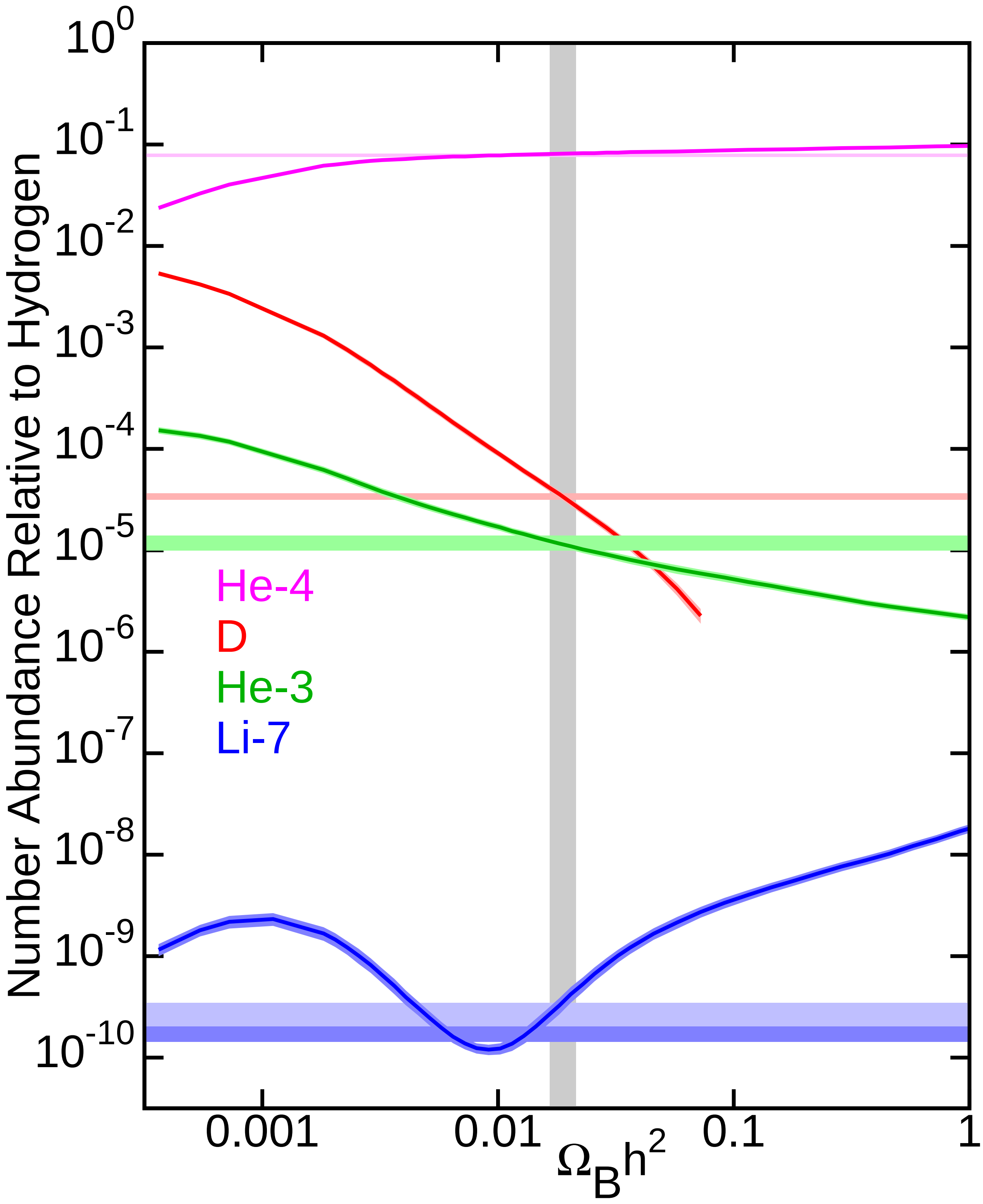

Fig. 22.— Final abundances of the light elements as a function of the baryon density. The horizontal bands show the observed values, and the vertical band shows the best fit value of the baryon density.

rate of the Universe is H so the weak interactions will freeze out when $\langle n\sigma v\rangle < H$. We know that the density will be essentially equal to the critical density, so $H = \sqrt{8\pi G\rho/3}$. The density will be determined by the thermal equilibrium density of γ's, ν's and e^+e^-'s. This gives

$$H = \sqrt{\frac{8\pi G}{3}\,\frac{aT^4 + 3\times(7/8)aT^4 + 2\times(7/8)aT^4}{c^2}} = 0.5\left(\frac{T}{10^{10}\ \mathrm{K}}\right)^2\ \mathrm{sec}^{-1} \tag{207}$$

Thus the weak interactions freeze out when $T_f \approx 10^{10}$ 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 878.5 ± 1 seconds (Mathews *et al.*, astro-ph/0408523) 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 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 21 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 $d + d \to \mathrm{He} + \gamma$. 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:

$$\frac{dX_d}{dt} = -2\alpha(T)n_B X_d^2 = -2\alpha(T_1(t/t_1)^{1/2})n_B(t_1)(t/t_1)^{-3/2}X_d^2 \tag{208}$$

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

$$X_d^{-1} = 2n_B(t_1)\int_{t_1}^{\infty} \alpha(T_1(t/t_1)^{-1/2})(t/t_1)^{-3/2}dt + X_d(t_1)^{-1} \tag{209}$$

The time versus temperature is given by

$$\begin{aligned} \frac{1.68aT^4}{c^2} &= \frac{3}{32\pi G t^2} \\ t &= \frac{1.78\times 10^{20}\ \mathrm{K}^2}{T^2} \end{aligned} \tag{210}$$

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

$$X_d = \frac{1}{n_B(t_1)}\,\frac{T_1}{4t_1\int_0^{T_1}\alpha(T)dT} = \frac{T_1^3}{n_B(t_1)}\,\frac{1}{7.1\times10^{20}\ \mathrm{K}^2\int_0^{T_1}\alpha(T)dT} \tag{211}$$

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 $n_\gamma = 411\ \mathrm{cm}^{-3}$ for $T_\circ = 2.725$ K, we find $n_B(t_\circ) = 0.411\eta_{10}\times10^{-7}$/cc. The critical density for $H_\circ = 100$ km/sec/Mpc corresponds to $n_B = 1.12\times10^{-5}$/cc, so

$$\Omega_B h^2 = \frac{0.411\eta_{10}\times10^{-7}}{1.12\times10^{-5}} = 0.00367\eta_{10} \tag{212}$$

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 $2.5 < \eta_{10} < 6$ (Copi *et al.*, 1995), which corresponds to $\Omega_B h^2 = 0.0142\pm44\%$. Schramm & Turner (1997, astro-ph/9706069) give $\eta_{10} = 6\pm1$ or $\Omega_B h^2 = 0.022\pm0.004$. Burles, Nollett, Truran & Turner (1999, PRL, 82, 4176) give $\eta_{10} = 5.1\pm0.5$ or $\Omega_B h^2 = 0.019\pm0.024$ so the baryon density is well constrained. Figure 22 shows the abundances of various isotopes as a function of the baryon density. For $H_\circ = 65$ this gives $\Omega_B = 0.045$ 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.

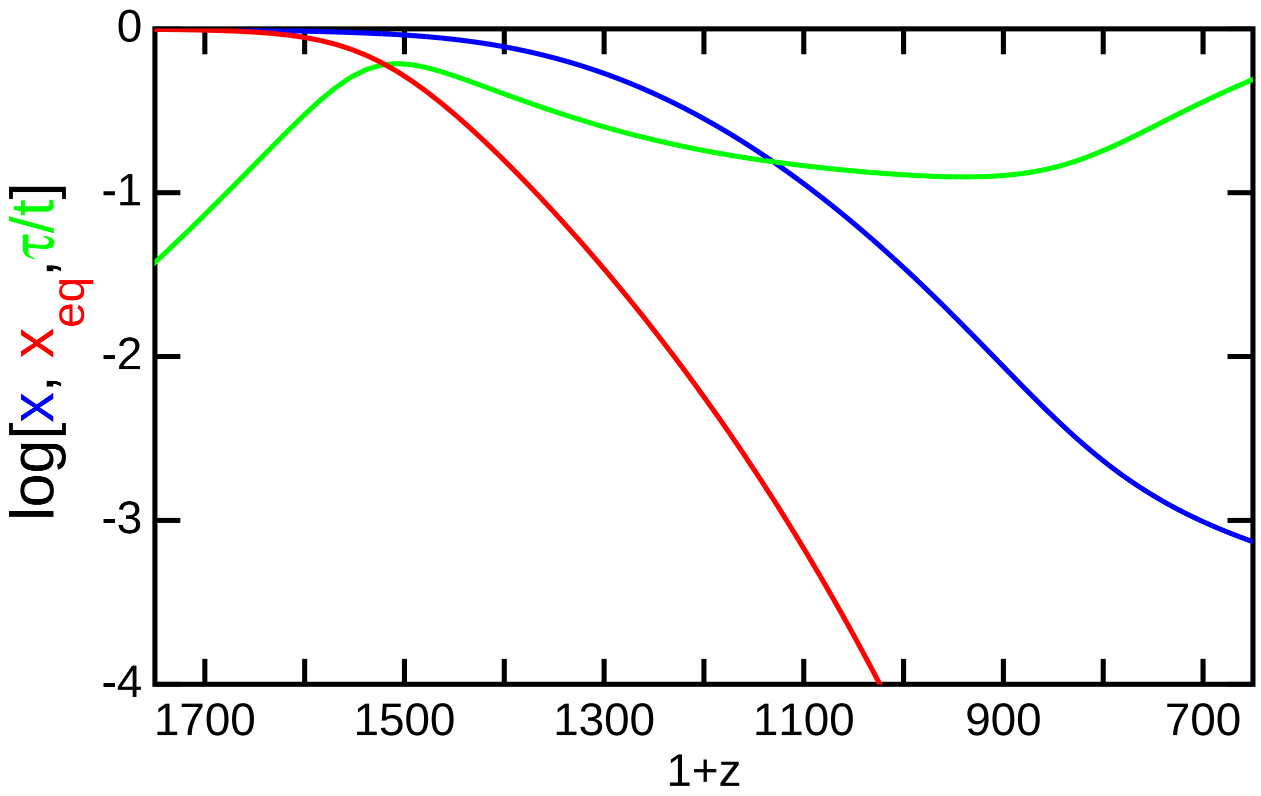

Fig. 23.— Saha equilbrium and actual values of the ionization fraction x for a ΛCDM model with $H_\circ = 70$, $\Omega_m = 0.3$ and $\Omega_b h^2 = 0.20$. The ratio of equilibration time constant to the age of the Universe is also shown.

13. Last Scattering

In Peacock's *Cosmological Physics* we find

$$\frac{d(nx)}{dt} = -R(T)(nx)^2 \frac{\Lambda}{\Lambda + \beta_e(T)} \tag{213}$$

where $R = 3{\times}10^{-11} T^{-1/2}$ cm^3/s is the recombination coefficient into excited levels. Recombinations into the ground state produce an ionizing photon which immediately ionizes another atom and are thus not effective. This neglects photoionizations from the ground state but these turn off well before the last scattering surface. Also

$$\frac{dt}{dz} = -3.09 \times 10^{17} (\Omega h^2)^{-1/2} z^{-5/2} \text{ sec} \tag{214}$$

Combining these gives

$$d(x^{-1}) = -f(T_\circ(1+z))(\Omega h^2)^{-1/2}(\Omega_b h^2) z^{1/2} dz \tag{215}$$

so $x(z) \propto (\Omega h^2)^{1/2}/(\Omega_b h^2)$. The integral for optical depth is $\int nx dt$ and since $n(z) \propto (\Omega_b h^2)$ and $dt \propto (\Omega h^2)^{-1/2}$ the optical depth *vs.* redshift is independent of cosmological parameters. Peacock gives

$$\tau(z) \approx 0.37 \left(\frac{z}{1000}\right)^{14.25} \tag{216}$$

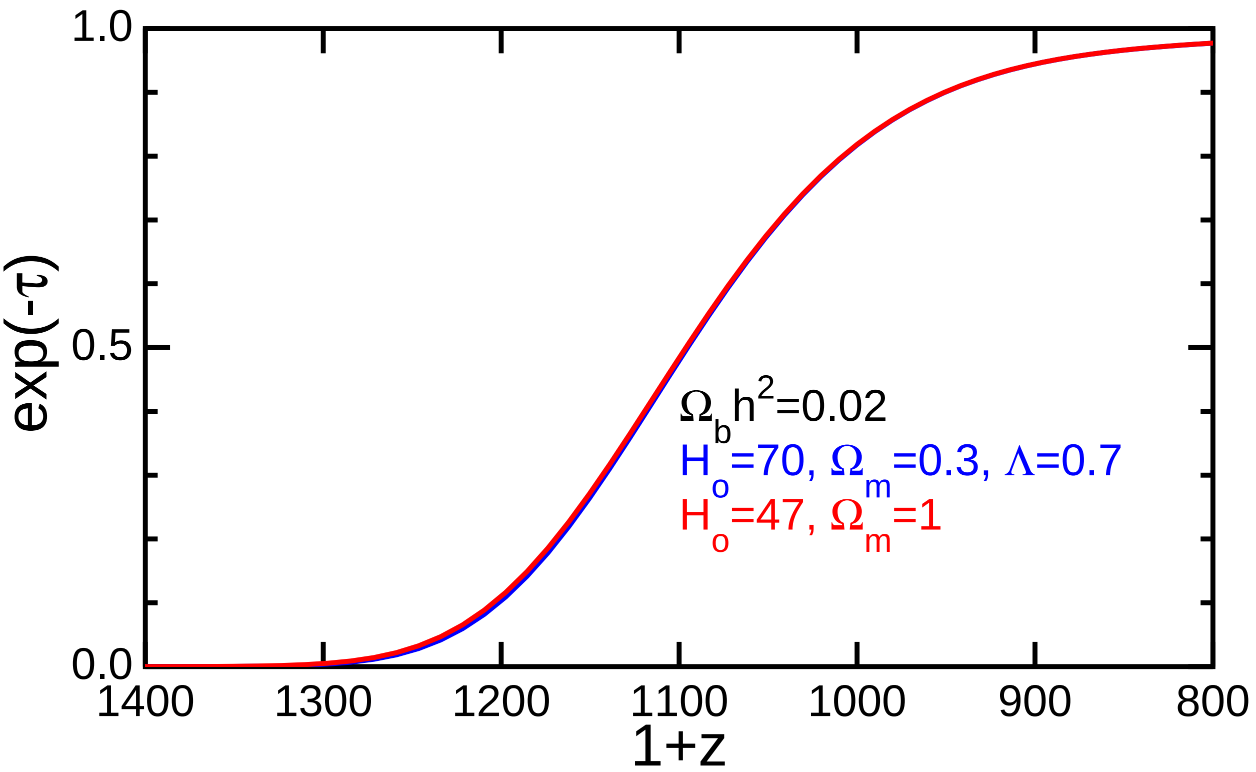

Fig. 24.— The photon survival fraction $\exp(-\tau)$ *vs.* redshift for two cosmological models.

For this approximation $e^{-\tau} = 0.5$ when $z = 1045$. Using the equations in Peebles' *Principles of Physical Cosmology* I have computed the curves shown in Figure 23. The Saha equilibrium value of x_{eq} falls at redshifts a few hundred units higher than the actual x. The ratio of the time constant for establishing ionization equilbrium to the age of the Universe, τ/t, varies from few tenths to slightly less than 1, which explains why the ionization is close to but not quite in equilibrium. The survival fraction $e^{-\tau}$ is shown in Figure 24, which indeed shows very little variation in the $\tau(z)$ curve for different cosmological models. $e^{-\tau} = 0.5$ occurs at $1 + z = 1090$. The width of the surface of last scattering computed from the $e^{-\tau} = 0.25$ and 0.75 points is $\sigma_z = 93$.

The width in comoving radial distance is $c(1 + z)(dt/dz)\sigma_z$ which scales like $(\Omega h^2)^{-1/2}$. The angular equivalent of this width is

$$\theta_d = \frac{c(1 + z)(dt/dz)\sigma_z}{(1 + z)D_A(z)} \tag{217}$$

which scales in almost the same way with parameters as the angular scale for the first Doppler peak. However, the sound speed in the photon-baryon gas before recombination is not involved. The value is $\theta_d = 4.5'$ for the ΛCDM model. One expects that C_ℓ will be reduced for $\ell > 1/\theta_d = 760$ by the interference between the front and back sides of the "surface" of last scattering.

14. 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

$$ds^2 = c^2 dt^2 - a(t)^2 \left[d\rho^2 + R_\circ^2 \left\{ \begin{array}{c} \sinh^2\left(\rho R_\circ^{-1}\right) \\ \left(\rho R_\circ^{-1}\right)^2 \\ \sin^2\left(\rho R_\circ^{-1}\right) \end{array} \right\} \left(d\theta^2 + \sin^2\theta d\phi^2\right) \right] \tag{218}$$

and the proper radial distance to an object at $z = \infty$ is just

$$D_H = \rho = \int \frac{cdt}{a(t)} = \int_0^\infty (1+z) c \frac{dt}{dz} dz \tag{219}$$

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

$$D_H(z) = \int_z^\infty (1+z) c \frac{dt}{dz} dz = \frac{2c}{H_\circ \sqrt{1+z}} \tag{220}$$

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 ρ *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

$$ds^2 = a(t)^2 \left(d\eta^2 - \left[d\rho^2 + R_\circ^2 \left\{ \begin{array}{c} \sinh^2\left(\rho R_\circ^{-1}\right) \\ \left(\rho R_\circ^{-1}\right)^2 \\ \sin^2\left(\rho R_\circ^{-1}\right) \end{array} \right\} \left(d\theta^2 + \sin^2\theta d\phi^2\right) \right] \right) \tag{221}$$

and the paths of light rays are obviously the $\pm 45°$ lines $\rho = \rho_\circ \pm (\eta - \eta_\circ)$. 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 $\pm 45°$.

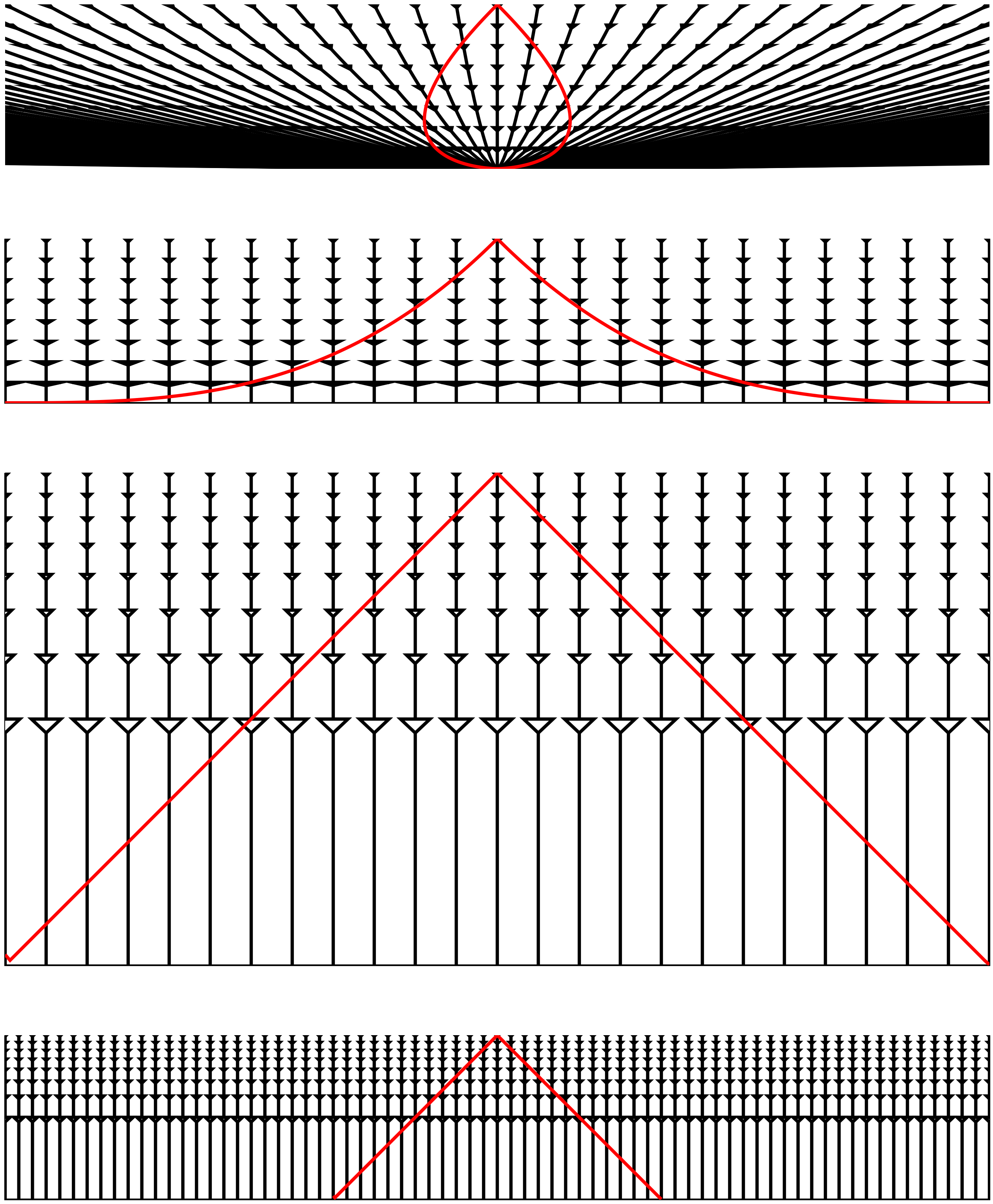

Fig. 25.— From top to bottom: a) $\Omega = 1$ S-T in standard form; b) with distances divided by $a(t)$; c) time axis "stretched" into conformal time; d) a wider view showing the Universe is much bigger than the observable Universe.

15. Inflationary Scenario

15.1. 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 behaviors: 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

$$V(\phi_1, \phi_2) = \lambda(\sigma^2 - (\phi_1^2 + \phi_2^2))^2 \tag{222}$$

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

$$V(\psi, \phi_2) \approx 4\lambda\sigma^2\psi^2 + \ldots \tag{223}$$

The Lagrangian density is then

$$\begin{aligned} \mathcal{L} &= \partial_\mu\psi\partial^\mu\psi - 4\lambda\sigma^2\psi^2 \\ &+ \partial_\mu\phi_2\partial^\mu\phi_2 \end{aligned} \tag{224}$$

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.

15.2. 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

$$\begin{aligned} \phi_1 &= \sigma f(r)\frac{x}{\sqrt{x^2+y^2}} \\ \phi_2 &= \sigma f(r)\frac{y}{\sqrt{x^2+y^2}} \end{aligned} \tag{225}$$

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

$$T = 4\pi \int \left(\sigma^2 \left[\left(\frac{df}{dr}\right)^2 + \left(\frac{f}{r}\right)^2 \right] + \lambda\sigma^4(1 - f(r)^2)^2 \right) r dr \tag{226}$$

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.

15.3. 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 $ct = 3{\times}10^{-26}$ 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.

15.4. 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

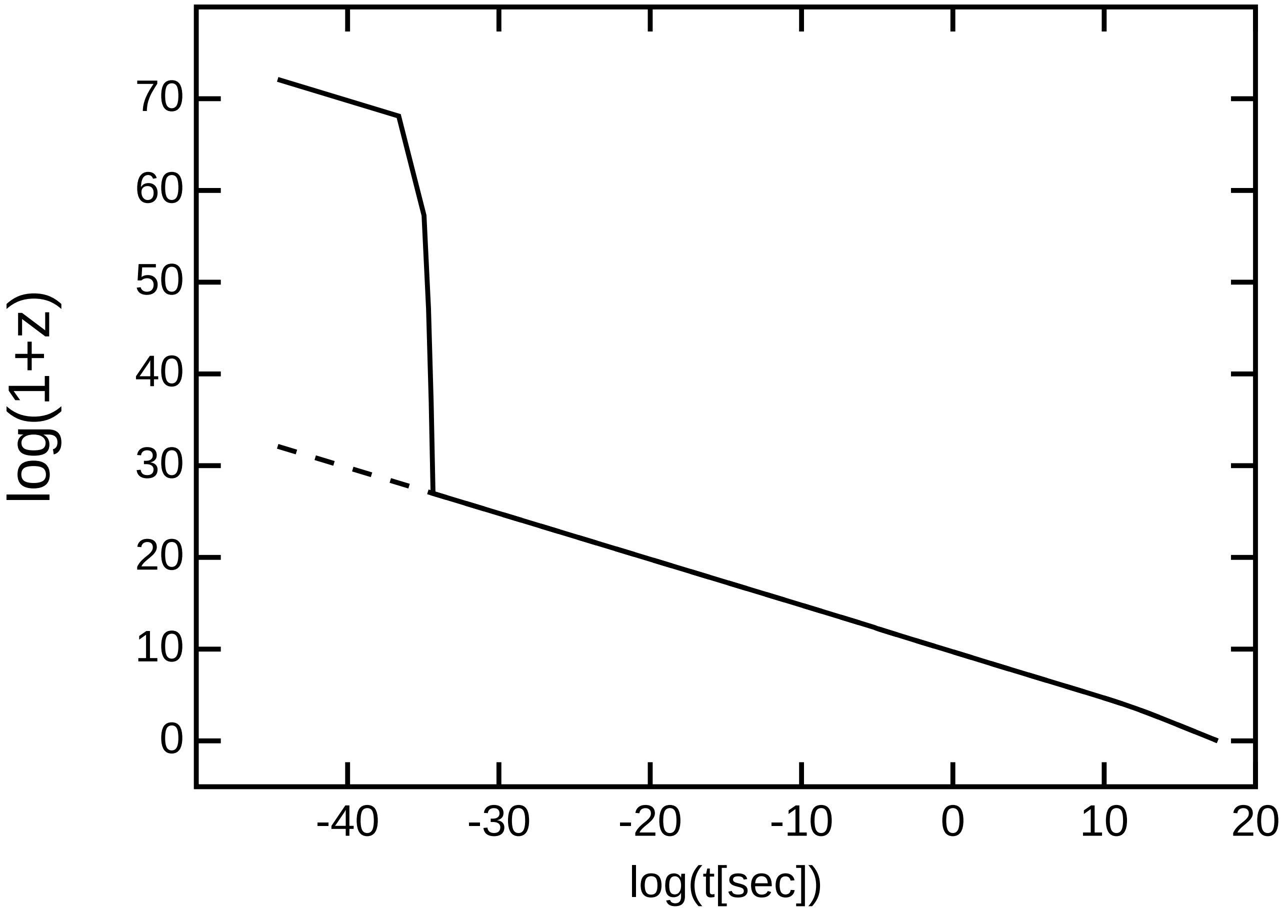

Fig. 26.— The redshift *vs.* time for the standard hot Big Bang model (dashed) and for a model with inflation.

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 26 shows $1 + z = 1/a$ *vs.* time with and without inflation.

This expansion will solve the three problems of the Big Bang:

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

$$\left(\frac{1}{\Omega(t)} - 1\right) = \left(\frac{1}{\Omega_\circ} - 1\right)\left(\frac{\rho_\circ a(t_\circ)^2}{\rho(t)a(t)^2}\right) \tag{227}$$

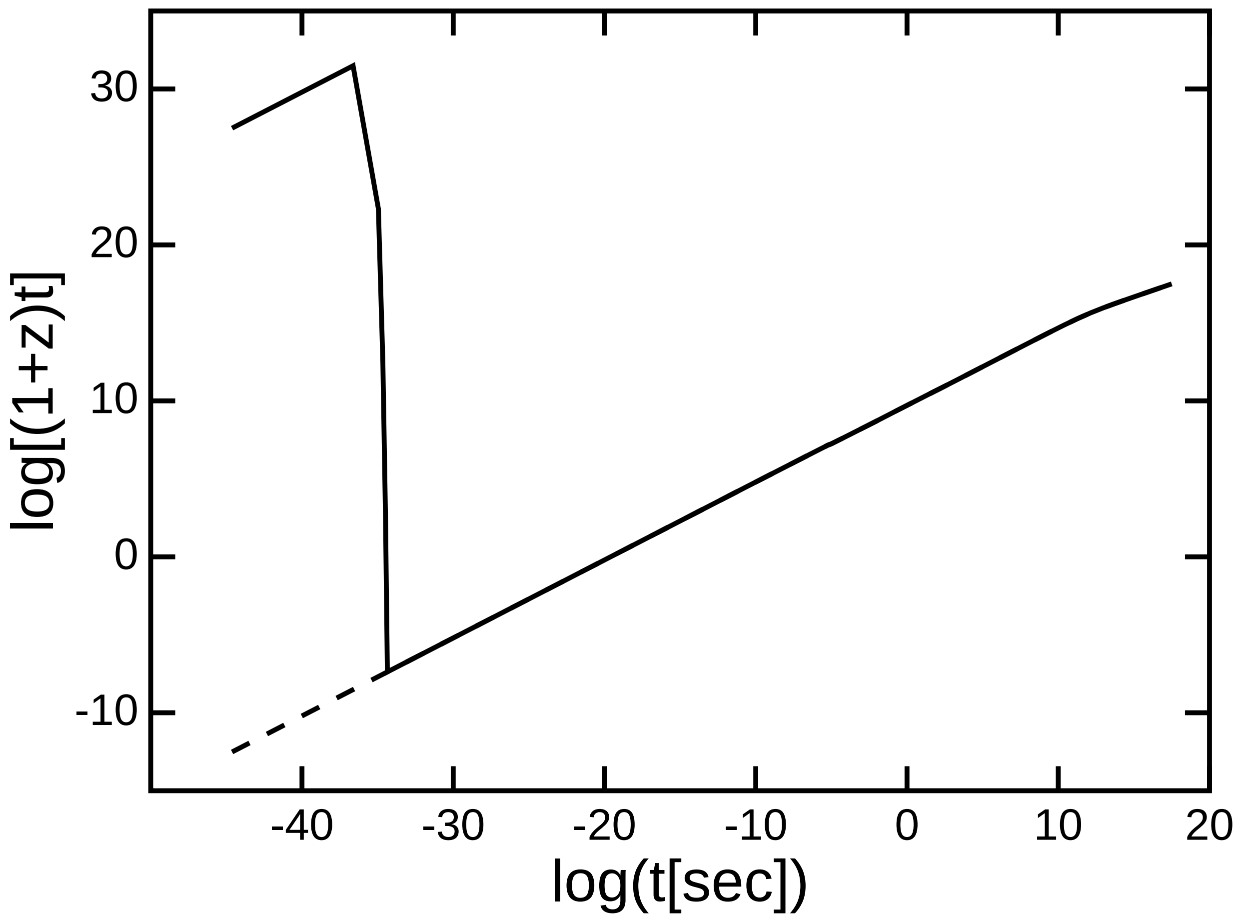

Fig. 27.— The contribution to the conformal time per octave in time, $(1+z)t$, for the standard hot Big Bang model (dashed) and for a model with inflation.

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

$$\eta = \int (1+z)cdt = \int ct(1+z)d\ln t \tag{228}$$

so the contribution per logarithmic time interval is given by $t \times (1+z)$, shown in Figure 27. This peaks after recombination in the standard model but before recombination in the 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 definite 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.

16. Fluctuations from Inflation

Inflation produces such a huge expansion that quantum fluctuations on the microscopic scale can grow to be larger than the observable Universe. These perturbations can be the seeds of structure formation and also will create the anisotropies seen by COBE for spherical harmonic indices $\ell \geq 2$. For perturbations that are larger than $\approx c_s t$ (or $\approx c_s/H$) we can ignore pressure gradients, since pressure gradients produce sound waves that are not able to cross the perturbation in a Hubble time. In the absence of pressure gradients, the density perturbation will evolve in the same way that a homogeneous Universe does, and we can use the equation

$$\rho a^2 \left(\frac{1}{\Omega} - 1\right) = \text{const} \tag{229}$$

along with the assumptions that $\Omega \approx 1$ for early times, and $\Delta\rho << \rho$ as indicated by the smallness of the ΔT's seen by COBE, to derive

$$-\rho a^2 \left(\frac{1}{\Omega} - 1\right) \approx \rho_{crit} a^2 \Delta\Omega \approx \Delta\rho a^2 = \text{const} \tag{230}$$

and hence

$$\Delta\phi = \frac{G\Delta M}{R} = \frac{4\pi}{3}\frac{G\Delta\rho_\circ (aL)^3}{aL} = \frac{1}{2}\frac{\Delta\rho_\circ}{\rho_{crit}}(H_\circ L)^2 \tag{231}$$

where L is the comoving size of the perturbation. This is independent of the scale factor so it doesn't change due to the expansion of the Universe.

During inflation the Universe is approximately in a steady state with constant H. Thus the magnitude of $\Delta\phi$ for perturbations with physical scale c/H will be the same for all times during the inflationary epoch. But since this constant physical scale is aL and the scale factor a changes by more than 30 orders of magnitude during inflation, this means that the magnitude of $\Delta\phi$ will be the same over 30 decades of comoving scale L. Thus we get a strong prediction that $\Delta\phi$ will be the same on all observable scales from $c/H_\circ$ down to the scale which is no longer always larger than the sound speed horizon. This means that

$$\frac{\Delta\rho}{\rho} \propto L^{-2} \tag{232}$$

so the Universe becomes extremely homogeneous on large scales even though it is quite inhomogeneous on small scales.

This behaviour of $\Delta\phi$ being independent of scale is called *equal power on all scales.* It was originally predicted by Harrison (1970, PRD, 1, 2726-2730) and Zel'dovich (1972, MNRAS, 160, 1p) and Peebles & Yu (1970, ApJ, 162, 815) based on a very simple argument: There is no scale length provided by the early Universe, and thus the perturbations should be *scale-free*: a power law. Therefore $\Delta\phi \propto L^m$. The gravitational potential divided by c^2 is a component of the metric, and if it gets comparable to unity then wild things happen. If $m < 0$ then $\Delta\phi$ gets large for small L, and many black holes would form. But we observe that this did not happen. Therefore $m \geq 0$. But if $m > 0$ then $\Delta\phi$ gets large on large scales, and the Universe would be grossly inhomogeneous. But we observe that this is not the case, so $m \leq 0$. Combining both results requires that $m = 0$, which is a *scale-invariant* perturbation power spectrum. This particular power law power spectrum is called the Harrison-Zel'dovich spectrum. It was expected that the primordial perturbations should follow a Harrison-Zel'dovich spectrum because all other answers were wrong, but the inflationary scenario provides a good mechanism for producing a Harrison-Zel'dovich spectrum.

Sachs & Wolfe (1967, ApJ, 147, 73) show that a gravitational potential perturbation produces an anisotropy of the CMBR with magnitude

$$\frac{\Delta T}{T} = \frac{1}{3}\frac{\Delta\phi}{c^2} \tag{233}$$

where $\Delta\phi$ is evaluated at the intersection of the line-of-sight and the surface of last scattering (or recombination at $z \approx 1100$). The (1/3) factor arises because clocks run faster by factor $(1 + \phi/c^2)$ in a gravitational potential, and we can consider the expansion of the Universe to be a clock. Since the scale factor is varying as $a \propto t^{2/3}$ at recombination, but faster expansion leads to a decreased temperature by $\Delta T/T = -(2/3)\Delta\phi/c^2$ which when added to the normal gravitational redshift $\Delta T/T = \Delta\phi/c^2$ yields the (1/3) factor above. The anisotropy is usually expanded in spherical harmonics:

$$\frac{\Delta T(\hat{n})}{T} = \sum_\ell \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{n}) \tag{234}$$

Because the Universe is approximately isotropic the probability densities for all the different m's at a given ℓ are identical. Furthermore, the expected value of $\Delta T(\hat{n})$ is obviously zero, and thus the expected values of the $a_{\ell m}$'s is zero. But the variance of the $a_{\ell m}$'s is a measurable function of ℓ, defined as

$$C_\ell = \langle |a_{\ell m}|^2 \rangle \tag{235}$$

The harmonic index ℓ associated with an angular scale θ is given by $\ell \approx 180°/\theta$, but the total number of spherical harmonics contributing to the anisotropy power at angular scale θ is given by $\Delta\ell \approx \ell$ times $2\ell + 1$. Thus to have equal power on all scales one needs to have approximately $C_\ell \propto \ell^{-2}$. Given that the square of the angular momentum operator is actually $\ell(\ell + 1)$, it is not surprising that the actual angular power spectrum of the CMBR predicted by "equal power on all scales" is

$$C_\ell = \frac{4\pi\langle Q^2\rangle}{5T_\circ^2}\frac{6}{\ell(\ell+1)} \tag{236}$$

where $\langle Q^2 \rangle$ or Q^2_{rms-PS} is the expected variance of the $\ell = 2$ component of the sky, which must be divided by $T_\circ^2$ because the $a_{\ell m}$'s are defined to be dimensionless. The "4π" term arises because the mean of $|Y_{\ell m}|^2$ is $1/(4\pi)$, so the $|a_{\ell m}|^2$'s must be 4π times larger to compensate. Finally the quadrupole has 5 components while C_ℓ is the variance of a single component, giving the "5" in the denominator. The COBE DMR experiment determined $\sqrt{\langle Q^2 \rangle} = 18\ \mu$K, and that the C_ℓ's from $\ell = 2$ to $\ell = 20$ were consistent with Equation 236.

The angular correlation function of the anisotropy is given by

$$C(\theta) = \frac{\langle \Delta T(\hat{n})\Delta T(\hat{n}')\rangle}{T_\circ^2} = \frac{1}{4\pi}\sum_\ell (2\ell+1) C_\ell P_\ell(\cos\theta) \tag{237}$$

where P_ℓ is a Legendre polynomial. If the temperature is measured in a beam with a finite beamwidth, then there will be reduction in the effect of high ℓ terms. For example, if the beam is approximated by a Gaussian with FWHM given by $\sigma\sqrt{8\ln 2}$, then the response to $Y_{\ell m}$ will be reduced by a factor of $\exp(-\ell(\ell+1)\sigma^2/2)$, giving an effective correlation function

$$C_\sigma(\theta) = \frac{\langle \Delta T(\hat{n})\Delta T(\hat{n}')\rangle}{T_\circ^2} = \frac{1}{4\pi}\sum_\ell (2\ell+1)C_\ell \exp(-\ell(\ell+1)\sigma^2)P_\ell(\cos\theta) \tag{238}$$

This low-pass filtering stops the logarithmic divergence that would otherwise exist as $\theta \to 0$ for $C_\ell \propto 1/(\ell(\ell+1))$. Most experiments (other than COBE and WMAP) do not produce a map of the whole sky. Instead they measure temperature differences between points separated by some chopping angle on the sky. A "single-subtracted" experiment with chopper throw θ measures $\Delta T/T_\circ = [T(\hat{n}) - T(\hat{n}+\theta)]/T_\circ$. The variance of this measured value is

$$\begin{aligned}\left\langle\left(\frac{\Delta T}{T_\circ}\right)^2\right\rangle &= \left\langle\left(\frac{\Delta T(\hat{n})}{T_\circ}\right)^2\right\rangle + \left\langle\left(\frac{\Delta T(\hat{n}+\theta)}{T_\circ}\right)^2\right\rangle - 2\left\langle\frac{\Delta T(\hat{n})}{T_\circ}\frac{\Delta T(\hat{n}+\theta)}{T_\circ}\right\rangle \\ &= 2(C(0) - C(\theta)) \end{aligned}\tag{239}$$

Including both the effect of chopping and the beamsize gives

$$\left\langle\left(\frac{\Delta T}{T_\circ}\right)^2\right\rangle = 2(C_\sigma(0) - C_\sigma(\theta)) = \frac{1}{4\pi}\sum_\ell (2\ell+1)C_\ell \exp(-\ell(\ell+1)\sigma^2)(2 - 2P_\ell(\cos\theta)) \tag{240}$$

More complicated chopping patterns such as double subtraction with weights -0.5, 1 and -0.5 and easily be used instead of single subtraction, and give modifications of the θ dependence.

16.1. Relation of Potential and Density Perturbations

The density perturbations are usually described in terms of

$$\delta(\vec{r}) = \frac{\Delta\rho}{\overline{\rho}} = \frac{(2\pi)^{3/2}}{V_u^{1/2}}\sum \delta_k e^{i\vec{k}\cdot\vec{r}} \tag{241}$$

where V_u is the volume of a box with periodic boundary conditions. See Equation (21.39) in Peebles. Using a normalizing box makes it easier to keep track of the dimensions of variables. Note that $\delta(\vec{r})$ is dimensionless but δ_k has units of length$^{3/2}$. The two point correlation function of the density is given by

$$\begin{aligned}\xi(\vec{r}) &= \langle\delta(\vec{r}')\delta(\vec{r}'+\vec{r})^*\rangle = \frac{(2\pi)^3}{V_u}\sum\sum \delta_k\delta_{k'}^*\left\langle e^{i\vec{k}\cdot\vec{r}'}e^{-i\vec{k}'\cdot(\vec{r}'+\vec{r})}\right\rangle \\ &= \frac{(2\pi)^3}{V_u}\sum|\delta_k|^2 e^{-i\vec{k}\cdot\vec{r}}\end{aligned}\tag{242}$$

But since the spacing of the k's in the sum is given $\Delta k_x = 2\pi/L_x$, where L_x is the length of the box in x, we have $\Delta k_x\Delta k_y\Delta k_z = (2\pi)^3/V_u$. We define the power spectrum of the density perturbations as

$$P(k) = \langle|\delta_k|^2\rangle \tag{243}$$

which has the units of length^3, and get

$$\begin{aligned}\xi(r) &= \int P(k)e^{-i\vec{k}\cdot\vec{r}}d^3\vec{k} = \int P(k)k^2 \int e^{-ikr\cos\theta}d\Omega dk \\ &= \int P(k)k^2\, 2\pi \int e^{-ikr\mu}d\mu = 4\pi \int k^2 P(k)\frac{\sin(kr)}{kr}dk \end{aligned} \tag{244}$$

When dealing with perturbations in an expanding Universe it is most convenient to work in comoving coordinates. Poisson's equation involves derivatives that must be taken in physical coordinates given by ax where $a(t)$ is the scale factor, normalized to $a(t_\circ) = 1$. If $\phi = e^{ikx}$ we see that

$$\nabla^2\phi = \frac{\partial^2 e^{ikx}}{\partial(ax)^2} = -\frac{k^2}{a^2}e^{ikx} = 4\pi G\overline{\rho}\delta \tag{245}$$

Letting

$$\Delta\phi(\vec{r}) = \frac{(2\pi)^{3/2}}{V_u^{1/2}}\sum \phi_k e^{i\vec{k}\cdot\vec{r}} \tag{246}$$

we see that

$$\phi_k = -4\pi G\overline{\rho_m}a^2\delta_k k^{-2} \tag{247}$$

This $\overline{\rho_m}$ only includes the matter density and the density contrast δ grows like $D(t)$, the linear growth function normalized to $D(t_\circ) = 1$. For an EdS model $D(t) = a(t)$ but this is not true for a ΛCDM model. Since $\overline{\rho_m}a^2\delta$ is independent of time for linear perturbations larger than the sound speed horizon while $\Omega_m \approx 1$, we see again that ϕ is independent of scale during the expansion of the Universe. We can also derive the correlation function of the potential fluctuations:

$$C_\phi(r) = 64\pi^3 G^2\overline{\rho_m}^2 a^4 \int P(k)k^{-2}\frac{\sin(kr)}{kr}dk \tag{248}$$

but since $\overline{\rho_m}a^2\delta$ is $\overline{\rho_m}a^3(D(t)/a)\delta(t_\circ)$ we can write this as $\overline{\rho_m}a^2\delta = \rho_{crit}(t_\circ)\Omega_{m\circ}(D(t)/a)\delta(t_\circ)$. Using $\Delta T/T = \phi/(3c^2)$ gives the $\Delta T/T$ correlation function in terms of the density power spectrum at $t_\circ$:

$$C(\theta) = \frac{C_\phi(r)}{9c^4} = \pi\left(\frac{H_\circ}{c}\right)^4\left(\Omega_{m\circ}\left.\frac{D(t)}{a(t)}\right|_{LS}\right)^2\int P(k)k^{-2}\frac{\sin(kr)}{kr}dk \tag{249}$$

and the relationship between r and θ is given by

$$r = 2R_{LS}\sin(\theta/2) \tag{250}$$

with the comoving circumference over 2π of the surface of last scattering given by

$$R_{LS} = (1+z_{LS})D_A(z_{LS}) \tag{251}$$

Since the Legendre polynomials are orthogonal in $[-1, 1]$, we can find the C_ℓ's from $C(\theta)$ using

$$\int P_\ell(\cos\theta)C(\theta)d[\cos\theta] = \frac{2\ell+1}{4\pi}C_\ell\int_{-1}^{+1}P_\ell(\mu)^2d\mu = \frac{C_\ell}{2\pi} \tag{252}$$

Thus

$$C_\ell = 2\pi^2\left(\frac{H_\circ}{c}\right)^4\left(\Omega_{m\circ}\left.\frac{D(t)}{a(t)}\right|_{LS}\right)^2\int k^{-2}P(k)\int\frac{\sin(2kR_{LS}\sin(\theta/2))}{2kR_{LS}\sin(\theta/2)}P_\ell(\cos\theta)d[\cos\theta]dk \tag{253}$$

If we use the expansion of a plane wave into spherical harmonics,

$$
\begin{aligned}
e^{i\vec{k}\cdot\vec{r}} &= 4\pi \sum_{\ell=0}^{\infty} i^{\ell} j_{\ell}(kr) \sum_{m=-\ell}^{\ell} Y_{\ell m}^{*}(\theta,\phi) Y_{\ell m}(\theta',\phi') \\
&= \sum_{\ell=0}^{\infty} i^{\ell}(2\ell+1) j_{\ell}(kr) P_{\ell}(\cos\gamma) \qquad (254)
\end{aligned}
$$

where $j_\ell(x)$ is a spherical Bessel function, (θ,ϕ) is the direction of $\vec{r}$, (θ',ϕ') is the direction of $\vec{k}$, and γ is the angle between $\vec{k}$ and $\vec{r}$, we can find a more convenient form for the relationship between $P(k)$ and C_ℓ. The density correlation between two points $\vec{r}_1$ and $\vec{r}_2$, which are both situated on the last scattering surface ($|\vec{r}| = R_{LS}$) is given by

$$
\begin{aligned}
\xi &= \langle \delta(\vec{r}_1)\delta(\vec{r}_2)^{*} \rangle = \frac{(2\pi)^3}{V_u} \sum_{k} \sum_{k'} \langle \delta_k \delta_{k'}^{*} \rangle \, e^{i\vec{k}\cdot\vec{r}_1} e^{-i\vec{k}'\cdot\vec{r}_2} \\
&= \frac{(2\pi)^3}{V_u} \sum_{k} |\delta_k|^2 e^{i\vec{k}\cdot\vec{r}_1} e^{-i\vec{k}\cdot\vec{r}_2} \\
&= \frac{(2\pi)^3}{V_u} \sum_{k} |\delta_k|^2 4\pi \sum_{\ell=0}^{\infty} i^{\ell} j_{\ell}(kR_{LS}) \sum_{m=-\ell}^{\ell} Y_{\ell m}^{*}(\theta_1,\phi_1) Y_{\ell m}(\theta',\phi') \\
&\times \; 4\pi \sum_{\ell'=0}^{\infty} (-i)^{\ell'} j_{\ell'}(kR_{LS}) \sum_{m'=-\ell'}^{\ell'} Y_{\ell' m'}(\theta_2,\phi_2) Y_{\ell' m'}^{*}(\theta',\phi') \qquad (255)
\end{aligned}
$$

Now the sum over k will be converted into an integral over k that will include an integral over (θ',ϕ') which will force $\ell = \ell'$ and $m = m'$, since

$$
\int Y_{\ell m}(\theta',\phi') Y_{\ell' m'}^{*}(\theta',\phi') d\Omega' = \delta_{\ell\ell'} \delta_{mm'} \qquad (256)
$$

giving

$$
\xi = \int k^2 P(k) \left[4\pi \sum_{\ell=0}^{\infty} j_{\ell}^{2}(kR_{LS}) 4\pi \sum_{m=-\ell}^{\ell} Y_{\ell m}^{*}(\theta_1,\phi_1) Y_{\ell m}(\theta_2,\phi_2) \right] dk \qquad (257)
$$

Now the sum of spherical harmonics is

$$
4\pi \sum_{m=-\ell}^{\ell} Y_{\ell m}^{*}(\theta_1,\phi_1) Y_{\ell m}(\theta_2,\phi_2) = (2\ell+1) P_{\ell}(\cos\gamma) \qquad (258)
$$

where γ is the angle between $\vec{r}_1$ and $\vec{r}_2$, so

$$
\xi = 4\pi \sum_{\ell=0}^{\infty} (2\ell+1) P_{\ell}(\cos\gamma) \int k^2 P(k) j_{\ell}^{2}(kR_{LS}) dk \qquad (259)
$$

Multiplying by

$$
\left(\frac{4\pi G \overline{\rho_m} a^3}{3k^2 c^2 D(t)} \right)^2 = \frac{1}{4} \left(\frac{H_\circ}{c} \right)^4 \left(\Omega_{m\circ} \frac{D(t)}{a(t)} \bigg|_{LS} \right)^2 k^{-4} \qquad (260)
$$

to get the correlation function of the temperature gives

$$C(\gamma) = \pi \left(\frac{H_\circ}{c}\right)^4 \left(\Omega_{m\circ} \frac{D(t)}{a(t)}\bigg|_{LS}\right)^2 \sum_{\ell=0}^{\infty} (2\ell+1) P_\ell(\cos\gamma) \int k^{-2} P(k) j_\ell^2(kR_{LS}) dk \tag{261}$$

from which we can easily read off the angular power spectrum:

$$C_\ell = 4\pi^2 \left(\frac{H_\circ}{c}\right)^4 \left(\Omega_{m\circ} \frac{D(t)}{a(t)}\bigg|_{LS}\right)^2 \int k^{-2} P(k) j_\ell^2(kR_{LS}) dk \tag{262}$$

For the inflationary prediction $P(k) = Ak$ we get for the quadrupole

$$\begin{aligned}
C_2 &= 4\pi^2 \left(\frac{H_\circ}{c}\right)^4 \left(\Omega_{m\circ} \frac{D(t)}{a(t)}\bigg|_{LS}\right)^2 A \int j_2^2(kR_{LS}) \frac{dk}{k} \\
&= 4\pi^2 \left(\frac{H_\circ}{c}\right)^4 \left(\Omega_{m\circ} \frac{D(t)}{a(t)}\bigg|_{LS}\right)^2 A \int j_2^2(x) \frac{dx}{x} \\
&= \frac{\pi^2}{3} \left(\frac{H_\circ}{c}\right)^4 \left(\Omega_{m\circ} \frac{D(t)}{a(t)}\bigg|_{LS}\right)^2 A
\end{aligned} \tag{263}$$

Note that $\int j_\ell^2(x) x^{-1} dx = 2\ell(\ell+1)$ so $C_\ell = 6C_2/[\ell(\ell+1)]$.

Combining this with Equation 236 gives

$$A = \frac{3}{\pi^2} \left(\frac{c}{H_\circ}\right)^4 \left(\Omega_{m\circ} \frac{D(t)}{a(t)}\bigg|_{LS}\right)^{-2} \frac{4\pi\langle Q^2\rangle}{5T_\circ^2} \tag{264}$$

Since $P(k)$ has dimensions of length^3, it is reasonable to normalize it and talk about the dimensionless combination $P' = (H_\circ/c)^3 P(k)$. Furthermore it is good to use a dimensionless wavenumber $k' = (ck/H_\circ)$, which also puts distances on a redshift scale. Since we actually use redshifts to determine distances this is a very reasonable normalization. If we write

$$\left(\frac{H_\circ}{c}\right)^3 P(k) = A' \frac{ck}{H_\circ} \tag{265}$$

then we see that the CMBR anisotropy gives a value of

$$A' = (H_\circ/c)^4 A = \frac{12\langle Q^2\rangle}{5\pi T_\circ^2} \left(\Omega_{m\circ} \frac{D(t)}{a(t)}\bigg|_{LS}\right)^{-2} \tag{266}$$

that is dimensionless and independent of uncertainty about $H_\circ$.

17. Perturbations

The homogeneous and isotropic models we have discussed so far are a good representation of the Universe on the largest length scales, but on smaller length scales large inhomogeneities obviously exist. Therefore we need to study how density inhomogeneities grow under the influence of gravity, especially in an expanding Universe. But first we will write the equations for hydrodynamics combined with gravity in a stationary background: These are:

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) = 0 \tag{267}$$

for conservation of mass, and

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{v} = -\vec{\nabla}\phi - \frac{\vec{\nabla} P}{\rho} \tag{268}$$

for conservation of momentum (or $F = ma$), and Poisson's equation:

$$\nabla^2 \phi = 4\pi G \rho \tag{269}$$

We now linearize these equations using $\delta = \Delta\rho/\rho$, and remembering that the unperturbed $\vec{v} = 0$, and that the unperturbed pressure P is a constant. The term $(\vec{v} \cdot \vec{\nabla})\vec{v}$ is second order in the velocity and can be dropped for a linear analysis.

The linear version of Eqn(267) is

$$\frac{\partial \delta}{\partial t} = -\vec{\nabla} \cdot \vec{v} \tag{270}$$

and if we take another time derivative we get

$$\frac{\partial^2 \delta}{\partial t^2} = -\vec{\nabla} \cdot \frac{\partial \vec{v}}{\partial t} = \nabla^2 \phi + \frac{\nabla^2 P}{\rho} \tag{271}$$

when we substitute the linearized momentum equation on the RHS. Using Poisson's equation and $\Delta P = c_s^2 \Delta\rho$, where c_s is the sound speed, gives us

$$\frac{\partial^2 \delta}{\partial t^2} = 4\pi G \rho \delta + c_s^2 \nabla^2 \delta \tag{272}$$

This has the dispersion relationship (for waves of the form $\delta \propto \exp(ikx - i\omega t)$):

$$\omega^2 = c_s^2 k^2 - 4\pi G \rho \tag{273}$$

Thus short wavelength perturbations travel as sound waves, but waves with wavevectors $k < \sqrt{4\pi G\rho}/c_s$ are unstable growing perturbations. This is the Jeans instability, and the length $L_J = c_s/\sqrt{4\pi G\rho}$ is the Jeans length. The maximum growth rate of the Jeans instability is one e-folding in a time given by $t_J = 1/\sqrt{4\pi G\rho}$.

In an expanding Universe, with a background density given by the critical density, the expansion of the Universe during one Jeans time is

$$Ht_J = \frac{H}{\sqrt{4\pi G\rho}} = \sqrt{\frac{2}{3}\frac{3H^2}{8\pi G\rho}} = \sqrt{\frac{2}{3}} \tag{274}$$

so that before a perturbation can e-fold once the Universe has gotten about ten times less dense which increases the e-folding time by a factor of three. Therefore the expansion of the Universe has a large effect on the growth of small amplitude density perturbations and we must explicitly allow for it.

17.1. Expanding Background

The way to allow for the expansion is to linearize around a homologously expanding solution. Let us use new variables $\vec{x}$ and $\vec{u}$ defined by

$$
\begin{aligned}
\vec{r} &= R(t)\vec{x} \\
\vec{v} &= \dot{R}(t)\vec{x} + R(t)\vec{u}
\end{aligned}
\tag{275}
$$

In these new coordinates, every $\vec{\nabla}$ will have to be divided by R, and

$$
\frac{\partial \vec{v}}{\partial t} = \ddot{R}(t)\vec{x} + \dot{R}(t)\vec{u} + R(t)\frac{\partial \vec{u}}{\partial t}
\tag{276}
$$

The first term on the RHS will combine with the gravitational effect of the unperturbed background to give us the equation for $\ddot{R}$ that we have already discussed. So we will drop it and the gravitational effect of the unperturbed background, and just use $R(t)$ or $a(t)$ as derived earlier. The second term on the RHS introduces an effective velocity decay which is caused by the fact that a constant $\vec{v}$, which actually corresponds to zero force, leads to a $\vec{u}$ that decays like $1/R$. The other change we have to make is that since the unperturbed velocity $\vec{v}$ is no longer zero, the convective term $(\vec{v}\cdot\vec{\nabla})\vec{v}$ can no longer be dropped when linearizing the equation. This term gives the effect due to the fact that $\vec{u}$ at $\vec{x}$ is replaced by the value of $\vec{u}$ at $\vec{x} - \vec{u}\Delta t$. But when this fluid moves into $\vec{x}$, it is $\vec{v}$, not $\vec{u}$, that is constant for force free motion. The change in $\vec{x}$, $\Delta\vec{x} = \vec{u}\Delta t$, produces a change in $\vec{v}$ of $\dot{R}\Delta\vec{x}$, which then adds $(\dot{R}/R)\vec{u}$ to the convective term. This contribution is only first order in the perturbation and must be kept. The net result is a momentum conservation equation given by

$$
\frac{\partial \vec{u}}{\partial t} + \frac{1}{R}(\vec{u}\cdot\vec{\nabla})\vec{u} + 2\frac{\dot{R}}{R}\vec{u} = -\frac{\vec{\nabla}P}{\rho R} - \frac{\vec{\nabla}\phi}{R}
\tag{277}
$$

which is then combined with the mass conservation equation

$$
\frac{\partial \delta}{\partial t} + \frac{1}{R}\vec{\nabla}\cdot[(1+\delta)\vec{u}] = 0
\tag{278}
$$

and the Poisson equation

$$
\nabla^2\phi = 4\pi G R^2 \rho\delta
\tag{279}
$$

First take the time derivative of the linearized Eqn(278):

$$
\frac{\partial^2 \delta}{\partial t^2} = -\frac{1}{R}\vec{\nabla}\cdot\frac{\partial \vec{u}}{\partial t}
\tag{280}
$$

and then use the linearized version of Eqn(277)

$$
\frac{\partial^2 \delta}{\partial t^2} = -\frac{1}{R}\vec{\nabla}\cdot\left(-2\frac{\dot{R}}{R}\vec{u} - \frac{\vec{\nabla}P}{\rho R} - \frac{\vec{\nabla}\phi}{R}\right)
\tag{281}
$$

Use Eqn(278) to replace $\vec{\nabla}\cdot\vec{u}$ with $-\partial\delta/\partial t$, giving the linearized perturbation evolution equation

$$
\frac{\partial^2 \delta}{\partial t^2} + 2\frac{\dot{R}}{R}\frac{\partial \delta}{\partial t} = \left(\frac{c_s}{R}\right)^2\nabla^2\delta + 4\pi G\rho\delta
\tag{282}
$$

For simple cases we can analytically solve this equation. If the size of the perturbations is much greater than c_s/H then the $\nabla^2\delta$ term on the RHS can be dropped. This condition will be met for most perturbations in a matter-dominated Universe since the sound speed is small. For a radiation-dominated Universe it will only hold for perturbations bigger than the Hubble scale c/H because the sound speed will be $c/\sqrt{3}$. In a matter-dominated critical density Universe the Equation becomes

$$\frac{\partial^2\delta}{\partial t^2} + \frac{4}{3t}\frac{\partial\delta}{\partial t} = 4\pi G\rho_{crit}\delta = \frac{2}{3t^2}\delta \tag{283}$$

which has power law solutions $\delta \propto t^\alpha$ with $\alpha = 2/3$ and $\alpha = -1$. The first solution grows with time while the second decays with time. The growing mode has an amplitude δ that is proportional to the scale factor, so an 0.01% perturbation at $(1+z) = 1000$ will grow to be 10% at $z = 0$. For a low density Universe one gets

$$\frac{\partial^2\delta}{\partial t^2} + \frac{2}{t}\frac{\partial\delta}{\partial t} = 0 \tag{284}$$

which has power law solutions with $\alpha = 0$ and -1. So the growing mode does not grow when the density is well below the critical density. If there is a cosmological constant then when it dominates the density we have

$$\frac{\partial^2\delta}{\partial t^2} + 2H\frac{\partial\delta}{\partial t} = 0 \tag{285}$$

which has solutions $\delta =$ constant and $\exp(-2Ht)$. Thus the growing mode stops growing when the Universe becomes vacuum-dominated. The current growth rate of the growing mode ($\delta(t) = D(t)$) is often approximated as

$$\frac{\partial \ln D}{\partial \ln a} \approx \Omega_{m\circ}^{0.6} \tag{286}$$

This factor combined with the mass conservation equation allows one to compute peculiar velocities from density contrasts and vice versa.

17.2. Two Component Model

We can use the equation developed above to calculate the collapse of dark matter concentrations in a uniform background of radiation. In this case the relevant time is during matter-radiation equality at $z \approx 4,000$. As a result the time dependence of the scale factor cannot be described by a simple power law during this epoch. The variable we will use is $y = \rho_m/\rho_r$. Since $\rho_m \propto R^{-3}$ and $\rho_r \propto R^{-4}$, we see that $y \propto R$. We need to convert the time derivatives of δ to derivatives with respect to y using:

$$\frac{\partial\delta}{\partial t} = \frac{\partial\delta}{\partial y}\frac{\partial y}{\partial t} = \frac{\partial\delta}{\partial y}\dot{R} \tag{287}$$

and then

$$\frac{\partial^2\delta}{\partial t^2} = \frac{\partial}{\partial t}\left(\frac{\partial\delta}{\partial y}\dot{R}\right) = \frac{\partial\delta}{\partial y}\ddot{R} + \frac{\partial^2\delta}{\partial y^2}\dot{R}^2 \tag{288}$$

We assume that the total density is always the critical density, and use the gravitational force equation to set

$$\ddot{R} = -\frac{4\pi G}{3}\rho\left(1+\frac{3P}{\rho c^2}\right)R = -\frac{4\pi G}{3}\frac{3}{8\pi G}\left(\frac{\dot{R}}{R}\right)^2\left(1+\frac{1}{1+y}\right)R \tag{289}$$

The $4\pi G\rho\delta$ term on the RHS should only include the matter density since the radiation field doesn't participate in the collapse. The matter density is

$$\rho_m = \frac{3}{8\pi G}\left(\frac{\dot{R}}{R}\right)^2\left(\frac{y}{1+y}\right) \tag{290}$$

With all of these substitutions we have

$$\begin{aligned}
\dot{R}^2\frac{\partial^2\delta}{\partial y^2} + \left(2\frac{\dot{R}^2}{R} + \ddot{R}\right)\frac{\partial\delta}{\partial y} &= 4\pi G\rho_m\delta \\
\dot{R}^2\frac{\partial^2\delta}{\partial y^2} + \dot{R}^2\left(\frac{2}{y} - \frac{2+y}{2y(1+y)}\right)\frac{\partial\delta}{\partial y} &= \dot{R}^2\frac{3\delta}{2y(1+y)} \\
\frac{\partial^2\delta}{\partial y^2} + \frac{2+3y}{2y(1+y)}\frac{\partial\delta}{\partial y} &= \frac{3\delta}{2y(1+y)}
\end{aligned} \tag{291}$$

which has a growing mode given by

$$\delta \propto 1 + 3y/2 \tag{292}$$

Thus while the Universe is radiation dominated the expansion is so fast that matter has little chance to collapse. By the time matter and radiation are equal in density, initial perturbations in the matter density have grown by a factor of 2.5. After $z_{eq} \approx 4,000$, the matter perturbations grow by another factor of $4,000$.

17.3. Scales larger than c/H

Very large scale perturbations can be treated as separate homogeneous Universes. The equation we derived for the evolution of $(\Omega^{-1} - 1)$ can be used to determine their behavior. We know that

$$\frac{1}{\Omega} - 1 \propto \frac{1}{\rho a^2} \propto \begin{cases} (1+z)^{-1} & \text{for matter dominated} \\ (1+z)^{-2} & \text{for radiation dominated.} \end{cases} \tag{293}$$

For the matter dominated case this gives growth proportional to the scale factor just like the previous analysis in this section. For radiation dominated cases we get *faster* growth: $\delta \propto a^2 \propto t$. This is because for scales larger than c/H, photons are not able to stream out of a density enhancement, and their gravitational field contributes to the growth of the perturbation. While the large radiation pressure leads to a high sound speed, the pressure gradients are still low because of the large linear size of the perturbations. The modifications to the equations for δ are a factor of $4/3$ in the mass conservation equation to allow for the effect of pressure:

$$\frac{\partial\delta}{\partial t} + \frac{4}{3R}\vec{\nabla}\cdot[(1+\delta)\vec{u}] = 0 \tag{294}$$

and a factor of two increase in the gravitational force to allow for pressure in

$$\frac{\partial\vec{u}}{\partial t} + \frac{1}{R}(\vec{u}\cdot\vec{\nabla})\vec{u} + 2\frac{\dot{R}}{R}\vec{u} = -\frac{\vec{\nabla}P}{\rho R} - \frac{2\vec{\nabla}\phi}{R} \tag{295}$$

The combined linearized equation is then

$$\frac{\partial^2 \delta}{\partial t^2} = -\frac{4}{3R}\vec{\nabla} \cdot \left(-2\frac{\dot{R}}{R}\vec{u} - \frac{\vec{\nabla} P}{\rho R} - \frac{2\vec{\nabla}\phi}{R}\right) \tag{296}$$

Use Eqn(294) to replace $\vec{\nabla} \cdot \vec{u}$ with $-(3/4)\partial\delta/\partial t$, giving the linearized perturbation evolution equation

$$\frac{\partial^2 \delta}{\partial t^2} + 2\frac{\dot{R}}{R}\frac{\partial \delta}{\partial t} = \frac{4}{3}\left[\left(\frac{c_s}{R}\right)^2 \nabla^2 \delta + 8\pi G \rho \delta\right] \tag{297}$$

for the radiation dominated case. For scales $>> c_s t$, with $\rho = \rho_{crit} = 3/(32\pi G t^2)$ and $\dot{R}/R = 1/(2t)$ this becomes

$$\frac{\partial^2 \delta}{\partial t^2} + \frac{1}{t}\frac{\partial \delta}{\partial t} = \frac{1}{t^2}\delta \tag{298}$$

which has power law solutions $\delta \propto t^\alpha$ for $\alpha = -1$ (a decaying mode) and $\alpha = 1$: a mode growing like the scale factor squared.

17.4. CDM Model

We now have the pieces necessary to put together the *Cold Dark Matter* (CDM) model. In the simplest version of this model the only components of the Universe are the CMBR (and thermal massless neutrino) radiation and pressureless matter (hence *cold*) which does not scatter, absorb or emit electromagnetic radiation (hence *dark*). The initial perturbations are *adiabatic*, which means that the radiation and matter density perturbations are in phase, and that $n_\gamma/n_{DM} =$ const. The initial power spectrum is a Harrison-Zel'dovich spectrum; $P(k) = Ak$.

If a perturbation has a scale that is larger than c/H at the time that the Universe goes from radiation-dominated to matter-dominated (this is $z_{eq} \approx 2.4 \times 10^4 \Omega_M h^2$), then it will always be outside the sound speed horizon. This happens because once the Universe is matter-dominated the sound speed drops to zero. Because this perturbation is always larger than c_s/H it will grow as $\delta \propto (\rho a^2)^{-1}$ at all times, which is $\propto a^2$ before z_{eq} and $\propto a$ after z_{eq}. Even if it does cross the light speed horizon after z_{eq} and the photon density perturbation goes away, the matter density perturbation still grows like $\delta \propto a$ (see §17.2).

But if a perturbation has a scale that is equal to c/H before z_{eq}, then after the photon density perturbation has streamed away, the matter density perturbation does not grow until z_{eq}. Thus if a perturbation has a scale that is equal to c/H at $z_{cross} > z_{eq}$, then it loses a factor of $(z_{cross}/z_{eq})^2$ of growth. Since the power spectrum goes like δ^2, we get the final $P(k)$ is

$$P(k) = Ak \begin{cases} 1 & z_{cross} < z_{eq} \\ \left(\frac{1+z_{eq}}{1+z_{cross}}\right)^4 & z_{cross} > z_{eq} \end{cases} \tag{299}$$

So now we need to evaluate z_{cross} as a function of the scale $1/k$. This is done using

$$\frac{a(t)}{k} = \frac{1}{(1+z)k} = ct = \frac{c}{(1+z)^2}\sqrt{\frac{3}{32\pi G \rho_{r\circ}}} \tag{300}$$

since we are only interested in times before z_{eq} so we use the radiation dominated formula. This gives

$$1+z_{cross} \propto \sqrt{\frac{3}{32\pi G\rho_{r\circ}}}ck \tag{301}$$

so the power spectrum has a 4 unit break in the power law index at the k where $z_{cross} = z_{eq}$. The wavenumber at this break, which is the peak of $P(k)$, is given by

$$k_{pk} \propto \frac{\rho_{m\circ}}{\rho_{r\circ}}c^{-1}\sqrt{\frac{32\pi G\rho_{r\circ}}{3}} \propto \Omega_{m\circ}h^2 \tag{302}$$

since the radiation density $\rho_{r\circ}$ is well known from the CMBR.

When connecting the CMBR to $P(k)$, the simple Sachs-Wolfe formulae only work for $\Omega_m = 1$, because in other cases the gravitational potential ϕ is a function of time, leading to what is called the integrated Sachs-Wolfe effect. If ϕ changes with time, the energy gained by a photon falling into a potential well is different than the energy it loses climbing out of the potential well, which creates a ΔT. To avoid considering this effect, we will only consider $\Omega_m = 1$. Since the CMBR large scale anisotropy is given by $Q^2 \propto Ah^4$ where $P(k) = Ak$ for small k, the power spectrum for small scales (large k) is given by

$$P(k) \propto \frac{Q^2}{h^4}k_{pk}^4 k^{-3} \propto \frac{Q^2h^4}{k^3} \tag{303}$$

Since the fractional density variance on small scales is determined by $k^3P(k)$, we see that the density contrast is independent of scale (for small enough scales) but the ratio of density contrast to the CMBR anisotropy Q varies like h^2. This turns out to be an oversimplification. The actual break in $P(k)$ is quite gradual and rounded, and the actual slope of $P(k)$ on scales relevant to clustering of galaxies is close to $P(k) \propto k^{-1}$ or k^{-2}. As a result the ratio of the CMBR Q to the amplitude of clustering in $8/h$ Mpc spheres is approximately $\sigma_8/Q \propto h$.

The form for the power CDM power spectrum following Peebles (1982, ApJ, 263, L1-L5) as corrected in Equation 25.22 of *Principle of Physical Cosmology* is

$$P(k) = \frac{Ak}{(1+\alpha k+\beta k^2)^2} \tag{304}$$

with $\alpha = 8/(\Omega h^2)$ Mpc and $\beta = 4.7/(\Omega h^2)^2$ Mpc^2. Note that the intersection of the large and small scale asymptotic branches of this $P(k)$ occurs at $k_{pk} = 1/\sqrt{(\beta)} = 0.46\,\Omega_{m\circ}h^2$ Mpc^{-1}. In terms of the dimensionless quantities in Equation 266, this is

$$P'(k') = \frac{12\langle Q^2\rangle}{5\pi T_\circ^2}\frac{k'}{(1+\alpha'[k'/(\Omega h)]+\beta'[k'/(\Omega h)]^2)^2} \tag{305}$$

with $\alpha' = \alpha\Omega h^2/3000 = 0.00267$ and $\beta' = \beta(\Omega h^2)^2/3000^2 = 5.22\times10^{-7}$. The parameter combination Ωh that controls the shape of P' is often called Γ, and the best fits to observations of large scale structure suggest that $\Gamma \approx 0.3$ If we calculate the correlation function using P', it naturally comes out in redshift units:

$$\xi(z) = 4\pi\int k'^2P'(k')\frac{\sin(k'z)}{k'z}dk'$$

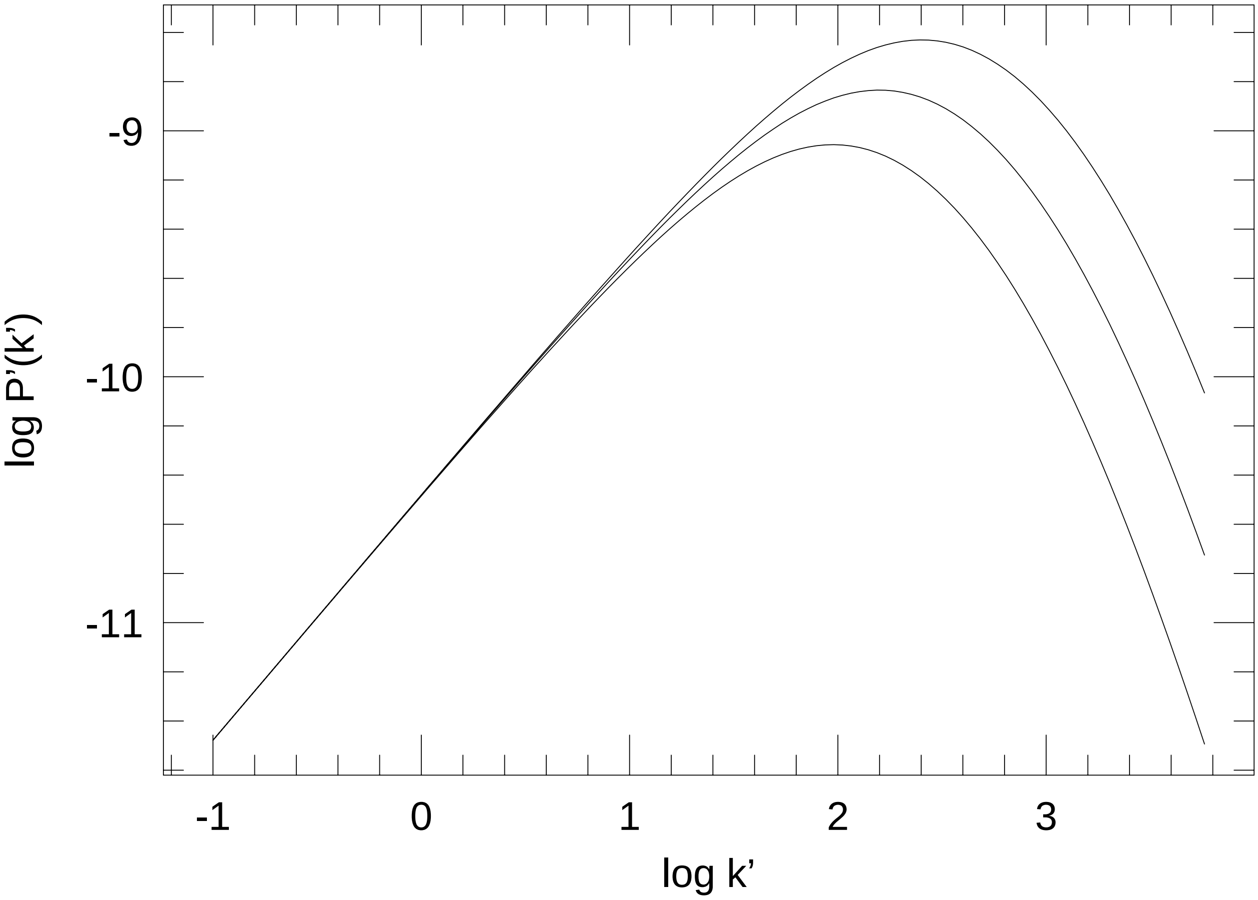

Fig. 28.— CDM power spectrum for $h = 0.8$, 0.5 and 0.3 from top to bottom. $k' = (c/H_\circ)k$ and $P' = (H_\circ/c)^3 P$.

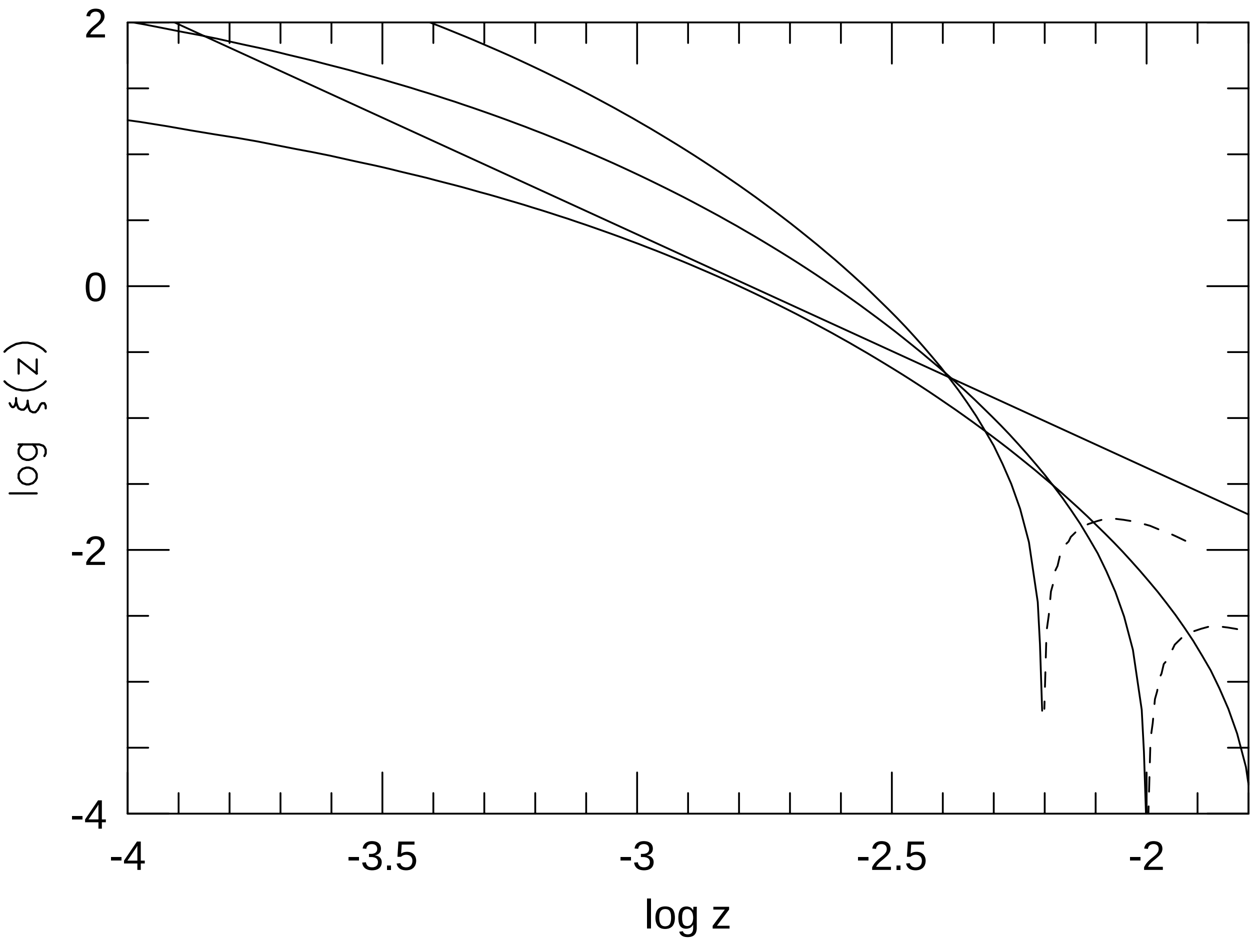

Fig. 29.— CDM 2-point correlation function for $h = 0.8$, 0.5 and 0.3 from top to bottom at $\log z = -3$. $z = H_\circ r/c$, and the straight line is the observed $\xi = (cz/500 \text{ kms})^{-1.77}$. The curves are dashed for negative ξ.

$$\begin{aligned}
&= 4\pi\frac{12\langle Q^2\rangle}{5\pi T_\circ^2}\Gamma^4\int\frac{\kappa^3}{(1+\alpha'\kappa+\beta'\kappa^2)^2}\frac{\sin(\kappa\Gamma z)}{\kappa\Gamma z}d\kappa \\
&= 4\pi\frac{12\langle Q^2\rangle}{5\pi T_\circ^2}\Gamma^4 F(\Gamma z) \qquad (306)
\end{aligned}$$

with $k' = \kappa\Gamma$ and

$$F(x) = \int\frac{\kappa^3}{(1+\alpha'\kappa+\beta'\kappa^2)^2}\frac{\sin(\kappa x)}{\kappa x}d\kappa \qquad (307)$$

Observations suggest that $\xi(r) = 1$ for $r = 5h^{-1}$ Mpc or that $\xi(z) = 1$ for $z = 0.00167$, and that $\xi(r) \propto r^{-1.77}$. We note that $F(x) \propto x^{-1.77}$ for $x \approx 0.0005 = 0.3 \times 0.00167$, indicating that the slope of the correlation function is about right for $\Gamma = 0.3$, and that $F(0.0005) \approx 2.77 \times 10^{11}$. This gives $\xi(z) = 0.94$ at $z = 0.00167$ if $\Gamma = 0.3$. So if $H_\circ$ were equal to 30 km/sec/Mpc, then both the amplitude of the clustering and the location of the peak of $P'(k')$ would agree with observations. This is the *Super-Sandage* model. But since $H_\circ$ appears to be 71 km/sec/Mpc, the amplitude of small scale perturbations appears to be too high, and the peak of $P'(k')$ appears to occur at too small a scale. Both problems could be solved by lowering k_{pk} which requires lowering $\Omega_{m\circ}$ if h is fixed. So an open Universe with $\Omega = 0.3$ and a flat vacuum-dominated ΛCDM model with $\Omega_{m\circ} = 0.3$ and $\Omega_{v\circ} = 0.7$ were proposed as variants of the standard CDM model at the time of the COBE anisotropy results in 1992. Finally, a mixed dark matter model with about 25% of the dark matter being massive neutrinos has been proposed. The neutrinos free stream out of density perturbations and thus lower the small scale density contrast for a given Q. Currently the ΛCDM is the accepted concordance model.

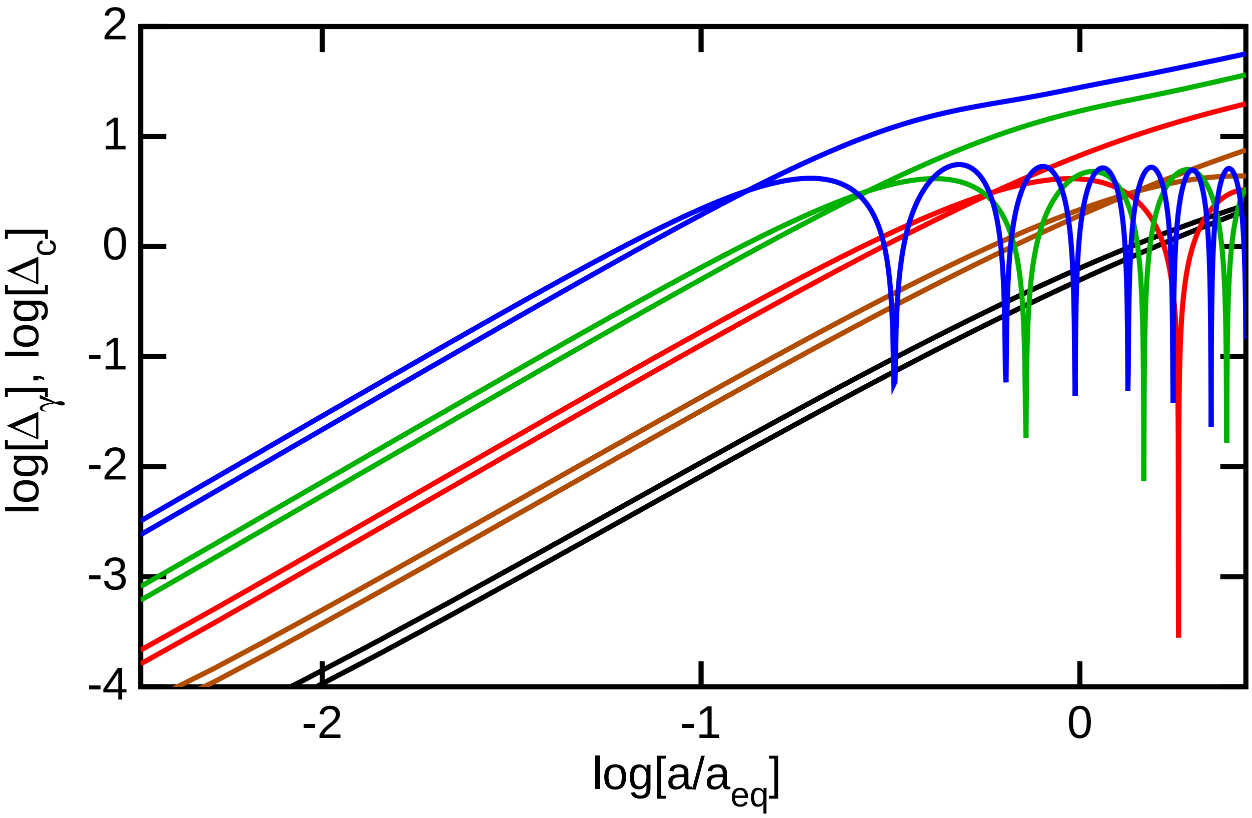

Fig. 30.— The photon & cold dark matter density perturbations Δ_γ & Δ_c for five different scales $\kappa = 5, 10, 20, 40\&80$ from bottom to top at the left.

18. Two Fluid Approximation

The section is based on the two fluid approximation described by Seljak (1994, astro-ph/9406050) in which the baryons and photons are treated as a tightly coupled fluid with pressure, while the cold dark matter is treated as a pressureless fluid. The two fluids interact only gravitationally. Let η be the conformal time, $d\eta = dt/a$ where $a(t)$ is the scale factor normalized to unity at the current time, $t_\circ$. Let $\eta_\circ$ be the current conformal time. Working in the Newtonian gauge gives the temperature fluctuation in direction $\hat{n}$ as

$$\frac{\Delta T(\hat{n})}{T_\circ} = \int_0^{\eta_0} [\dot{\tau}(\phi + \frac{\delta_\gamma}{4} + \hat{n} \cdot \vec{v_b}) + 2\dot{\phi}]e^{-\tau} d\eta. \tag{308}$$

Here ϕ is the gravitational potential, δ_γ is the photon density perturbation and $\vec{v_b}$ is the electron velocity. Note that the temperature fluctuation is 1/4 of the photon density fluctuation. τ is the Thomson scattering optical depth along the line-of-sight, given by $d\tau = a n_e \sigma_T d\eta$. In the limit of an infinitely thin LSS $\dot{\tau}e^{-\tau}$ reduces to a Dirac δ-function at η_{rec}, the conformal time at recombination. Equation 308 then reduces to

$$\frac{\Delta T(\hat{n})}{T_\circ} = \phi(\eta_{rec}) + \frac{\delta_\gamma(\eta_{rec})}{4} + \hat{n} \cdot \vec{v_b}(\eta_{rec}) + 2\int_{\eta_{rec}}^{\eta_\circ} \dot{\phi}(\eta) d\eta. \tag{309}$$

The resulting angular power spectrum is given by

$$\begin{aligned}
C_\ell &= 4\pi \int_0^\infty k^2 P_\phi(k) T(k) D_\ell^2 dk \\
D_\ell &= (\phi + \frac{\delta_\gamma}{4}) j_\ell(k\eta_\circ - k\eta_{rec}) + v_b j'_\ell(k\eta_\circ - k\eta_{rec}) \\
&+ 2\int_{\eta_{rec}}^{\eta_\circ} j_\ell(k\eta_\circ - k\eta)\dot{F}(\eta) d\eta,
\end{aligned} \tag{310}$$

where j_ℓ is the spherical Bessel functions and j'_ℓ its derivative. The perturbed quantities are evaluated in k-space at η_{rec}. $P_\phi(k)$ denotes the primordial power spectrum of the potential ϕ, usually expressed as a power law $P_\phi(k) \propto k^{n-4}$. The function $T(k)$ incorporates the damping effects due to the thickness of the LSS.

The photon evolution equations in k-space are given by

$$\dot{\delta}_\gamma = -\frac{4}{3} k v_\gamma + 4\dot{\phi}, \quad \dot{v}_\gamma = \frac{k\delta_\gamma}{4} + \dot{\tau}(v_b - v_\gamma) + k\phi. \tag{311}$$

The baryon and CDM perturbations follow:

$$\begin{aligned}
\dot{\delta}_b &= -kv_b + 3\dot{\phi}, \quad \dot{v}_b = -\frac{\dot{a}}{a} v_b + \frac{4\bar{\rho}_\gamma}{3\bar{\rho}_b}\dot{\tau}(v_\gamma - v_b) + k\phi \\
\dot{\delta}_c &= -kv_c + 3\dot{\phi}, \quad \dot{v}_c = -\frac{\dot{a}}{a} v_c + k\phi,
\end{aligned} \tag{312}$$

The energy and momentum constraint equations give the equations for ϕ and $\dot{\phi}$:

$$\begin{aligned}
\phi &= -\frac{4\pi G a^2}{k^2}(\rho + \frac{3\dot{a}f}{ak}) \\
\dot{\phi} &= -\frac{\dot{a}}{a}\phi + \frac{4\pi G a^2 f}{k},
\end{aligned} \tag{313}$$

where $\rho = (\bar{\rho}_\gamma + \bar{\rho}_\nu)\delta_\gamma + \bar{\rho}_b\delta_b + \bar{\rho}_c\delta_c$ and $f = \frac{4}{3}(\bar{\rho}_\gamma + \bar{\rho}_\nu)v_\gamma + \bar{\rho}_b v_b + \bar{\rho}_c v_c$. Here $\bar{\rho}_\nu$ and $\bar{\rho}_c$ are the neutrino and CDM mean densities, respectively. Seljak replaced neutrino density and velocity perturbations with the corresponding photon perturbations. This becomes invalid on small scales due to the free-streaming of neutrinos, but does not affect significantly the final results. Seljak also neglected the anisotropic shear and possible curvature terms.

The above equations are supplemented by the Friedmann equation, which at early times (when a possible cosmological constant or curvature term can be neglected) is given by

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G a^2}{3}(\bar{\rho}_\gamma + \bar{\rho}_\nu + \bar{\rho}_b + \bar{\rho}_c). \tag{314}$$

The solution to this equation is

$$y \equiv \frac{a}{a_{eq}} = (\alpha x)^2 + 2\alpha x, \quad x = \left(\frac{\Omega_m}{a_{rec}}\right)^{1/2} \frac{H_0 \tau}{2} \equiv \frac{\eta}{\eta_{rec}}, \tag{315}$$

where $a_{eq} = (\bar{\rho}_\gamma + \bar{\rho}_\nu)/(\bar{\rho}_b + \bar{\rho}_c) \approx 4.2 \times 10^{-5}\Omega_m^{-1}h^{-2}$ (assuming three flavors of massless neutrinos), $a_{rec}^{-1} \approx 1100$ for the standard recombination, $\alpha^2 \equiv a_{rec}/a_{eq}$, $\Omega_m = \Omega_b + \Omega_c$ is the value of matter

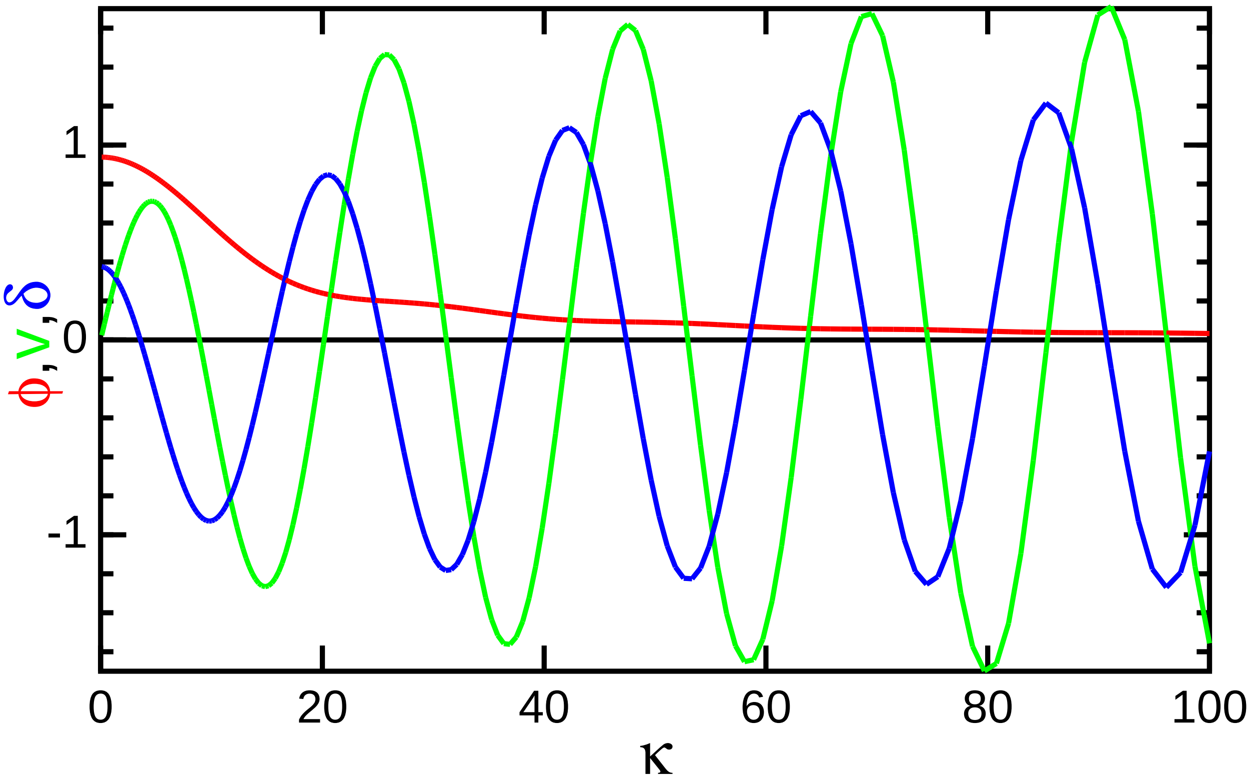

Fig. 31.— Plot of the potential perturbation ϕ, the apparent temperature perturbation $\delta = \phi + \delta_\gamma/4$, and the velocity perturbation v at the LSS as a function of the wavenumber of perturbation κ. The first Doppler peak corresponds to the large negative peak in δ at $\kappa = 6.4$, which gives $\ell \approx \kappa\sqrt{1+z_{rec}} = 200$ on the sky.

density today in units of critical density and h is the value of Hubble constant today in units of 100 km/s/Mpc.

Assuming the tight coupling limit $\tau \gg 1$, which is a good approximation on scales larger than the Silk damping scale, couples the photons and baryons into a single fluid with $\delta_b = \frac{3}{4}\delta_\gamma$ and $v_b = v_\gamma$. The above equations rewritten in terms of dimensionless time x and dimensionless wavevector $\kappa = k\eta_{rec}$ become

$$
\begin{aligned}
\dot{\delta}_c &= -\kappa v_c + 3\dot{\phi} \\
\dot{v}_c &= -\zeta v_c + \kappa\phi \\
\dot{\delta}_\gamma &= -\frac{4}{3}\kappa v_\gamma + 4\dot{\phi} \\
\dot{v}_\gamma &= (\frac{4}{3} + y_b)^{-1}\left[-\zeta y_b v_\gamma + \frac{\kappa\delta_\gamma}{3} + \kappa\phi(\frac{4}{3} + y_b)\right] \\
\phi &= -\frac{3}{2}(\zeta/\kappa)^2(\delta + 3\zeta v/\kappa) \\
\dot{\phi} &= -\zeta\phi + \frac{3\zeta^2 v}{2\kappa}
\end{aligned}
$$

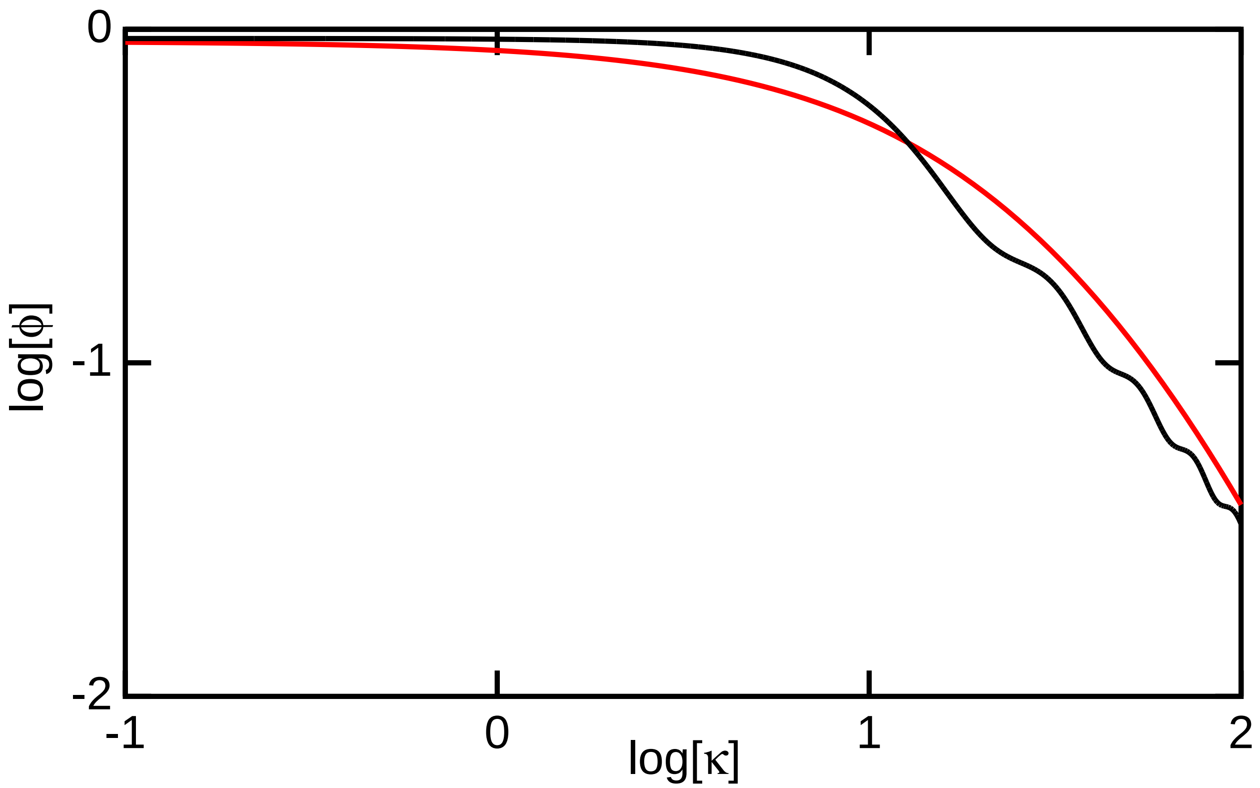

Fig. 32.— Potential perturbation ϕ at recombination *vs.* wavenumber κ, compared to Peebles' approximation to the CDM transfer function in red.

$$\begin{aligned}
\delta &= \frac{\delta_\gamma[1+\frac{3}{4}(y-y_c)]+y_c\delta_c}{1+y} \\
v &= \frac{v_\gamma(\frac{4}{3}+y-y_c)+y_c v_c}{1+y}
\end{aligned} \tag{316}$$

where the derivatives are taken with respect to x,

$$y_b \equiv \frac{\bar{\rho}_b}{\bar{\rho}_\gamma} = [1+\frac{3\times 7}{8}(\frac{4}{11})^{4/3}]\frac{\Omega_b}{\Omega_m}y = 1.68\frac{\Omega_b}{\Omega_m}y,$$

$$y_c = \frac{\Omega_c}{\Omega_m}y = (1-\frac{\Omega_b}{\Omega_m}/1.68)y$$

and "Hubble" parameter with respect to x, $\zeta = d\ln[a]/dx = 2\alpha(\alpha x+1)/(\alpha^2x^2+2\alpha x)$.

Note that y is the ratio of matter density to radiation density, y_c is the ratio of cold dark matter density to radiation density, but y_b is the ratio of baryon density to photon density which enters into the sound speed.

The initial conditions at $x << 1$ (when the universe is radiation dominated) and $\kappa\zeta << 1$ (when the mode is larger than the Hubble sphere radius) are $\phi = 1$ so that $P_\phi(k)$ in Eqn(310) is the only place where the actual perturbation amplitude is specified, and $\delta_\gamma \approx -2\phi$. The latter arises because the positive ϕ causes clocks to run fast, and the expansion of the Universe in the radiation-dominated initial condition goes like proper time to the 1/2 power, so the temperature perturbation is $-\phi/2$ leading to $\delta_\gamma \approx -2\phi$. However, a small additional density perturbation is needed to give the true density deficit needed to drive a positive potential perturbation. This gives:

$$\begin{aligned}
\phi &= 1 \\
\delta_\gamma &= -2\phi(1+\frac{3y}{16}) \\
\delta_c &= \frac{3}{4}\delta_\gamma \\
v_\gamma = v_c &= -\frac{\kappa}{\zeta}\left[\frac{\delta_\gamma}{4} + \frac{2\kappa^2(1+y)\phi}{9\zeta^2(\frac{4}{3}+y)}\right].
\end{aligned} \tag{317}$$

In order to keep track of the real, as opposed to the gauge induced, density perturbations we also need

$$\frac{\Delta a}{a} = H\int \phi dt = \frac{\zeta}{y}\int \phi y dx. \tag{318}$$

Then the real density perturbations are given by:

$$\begin{aligned}
\Delta_c &= \delta_c + 3\frac{\Delta a}{a} \\
\Delta_\gamma &= \delta_\gamma + 4\frac{\Delta a}{a}
\end{aligned} \tag{319}$$

These equations need to be evolved until $x_{rec} = [(\alpha^2+1)^{1/2}-1]/\alpha \approx 1$. The time evolution of these quantities is plotted in Figure 30 for $\kappa = 5, 10, 20, 40\&80$ in a model with $\Omega_m = 0.3$, $h = 0.65$ and $\Omega_b = 0.05$. The vacuum energy density Ω_v is irrelevant because the vacuum energy density is negligible prior to recombination.

The terms that contribute to the CMB anisotropy are the apparent temperature perturbation $\phi + \delta_\gamma$ and the velocity v at the LSS. These are plotted along with ϕ as a function of κ at x_{rec} in Figure 31.

The potential perturbation surviving to recombination is a good proxy for the matter transfer function. Figure 32 shows this potential compared to Peebles' $(1+\alpha k+\beta k^2)^{-2}$ approximation to the transfer function.

The temperature anisotropy expressed with the dimensionless variables is given by

$$\begin{aligned}
C_l &= 4\pi A\int_0^\infty \kappa^n T(\kappa)\left[\left(\phi+\frac{\delta_\gamma}{4}+2\Delta\phi\right)j_\ell(\kappa x_\circ)\right. \\
&+ \left. v_\gamma j'_\ell(\kappa x_\circ)\right]^2 d\ln\kappa,
\end{aligned} \tag{320}$$

where $x_\circ$ is the angular distance to the LSS in units of η_{rec} and the potential power spectrum is assumed to be $P_\phi(k) = Ak^{-3}\kappa^{n-1}$. The term $\Delta\phi = [2 - 8/y(x_{rec}) + 16x_{rec}/y^3(x_{rec})]/10y(x_{rec})$ arises from the ISW effect due to the potential varying with time during the transition period from the radiation dominated to the matter dominated universe. With $\Omega_m = 0.3$ and $h = 0.65$, $y(x_{rec}) = 2.74$ and $x_{rec} = 0.564$ and $\Delta\phi = -0.02$.

Note that the spherical Bessel functions $j_\ell(x)$ are asymptotically $\propto \sin(x-\ell\pi/2)/x$ so if nothng is varying too rapidy then the integrand for C_ℓ is just $\propto [(\phi+\frac{\delta_\gamma}{4}+2\Delta\phi)^2+v_\gamma^2]$ if one averages over a band of $\Delta\ell = 4$. Since $\phi+\frac{\delta_\gamma}{4}$ and v_γ are varying like sine and cosine of some angle for large κ, one would get anisotropy but not oscillations at large ℓ without the $\Delta\phi$ term.

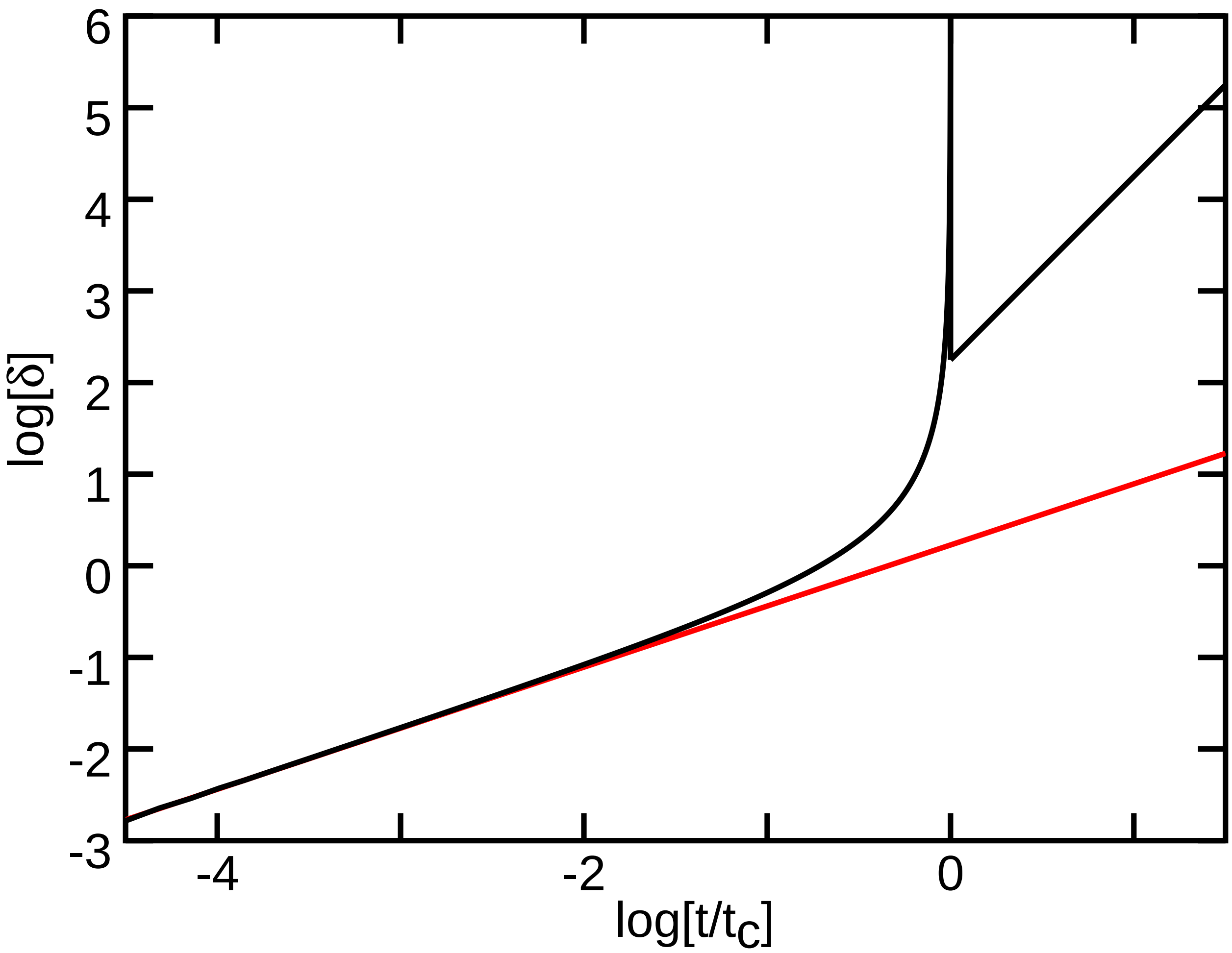

Fig. 33.— Density contrast of a spherical overdense region in a critical density Universe.

Non-linear growth of perturbations is hard to follow analytically. But there is a simple model that does a remarkably good job. This is the Press-Schechter method. The basic idea is that a spherical symmetric overdense region will reach a maximum radius, and then collapse almost to a point. When the region is at maximum radius, its kinetic energy is zero and thus too small for a system in virial equilibrium. When it collapses to nearly zero radius, the potential energy is much bigger in magnitude than it was at maximum radius, and thus the kinetic energy is essentially equal to the magnitude of the potential energy. This is now a factor of two too large for virial equilibrium If the system rebounds to a radius that is twice the radius at maximum expansion, the kinetic energy is one half the magnitude of the potential energy as desired by the virial theorem.

The density perturbations can be described by a function

$$\delta(\vec{x}, t) = \frac{\rho(\vec{x}, t)}{\overline{\rho(t)}} - 1 \tag{321}$$

As long as the conditions $\Omega_m \approx 1$ and $c_s t < L/(1+z)$ are met, δ grows at the same rate as the scale factor for the Universe, so $\delta \propto t^{2/3}$. This can be seen in Figure 33. However, overdense regions

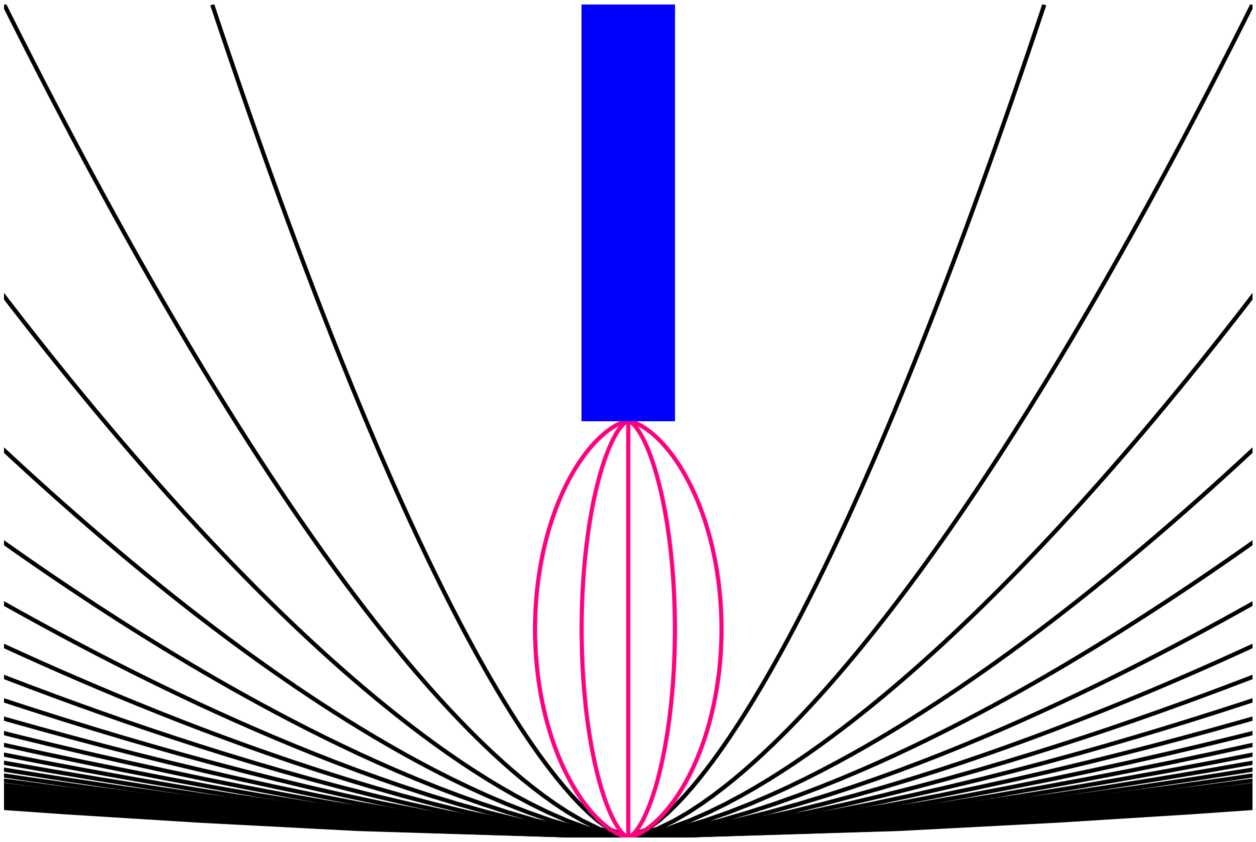

Fig. 34.— A space-time diagram showing a spherical overdense region collapsing and virializing.

will eventually stop expanding and recollapse, as seen in Figure 34. As the density increases, gravitational forces get larger, leading to a nonlinear enhancement of the density contrast. Eventually the overdense region collapses – to a point for a perfectly symmetric and homogeneous overdensity – but in general to a virialized cluster. At the point of maximum expansion, the overdensity has zero kinetic energy. In order to satisfy the virial theorem, with $-\mathrm{PE} = 2\mathrm{KE}$, the virialized cluster has to have a radius that is one-half of the radius at maximum expansion. The condition that $L/(1+z) >> c_s t$ means that pressure gradients are unimportant in the evolution of the overdense region, so the equations for a homogeneous Universe apply, giving

$$r = \frac{r_{max}}{2}(1-\cos\eta) \qquad t = \frac{t_c}{2\pi}(\eta - \sin\eta) \tag{322}$$

For small η this reduces to

$$R = \frac{r_{max}}{4}\left(\frac{12\pi}{t_c}\right)^{2/3} t^{2/3} \tag{323}$$

and this applies exactly to the inner boundary of the unperturbed region. At the collapse time, the ratio of the virialized radius $r_{vir} = r_{max}/2$ to the radius of the unperturbed region is a fixed value, and its cube gives the post-virialization density contrast

$$\frac{\rho_{vir}}{\overline{\overline{\rho(t_c)}}} = \left(\frac{R(t_c)}{r_{vir}}\right)^3 = 18\pi^2 = 177.7 \tag{324}$$

Once a region has collapsed, its density remains fixed, but the density of the Universe continues to decline so the density contrast increases like the cube of the scale factor.

Press & Schechter (1976) pointed out that the linearly extrapolated density contrast (the red line in Figure 33) has a fixed value at t_c. The density contrast at any time prior to the collapse can be found by noting that the density in the overdense region goes like r^{-3} while the density in the unperturbed region goes like t^{-2}. Thus

$$\delta(t) = 4.5\frac{(\eta - \sin\eta)^2}{(1-\cos\eta)^3} - 1 \approx \frac{3}{20}\eta^2 + \ldots \approx \frac{3}{20}\left(\frac{12\pi t}{t_c}\right)^{2/3} + \ldots \tag{325}$$

At the collapse time this gives the critical density contrast $\delta_c = 0.15(12\pi)^{2/3} = 1.68647$. These results are exact for spherically symmetric, homogeneous overdense regions embedded in a critical density Universe. Press & Schechter generalized this to say that whenever the linearly extrapolated density contrast averaged over a spherical region exceeds δ_c, then that region will have collapsed.

Note that the gravitational potential perturbation is constant as a function of time both before collapse in the linear region with $t << t_c$ and after collapse in the region with $t >> t_c$, but the two constants differ by a factor of 10/3, so the non-linear collapse and virialization make the potential well of the overdense region deeper.

One is now left with the simpler problem of identifying regions with a spherically averaged linearly extrapolated overdensity that is larger than δ_c. This can be done by convolving the density field with a spherical top-hat smoothing function that is constant for $r < R$ and zero for $r > R$. Like most convolution problems, this one becomes much simpler in the Fourier domain. The Fourier transform of the spherical top-hat, or its window function, is given by

$$\begin{aligned} W_R(k) &= (4\pi R^3/3)^{-1}\int_{r<R} \exp(i\vec{k}\cdot\vec{x})d^3\vec{x} \\ &= \frac{3[\sin(kR) - kR\cos(kR)]}{(kR)^3} \approx 1 - \frac{(kR)^2}{10} + \ldots \end{aligned} \tag{326}$$

The Fourier transform of the density fluctuation is defined by

$$\delta(\vec{r}) = \frac{\Delta\rho}{\overline{\rho}} = \frac{(2\pi)^{3/2}}{V_u^{1/2}}\sum \delta_k e^{i\vec{k}\cdot\vec{r}} \tag{327}$$

where V_u is a box used for discretizing and normalizing the Fourier transforms. Since the density of terms in the sum is proportional to V_u the relationship between the variance of $\delta(\vec{r})$ and the variance of δ_k is not changed as V_u is changed. The variance of δ_k is the *power spectrum*:

$$P(k) = \langle|\delta_k|^2\rangle \tag{328}$$

Since δ_k has units of Length$^{3/2}$, the power spectrum has units of volume. It is convenient to make the power spectrum dimensionless by dividing by the Hubble volume, and to measure the wavevector k in units of radians per Hubble radius. This gives $P'(k') = (H_\circ/c)^3 P(k)$ with $k' = ck/H_\circ$. For critical density CDM models, the power spectrum is given by Peebles (1982):

$$P'(k') = \frac{12\langle Q^2\rangle}{5\pi T_\circ^2}\frac{k'}{(1+\alpha'[k'/(\Omega h)] + \beta'[k'/(\Omega h)]^2)^2} \tag{329}$$

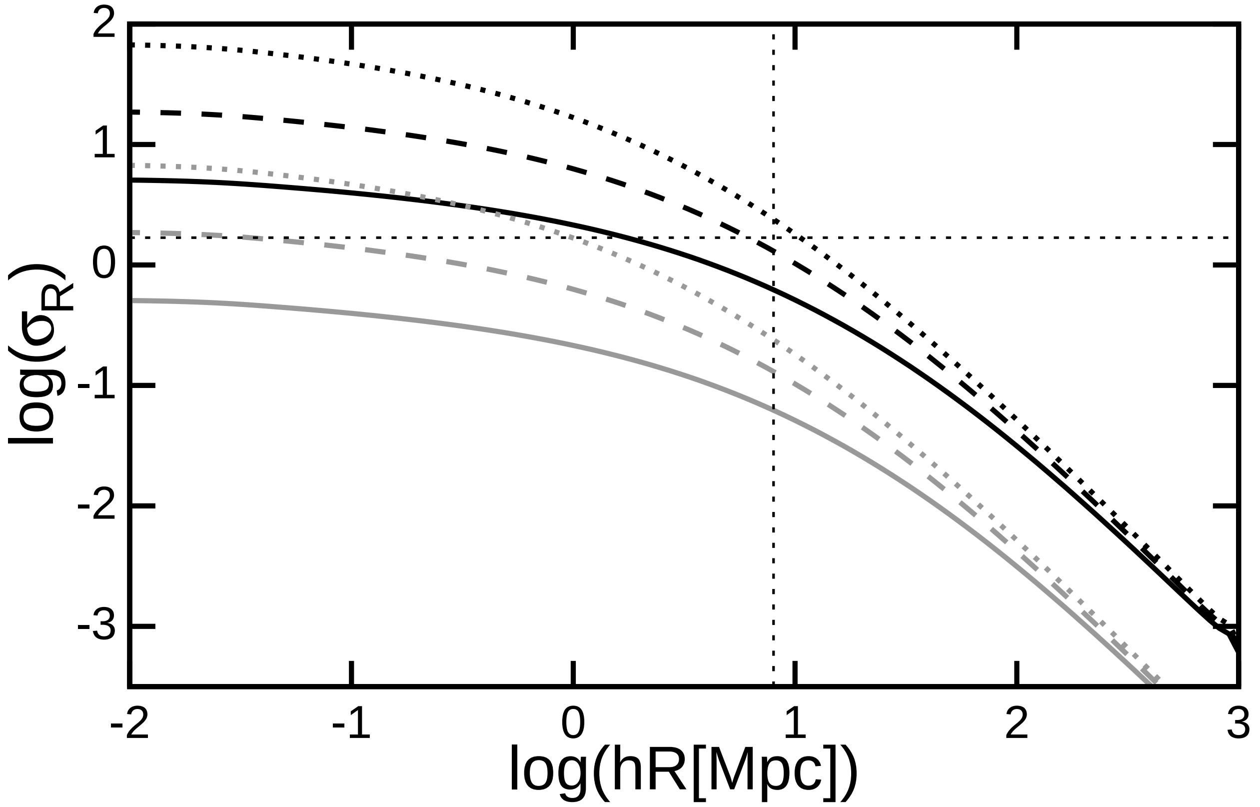

Fig. 35.— The relative density fluctuations within spheres of radius R for $\Omega =$ CDM models with $H_\circ = 25$ (solid), 50 (dashed) and 100 (dotted) at the current time (black) and at $z = 9$ (gray). The horizontal dashed line shows the critical overdensity for collapse, δ_c, while the vertical dashed line shows $8h^{-1}$ Mpc.

where Q is the CMB RMS quadrupole ($18 \pm 1.6\ \mu$K, Bennett *et al.* 1996), $\alpha' = 0.00267$ and $\beta' = 5.22 \times 10^{-7}$. For open or vacuum-dominated models the normalization in terms of the CMB changes due to the integrated Sachs-Wolfe effect, but the shape of the spectrum still depends only on the parameter Ωh, where h is the Hubble constant in units of 100 km/sec/Mpc. In Mixed Dark Matter (MDM) models the shape of the high-k cutoff is changed, and for Hot Dark Matter the high-k cutoff is much sharper. The variance within a spherical top-hat smoothing filter of radius R specified in h^{-1} Mpc, is given by

$$\sigma_R^2 = 4\pi \int k'^2 P'(k') \left[\frac{3[\sin(x) - x\cos(x)]}{x^3}\right]^2 dk' \tag{330}$$

with $x = k'R/3000$. Note that for $P(k) \propto k^{-1.23}$, which is approximately right near $R \approx 8h^{-1}$ Mpc, the effective wavenumber for σ_R is given by $k_{eff}R \approx 1.44$ so the effective ℓ for σ_8 is about 1800 for ΛCDM with $R_{LSS} \approx 10^4 h^{-1}$ Mpc. Figure 35 shows σ *vs.* R for several different critical density CDM models. A CDM model with $H_\circ \approx 30$ is consistent with all current CMB and large-scale structure data, but is not consistent with determinations of $H_\circ$, Ω from dynamics, or $q_\circ$ from supernovae. Also, while the solid black curve gives approximately the correct σ_8 now, the gray curve which shows σ_R at $z = 9$ is always well below δ_c, so very little structure would form at very high redshifts. But any model with the same value of Ωh gives the same shape for σ *vs.* R, so a

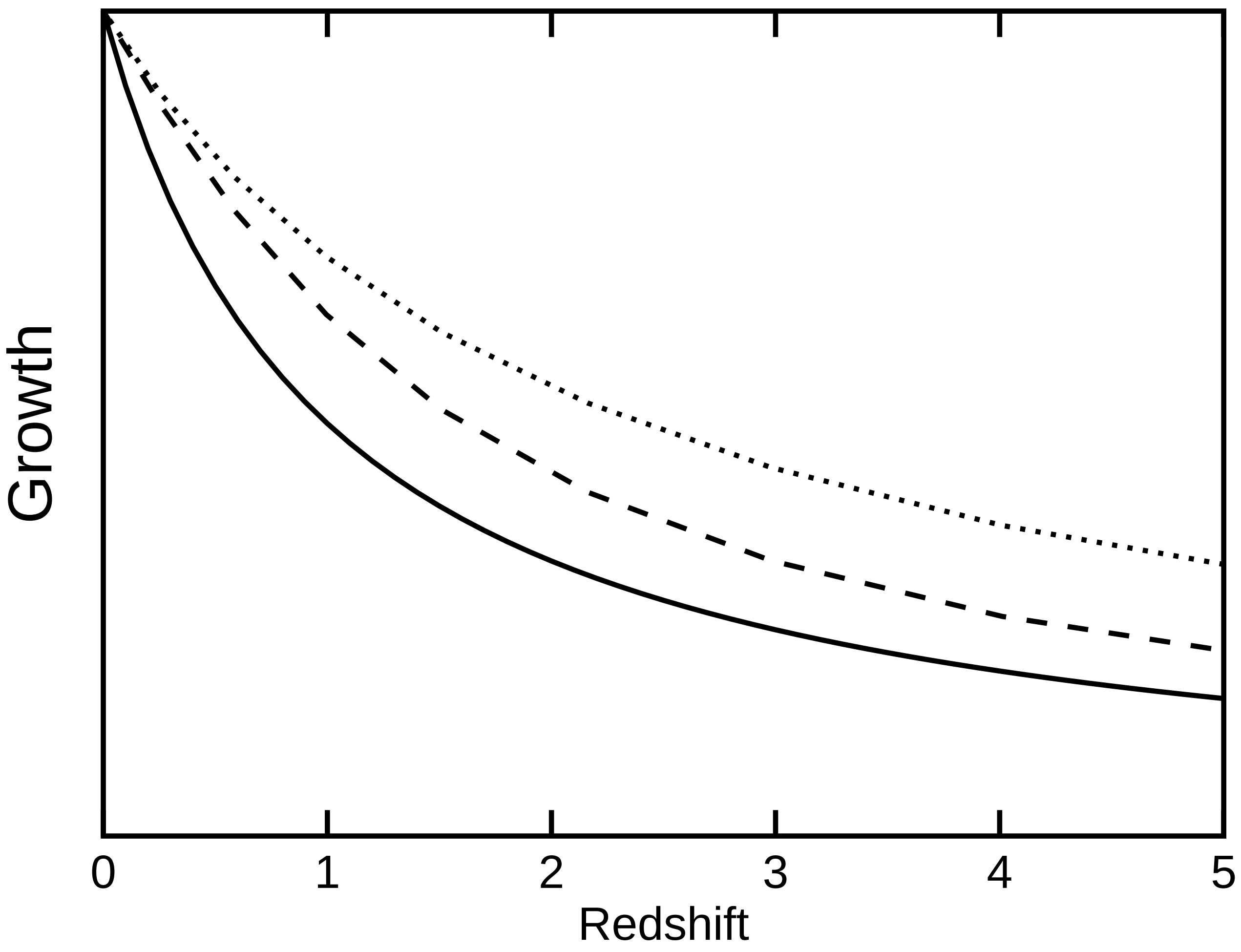

Fig. 36.— The growing mode *vs.* redshift for $\Omega_m = 1$ (solid), $\Omega_m = 0.25,\ \Omega_v = 0.75$ (dashed), and $\Omega_m = 0.25,\ \Omega_v = 0$ (dotted).

vacuum-dominated model with $H_\circ = 65$ and $\Omega = 0.4$ is a good fit.

One way to have more structure at higher redshifts while maintaining the same level of current structure is to reduce the late time growth of perturbations. The growing mode is given by (Peebles 1993)

$$D(t) = \frac{\sqrt{X}}{a} \int X^{-3/2} da \tag{331}$$

where $a = (1+z)^{-1}$ and $X = \Omega_m/a + \Omega_v a^2 + (1 - \Omega_m - \Omega_v)$. Here Ω_m is the density parameter for ordinary matter, usually just called Ω, and Ω_v is the vacuum density divided by the critical density, often denoted λ. For $\Omega_m = 1$, $\Omega_v = 0$ this function is just $D(t) \propto a(t)$, but for a vacuum-dominated model with $\Omega_m = 0.25$, $\Omega_v = 0.75$ the growth rate slows down for $z < 1$. This slower growth, combined with a normalization to σ_8 now, leads to larger density contrast at $z = 9$. For an open model with $\omega_m = 0.25$ and $\Omega_v = 0$ the effect is even larger, since growth slows for $z < 3$, as shown in Figure 36. The slowdown in the growth rate of perturbations in low Ω models may be the cause of the decline in the Madau curve between $z = 1.5$ and the present.

The Press-Schechter method is normally applied to the computation of the mass function of collapsed and virialized objects. The mass M defines a comoving smoothing radius via $M = (4\pi/3)\rho_\circ R^3$. The fraction of the matter that is in collapsed objects of mass M or greater is

$$P(>M) = (2\pi)^{-0.5} \int_{\delta_c/\sigma_R}^{\infty} \exp(-0.5x^2)dx \tag{332}$$

Thus the comoving differential mass function of objects is given by

$$n(M) = \frac{\rho_\circ}{M} \left| \frac{\partial P}{\partial M} \right| \tag{333}$$

20. Gravitational Lensing

Gravitational lensing is the bending of light by massive bodies along the line of sight. In the Newtonian "bullet" model for light, the deflection of a ray passing a mass M with impact parameter b is given by $\Delta\theta = \int a_\perp dt/v = c^{-1}\int[GMb/(c^2t^2+b^2)^{3/2}]dt = [2GM/bc^2]$. For $M = M_\odot$ and $b = R_\odot$, this is $0.87''$. But the full solution to the problem of light passing a point mass in General Relativity gives an answer that is twice as large:

$$\Delta\theta = \frac{4GM}{bc^2} = 1.75'' \frac{M}{M_\odot}\frac{R_\odot}{b} \tag{334}$$

Measurement of this bending during the 1919 total solar eclipse made Einstein a worldwide celebrity. In geometric optics this bending is caused by a time delay, just as the bending of light by a glass lens is caused by the time delay due to the index of refraction of the glass. GR gives a time delay for a ray starting and ending at radius r of

$$\Delta t = \frac{4GM}{c^3}\left[\ln\left(\frac{r+\sqrt{r^2-b^2}}{b}\right) + \sqrt{\frac{r-b}{r+b}}\right] \tag{335}$$

that creates the observed bending. This time delay was directly observed as the Mars Viking landers went behind the Sun by a time led by Irwin Shapiro, and is known as the Shapiro delay.

The deflection due to several masses is given by the sum of the deflections produced by each mass, and the time delay is the sum of the individual time delays. Usually the bending of light in gravitational lensing is very small, and the displacement of the light ray from its original path during its passge through the lensing mass is negligible. In this case we have a *thin lens*, and the mass density ρ of the lensing material can be collapsed into a surface density $\Sigma(x,y)$ on a plane perpendicular to our line of sight through the lens. Then the lensing deflection due to the total mass distribution is

$$\vec{\beta}(\vec{b}) = -\frac{4G}{c^2}\int\int \Sigma(\vec{b}')\frac{\vec{b}-\vec{b}'}{|\vec{b}-\vec{b}'|^2}d^2\vec{b}' \tag{336}$$

Here $\vec{\beta}$ is a two dimensional angle vector, and $\vec{b}$ is a two dimensional vector in the lensing plane.

The time delay is proportional to the two dimensional potential ψ derived from the surface mass density by

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} = 4\pi G\Sigma. \tag{337}$$

Note that ψ is equal to the integral along the line of sight of the ordinary gravitational potential, and that the time delay is $\tau_L = -2\psi/c^3$.

A source that would normally appear at position $\vec{\phi}$ on the sky will actually appear as an image at position $\vec{\theta}$, which means that the line of sight reaches the source plane offset by a physical distance $D_{OS}(\vec{\theta}-\vec{\phi})$, where D_{OS} is the angular size distance from the observer to the source. The deflection of the light ray by an angle $\vec{\beta}$ produces an offset in the source plane of $D_{LS}\vec{\beta}$ where D_{LS} is the angular size distance of the source as seen from the lens, and this deflection must cancel $D_{OS}(\vec{\theta}-\vec{\phi})$. Finally, the impact parameter in the lens plane is given by $D_{OL}\vec{\theta}$ where D_{OL} is the angular size distance of the lens. Therefore,

$$\vec{\phi} = \vec{\theta} - D_{LS}\vec{\beta}(D_{OL}\vec{\theta})/D_{OS}. \tag{338}$$

The angular size distance from the lens to the source is given by

$$D_{LS} = [R_\circ/(1+z_S)]S[S^{-1}([1+z_S]D_{OS}/R_\circ) - S^{-1}([1+z_L]D_{OL}/R_\circ)], \tag{339}$$

where $S() = \sin()$ for $k = 1$ or $S() = \sinh()$ for $k = -1$, and $R_\circ = (c/H_\circ)/\sqrt{|1-\Omega_{tot}|}$. For the Einstein-de Sitter model with $\Omega_{m\circ} = 1$, the distances are $D_{OL} = (2c/H_\circ)((1+z_L)^{-1} - (1+z_L)^{-1.5})$, $D_{OS} = (2c/H_\circ)((1+z_S)^{-1} - (1+z_S)^{-1.5})$, and $D_{LS} = (2c/H_\circ)((1+z_L)^{-0.5} - (1+z_S)^{-0.5})/(1+z_S)$.

The surface brightness or intensity of an image is not changed by gravitational lensing, so all of the change in the flux of the images is due to the change in the solid angle of the image compared to the unlensed source. This gives the magnification matrix $\mathbf{M}$:

$$\mathbf{M} = \frac{\partial\vec{\theta}}{\partial\vec{\phi}} = \left[\mathbf{I} - \frac{D_{LS}D_{OL}}{D_{OS}}\frac{\partial\vec{\beta}}{\partial\vec{b}}\right]^{-1} \tag{340}$$

The flux magnification is given by the determinant of the magnification matrix:

$$\frac{F_i}{F_\circ} = |\det(\mathbf{M})| \tag{341}$$

Since $\vec{\beta} = \nabla\tau_L/c$, we can also write the magnification matrix as

$$\mathbf{M} = \frac{\partial\vec{\theta}}{\partial\vec{\phi}} = \left[\mathbf{I} - \frac{D_{LS}D_{OL}}{D_{OS}}\begin{pmatrix} \partial^2\tau_L/\partial x^2 & \partial^2\tau_L/\partial x\partial y \\ \partial^2\tau_L/\partial x\partial y & \partial^2\tau_L/\partial y^2 \end{pmatrix}\right]^{-1} \tag{342}$$

20.1. Symmetric Cases

If the lensing mass is circularly symmetric, then the surface mass density Σ and the deflection are only functions of radius, $r = \sqrt{x^2+y^2}$, and the deflection is always in the radial direction. This requires that $\vec{\theta}$ for all the images and the unlensed position $\vec{\phi}$ all be on a line through the origin of the lens. If we rotate the coordinate system to make this line the x-axis then we can drop all the vector symbols.

The calculation of the deflection can be simplified for symmetric lenses because the deflection due to a uniform circular ring of mass vanishes for light rays that pass through the interior of the ring, and for rays that pass through the lens plane outside the ring the deflection is the same as that due to a point mass. Therefore the deflection is given by

$$\beta(b) = \frac{4GM_c(<b)}{bc^2} \tag{343}$$

where $M_c(<b)$ is the mass contained within a cylinder of radius b. This mass is projected into the circle of radius b in the lensing plane.

20.1.1. Point Mass Lens

The point mass lens gives

$$\phi = \theta - \frac{4GMD_{LS}}{\theta D_{OL}D_{OS}c^2} \tag{344}$$

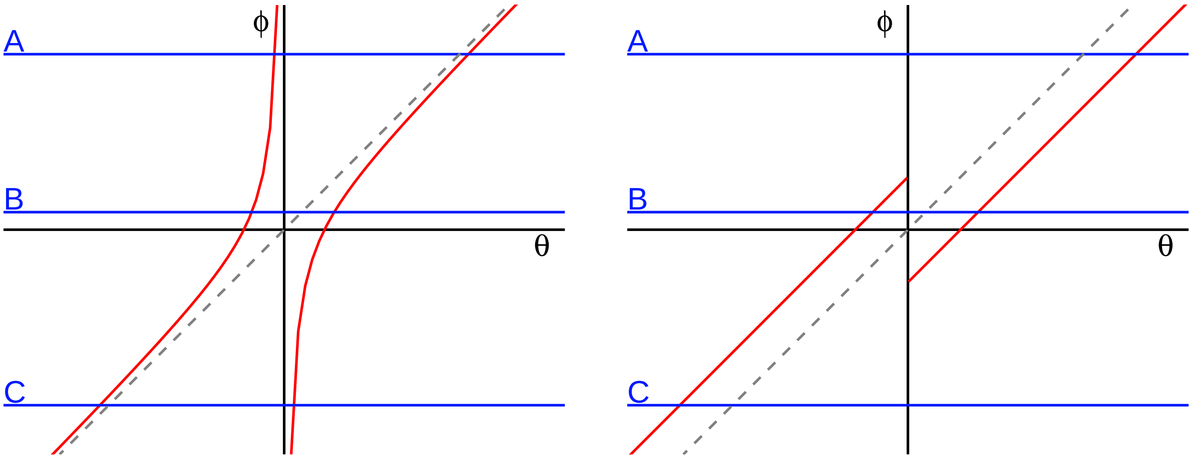

Fig. 37.— Left: Source position ϕ vs. image position θ for a point mass lens. For any source position there are two images. Right: A singular isothermal sphere lens. There is one image of distant sources (A and C) but two images of a nearby source (B).

which has the solutions

$$\theta = \frac{\phi \pm \sqrt{\phi^2 + 4\theta_E^2}}{2} \tag{345}$$

where the *Einstein ring* radius is given by $\theta_E = \sqrt{(4GMD_{LS})/(D_{OL}D_{OS}c^2)}$. If the source is directly centered behind the lens, then one gets a complete ring of radius θ_E for an image.

The magnification matrix for a symmetric lens is given by

$$\mathbf{M} = \begin{pmatrix} \partial\theta/\partial\phi & 0 \\ 0 & \theta/\phi \end{pmatrix} \tag{346}$$

so the flux magnification factor is given by

$$M = \left| \frac{\partial\theta}{\partial\phi} \frac{\theta}{\phi} \right| = \frac{(1 \pm 1/\sqrt{1 + 4\theta_E^2/\phi^2})(1 \pm \sqrt{1 + 4\theta_E^2/\phi^2})}{4} \tag{347}$$

If we define $u = x_\circ / r_E$, then this is

$$M = \frac{\left(1 \pm \sqrt{1 + 4/u^2}\right)^2}{4\sqrt{1 + 4/u^2}} \tag{348}$$

Note that two images always exist for a point mass lens, and that the ratio of the flux magnifications is given by the ratio of the squares of the separations from the lens. When the lens is far from the line of sight, one image is essentially unmagnified and nearly at the unlensed position, while the second image in very close to the lens on the opposite side, and highly demagnified. When the source nearly aligned with the lens, the θ/ϕ term in the magnification gets very large and both images are bright.

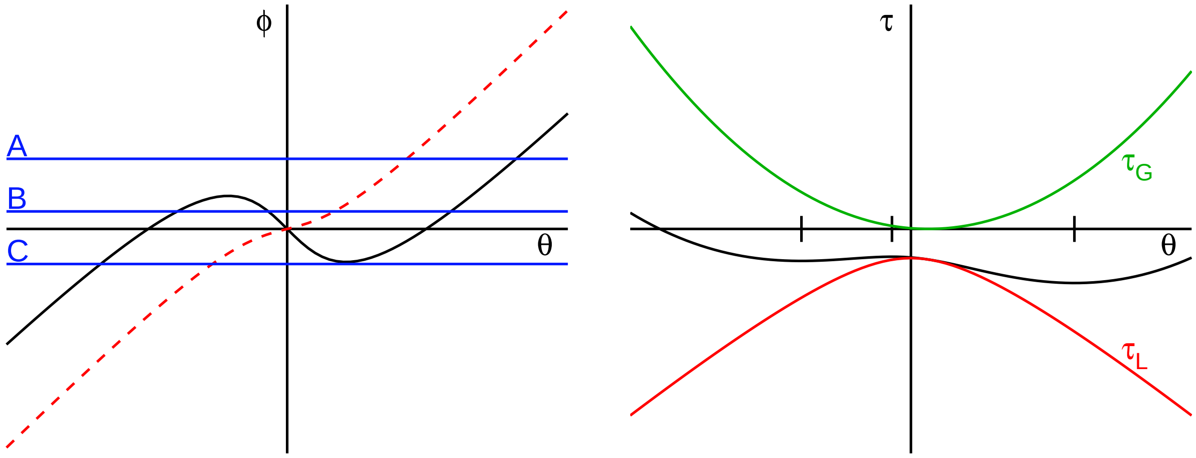

Fig. 38.— Left: Source position vs. image position for a non singular isothermal sphere. For a source at position A there is only one image. A source at position B gives three images. A source at position C is at the caustic, and has infinite magnfication. A lens with maximum surface density less than the critical surface density [dashed curve] never produces multiple images. Right: The geometric delay τ_G, the lensing delay τ_L, and the total delay for source position B. The minima and maxima of the total delay are ticked and correspond to the image positions.

In microlensing experiments like *MACHO*, *EROS* or *OGLE*, only the sum of the magnifications of the two images is measured, with a total magnification given by

$$M_{tot} = \frac{(1+1+4/u^2)}{2\sqrt{1+4/u^2}} = \frac{2+u^2}{u\sqrt{4+u^2}} \tag{349}$$

which is a factor of 1.34 when $u = 1$ and diverges as $u \to 0$. For a lens moving with constant angular speed in front of a background star, $u(t) = \sqrt{u_{min}^2 + [2(t-t_{min})/t_1]^2}$. In any given event, only u_{min}, t_{min} and Δt can be found. For a lens in the Milky Way halo with $D_{OL} = 10$ kpc, a source in the LMC with $D_{LS} = D_{LMC} - D_{OL} = 40$ kpc so $F = 8$ kpc, and $M = 0.5M_\odot$, then the Einstein ring radius is $r_E = 8.5 \times 10^{13}$ cm = 5.7 au, and it subtends an angle of 0.57 milliarcseconds. A lens in the Milky Way halo with a transverse velocity of 100 km/sec will traverse the Einstein diameter $2r_E$ in $t_1 = 200$ days so a microlensing event causes a star in the LMC to brighten for several months. The width of the Einstein ring depends on the lensing mass, so by studying the distribution of event durations the typical mass can be found to be $\approx 0.5\ M_\odot$. A much firmer mass estimate could be obtained if the Einstein ring radius were measured directly instead of inferred from the duration distribution. This can be done by having a telescope in a heliocentric orbit so it is close to 1 au away from the Earth. With two microlensing light curves from separated locations one can find r_E and the transverse velocity. One can also use the change in the Earth's velocity during the microlensing event to estimate r_E, because the events last for a good fraction of a year. Thus microlensing is a way of detecting Massive Compact Halo Objects, or MACHOs, which appear to make up about 50% of the density of the Milky Way halo.

20.1.2. *Singular Isothermal Sphere Lens*

A simple model for extended symmetric lens is the *singular isothermal sphere* (SIS), with velocity dispersion σ^2 and density

$$\rho(r) = \frac{\sigma^2}{2\pi G r^2} \tag{350}$$

While the SIS has infinite mass, its average density equals the critical density at radius $r = 2\sigma/H$. The surface mass density of the SIS is $\Sigma(b) = \sigma^2/(2Gb)$, and thus the mass contained within a cylinder of radius b is $M_c(< b) = \pi\sigma^2 b/G$. Therefore the deflection in the SIS lens is

$$\beta(b) = \frac{4GM_c(< b)}{bc^2} = 4\pi\left(\frac{\sigma}{c}\right)^2 \tag{351}$$

which is a constant, independent of the impact parameter. If $|\phi| > (D_{LS}/D_{OS})4\pi(\sigma/c)^2$, there is only one image, with $|\theta| = |\phi| + (D_{LS}/D_{OS})4\pi(\sigma/c)^2$. If $|\phi| < (D_{LS}/D_{OS})4\pi(\sigma/c)^2$, then there is another image on the opposite side of the lens with $\theta = \phi - (D_{LS}/D_{OS})4\pi(\sigma/c)^2$. If the source is perfectly aligned, with $\phi = 0$, then there is an Einstein ring with $\theta_E = (D_{LS}/D_{OS})4\pi(\sigma/c)^2$. For a cluster of galaxies with $\sigma = 10^3$ km/sec, at $z_L = 7/9$, and a source at $z_S = 3$ with $\Omega_{m\circ} = 1$, we have $D_{LS} = 0.125(c/H_\circ)$, $D_{OS} = 0.25(c/H_\circ)$, and the Einstein ring radius is $14.4''$.

For this mass distribution $\partial\theta/\partial\phi = 1$, so the magnification is $M = |\theta/\phi|$. Thus the ratio of the magnifications of the two images is given by the ratio of their separations from the lens.

20.1.3. *Non-Singular "Isothermal" Sphere*

The density law $\rho(r) = \rho_\circ a^2/(r^2 + a^2)$ is often called an isothermal sphere, even though it is not a solution of the Vlasov equation with an isothermal distribution function. This model gives

$$\begin{aligned}
\Sigma(b) &= \frac{\pi\rho_\circ a^2}{\sqrt{b^2 + a^2}} \\
M_c(< b) &= 2\pi^2\rho_\circ a^2\left(\sqrt{b^2 + a^2} - a\right) \\
\beta(b) &= \frac{8\pi^2 G\rho_\circ a^2}{c^2}\frac{\sqrt{b^2 + a^2} - a}{b}
\end{aligned} \tag{352}$$

We can find the image position(s) for a lens with this mass distribution graphically, by plotting the RHS of the equation

$$\phi = \theta - D_{LS}\beta(D_{OL}\theta)/D_{OS} \tag{353}$$

vs. θ, as shown in Figure 38. Different source positions ϕ correspond to different placements of horizontal lines, and can lead to 1 or 3 intersection with the solid curve, corresponding to 1 or 3 images. In general non-singular lenses always have an odd number of images, so the two images exhibited by the SIS and point mass lenses are anomalies.

When the source position is at C in the figure, the horizontal line is tangent to the curve. This leads to $\partial\theta/\partial\phi = \infty$ and hence infinite magnification for sources on the whole circle with radius ϕ_C. This circle is called a *caustic*, and caustics occur whenever there is a change in the number of images. But multiple images can only occur if the slope of the central part of the curve is negative, and this happens only if

$$\frac{D_{LS}D_{OL}}{D_{OS}}\frac{\partial\beta}{\partial b} = \frac{D_{LS}D_{OL}}{D_{OS}}\frac{4\pi G}{c^2}\Sigma(0) > 1 \tag{354}$$

Thus there is a critical surface mass density for multiple lensing:

$$\Sigma_{crit} = \frac{D_{OS}}{D_{LS}D_{OL}}\frac{c^2}{4\pi G} \tag{355}$$

If we take $D_{OS} = D_{LS} = D_{OL} = c/H_\circ$ to get a rough order of magnitude estimate for the cosmologically interesting critical surface density, we get $\Sigma_{crit} = (2/3)\rho_{crit}(c/H_\circ)$. Thus we need to have a surface mass density roughly equal to the critical density times the Hubble radius in order to get strong lensing with multiple images.

20.2. Time Delay and $H_\circ$

The deflection in a gravitational lens is related to gradient of the time delay through the lens. The lensing time delay due to the total mass distribution is

$$\tau_L(\vec{b}) = -\frac{4G}{c^3}\int\int \Sigma(\vec{b}')\ln\left(|\vec{b}-\vec{b}'|\right)d^2\vec{b}' \tag{356}$$

Note that this formula has the problem of putting a dimensionful quantity into the argument of the ln(), but the only result of this error is to add a constant to the time delay. Only time delay diffferences are necessary, so it is acceptable to drop the normalization scale that should go in the logarithm – it should be $\ln(b/r)$ where r is the distance of the observer, but since that distance is essentially constant over the whole lens, there is no point in including it. There is also a problem of an infinite delay for infinite mass objects like the isothermal sphere models. But the differences in delays between two points in the lensing plane are finite. Thus for the singular isothermal sphere model, we can write

$$\tau_L(\vec{b}) = -4\pi\frac{|\vec{b}|}{c}\frac{\sigma^2}{c^2} \tag{357}$$

by integrating the deflection instead of directly evaluating the integral formula. For cosmological lenses we need to multiply this delay, which is measured at the lens, by $(1+z_L)$ to allow for the dilation of light curves due to the expansion of the Universe.

The total delay along a ray is the sum of the lensing delay τ_L and the geometric delay τ_G due to the longer path length for a deflected ray. This delay is

$$\tau_G = \frac{1}{2}\left|\vec{\theta}-\vec{\phi}\right|^2\frac{(1+z_L)D_{OL}D_{OS}}{D_{LS}c} \tag{358}$$

For the simple case of the singular isothermal sphere with two images, $|\theta-\phi|$ is the same for both images. Thus the difference in time delays between the two images is given by

$$\tau_1 - \tau_2 = \frac{(1+z_L)D_{OL}}{c}\frac{4\pi\sigma^2}{c^2}(|\theta_2|-|\theta_1|) = \frac{(1+z_L)D_{OL}D_{OS}}{2D_{LS}c}(|\theta_2|-|\theta_1|)(|\theta_2-\theta_1|) \tag{359}$$

For example, if $z_L = 0.44$ and $z_S = 1.56$, then

$$\tau_1 - \tau_2 = 25h^{-1}\frac{|\theta_2|^2 - |\theta_1|^2}{(1'')^2} \text{ days.} \tag{360}$$

Note that the brighter image, the one further from the lens center, will have less delay and thus arrive first. The RHS of Eqn(359) uses the image separation to compute the velocity dispersion from the lensing equation, and requires knowledge of both z_L and z_S. The "middle side" of Eqn(359) requires a spectroscopic velocity dispersion, but the source redshift z_S is not needed. Either version can be used to find the Hubble constant $H_\circ$ if the time delay can be measured.

However, a uniform mass sheet added to the lens model can bias the derived $H_\circ$. The uniform mass sheet introduces a delay proportional to $-|b|^2$, which acts like a convex lens. This acts to cancel part of the geometric time delay, and increases the image separations. If not allowed for, these larger separation requires that the distances be smaller, and for known redshifts this means that $H_\circ$ is too big. Thus, the $H_\circ$ derived from a lens is an upper limit. But by measuring both the image separations and the spectroscopic velocity dispersion, one can measure and correct for the effect of a uniform mass sheet. The velocity dispersion, time delay and deflection angles of the lensing galaxy in 0957+61 have been measured, giving $H_\circ = 70 \pm 7$ km s^{-1} Mpc^{-1} (Tonry & Franx, astro-ph/9809064).

Other lenses have been used to determine $H_\circ$: PG1115+080 gives 51^{+14}_{-13} km s^{-1} Mpc^{-1} (Keeton & Kochanek, astro-ph/9611216) and B0218+357 gives 69^{+7}_{-10} km s^{-1} Mpc^{-1} (Biggs *et al.*, astro-ph/9811282). Combining these results gives $H_\circ = 67 \pm 5$ km s^{-1} Mpc^{-1}.

20.3. Asymmetric Lenses

When the lensing mass distribution is not symmetric, then very complicated image patterns can result. For binary point source lenses, extremely high magnifications on caustic lines can occur well off the center of the main lensing event. This has led to some spectacular light curves in the MACHO experiments, as discussed in Alcock *et al.* (1999, ApJ, TBD, TBD [astro-ph/9807163]).

20.4. Weak Lensing

Even when the surface mass density is always less than Σ_{crit}, so there are no multiple images, it is still possible to measure the mass distribution in the lens by studying the pattern of distortions in the images of numerous faint background galaxies. Because these galaxies are perceptibly extended, one can measure the second moments of their brightness distributions. The galaxies will be slightly magnified by an amount which generally cannot be measured, because their true brightness is unknown. But in general the images will be magnified more in one direction than another, and this *shear* will be the same over extended regions on the sky. If the background galaxies do not have any intrinsic preferred orientation, one can measure the shear by taking the average ellipticity of the faint background galaxies.

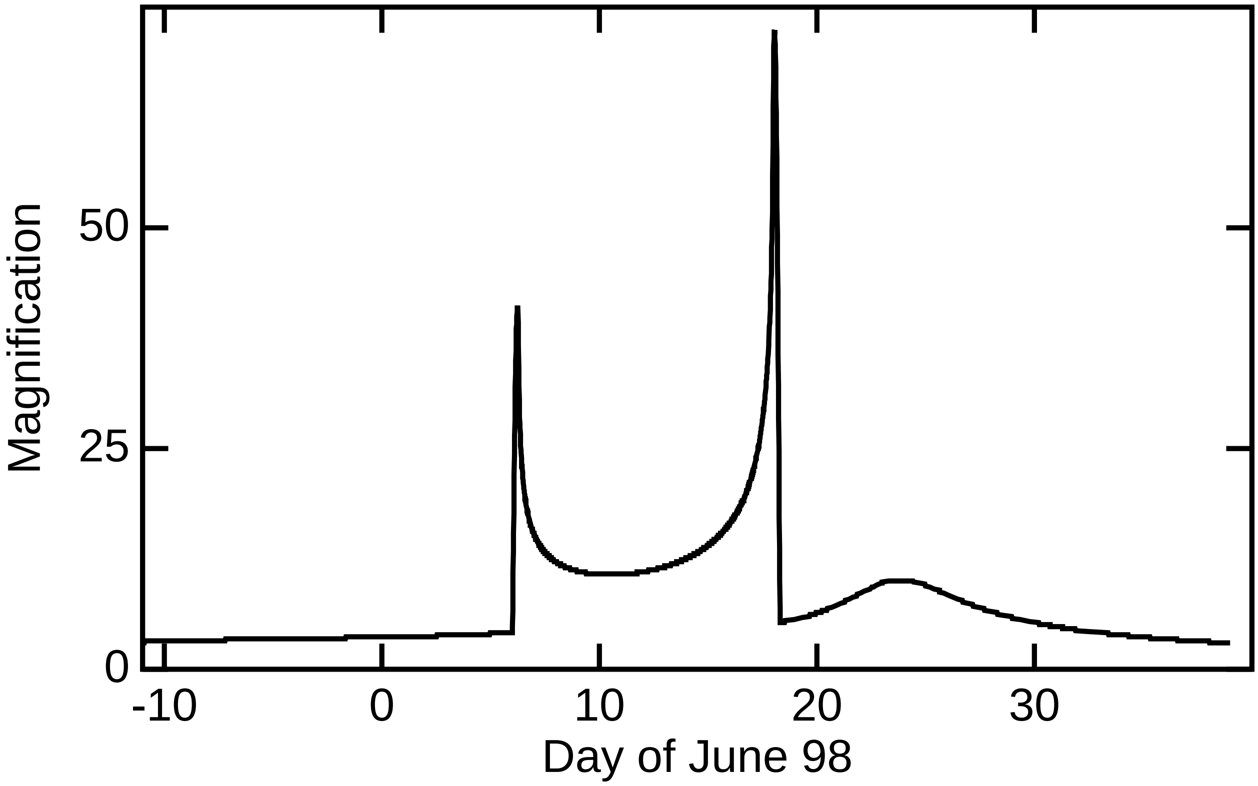

Fig. 39.— Light curve for the MACHO event 98-SMC-01, showing the sharp cusps in magnification produced by caustic crossings.

The magnification matrix is the inverse matrix of the gradient of the deflection with respect to image position, and the deflection is proportional to the gradient of the time delay. Since the second derivative of the time delay is a symmetric matrix, the magnification matrix is symmetric as well. This symmetric 2×2 matrix has three independent components, which we can take to be the angle of the rotation that diagonalizes the matrix, and then write the two eigenvalues of the magnification matrix $\mathbf{M}$ as $(1-\kappa \pm \gamma)^{-1}$. The trace, which is the sum of the eigenvalues, of $\mathbf{M}^{-1}$ gives $(1-\kappa)$. Because the trace is also the sum of the diagonal elements of the matrix, which gives the value of $\partial^2\tau_L/\partial x^2 + \partial^2\tau_L/\partial y^2$ which is proportional to the surface mass density, we find that $\kappa = \Sigma/\Sigma_{crit}$. So if we can determina κ we can determine the surface mass density. But κ usually cannot be determined because we don't know the unlensed brightness of the source. However, we can determine $\partial^2\tau_L/\partial x^2 - \partial^2\tau_L/\partial y^2$ and $\partial^2\tau_L/\partial x\partial y$ from the shear. The gradient of κ is given by

$$\begin{aligned}
\vec{\nabla}\kappa \propto \vec{\nabla}\Sigma &\propto \begin{pmatrix} \partial^3\tau_L/\partial x^3 + \partial^3\tau_L/\partial y^2\partial x \\ \partial^3\tau_L/\partial x^2\partial y + \partial^3\tau_L/\partial y^3 \end{pmatrix} \\
&= \begin{pmatrix} \partial\left[\partial^2\tau_L/\partial x^2 - \partial^2\tau_L/\partial y^2\right]/\partial x + 2\partial\left[\partial^2\tau_L/\partial x\partial y\right]/\partial y \\ 2\partial\left[\partial^3\tau_L/\partial x\partial y\right]/\partial x - \partial\left[\partial^2\tau_L/\partial x^2 - \partial^2\tau_L/\partial y^2\right]/\partial y \end{pmatrix}
\end{aligned} \tag{361}$$

Thus a linear combination of the gradients of the coherent ellipticity produced by the shear can be used to construct the gradient of the surface mass density and this can be integrated to determine the mass distribution modulo the uncertainty due to the possibility of adding a uniform mass sheet.

A very simple example of mass determination using shear would be to find the average tan-

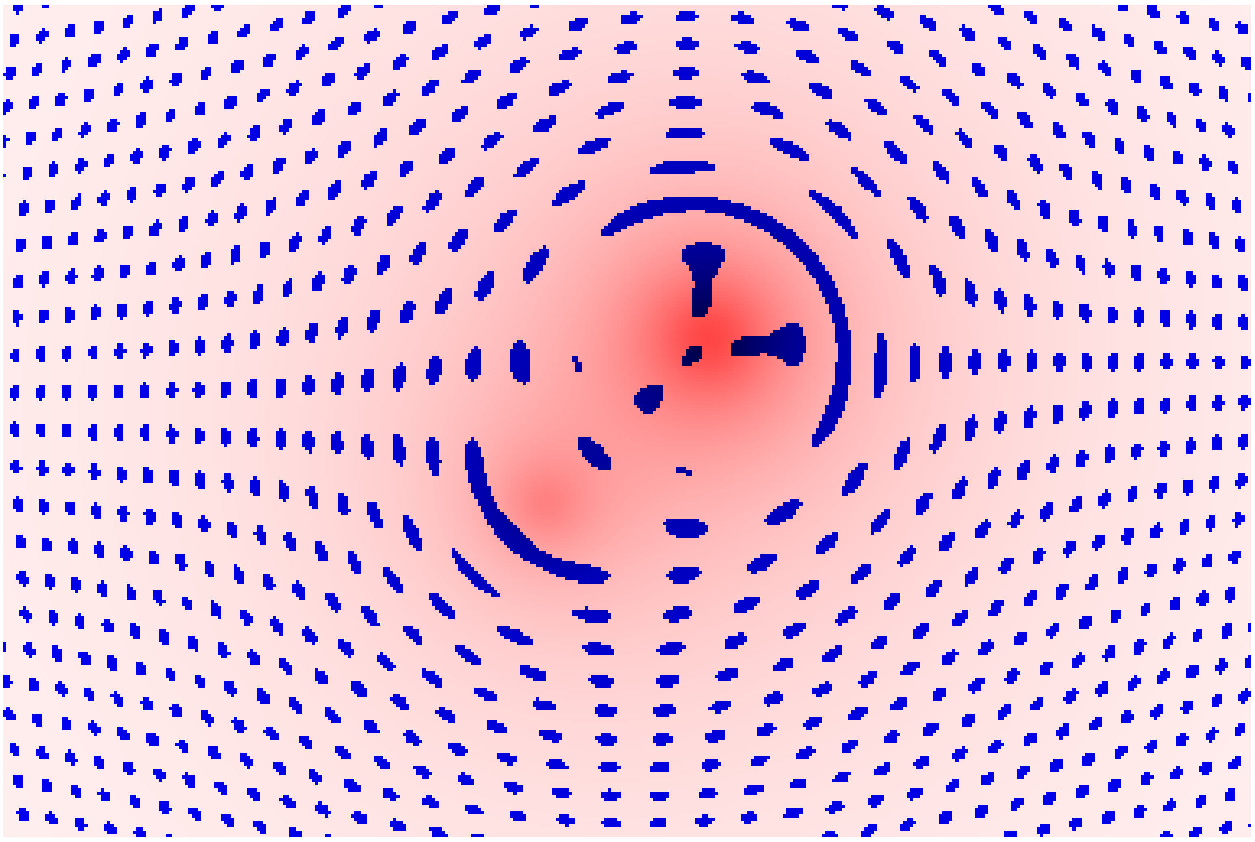

Fig. 40.— Distortion and magnification of faint blue background galaxies by a foreground cluster with surface mass desnity shown by the pink shading.

gential ellipticity of the images on a circle at radius θ arround an SIS lens. These images will be $1 + 2\langle\gamma_T\rangle = 1 + \theta_E/\theta$ times wider in the tangential direction than in the radial direction, so this shear measurement determines θ_E and hence the mass contained within the circle of radius θ:

$$M_c(<\theta) = \pi\sigma^2 D_{OL}\theta/G = \frac{c^2}{8G}\frac{D_{OL}D_{OS}}{D_{LS}}\langle\gamma_T\rangle\theta^2 \tag{362}$$

For a point mass lens, $\langle\gamma_T\rangle = \theta_E^2/\theta^2$, and the corresponding mass formula is

$$M = \frac{c^2}{4G}\frac{D_{OL}D_{OS}}{D_{LS}}\langle\gamma_T\rangle\theta^2 \tag{363}$$

which is different because the gradient of γ with respect to radius is different in the two cases. Both formulae can be written as

$$M_c(<\theta) = \frac{c^2}{8G}\frac{D_{OL}D_{OS}}{D_{LS}}\left|\frac{\partial\langle\gamma_T\rangle}{\partial\theta}\right|\theta^3 \tag{364}$$

Applying these techniques to the cluster MS1224.7+2007 Fahlman *et al.* (astro-ph/9402017)) obtained a mass to luminosity ratio of $M/L = 825h\ M_\odot/L_\odot$. But Kaiser *et al.* (astro-ph/9809268) get $M/L = 270h\ M_\odot/L_\odot$ for the supercluster MS0302+17, and Hoekstra *et al.*(astro-ph/9711096) get $M/L = 180hM_\odot/L_\odot$ for the cluster Cl 1358+62, so there is not enough total mass clumped into clusters to give $\Omega_{m\circ} = 1$.