Bull. Astr. Soc. India (1992) **20**, 1-155

Galaxy formation

T. Padmanabhan[1] and K. Subramanian[2]

[1] *Theoretical Astrophysics group, TIFR, Homi Bhabha Road, Bombay 400005, India*

[2] *National Centre for Radio Astrophysics, TIFR, Poona University Campus, Ganeshkind, Pune 411007, India*

Contents

1: Introduction

1.1. Prologue

The observed universe contains structures at different scales. At the same time it also exhibits a remarkable level of uniformity as regards the large scale properties. It has been a challenge to explain both these features in a consistent theory. The conventional wisdom attempts to do so in the following manner: We begin with a uniform universe with very small irregularities; then gravitational instability is expected to enhance these irregularities, eventually forming the currently observed structures.

The present review discusses several aspects of this picture of structure formation. The emphasis, throughout, is on theoretical aspects and physical basis for the models rather than on the detailed observational features or detailed examination of particular models. Many of the concepts are also developed in a self contained manner. This has made the review somewhat more pedagogical than is usual for a review of this type, but has the advantage that a non-expert astronomer or physicist will be able to understand and appreciate the contents. In order to keep the article pedagogical and within reasonable size, we have to make a ruthless selection in the topics: for example, we have *not* discussed the following important topics : The infrared astronomy satellite (IRAS) redshift survey and its use in probing both the large scale velocity and density fields, structure formation theories involving isocurvature perturbations or seeds from the early universe like strings and textures or non gravitational processes like explosions and radiation pressure.

We begin in this part with a discussion of the smooth universe. The linear theory of perturbations is spelled out in part 2 and its applications in part 3. Nonlinear evolution is taken up in part 4. The relevance of the high redshift universe to galaxy formation is discussed in part 5. The review ends with a consideration of the origin of density perturbations and some concluding remarks.

1.2. The parameters of the smooth universe

The evolution of the universe, like that of several other physical systems, is described by a second order differential equation. Such an evolution can be uniquely specified by specifying two independent constants at any chosen instant of time. One may choose these constants to be : (a) the mean energy density ρ present in the universe and (b) the expansion rate of the universe, parameterised by the Hubble 'constant' H_0; both ρ and H_0, of course, correspond to the values measured today.

It is convenient to parameterise H_0 by the relation

$$H_0 = 100h\ \mathrm{kms}^{-1}\,\mathrm{Mpc}^{-1} \tag{1.1}$$

Observations determine h to be in the range $0.5 \lesssim h \lesssim 1$, the uncertainty being due to the difficulty of measuring distances to remote galaxies. Given the value of H_0, one can immediately construct several other physical quantities of interest. Since H_0 has the dimensions of $(\mathrm{time})^{-1}$, we have

$$\begin{aligned}\text{time scale} &= H_0^{-1} = 3.1 \times 10^{17} h^{-1}\mathrm{sec} = 9.8 \times 10^{9} h^{-1}\mathrm{yr}\\ \text{length scale} &= cH_0^{-1} = 9.3 \times 10^{27} h^{-1} cms \cong 3000 h^{-1}\mathrm{Mpc}.\end{aligned} \tag{1.2}$$

Equally importantly, we can construct a parameter with the dimensions of mass (energy) density:

$$\text{critical density} = \rho_c \equiv \frac{3H_0^2}{8\pi G} = \begin{cases} 1.88 \times 10^{-29} h^2 & \mathrm{gcm}^{-3} \\ 1.05 \times 10^{4} h^2 & \mathrm{eVcm}^{-3}\end{cases} \tag{1.3}$$

[Here, as well as in what follows, we will set $c = 1$]. The significance of this value of the density will shortly become clear. It sets the scale of densities at which gravitational attraction significantly affects the Hubble expansion.

Since H_0 defines a natural scale for mass density, it is convenient to measure the mass density of the universe in terms of ρ_c. We will denote this ratio by the generic symbol Ω and add subscripts - like Ω_B (baryons), Ω_R (relativistic particles), Ω_{NR} (non-relativistic particles), Ω_γ (photons) etc. - to denote various contributions. In general, $\Omega_x \equiv (\rho_x/\rho_c)$.

The observational status of the value of Ω is not very certain. The following claims have been made in the literature: (for a review, see Binney & Tremaine 1987 Chapter 10; Trimble 1988; Peebles 1986) (i) Mean density in solar neighbourhood gives about $0.003h^{-1}$ (ii) Studies based on the Magellanic stream and timing arguments in local groups give a higher value of about $0.06h^{-1}$ (iii) Mass density in groups of galaxies contributes about 0.16 and that in large clusters give about 0.25. (iv) The Virgocentric fall also suggests a mass density of 0.25. (v) Lastly the constraints from primordial nucleosynthesis imply a constraint on the *baryonic* contribution to mass density: $\Omega_B = (0.014 - 0.026)h^{-2}$; or if we take $0.5 < h < 1$, we get $0.014 < \Omega_B < 0.104$. (authors differ somewhat on the upperbound on Ω_B and the cited values are in the range 0.1 to 0.2; see Kolb & Turner 1990)

Two features stand out in the above estimates if they are all correct. There seems to be a tendency for Ω to increase with the scale over which it is measured. (This conclusion is somewhat tentative and quite controversial.) If one uses gravitational effects occurring in a system of size L to measure Ω then we will miss out on matter distributed smoothly over sizes significantly larger than L; thus if the universe has significant fraction of mass distributed smoothly over scales larger than, say 50 Mpc, we can still reconcile the above observations with $\Omega = 1$. Secondly, the observations are just marginally consistent with a fully baryonic universe with $\Omega_B \approx 0.2$ if (and only if) $h = 0.4$. Thus these observations alone, probably, do not rule out a completely baryonic universe (yet!).

Several popular models for the early universe (the so called 'inflationary models') predict that Ω for the universe must be unity (to a high degree of accuracy). This suggests that the dark matter in the universe should be nonbaryonic. If we assume that the dark matter is made of 'weakly interacting massive particles' ('WIMP's) then one can broadly classify them depending on their mass as 'Hot' or 'Cold' dark matter.

Candidates for dark matter with a mass $m_x \lesssim 100$eV are called 'hot' dark matter or HDM in short. This terminology arises due to the fact that such particles have large random velocities and in fact are still relativistic when galactic mass scales are first encompassed within the horizon scale (see below). A typical example of HDM is a massive neutrino with mass $m_\nu \sim 30$eV. At the other extreme dark matter candidates with small random velocities are called 'cold' dark matter, CDM in short. Possible candidates include WIMPS with $m_x \sim 1$GeV, axions or primordial black holes. As we shall see structure formation proceeds very differently depending on whether the DM is hot or cold.

We may summarise the observations as indicating

$$\Omega \equiv \Omega_{\rm total} > 0.2;\ 0.011 < \Omega_B < 0.21;\ \Omega_\gamma h^2 = 2.5 \times 10^{-5} \tag{1.4}$$

Note that Ω_γ is contributed mainly by the microwave background photons while Ω_R includes contributions from all massless species of particles; for example, if there are 3 massless neutrino species then $\Omega_R h^2 = 4.31 \times 10^{-5}$. This can be seen as follows: For any relativistic particle species $-x$ with temperature T_x, the energy density will be given by the relation:

$$\rho_x = g_x \cdot \frac{\pi^2}{30} T_x^4 = \left(\frac{g_x}{2}\right) 4.8 \times 10^{-34} \left(\frac{T_x}{2.75K}\right)^4 \text{gcm}^{-3} \tag{1.5}$$

The g-factor is given by the relation

$$g_x = \begin{cases} g_{\rm spin}(T_x/T_\gamma)^4 & \text{(bosons)} \\ \frac{7}{8} g_{\rm spin}(T_x/T_\gamma)^4 & \text{(fermions)} \end{cases} \tag{1.6}$$

For a relativistic soup consisting of 3 species of neutrinos and photons, the total g-factor will be

$$g_{\rm total} = 2 + 3 \times 2 \times \frac{7}{8} \times \left(\frac{T_\nu}{T_\gamma}\right)^4 = 2 + \frac{42}{8}\left(\frac{4}{11}\right)^{4/3} \simeq 3.36 \tag{1.7}$$

where we have used the fact that $(T_\nu/T_\gamma)^3 = (4/11)$, a relation derived in standard texts in cosmology. This implies that $\rho_R = (3.36/2)\,\rho_\gamma \simeq 1.68\rho_\gamma$, as advertised earlier.

Taking the MBR-temperature to be 2.75K, we can also determine the number density of photons to be $n_\gamma = 422(T_\gamma/2.75K)^3 \mathrm{cm}^{-3}$. The ratio between the number densities of baryons and photons will then work out to be

$$\left(\frac{n_B}{n_\gamma}\right) = 2.68 \times 10^{-8}(\Omega_B h^2) \tag{1.8}$$

1.3. The Friedmann model of the universe

Observations suggest that our universe is homogeneous and isotropic at sufficiently large scales. Such an idealised universe can be described - in Einstein's theory of gravity - by the metric (Friedmann 1922, 1924)

$$ds^2 = dt^2 - a^2(t)\left[d\chi^2 + f^2(\chi)(d\theta^2 + \sin^2\theta d\phi^2)\right] \tag{1.9}$$

where the function $f(\chi)$ is determined by the value of $\Omega = \Omega_{\mathrm{total}}$:

$$f(\chi) = \begin{cases} \sin\chi & (\text{for } \Omega > 1; \text{`closed'}) \\ \chi & (\text{for } \Omega = 1; \text{`flat'}) \\ \sinh\chi & (\text{for } \Omega < 1; \text{`open'}) \end{cases} \tag{1.10}$$

The function $a(t)$ is called the scalefactor. It describes the dynamical evolution of the universe and its specific form can be determined using Einstein's equations, if the matter content is known.

Several important conclusions can be drawn from the form of the metric in (1.9), even without knowing the form of $a(t)$. We list below some of these results:

(i) The action for a free particle moving in this metric is given by

$$A = -m\int ds = -m\int dt\sqrt{1-v^2} \tag{1.11}$$

where $v^2 = g_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta$ is the three- velocity measured with respect to a coincident observer moving along a worldline $\mathbf{x} = \mathit{constant}$. Such observers, sometimes called fundamental, play an important role in cosmological models. Since the action is independent of $\mathbf{x}$, we have the conserved momentum

$$P_\alpha = \frac{\partial L}{\partial\dot{x}^\alpha} = \frac{mg_{\alpha\beta}\dot{x}^\beta}{\sqrt{1-v^2}} \tag{1.12}$$

which in turn implies that

$$P_\alpha P^\alpha \equiv P^2 = g^{\alpha\beta}P_\alpha P_\beta = \frac{m^2v^2}{(1-v^2)} = \frac{constant}{a^2}. \tag{1.13}$$

In other words the magnitude of the 3- momentum decreases as a^{-1} due the expansion. If the particle is nonrelativistic, then $v \propto P$ and the "peculiar" velocity $a|(d\mathbf{x}/dt)|$ itself decays as $a(t)^{-1}$ during expansion.

(ii) By a similar analysis, one can conclude that the wavelength of light is 'redshifted' during expansion: $\lambda \propto a(t)$; $\omega \propto a^{-1}$. It is convenient to define the redshift $z(t)$ by $1 + z(t) \equiv (a_0/a(t))$.
(iii) It follows from the previous result that the Planck spectrum retains its form during the expansion of the universe if we rescale the temperature-parameter T, appearing in the spectrum, by the law $T \propto a^{-1}$. The net energy density of the relativistic particles, $\rho_R \propto T^4$ will fall as a^{-4} under expansion of the universe.
(iv) The Hubble constant is related to the expansion factor by the relation $H = (\dot{a}/a)$ evaluated at present.

Let us now turn to the question of determining the form of $a(t)$. Einstein's equations reduce to the following set in this case:

$$\ddot{a} = -\frac{G}{a^2}\left(\frac{4\pi}{3}a^3\right)(\rho + 3p) \tag{1.14}$$

$$\frac{1}{2}\dot{a}^2 - \frac{G}{a}\left(\frac{4\pi}{3}a^3\right)\rho = \text{constant} = -\frac{k}{2} \tag{1.15}$$

where p is the pressure. [They are cast in a form which is easy to remember in terms of a Newtonian picture but we warn the reader against elevating the mnemonic to a derivation]. In the second equation k is $+1, 0$ or -1 depending on Ω is greater than, equal to or less than unity. These two equations can be combined to give

$$\frac{d}{dt}(\rho a^3) = -p\frac{da^3}{dt}. \tag{1.16}$$

which determines the function $\rho = \rho(a)$ if the equation of state $p = p(\rho)$ is given. Consider matter sources for which $p = w\rho$ with some constant w [e.g. $w = \frac{1}{3}$ for radiation and $w \approx 0$ for non- relativistic matter]. Integrating (1.16), we find that $\rho \propto a^{-3(1+w)}$; in particular,

$$\rho_{rad} \propto a^{-4}; \quad \rho_{NR} \propto a^{-3} \tag{1.17}$$

Equation (1.17)shows that, in the past when $a(t)$ was smaller, ρ_{rad} would have dominated over matter [even though, at present, $\rho_{rad} \ll \rho_{NR}$]. Let $t = t_{eq}$ denote the epoch at which $\rho_{rad} = \rho_{NR}$, with $\rho_{rad} > \rho_{NR}$ for $t < t_{eq}$. Clearly,

$$(1 + z_{eq}) \equiv \frac{a_0}{a_{eq}} = \frac{\Omega_{NR}}{\Omega_R} = 2.32 \times 10^4(\Omega h^2)\theta^4 \tag{1.18}$$

where we have used the notation $\theta \equiv (T/2.75K)$ and assumed that Ω_R is contributed by photons and 3 species of massless neutrinos. At the time $t = t_{eq}$, the temperature of the universe will be

$$T_{eq} \equiv T_{now}(1 + z_{eq}) = 5.5(\Omega h^2)\text{eV} \tag{1.19}$$

[If only photons contribute to Ω_R, $z_{eq} = 3.9 \times 10^4$ $(\Omega h^2)\theta^4$ and $T_{eq} = 9.24(\Omega h^2)$eV.]

The epoch $t = t_{eq}$ is quite close to another important epoch in the history of the universe, denoted by t_{dec}. At very early times, the temperature of the universe was much higher than the ionisation potential of atomic systems; thus matter would have existed in a fully ionised form of positively charged nuclei and free electrons. As the universe expands and cools more and more electrons will combine with nuclei to form bound neutral atoms. The epoch $t = t_{dec}$ typically denotes this instant. More precisely, we may characterise this epoch as the time at which the mean-free-path of the photons is of the order of the scale length ("size") of the universe. Once this happens, photons will effectively decouple from the rest of the matter and will move freely through the universe, experiencing negligible collisions.

Detailed calculation (see e.g. Jones & Wyse 1985; Kolb & Turner 1990) shows that the value of t_{dec} (or, equivalently, z_{dec} and T_{dec}) depends weakly on Ω_B and can be fitted to the form :

$$(1+z_{dec}) \approx 1100 \left(\frac{\Omega_{total}}{\Omega_B}\right)^{0\ 018} \approx 1100; \qquad 1 < \frac{\Omega}{\Omega_B} < 100. \tag{1.20}$$

We will take $z_{dec} = 1100$ for the purposes of calculation; this corresponds to $T_{dec} = 0.26\text{eV}$. These calculations also show that a small fraction of matter remains ionised well after t_{dec}; the surviving, asymptotic fractional ionisation is given by the formula,

$$X_e = 2.4 \times 10^{-3} \frac{(\Omega h^2)^{1/2}}{(\Omega_B h^2)} \left(\frac{z_{dec}}{1000}\right)^{12.75} \tag{1.21}$$

We can now return to the task of determining the form of $a(t)$. The above analysis shows that $\rho \propto a^{-4}$ for $t \ll t_{eq}$ while $\rho \propto a^{-3}$ for $t \gg t_{eq}$. The explicit solutions to (1.15) are complicated and are not of much use. These solutions, however, can be approximated very well by the following analytic expressions: For definiteness, consider a universe with $\Omega > 1$. For z less than about 30 or so, $a(t)$ is implicitly given by

$$\begin{aligned} a(z) &= a_0(1+z)^{-1} = H_0^{-1}(\Omega_0 - 1)^{-\frac{1}{2}}(1+z)^{-1} \\ H_0 t(z) &= \frac{\Omega_0}{2(\Omega_0-1)^{3/2}}\left[\cos^{-1}\left(\frac{\Omega_0 z - \Omega_0 + 2}{\Omega_0(1+z)}\right) - \frac{2(\Omega_0-1)^{1/2}(\Omega_0 z+1)^{1/2}}{\Omega_0(1+z)}\right] \end{aligned} \tag{1.22}$$

while for larger z, $a(t)$ is given by

$$H_{eq}t = \left(\frac{2\sqrt{2}}{3}\right)\left[\left(\frac{a}{a_{eq}} - 2\right)\left(\frac{a}{a_{eq}} + 1\right)^{1/2} + 2\right] \tag{1.23}$$

Here $a_{eq} = a_0(1+z_{eq})^{-1} = H_0^{-1}(\Omega_0 - 1)^{-\frac{1}{2}}(1+z_{eq})^{-1}$ and $H_{eq} = H(t_{eq})$ is the Hubble 'constant' at the epoch corresponding to z_{eq}. Putting $a = a_{eq}$ in (1.23) gives the relation $H_{eq}t_{eq} \approx 0.553$. H_{eq} itself can be estimated from $H_{eq}^2 = (8\pi G/3)[2\rho_c\Omega(1+z_{eq})^3]$. So the explicit value of t_{eq} is given by

$$t_{eq} = 3.4 \times 10^{10}(\Omega h^2)^{-2} sec \tag{1.24}$$

In fact, for z larger than about 100 or so, one can use much simpler expressions

$$\frac{t}{t_{eq}} = \begin{cases} 1.7(a/a_{eq})^{3/2}; & a \gg a_{eq}(\text{matter dominated}) \\ 1.3(a/a_{eq})^2; & a \ll a_{eq}(\text{radiation dominated}) \end{cases} \tag{1.25}$$

Near $t \cong t_{eq}$, we need to use (1.23). All these results can be easily derived by inspecting the various terms in (1.15). These expressions also show that $a(t) \propto t^{1/2}$ in the radiation dominated phase and $a(t) \propto t^{2/3}$ in the matter dominated phase, for larger redshifts.

In the above discussion, it was assumed that $\Omega > 1$. For discussion of the early universe, it really does not matter whether Ω is less than or greater than unity. However, if $\Omega < 1$, then the right hand side of (1.15) will dominate over ρ_{NR} for $z < (1 - 2\Omega)\Omega^{-1}$; such an epoch is called "curvature dominated" and will be of importance in the theory of structure formation to be discussed later. Using the value of t_{eq}, we can compute the explicit value for t_{dec}. We get:

$$t_{dec} \cong 5.6 \times 10^{12}(\Omega h^2)^{-\frac{1}{2}}\, sec \tag{1.26}$$

1.4. The length scales of the universe

Consider two fiducial locations in the universe - say, the positions of two well-separated galaxies, one located at the origin and the other at (r, θ, ϕ). The coordinate distance between these two points will be a constant (r) as long as the galaxies have no peculiar motion. However, due to the overall expansion of the universe, the *proper* distance between these two points will keep increasing:

$$\text{proper distance} = l(t) = \left(\frac{a(t)}{a_0}\right) l_0 \propto a(t) \tag{1.27}$$

In other words, all proper distances in the universe scale with the expansion factor; hence they grow as $t^{\frac{1}{2}}$ in radiation dominated (RD) phase and as $t^{2/3}$ in the matter dominated (MD) phase. More realistically, they grow as t^n with $0.5 \lesssim n \lesssim 0.66$.

The dynamics of the expansion, on the other hand, is determined by another length scale, called the Hubble radius:

$$d_H(t) \equiv \left(\frac{\dot{a}}{a}\right)^{-1} \tag{1.28}$$

which is proportional to t if $a(t) \propto t^n$. It follows that the Hubble radius grows at a faster rate than the proper distance. This length $d_H(t)$ is typically the size over which physical processes operate coherently. Given the cosmological evolution of the model - that is, the function $a(t)$ - we can uniquely determine $d_H(t)$. Consider, for example, the Hubble radius at t_{eq}. Using (1.24) and the relation $H_{eq}t_{eq} \approx 0.553$, derived in the previous section we have

$$d_H(t_{eq}) = (H_{eq})^{-1} \cong 1.85 \times 10^{21}(\Omega h^2)^{-2} cms \tag{1.29}$$

A region with size $d_H(t_{eq})$ at $t = t_{eq}$ would have expanded by the factor $[a_0/a(t_{eq})] = (1+z_{eq})$ from $t = t_{eq}$ till today. Therefore, the proper size today for that region-which was as big as the Hubble radius at $t = t_{eq}$ - will be

$$l_{eq}(now) = d_H(t_{eq})(1+z_{eq}) \cong 13\text{Mpc}(\Omega h^2)^{-1} \tag{1.30}$$

This size, of course, is much smaller than the Hubble radius today: d_H (today) $\simeq 3000h^{-1}$ Mpc. This length scale will play an important role in future discussions.

Reversing the above argument, we can draw an important conclusion: Consider a region of proper size λ today [with $\lambda < d_H$ (today)]. As we go back in time, this (proper) length will shrink as $a(t) \propto t^n$ with $n < 1$; but the Hubble radius of the universe decreases *faster*, as t. Therefore, there will be some time $t = t_{\text{enter}}$ (λ) in the past, when the proper size of this region will equal the Hubble radius of the universe. For $t < t_{\text{enter}}(\lambda)$, the proper size will be bigger than the Hubble radius. It is usual to say that 'the length scale λ enters the Hubble radius' at $t = t_{\text{enter}}(\lambda)$. This feature is illustrated in fig. 1.1.

It follows from the earlier analysis that a length of $\lambda_{eq} \equiv 13\text{Mpc}(\Omega h^2)^{-1}$ enters the Hubble radius at $t = t_{eq}$, the time of transition from RD-phase to MD-phase. Smaller regions ($\lambda < \lambda_{eq}$) will enter the Hubble radius earlier, in the RD-phase while larger sizes ($\lambda > \lambda_{eq}$) will enter later, in the MD-phase. Given the explicit form of $a(t)$, we can easily compute the time $t_{\text{enter}}(\lambda)$ by solving the equation

$$\left(\frac{\dot{a}}{a}\right)^{-1}_{t=t_{\text{enter}}} = \lambda \left(\frac{a(t)}{a_0}\right)_{t=t_{\text{enter}}} \tag{1.31}$$

This leads to the following result:

$$t_{\text{enter}}(\lambda) = \begin{cases} 2.6 \times 10^7 (\Omega h^2) \left(\frac{\lambda}{1\text{Mpc}}\right)^3 ; \text{sec} & \lambda > \lambda_{eq} \\ 6.1 \times 10^8 g^{1/2} g_s^{-2/3} \left(\frac{\lambda}{1\text{Mpc}}\right)^2 ; \text{sec} & \lambda < \lambda_{eq} \end{cases} \tag{1.32}$$

The factors g and g_s take into account the contribution of various particle species (relative to a spin-zero boson) to the energy and entropy densities of the universe respectively; they do not change the results appreciably except at very early phases of the universe when $T > 1$GeV or so. The numerical values in these expressions are different because we have scaled them to 1Mpc; they, of course, match at $\lambda = \lambda_{eq}$. Given t_{enter}, one can also compute the temperature of the universe at that epoch. We get:

$$T_{\text{enter}}(\lambda) = \begin{cases} 948\text{eV}(\Omega h^2)^{-1} \left(\frac{\lambda}{1\text{Mpc}}\right)^{-2} ; & \lambda > \lambda_{eq} \\ 63\text{eV}\, g_s^{1/3} g^{-1/2} \left(\frac{\lambda}{1\text{Mpc}}\right)^{-1} ; & \lambda < \lambda_{eq} \end{cases} \tag{1.33}$$

It should be noted that, in the above formulas, λ refers to proper distance *today*. Since proper distances scale with expansion, it is somewhat an inconvenient description [e.g. it is difficult to visualize which physical processes are important at proper sizes of 100 kpc at a redshift of 70, say]. It will be more useful to parameterise length

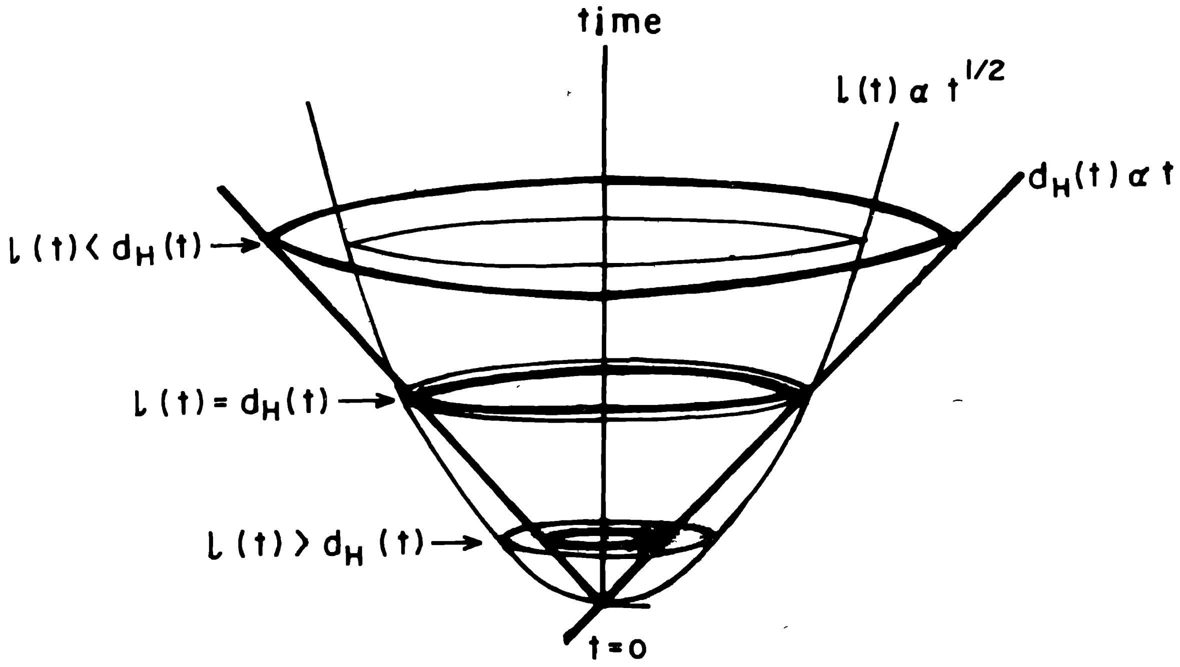

Figure 1.1. The Hubble radius and a wavelength entering the Hubble radius.

scales by some quantity which does not change with expansion. The amount of nonrelativistic mass $M(\lambda)$, contained inside a sphere of proper radius $(\lambda/2)$,

$$M(\lambda) \equiv \frac{4\pi}{3}\rho_{NR}\left(\frac{\lambda}{2}\right)^3 = 1.45 \times 10^{11} M_\odot (\Omega h^2)\left(\frac{\lambda}{1\text{Mpc}}\right)^3 \qquad (1.34)$$

will be such a quantity. As the universe expands, $\lambda \propto a(t)$ while $\rho_{NR} \propto a^{-3}$ keeping $M(\lambda)$ constant. Thus we can specify λ by just quoting the equivalent mass associated with it; and, we don't have to specify when this quantity is measured.

The previous formulas for $t_{\rm enter}$ and $T_{\rm enter}$ can be easily reexpressed in terms of M rather than λ. A region containing a mass

$$M_{eq} \equiv M(\lambda_{eq}) = 3.2 \times 10^{14} M_\odot \theta^6 (\Omega h^2)^{-2} \qquad (1.35)$$

will come into the Hubble radius at $t = t_{eq}$. Smaller regions will enter earlier and larger masses later. The relation

$$(1 + z_{\rm enter}) = \begin{cases} 1.1 \times 10^6(\Omega h^2)^{-1/3}(M/10^{12}M_\odot)^{-2/3}; & M > M_{eq} \\ 1.41 \times 10^5(\Omega h^2)^{1/3}(M/10^{12}M_\odot)^{-\frac{1}{2}}; & M < M_{eq} \end{cases} \tag{1.36}$$

gives the redshift at which a region containing mass M enters the Hubble radius.

Notice that the quantity $M(\lambda)$ is computed using the smoothed-out density of the homogeneous universe. According to (1.34), a typical galactic mass ($\sim 10^{11}M_\odot$) will correspond to a proper size of about 1 Mpc; actual galaxies are much smaller because they ceased to expand with the cosmic medium sometime in the past, and are now dominated by self-gravity. This fact, of course, is irrelevant to the scaling arguments given above which deal with a (hypothetical) smooth universe.

We conclude this section with a comment on another important length scale in cosmology, viz. the horizon siz Suppose for a moment that $a(t) = a_0 t^n$ with $n < 1$ for all $t \geq 0$. Then, a photon can travel a maximum coordinate distance of

$$r(t) = \int_0^t \frac{dx}{a(x)} = \frac{1}{a_0}\frac{t^{1-n}}{(1-n)} \tag{1.37}$$

which corresponds to the proper distance.

$$h(t) = a(t)r(t) = (1-n)^{-1}t \tag{1.38}$$

This differs from the Hubble radius $(\dot{a}/a)^{-1} = n^{-1}t$ only by a constant factor of order unity. This has led to considerable confusion in nomenclature with several publications calling the quantity Hubble radius $d_H(t)$ as "horizon" [and, sometimes, even attributing to it the causal properties of the horizon !]. Notice that $d_H(t)$ is a local quantity and its value at t is essentially decided by the behaviour of $a(t)$ near t; in contrast, the value of $h(t)$ depends on the entire past history of the universe. In fact, $h(t)$ depends very sensitively on the behaviour of $a(t)$ near $t = 0$ - something which we know nothing about! [If $a(t) \propto t^m$ with $m \geq 1$ near $t = 0$, then $h(t)$ is infinite for all $t \geq 0$!]. Thus, when $h(t)$ and $d_H(t)$ differ widely, it is the latter quantity which is usually relevant.

2: Linear theory of perturbations

2.1. Growth of inhomogeneities - general comments

If there were no inhomogeneities in the universe, then we would have no difficulty in explaining that observation! However, since our universe contains galactic and other structures, it is necessary to modify the formalism of part 1 to account for these inhomogenities.

Since galaxies provide a convenient unit in the cosmic mass ladder, it is natural to begin by asking how they are distributed in the Universe. Are they distributed randomly or do they cluster in any significant manner?

To answer such a question reliably one needs a good survey of the universe giving the coordinates of galaxies in the sky. Of the 3-coordinates needed to specify the position of the galaxy, the two angular coordinates are easy to obtain. There exists today several galaxy catalogues, containing the angular positions of galaxies in particular regions of the sky, complete up to a chosen depth. The APM Galaxy survey has about 5×10^6 galaxies out to a depth of $600h^{-1}$ Mpc; the Lick catalogue has about 1.6×10^6 galaxies and depth of $200h^{-1}$ Mpc; the IRAS catalogue has more than 14000 galaxies which are prominent in the infrared band; these are few major catalogues available today. If we know the redshift z of these galaxies as well then we can attribute to it a line-of-sight velocity $v \cong zc$. If we further assume that this velocity is due to cosmic expansion, then we can assign to the galaxy a radial distance of $r \cong H_0^{-1}v$. This will provide us with the galaxy position (r, θ, ϕ) in the sky.

The main difficulty in completing the survey lies in obtaining telescope time to make a systematic measurement of redshifts for the galaxies which are members of a catalogue. We know the redshifts to only about 30,000 or so galaxies [out of millions which exist in catalogues] and the largest systematic survey - the centre for astrophysics (CFA) survey - has only 9000 redshifts or so. The recently completed (partial) survey of IRAS galaxies has improved the situation somewhat; in a decade or so, the observations will be in far better shape.

Even the limited amount of data we have today points to a perplexing pattern in galaxy distribution. The single most useful function characterising the galaxy distribution is what is called the 'two-point-correlation function': $\xi_{GG}(r)$. This function is defined via the relation

$$dP = \overline{n}^2(1 + \xi_{GG}(\mathbf{r}_1 - \mathbf{r}_2))d^3\mathbf{r}_1 d^3\mathbf{r}_2 \qquad (2.1)$$

where dP is the probability to find two galaxies simultaneously in the regions $(\mathbf{r}_1, \mathbf{r}_1 + d^3\mathbf{r}_1)$ and $(\mathbf{r}_2 + d^3\mathbf{r}_2)$ and $\overline{n}$ is the mean number density of galaxies in space. The homogeneity of the background universe guarantees that $\xi_{GG}(\mathbf{r}_1, \mathbf{r}_2) = \xi_{GG}(\mathbf{r}_1 - \mathbf{r}_2)$ and isotropy will further make $\xi_{GG}(\mathbf{r}) = \xi_{GG}(|\mathbf{r}|)$. From (2.1) it follows that $\xi_{GG}(\mathbf{r})$ measures the excess probability (over random) of finding a pair of galaxies separated by a distance $\mathbf{r}$; so if $\xi_{GG}(\mathbf{r}) > 0$, we may interpret it as clustering of galaxies over and above the random Poisson distribution.

Considerable amount of effort was spent in the past decades in determining $\xi_{GG}(r)$ from observations. These studies show that

$$\xi_{GG}(r) \simeq \left(\frac{r}{5h^{-1}\text{Mpc}}\right)^{-1\ 8} \qquad (2.2)$$

in the range $0.1h^{-1}$ Mpc$\lesssim r \lesssim$ $20h^{-1}$Mpc. (Davies & Peebles 1983; Peebles 1980). This simple power law has been a challenge for theoreticians over the ages!

Nearly ten percent of all galaxies are found in rich clusters containing anything from hundred to thousands of galaxies. It is also possible - using the catalogues of rich clusters, like Abell catalogue which contains 4076 clusters - to compute the correlation function between galaxy clusters. The result turns out to be

$$\xi_{CC}(r) \simeq \left(\frac{r}{25h^{-1}\text{Mpc}}\right)^{-1.8} \qquad (2.3)$$

suggesting that clusters are more strongly correlated than the individual galaxies. If firmly established, this result implies that, the visible matter does not faithfully trace the mass distribution of the universe. Unfortunately, $\xi_{CC}(r)$ is not as well established as ξ_{GG} and hence one has to be cautious in interpreting results which depend on ξ_{CC}.

In recent years, researchers have also resorted to less quantitative - but more appealing - diagnostics to demonstrate clustering of galaxies. Several recent surveys present striking *visual* patterns in the redshift-angle space. The patterns are consistent with the interpretation that the universe contains several voids of size about $(20h^{-1}$ to $50h^{-1})$Mpc. The CFA slices, in fact, suggest that the galaxies are concentrated on sheet like structures surrounding nearly empty voids. (Bachall *et al.* 1983; DeLapparent *et al.* 1986; Giovanelli 1982; Koo *et al.*. 1986; Strauss *et al.*. 1988)

More recently, a redshift survey of randomly selected sample of 2163 galaxies from the IRAS catalogue has been completed, allowing one to construct the density field of the universe upto about $140h^{-1}$Mpc. Preliminary investigations of the clustering, based on this survey, indicate that density field of the universe has lot more power on large scales than anticipated before (Rowan-Robinson *et al.*. 1990; Saunders *et al.* 1990). This conclusion is also confirmed by the measurement of 2-dimensional angular correlation function of galaxies based on the machine scans of 185 UK Schmidt plates covering more than 2 million galaxies. (Maddox *et al.* 1990)

The conventional wisdom tries to account for the observed matter distribution in the universe in the following manner: We assume that, at some time in the past, there were small deviations from the homogeneity in our universe. These deviations grow due to gravitational instability over a period of time. As long as these deviations are small, we can linearise the equations and study the growth of these perturbations. Once the deviations from the smooth universe become large, we have to use different techniques to understand the non-linear evolution. Lastly, we have to develop some physical mechanism capable of generating the initial inhomogenity. In this part of the review we shall study the linear regime.

One can attempt a linear perturbation theory along the following lines: (i) Perturb the metric $g_{ik}(x)$ and the source T_{ik} into the form $(g_{ik}+\delta g_{ik})$ and $(T_{ik}+\delta T_{ik})$. The set (g_{ik}, T_{ik}) corresponds to the smooth background universe, while the set $(\delta g_{ik}, \delta T_{ik})$ denotes the perturbation. (ii) Assuming the latter to be 'small', we can linearise Einstein's equations to obtain a second-order-differential equation of the form

$$\hat{\mathcal{L}}(g_{ik})\delta g_{ik} = \delta T_{ik} \tag{2.4}$$

where $\hat{\mathcal{L}}$ is a linear differential operator depending on the background space-time (iii) Being a linear equation, it is convenient to Fourier transform the variables and obtain a separate equation $\hat{\mathcal{L}}_{(k)}\delta g_{(k)} = \delta T_{(k)}$ for each mode labeled by a wave vector $\mathbf{k}$. (iv) Solving this equation, we can determine the evolution of each mode separately.

There is, however, one major conceptual difficulty in carrying out this programme. In general relativity, the form (and numerical value) of the metric coefficients g_{ik} (or the stress-tensor components T_{ik}) can be changed by a relabelling of coordinates $x^i \to x^{i'}$. By such a trivial change we can make a small δT_{ik} large or even generate a component which was originally absent. Thus the perturbations may grow at different rates – or even decay! – when we relabel coordinates. It is necessary to tackle this difficulty before we can meaningfully talk about the growth of inhomogenities.

There is a simple way of handling this problem for modes which have proper wavelengths which are much smaller than the Hubble radius. The general relativistic effects due to the curvature of the space-time will be negligible at sizes far smaller than the Hubble radius. In such regions, there exists a natural choice of coordinates in which Newtonian gravity is applicable. All physical quantities can be unambiguously defined in this context. (see e.g. Weinberg 1972; Peebles 1980) As we will see, such a Newtonian analysis provides valuable insight into the behaviour of inhomogenities.

The trouble with the above idea is that the proper wavelength of any mode will be bigger than the Hubble radius at sufficiently early epochs. We saw in the last section that any proper length λ (as measured today) with $\lambda \ll H_o^{-1}$ today would have entered the Hubble radius at some time $t_{enter}(\lambda)$ in the past. Newtonian analysis can be used to study a mode labeled by λ only for times $t \gg t_{enter}(\lambda)$, when the mode is well within the Hubble radius. Thus, the early evolution of any mode needs to be tackled by general relativity and the coordinate ambiguities again rear their ugly head.

There are two different ways of handling such difficulties in general relativity and both have been tried out in the cosmological context. The first method is to resolve the problem by force: We choose a particular coordinate system and compute everything in that coordinate system. If the coordinate system is physically well motivated, then the quantities computed in that system can be interpreted easily; for example, we will treat δT_o^o to be the perturbed mass (energy) density even though it is, of course, coordinate dependent. The trouble with this method is that one cannot fix the gauge completely by simple, physical arguments; the residual gauge ambiguities create some headache.

The second approach is to construct quantities – linear combinations of various perturbed physical variables – which are scalars under the coordinate transformations. (Bardeen 1980). Einstein's equations are then rewritten as equations for these gauge invariant quantities. This approach, of course, is manifestedly gauge invariant from start to finish. However it is more complicated than the first one; besides, the gauge invariant objects are quite wierd and possess no straightforward interpretation.

In principle, therefore, the perturbation theory should proceed in two steps: (i) Given a mode λ, we know $t_{enter}(\lambda)$. For $t < t_{enter}(\lambda)$, $\lambda > d_H$ we use a general relativistic perturbation theory to evolve $\delta\rho_\lambda(t)$ from some $t = t_i$ to $t = t_{enter}(\lambda)$. (ii) For $t > t_{enter}(\lambda)$, $\lambda < d_H$ and we can study the evolution of $\delta\rho_\lambda$ using Newtonian theory.

It turns out that most of the results can be understood in terms of simple scaling arguments. Therefore, we will first discuss a simplified analysis of perturbation growth in the next section. A more rigorous – and sophisticated – analysis will be presented in section 2.3.

2.2. Suppression and growth of perturbations

The material content of the smooth universe has three main components – baryons (ρ_B), darkmatter (ρ_{DM}) and relativistic matter like photons (ρ_R). To characterise these sources, we specify the equations of state for each of them connecting the pressure p_x of component x with the density ρ_x. We usually take $p_{DM} = p_B \approx 0$, when matter is non-relativistic and $p_R = \frac{1}{3}\rho_R$, In studying the perturbations individually, we can set:

$$\begin{aligned} \delta p_{\text{matter}} &= \left(\frac{\partial p}{\partial \rho}\right)_{\text{matter}} \delta\rho_{\text{matter}} = \left(\frac{\dot{p}}{\dot{\rho}}\right)\delta\rho \equiv v^2(\delta\rho)_{\text{matter}} \\ \delta p_R &= \left(\frac{\delta p_R}{\delta\rho_R}\right)\delta\rho_R = \left(\frac{\dot{p}_R}{\dot{\rho}_R}\right)\delta\rho_R = v_R^2\delta\rho_R \end{aligned} \tag{2.5}$$

In general, there is no relation between $(\delta\rho)_{\text{matter}}$ and $(\delta\rho)_R$. Any fundamental theory explaining the physical origin of fluctuations will, however, provide such a relation (In the absence of such a theory, we may simply assume this relation). Inflationary models, for example, predict that

$$\delta\rho_R \cong \left(\frac{4\rho_R}{3\rho_{\text{matter}}}\right)\delta\rho_{\text{matter}} \tag{2.6}$$

This is equivalent to the statement that

$$\delta\left(\frac{T^3}{\rho_{\text{matter}}}\right) = \delta\left(\frac{n_R}{\rho_{\text{matter}}}\right) = 0 \tag{2.7}$$

[we have used the relations $\rho_R \propto T^4$, $n_R \propto T^3$]. Since the entropy of the radiation also scales as T^3, the above relation keeps the relative entropy constant; it is usual to say that such fluctuations are adiabatic.

Having fully characterised the fluctuations, we can study their evolution. Consider first the situation in which the wavelength λ of the perturbation is much larger than the Hubble radius d_H. Since processes like pressure, viscosity etc act at scales much smaller than d_H, they do not affect the evolution of the super-horizon-modes. Even though a rigorous study of such a mode requires general relativity, the final result can be obtained by the following trick (Zeldovich & Novikov 1983):

Consider a spherical region of radius $\lambda(> d_H)$ containing matter with a mean density ρ_1, embedded in a $k = 0$ Friedmann universe of density ρ_0 (with $\rho_1 = \rho_0 + \delta\rho$; $\delta\rho$ small and positive). It can be shown that the inside region is not affected by the matter outside and evolves as a $k = +1$ Friedmann universe. Therefore, we can write

$$H_1^2 + \frac{1}{a_1^2} = \frac{8\pi G}{3}\rho_1; \quad H_0^2 = \frac{8\pi G}{3}\rho_0; \quad \left(H_0 = \frac{\dot{a}_0}{a_0}; H_1 = \frac{\dot{a}_1}{a_1}\right) \tag{2.8}$$

We will compare the perturbed universe with the background universe ***when their expansion rates are equal***; i.e. we compare their densities at a time t when $H_1 = H_0$. We then get

$$\frac{8\pi G}{3}(\rho_1 - \rho_0) = \frac{1}{a_1^2} \tag{2.9}$$

or

$$\left(\frac{\rho_1 - \rho_0}{\rho_0}\right) = \frac{\delta\rho}{\rho_0} = \frac{3}{8\pi G(\rho_0 a_1^2)} \tag{2.10}$$

In general, if $H_0 = H_1$, at some time, then $a_0 \neq a_1$ at that time. But, if $(\delta\rho/\rho_0)$ is small, then a_1 and a_0 will differ by only a small quantity and we can set $a_1 \approx a_0$ in the right hand side of (2.10). This allows one to find how $(\delta\rho/\rho_0)$ scales with a. Since $\rho_0 \propto a^{-4}$ in RD-phase and $\rho_0 \propto a^{-3}$ in MD-phase, we get

$$\left(\frac{\delta\rho}{\rho}\right) \propto \begin{cases} a^2 & \text{(RD phase)} \\ a & \text{(MD phase)} \end{cases} \tag{2.11}$$

Thus, the amplitude of the super horizon mode always grows; as a^2 in RD-phase and as a in the MD-phase. [It is possible to prove this result rigorously using general relativity. Our choice of comparing ρ_1 and ρ_0 when $H_1 = H_0$ corresponds to choice of gauge].

Consider now what happens to this mode when it enters the horizon and becomes a subhorizon mode. $(\lambda < d_H)$. There are two processes which can prevent its amplitude from growing.

The first one is the familiar pressure support. If the pressure distribution of matter can readjust itself fast enough - i.e. if sufficient pressure can build up before gravity crushes the perturbed region under its own weight - then the pressure will prevent gravitational enhancement of the density contrast. The condition for this is

$$\{\text{timescale for the pressure readjustment}\} < \{\text{timescale for gravitational collapse}\} \tag{2.12}$$

That is

$$t_{\text{pressure}} \simeq \frac{\text{wavelength}}{\text{vel.dispersion}} = \frac{\lambda}{v} < \frac{1}{\sqrt{G\rho}} = \text{free fall time} \simeq t_{\text{collapse}}. \tag{2.13}$$

This condition for stability implies that growth is suppressed in modes with wavelengths λ less than a critical wavelength $\lambda_J \sim v(G\rho)^{-\frac{1}{2}}$ It is conventional to define this "Jeans length" with an extra $\sqrt{\pi}$ factor:

$$\lambda_J \equiv \sqrt{\pi}\frac{v}{\sqrt{G\rho}} \tag{2.14}$$

If the universe contains only one species of particle, then the v and ρ will both correspond to that species. In a multi-component medium, v will be the velocity dispersion of the perturbed component (it is the perturbed component that provides the pressure support), but ρ will be the density of the component which is most dominant gravitationally (it is this component which is contracting the perturbation; think of the atmosphere above earth where *gas* pressure works against the *earth's* gravity). In general, of course, these two components will not be the same.

The pressure in a baryonic gas is essentially provided by collisions. But in the DM-component, the collisions are usually quite ignorable. The "pressure" support in a collisionless system arises from the readjustment of orbits. In both the cases, however, the timescale t_{pressure} is set by the velocity dispersion, v.

There is a second process which can prevent the growth of perturbations. This occurs when (i) the perturbed species is *not* the dominant species (which governs the expansion rate) ***and*** (ii) the dominant species is smoothly distributed. (Mezzaros 1975). Suppose that $t_{\rm grav}$ for the perturbed species (say, DM) is indeed less than $t_{\rm pressure}$; thus the condition $\lambda > \lambda_J$ is satisfied and pressure cannot prevent the collapse. But suppose that we are in a RD-phase, when the expansion timescale $t_{exp} \sim (G\rho_{\rm dominant})^{-\frac{1}{2}} \sim (G\rho_R)^{-\frac{1}{2}}$ is smaller than $t_{\rm grav}$. Then the universe will be expanding too fast for the collapsing region to condense out. We have here a situation with $t_{exp} < t_{\rm grav} < t_{\rm pressure}$; that is,

$$\frac{1}{\sqrt{G\rho_R}} < \frac{1}{\sqrt{G\rho_{DM}}} < \frac{\lambda}{v} \tag{2.15}$$

It is rapid background expansion rather than pressure support which prevents the growth.

If neither of these processes are operational, then the amplitude will grow. It is clear that the second process will prevent growth in all *subhorizon* modes in the RD-phase. Thus, in RD-phase, only superhorizon modes grow and they grow as a^2 (see (2.11)). In the MD-phase, for all $\lambda \gg \lambda_J$, we can ignore pressure effects; thus the analysis leading to (2.11) is valid even for subhorizon modes with $\lambda \gg \lambda_J$ (i.e for $d_H > \lambda \gg \lambda_J$ as well). So we conclude that, for all $\lambda \gg \lambda_J$ in the MD-phase, the amplitude grows as a. Modes with $\lambda \gtrsim \lambda_J$ also grow, but in a more complicated manner (because of pressure corrections).

We can now put all the pieces together and study the life of a perturbation with wavelength λ. The relevant scalings are shown in fig. 2.1. Suppose that this mode (with proper wavelength $\lambda \propto a$) enters the Hubble radius in the RD-phase at some $a = a_{\rm enter}$. Let us consider the perturbations in DM-component at different epochs first. For DM, the velocity dispersion $v \simeq 1$ when the particles are relativistic ($a < a_{nr}$) and decays as $v \propto a^{-1}$ when the particles are non-relativistic ($a > a_{nr}$; see the discussion in section 2.2.) Since $\rho_{\rm dom} = \rho_R$ for $a < a_{eq}$ and $\rho_{\rm dom} = \rho_{DM}$ for $a > a_{eq}$, the quantity $\rho_{\rm dom}^{-\frac{1}{2}}$ will scale as $\rho_{\rm dom}^{-\frac{1}{2}} = \rho_R^{-\frac{1}{2}} \propto a^2$ for $a > a_{eq}$ and as $\rho_{dom}^{-\frac{1}{2}} \propto \rho_{DM}^{-\frac{1}{2}} \propto a^{3/2}$ for $a > a_{eq}$. Combining, we find that, for dark matter,

$$\lambda_J \propto \frac{v}{\rho_{\rm dom}^{1/2}} \propto \begin{cases} a^2 & a < a_{nr} \\ a & a_{nr} < a < a_{eq} \\ a^{1/2} & a_{eq} < a \end{cases} \tag{2.16}$$

There are 3 essential stages in the evolution of a mode which enters the Hubble radius between a_{nr} and a_{eq} [as we shall see later, these are the modes most relevant to astrophysics]:

(a) Stage 1 ($a < a_{\rm enter}$): The wavelength of the perturbation is bigger than the Hubble radius; from our earlier discussion we know that

$$\left(\frac{\delta\rho}{\rho}\right) \propto a^2 \tag{2.17}$$

(b) Stage 2 ($a_{\rm enter} < a < a_{eq}$): The wavelength is inside the Hubble radius and bigger than λ_J; so, pressure support cannot stop the collapse. However, the

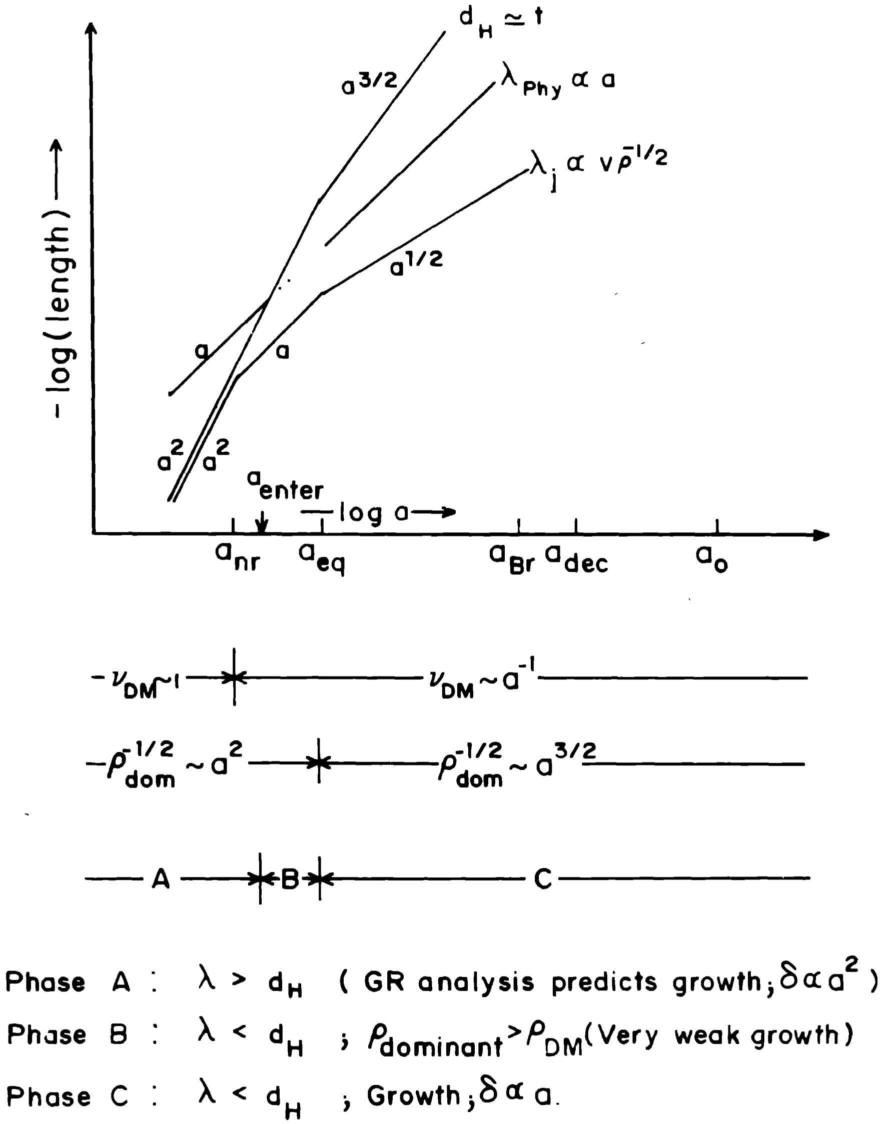
$d_H \simeq t$
$a^{3/2}$
$\lambda_{Phy} \propto a$
$\lambda_j \propto v\rho^{-1/2}$
$a^{1/2}$
a
a
a^2
a^2
a_{enter}
- log(length) →
—log a→
a_{nr}
a_{eq}
a_{Br}
a_{dec}
a_0
$-\nu_{DM} \sim 1$
$\nu_{DM} \sim a^{-1}$
$-\rho_{dom}^{-1/2} \sim a^2$
$\rho_{dom}^{-1/2} \sim a^{3/2}$
A
B
C
Phase A : $\lambda > d_H$ (GR analysis predicts growth; $\delta \propto a^2$)
Phase B : $\lambda < d_H$; $\rho_{dominant} > \rho_{DM}$ (Very weak growth)
Phase C : $\lambda < d_H$; Growth; $\delta \propto a$.

Figure 2.1. Jeans length for dark matter.

dominant component driving the expansion is radiation, and since $\rho_R > \rho_{DM}$ the $t_{exp} < t_{\text{collapse}}$. Thus, rapid expansion prevents the growth of perturbations in this stage:

$$\left(\frac{\delta\rho}{\rho}\right) = \text{constant}. \tag{2.18}$$

(c) Stage $3(a_{eq} < a)$: The wavelength is inside the Hubble radius and bigger than

λ_J; further $\rho_{\rm dom}$ now *is* ρ_{DM} itself. Neither process described before can prevent the growth. For $\lambda \gg \lambda_J$ (so that pressure corrections are ignorable), the growth is as described by (2.11):

$$\left(\frac{\delta\rho}{\rho}\right) \propto a \tag{2.19}$$

It is conventional to present the above results in terms of a quantity called Jean's mass:

$$M_J \equiv \frac{4\pi}{3}\rho\left(\frac{\lambda_J}{2}\right)^3 \tag{2.20}$$

where ρ is the component under discussion. Since $\rho_{DM} \propto a^{-4}$ for $a < a_{nr}$ and $\rho_{DM} \propto a^{-3}$ when $a > a_{nr}$, we see that

$$M_J \propto \begin{cases} a^2 & a < a_{nr} \\ \text{constant} & a_{nr} < a < a_{eq} \\ a^{-\frac{3}{2}} & a_{eq} < a \end{cases} \tag{2.21}$$

The mass inside Hubble radius M_H is similarly defined to be

$$M_H = \frac{4\pi}{3}\rho_{DM}\left(\frac{d_H}{2}\right)^3 \propto \begin{cases} a^2 & a < a_{nr} \\ a^4 & a_{nr} < a < a_{eq} \\ a^{3/2} & a_{eq} < a \end{cases} \tag{2.22}$$

The perturbation, of course, has a constant mass $M(\lambda)$. All the previous 3 stages can be reexpressed in terms of these masses, as shown in fig. 2.2.

Let us next consider the perturbations in the baryonic component. Since baryons and photons are tightly coupled for $a < a_{dec}$, the pressure and density of the baryon-photon soup is well correlated. Let a_{Br} be the epoch at which $\rho_R = \rho_B$; usually, $a_{Br} \lesssim a_{dec}$. For $a < a_{Br}$, $P_R \gg P_B$ and $\rho_R \gg \rho_B$; so

$$v^2 = \left(\frac{\partial P}{\partial\rho}\right) \simeq \frac{P}{\rho} = \frac{P_R + P_B}{\rho_R + \rho_B} \approx \frac{P_R}{\rho_R} = \frac{1}{3} \qquad a < a_{Br} \tag{2.23}$$

For $a_{Br} < a < a_{dec}$, baryons are still coupled to photons maintaining pressure equilibrium ($P = P_R + P_B \approx P_R$) but the dominant density is $\rho = \rho_B + \rho_R \approx \rho_B$. So

$$v^2 \simeq \frac{P}{\rho} \approx \frac{P_R}{\rho_B} \propto a^{-1} \qquad a_{Br} < a < a_{dec} \tag{2.24}$$

For $a_{dec} < a$, baryons are decoupled from photons and there is no pressure equilibrium. The $v^2 \propto (P_B/\rho_B)$ now is just the velocity dispersion of Hydrogen-Helium gas. So

$$v^2 \propto a^{-2} \qquad a_{dec} < a \tag{2.25}$$

Notice that, at decoupling, v^2 drops from (P_R/ρ_B) to (P_B/ρ_B). Since $P_R \propto n_R kT$ while $P_B \propto n_B kT$ with $(n_R/n_B) \simeq 10^8 \gg 1$, this is a large drop in v^2 and -

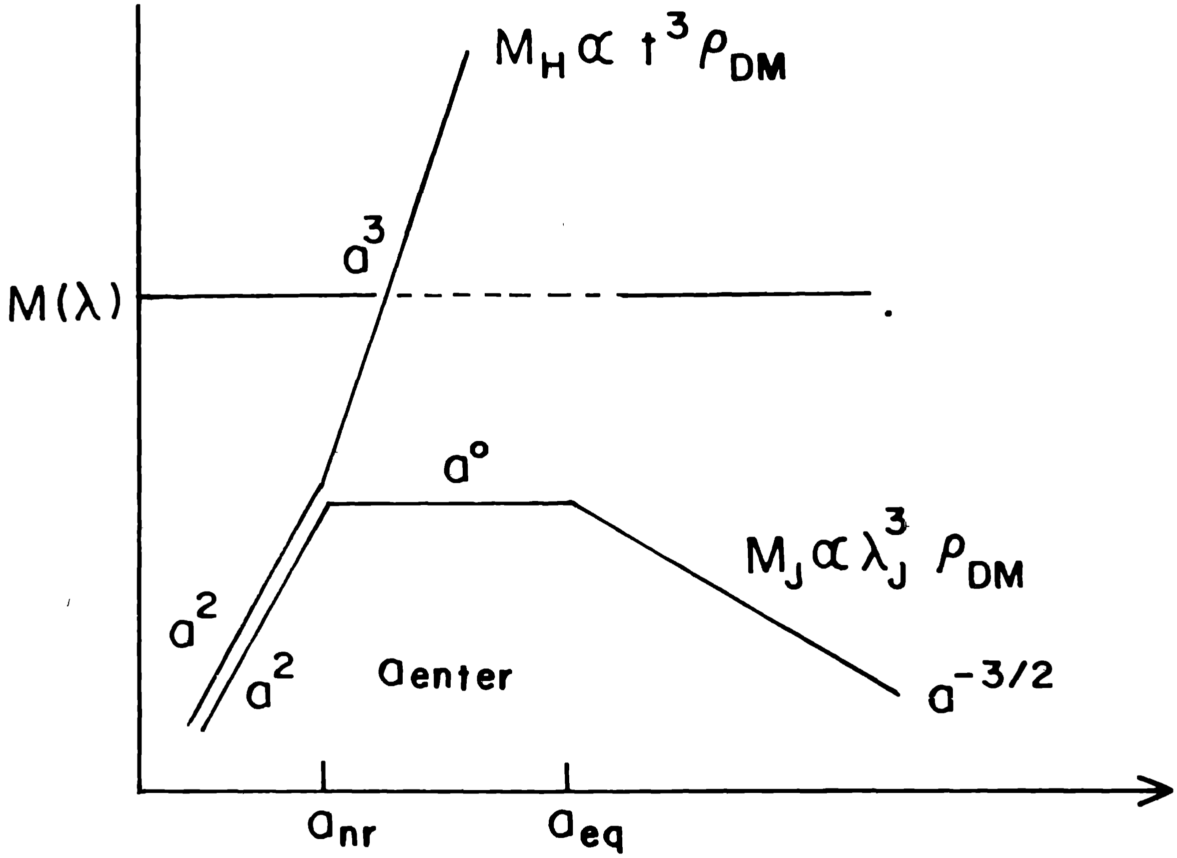

Figure 2.2. Jeans mass for dark matter.

consequently - in λ_J. λ_J and M_J are shown in fig. 2.3, 2.4. By repeating the previous analysis, we can see that,

$$\delta_B \equiv \left(\frac{\delta\rho}{\rho}\right)_B \propto \begin{cases} a^2 & a < a_{\rm enter} \\ constant & a_{\rm enter} < a < a_{dec} \\ a & a_{dec} < a \end{cases} \tag{2.26}$$

The last part deserves closer analysis. Notice that perturbations in DM can grow from a_{eq} onwards while perturbations in baryons grow only from a_{dec}. During the time from a_{eq} to a_{dec}, the perturbations in dark matter would have grown by a factor

$$\frac{a_{dec}}{a_{eq}} \cong 21\Omega h^2 \tag{2.27}$$

When the baryons decouple, their pertubration will feel the *perturbed* gravitational potential of dark matter and will be driven by it. (We may say that the baryons "fall into" the potential wells created by the DM). This implies that δ_B will grow *rapidly*

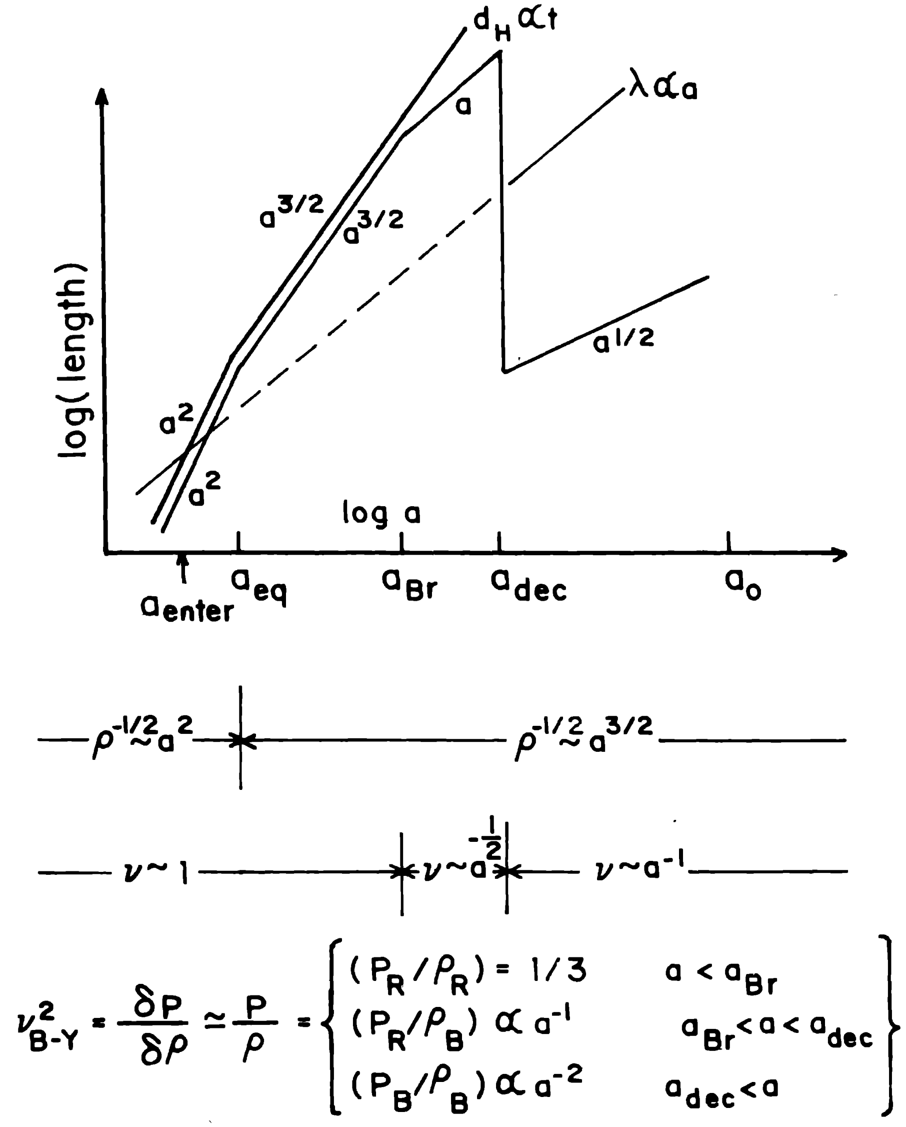
$d_H \propto t$
a
$\lambda \propto a$
$a^{3/2}$
$a^{3/2}$
$a^{1/2}$
a^2
a^2
log(length)
log a
a_{enter}
a_{eq}
a_{Br}
a_{dec}
a_o
$\rho^{-1/2} \sim a^2$
$\rho^{-1/2} \sim a^{3/2}$
$\nu \sim 1$
$\nu \sim a^{-\frac{1}{2}}$
$\nu \sim a^{-1}$
$\nu^2_{B\text{-}Y} = \frac{\delta P}{\delta \rho} \simeq \frac{P}{\rho} = \begin{cases} (P_R/\rho_R) = 1/3 & a < a_{Br} \\ (P_R/\rho_B) \propto a^{-1} & a_{Br} < a < a_{dec} \\ (P_B/\rho_B) \propto a^{-2} & a_{dec} < a \end{cases}$

Figure 2.3. Jeans length for baryons.

for a short time after a_{dec} and will equalise with the value of δ_{DM} ; after that, both δ_B and δ_{DM} will grow as a.

Throughout the above discussion, we have assumed that the background universe is a high density universe with Ω close to unity. If, instead, Ω is small (say $\Omega = 0.1$) then we have to worry about another additional complication. To see this, consider the evolution of the scale factor for a $k = -1$ universe. We can rewrite the Friedmann equation as

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi G}{3}\rho + \frac{1}{a^2} \tag{2.28}$$

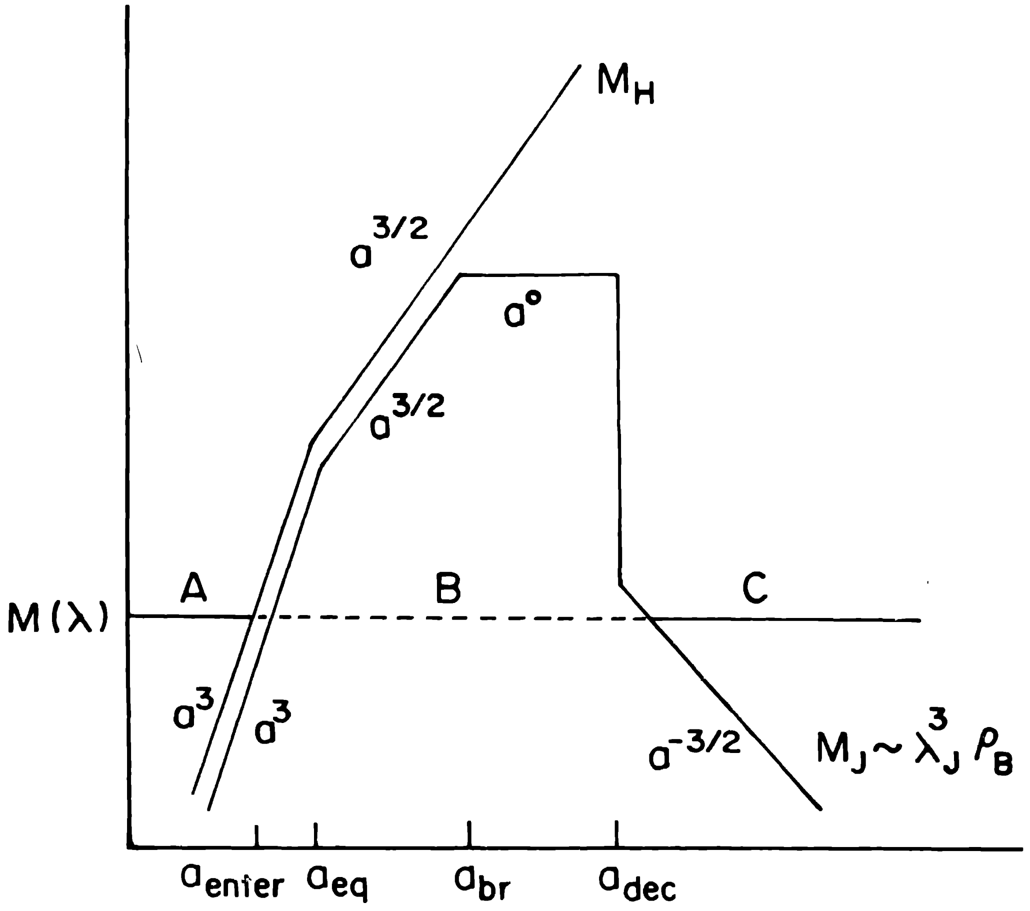

Figure 2.4. Jeans mass for baryons.

The matter density falls as a^{-3} while the second term on the right hand side, the curvature term, falls only as a^{-2}. So it is possible for the curvature to have dominated over matter at and after some epoch in the past. The two terms on the right will be equal in magnitude at a redshift z_c where

$$\frac{8\pi G}{3}\rho_c\Omega(1+z_c)^3 = \frac{1}{a_0^2}(1+z_c)^2 \tag{2.29}$$

or,

$$(1+z_c) = \frac{3}{8\pi G}\frac{1}{a_0^2}\cdot\frac{1}{\rho_c\Omega} = \frac{1}{H_0^2}\frac{1}{\Omega a_0^2} = \left(\frac{1}{\Omega}-1\right) \tag{2.30}$$

[In arriving at the last step we have used the relations $\rho_c = (3H_0^2/8\pi G)$ and $a_0^{-2} = H_0^2$ $(1-\Omega)$.]. Such a transition to curvature dominated universe could have occurred in the past ($z_c > 0$) only if $\Omega < 0.5$; for example, if $\Omega = 0.2$, $z_c \simeq 3$. If this transition occurs, then the growth of perturbations stops at z_c. Tnis suppression occurs for the same reason as the suppression in the radiation dominated phase: the expansion, dominated by a *smooth* a^{-2} term, is too rapid for the growth of perturbations: $t_{exp} < t_{\text{collapse}}$. Thus, in a low density ($\Omega < 0.5$) universe there is no growth of perturbation after $z = z_c$.

2.3. Aspects of general relativistic perturbation theory

We shall now take up the issue of perturbation growth in the general relativistic phase, i.e. when the perturbation scale is bigger than the Hubble radius. Though the actual derivation of the perturbation equation is complicated, the final result can be easily stated and understood physically. We will therefore, skip the derivation and discuss the final result. (The essential growth laws have, anyway, been derived by a different route in the last section. A full derivation in a particular gauge is given in Appendix 2.1 and 2.2).

Consider a perturbation, labeled by a wave number k, in the linear approximation. The second order differential equation governing the growth of the amplitude can be written as

$$\frac{d^2Q_A}{dt^2} + \alpha_A \left(\frac{\dot{a}}{a}\right) \frac{dQ_A}{dt} + \left(\frac{k^2 v_A^2}{a^2}\right) Q_A = \mu_A \left(\frac{\dot{a}}{a}\right)^2 Q_A \tag{2.31}$$

where the various quantities have the following physical meaning:

$$Q_A = \left\{ \begin{array}{l} \text{gauge invariant variable related to the} \\ \text{density contrast of the species labeled by A;} \\ \text{A could be baryons, dark matter or photons} \end{array} \right\} \tag{2.32}$$

$$v_A^2 = \left(\frac{\partial P_A}{\partial \rho_A}\right) = \left\{ \begin{array}{l} \text{`sound' speed at which pressure readjustments} \\ \text{can take place in the species} - \text{A} \end{array} \right\} \tag{2.33}$$

$$\alpha_A, \mu_A = \left\{ \begin{array}{l} \text{functions of } a\text{, which can be specified} \\ \text{in terms of the background matter.} \end{array} \right\} \tag{2.34}$$

Each of these quantities deserves comment.

We mentioned previously that, in general relativity, it is necessary to either choose a specific gauge or to deal with gauge invariant quantities. The dependent variable Q_A is a gauge invariant scalar related to the density contrast $(\delta\rho/\rho)_A \equiv \delta_A$ in the species labeled by A. In general, this quantity has no simple physical meaning; however, one can choose a coordinate system in such a way that

$$Q_A = (a^3\rho_A).\left(\frac{\delta\rho}{\rho}\right)_A = (a^3\rho_A)\delta_A \tag{2.35}$$

This is a convenient relation for interpreting Q_A.

In the adiabatic scenario, the velocity dispersion v_A decides the relation between perturbed pressure and perturbed density. Notice, however, that the label A can also include the relativistic component; in that case, $v_A \sim 1$, the speed of light.

Lastly, the variables μ_A, α_A are quantities which depend on the equation of state of the background matter. For the single component medium we are considering these turn out to be:

$$\mu_A = \frac{3}{2}(1 + w_A) \tag{2.36}$$

$$\alpha_A = 2(1 + \frac{3}{2}v_A^2) \tag{2.37}$$

where $w = (p/\rho)$ and $v_A^2 = (\dot{p}/\dot{\rho})$.

The specific form of these functions, of course, is crucial in deciding the behaviour of δ_A as a function of a. Notice that, if (P/ρ) is a constant, then these functions reduce to constants independent of time.

To understand the growth (or suppression) of Q_A's it is better to rewrite (2.31) in the following way:

$$\frac{d^2Q_A}{dt^2} = -\alpha_A H \frac{dQ_A}{dt} + \left(\mu_A H^2 - \frac{k^2 v_A^2}{a^2}\right) Q_A \tag{2.38}$$

where $H = (\dot{a}/a)$. The first term in the right hand side always dampens the growth; this is purely an effect of expansion. [The expansion rate $(\dot{a}/a)$, of course, is contributed by *all* matter in a multicomponent medium. If the dominant, smooth component is not the species A under consideration, then the damping due to expansion can suppress the growth; this is what happens in radiation dominated or curvature dominated phases].

The second term represents the conflict between pressure support and gravity. This term will act as a 'restoring force' and prevent growth if (ignoring the numerical factor μ_A)

$$a^2H^2 < k^2v_A^2 \tag{2.39}$$

Since $H^2 = (8\pi G\rho_{\text{dominant}}/3)$ and $k^2 = (2\pi/\lambda)^2$ we can write this as

$$\frac{(\lambda a)}{v} = \frac{\lambda_{\text{proper}}}{v} < \frac{1}{\sqrt{G\rho}}; \qquad \lambda < \lambda_J \equiv \frac{v}{\sqrt{G\rho}} \tag{2.40}$$

which is precisely the condition $t_{\text{pressure}} < t_{collapse}$ discussed before. In a single component medium, we have only two terms on the right hand side. The first term (expansion) always dampens the growth; the second term (pressure-gravity) may assist the growth or suppress it depending on the values of λ and λ_J. [In the multicomponent medium, Q_A will be driven by the species with largest value for (μQ); the restoring force, of course, comes from $k^2v_A^2$ term.]

Let us look at the solutions to this equation in some simple cases. To do this, it is better to choose a gauge such that $Q = (\rho a^3)\delta$. By substituting this relation, carrying out the differentiations, and using (1.16) , we can write down the equation satisfied by δ; we find that

$$\ddot{\delta} + [2 - 3(2w - v^2)]H\dot{\delta} - \frac{3H^2}{2}(1 - 6v^2 + 8w - 3w^2)\delta = -\frac{k^2}{a^2}v^2\delta \tag{2.41}$$

where $H = (\dot{a}/a)$, $w = (p/\rho)$ and $v^2 = (\dot{p}/\dot{\rho})$, and we have suppressed the subscript A. Since

$$\frac{d}{dt} = \dot{a}\frac{d}{da} = H(a)a\frac{d}{da} \tag{2.42}$$

we can now change the independent variable from t to a in the above equation, leading to

$$a^2\frac{d^2\delta}{da^2} + Aa\frac{d\delta}{da} + \left(B + \frac{k^2v^2}{H^2a^2}\right)\delta = 0 \tag{2.43}$$

where

$$A = \frac{3}{2}(1 - 5w + 2v^2); \qquad B = -\frac{3}{2}\left[1 - 6v^2 + 8w - 3w^2\right] \tag{2.44}$$

In the radiation dominated case $w = v^2 = (1/3)$, giving $A = 0, B = -2$; the equation becomes

$$a^2\frac{d^2\delta}{da^2} + \left(\frac{k^2v^2}{H^2a^2} - 2\right)\delta = 0 \tag{2.45}$$

while for matter dominated case $w = v^2 = 0$, giving $A = (3/2)$, $B = -(3/2)$; then the equation is

$$a^2\frac{d^2\delta}{da^2} + \frac{3}{2}a\frac{d\delta}{da} + \left(\frac{k^2v^2}{H^2a^2} - \frac{3}{2}\right)\delta = 0 \tag{2.46}$$

These equations have fairly simple solutions. Consider first the modes for which the quantity (k^2v^2/H^2a^2) is far less than unity. Then the equations become

$$\begin{aligned} &a^2\delta'' - 2\delta \approx 0 \qquad (\text{RD} - \text{phase}) \\ &a^2\delta'' + \frac{3}{2}a\delta' - \frac{3}{2}\delta \approx 0 \quad (\text{MD} - \text{phase}) \end{aligned} \tag{2.47}$$

The solutions can be written down by inspection; the growing modes are

$$\delta = \begin{cases} a^2 & (\text{RD} - \text{phase}) \\ a & (\text{MD} - \text{phase}) \end{cases} \tag{2.48}$$

These results were obtained earlier by a different method. The condition that $(k^2v^2/H^2a^2) \ll 1$ selects modes which are bigger than the horizon in the RD-phase and bigger than the Jean's length in the MD-phase.

In the opposite case, $(k^2v^2/H^2a^2) \gg 1$, the amplitudes do not grow but oscillate as a wave. In the RD-phase the equation becomes

$$\frac{d^2\delta}{da^2} + \frac{k^2v^2}{H^2a^4}\delta \cong 0 \tag{2.49}$$

Notice that $H^2 \propto \rho \propto a^{-4}$ in the RD-phase, making the coefficient of δ a constant. The solution is therefore

$$\delta = \exp i\left(\frac{kv}{Ha^2}\right)a = \exp i\left[\frac{d_H(a)}{\lambda_{phy}(a)}\right] \tag{2.50}$$

which oscillates rapidly because $d_H \gg \lambda_{phy}$. [We have set $v \approx 1$, which is valid in RD-phase].

In the MD-phase, we can rewrite (2.46)as

$$a\frac{d^2\delta}{da^2}+\frac{3}{2}\frac{d\delta}{da}+\frac{k^2v^2}{H^2a^3}\delta \approx 0 \tag{2.51}$$

Using the fact that H^2a^3 = constant and $v^2 \propto a^{-1}$ and changing the independent variable to $x \equiv a^{1/2}$, we get

$$\frac{d^2\delta}{dx^2}+\frac{2}{x}\frac{d\delta}{dx}+\frac{\omega^2\delta}{x^2}=0; \quad \omega^2 \equiv \frac{4k^2v^2}{H^2a^3} \tag{2.52}$$

where ω is a constant. This equation has the solution $\delta \propto x^n$ with $n \simeq [(-1/2) \pm i\omega]$ for $\omega \gg 1$. This means that

$$\delta = t^{-1/6}\exp\left(\pm\frac{i\omega}{3}\ln t\right) \tag{2.53}$$

which is a slow decay int he amplitude.

2.4. Aspects of Newtonian perturbation theory

If $\lambda \ll d_H$, then the perturbations can be analysed by Newtonian theory. The non-relativistic, Newtonian limit of (2.41)can be easily obtained by setting $w \approx 0; v \approx 0$ in that equation. Using further the fact that $H^2 = (8\pi G\rho/3)$ we get

$$\ddot{\delta}_A+\frac{2\dot{a}}{a}\dot{\delta}_A+\frac{k^2v_A^2}{a^2}\delta_A = 4\pi G\rho_A\delta_A \tag{2.54}$$

Since there are no gauge ambiguities, one can keep the density contrast δ as the dependent variable and t as the independent variable.

If there is more than one species populating the universe, then the right hand side of (2.54) contains contributions from all the perturbed species; the equation gets modified to

$$\ddot{\delta}_A+\frac{2\dot{a}}{a}\dot{\delta}_A+\frac{k^2v_A^2}{a^2}\delta_A = \sum_{all\,B} 4\pi G\rho_B\delta_B \tag{2.55}$$

The structure of (2.55) is quite similar to that of (2.31); there is one term giving dilution due to expansion $(2\dot{a}\dot{\delta}/a)$, one representing pressure support $(k^2v^2/a^2)\delta$ and the driving force due to the gravitational field of all perturbed matter $(4\pi G\sum\rho\delta)$. All the qualitative considerations mentioned in the previous two sections can be easily seen to hold in this case as well.

As an illustration of Newtonian limit of the perturbation theory, let us consider two situations mentioned in 2.5: (a) Suppression of δ_{DM}- growth in RD-phase and (b) Rapid growth of δ_B just after t_{dec}.

Consider a mode with $\lambda_J \ll \lambda \ll l_H$ in the RD-phase. Since $\lambda_J \ll \lambda$, we can ignore the pressure support. Further, in the right hand side of (2.55), we ignore δ_R

because $< \delta_R > \approx 0$ at subhorizon scales ($\lambda \ll ct$) due to rapid oscillations in (2.50). Then we get

$$\ddot{\delta}_{DM} + \frac{2\dot{a}}{a}\dot{\delta}_{DM} \cong 4\pi G\rho_{DM}\delta_{DM} \tag{2.56}$$

where the background universe is governed by the equation

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi G}{3}(\rho_R + \rho_{DM}) \tag{2.57}$$

Introducing the variable $x \equiv (a/a_{eq})$ and using (2.57) in (2.56), we can recast the equation in the form

$$2x(1+x)\frac{d^2\delta_{DM}}{dx^2} + (2+3x)\frac{d\delta_{DM}}{dx} = 3\delta_{DM}; \qquad x = \frac{a}{a_{eq}} \tag{2.58}$$

The growing solution to this equation can again be written down by inspection:

$$\delta_{DM} = 1 + \frac{3}{2}x \tag{2.59}$$

In other words $\delta_{DM} \approx$ constant for $a \ll a_{eq}$ (no growth in the RD-phase) and $\delta_{DM} \propto a$ for $a \gg a_{eq}$ (growth proportional to a in the MD-phase). Thus δ_{DM} does not grow in the RD phase even though $\lambda > \lambda_J$.

This statement, however, needs to be qualified. The above equations - and infact all perturbation equations considered so far - are second order differential equations having two linearly independent solutions. A general solution can be found only when two initial conditions are given. We have been avoiding this problem so far by just choosing the "growing" solution as *the* solution to the equation. (This procedure is justified as long as we are not interested in any transient phenomena.). In this particular case a proper analysis will require matching both δ and $\dot{\delta}$ at the instant $t = t_{\text{enter}}$. This would force us to choose a linear combination of solutions during the epoch $t_{\text{enter}} < t < t_{eq}$ rather than the purely growing mode given above. It turns out that such a more complicated analysis changes the result only slightly: The modes can grow very weakly - in fact, logarithmically - during the period $t_{\text{enter}} < t < t_{eq}$. (see e.g. Peebles 1980).

Finally consider the perturbation in the baryon-dark matter system just after decoupling. This is governed by the equations

$$\ddot{\delta}_{DM} + \frac{2\dot{a}}{a}\dot{\delta}_{DM} = 4\pi G(\rho_B\delta_B + \rho_{DM}\delta_{DM}) \approx 4\pi G\rho_{DM}\delta_{DM} \tag{2.60}$$

$$\ddot{\delta}_B + \frac{2\dot{a}}{a}\dot{\delta}_B = 4\pi G(\rho_B\delta_B + \rho_{DM}\delta_{DM}) \approx 4\pi G\rho_{DM}\delta_{DM} \tag{2.61}$$

where we have used the fact that, just after a_{dec}, $\rho_{DM}\ \delta_{DM} \gg \rho_B\delta_B$. Equation (2.60) represents the growth of perturbations in the DM- component. From the previous analysis we know that it has the solution

$$\delta_{DM} = (\text{constant})a \equiv \alpha a \tag{2.62}$$

(This can also be verified directly by using the facts that $a \propto t^{(2/3)}; \rho \propto t^{-2}$). Substituting this in (2.61), we can rewrite it as

$$a^{3/2}\frac{d}{da}\left(\frac{1}{a^{1/2}}\frac{d\delta_B}{da}\right)+2\frac{d\delta_B}{da}=\frac{3}{2}\alpha \tag{2.63}$$

[where we have again used the relation $a \propto t^{2/3}$]. This has the growing solution

$$\delta_B \propto (a-\alpha)=\delta_{DM}(a)\left(1-\frac{\alpha}{a}\right) \tag{2.64}$$

where α is a constant. This solution shows that $\delta_B \longrightarrow \delta_{DM}$ for $a \gg \alpha$, even if $\delta_B \approx 0$ at some $a=\alpha=a_{dec}$ (say). In other words, baryonic perturbations "catch up" with DM perturbations after the decoupling.

2.5. Free streaming in collisionless dark matter

In the discussion of perturbations so far, we have taken the matter content of the universe to be an ideal fluid. This approximation breaks down for wavelengths smaller than a particular critical value. At smaller wavelengths, all power is drained away by certain dissipative processes which we will now discuss.

The physical origin of dissipation is different in baryons and dark matter. In baryons, it arises due to the coupling between radiation and matter, which we will study in section 2.9. In collisionless dark matter, the dissipation occurs due to a process called 'free streaming' which we will discuss in this section.

If the dark matter is made of weakly interacting particles, then they do not feel each other's presence via collisions (unlike an ordinary gas where collisions are significant). Each dark matter particle, therefore, moves along a geodesic in the space-time. Perturbations modify the space-time metric and - consequently - the geodesic orbits. One can study the response of dark matter particles to such perturbations by invoking an "effective pressure" and treating darkmatter as an ideal fluid. Such an approximation is valid only for sufficiently large wavelengths. At small scales, the "free", geodesic motion of the particles will wipe out any structure, because the particles can freely propagate from an overdense region to an underdense region equalising the densities. (Bond *et al.* 1980, Peebles 1982, Bond & Szalay 1983).

Let $l_{FS}(t)$ be the proper distance which a darkmatter particle can travel in time t in the background space-time; and let $\lambda(t)$ be the proper wavelength of a perturbation at time t. Then all modes, for which $l_{FS}(t) > \lambda(t)$, will suffer due to free streaming. We know that $\lambda(t) \propto a(t)$; so we only need to compute $l_{FS}(t)$ to compare the two. This can be done as follows: The proper distance traveled by a particle in time t can be written as

$$l_{FS}(t)=a(t)\int_0^t \frac{v(t')}{a(t')}dt' \tag{2.65}$$

[Since $adx = vdt$ defines the proper velocity $v(t)$]. During $0 < t < t_{nr}$, the dark matter particles are relativistic and $v \simeq 1$; since $a(t) \propto t^{1/2}$, this gives

$$l_{FS}(t)=a\int_0^t \frac{dt'}{a_{nr}}\cdot\left(\frac{t_{nr}}{t'}\right)^{1/2}=a(t)\left[\frac{2t_{nr}^{1/2}t^{1/2}}{a_{nr}}\right]=2t \propto a^2 \quad (\text{for}\, t<t_{nr}) \tag{2.66}$$

For $t < t_{nr} < t_{eq}$, $v \propto a^{-1}$ and we get

$$\begin{aligned} l_{FS}(t) &= \left[\frac{l_{FS}(t_{nr})}{a_{nr}} + \int_{t_{nr}}^{t} \frac{dt'}{a(t')}.\frac{a_{nr}}{a(t')}\right] a(t) \\ &= \left[\frac{2t_{nr}}{a_{nr}} + \frac{2t_{nr}}{a_{nr}} \ln\frac{a}{a_{nr}}\right] a = \frac{2t_{nr}a}{a_{nr}}\left[1 + \ln\frac{a}{a_{nr}}\right] (t_{nr}, t < t_{eq}) \end{aligned} \tag{2.67}$$

For $t > t_{eq}$, $a(t) \propto t^{2/3}$. So

$$\begin{aligned} l_{FS}(t) &= \left[\frac{l_{FS}(t_{eq})}{a_{eq}} + \int_{t_{eq}}^{t} \frac{a_{nr}}{a_{eq}^2}\left(\frac{t_{eq}}{t'}\right)^{4/3} dt'\right] a(t) \\ &= \left[\frac{2t_{nr}}{a_{nr}}\left(1 + \ln\frac{a_{eq}}{a_{nr}}\right) + \frac{3t_{nr}}{a_{nr}}\left(1 - \frac{a_{eq}^{1/2}}{a^{1/2}}\right)\right] a(t) \end{aligned} \tag{2.68}$$

Thus we find that

$$\frac{l_{FS}(t)}{a(t)} = \begin{cases} (2t_{nr}/a_{nr}^2)a = (2t/a) & t < t_{nr} \\ (2t_{nr}/a_{nr})\left[1 + \ln(a/a_{nr})\right] & t_{nr} < t < t_{eq} \\ (2t_{nr}/a_{nr})\left[\frac{5}{2} + \ln(a_{eq}/a_{nr})\right] & t_{eq} \ll t \end{cases} \tag{2.69}$$

We have to now determine the range of wavelengths for which the condition $\lambda(t) \le l_{FS}(t)$ is satisfied; or equivalently, the range for which $(\lambda/a) \le (l_{FS}(t)/a)$. At $t < t_{nr}$, $l_{FS} \approx d_H(t)$ the Hubble radius, $(\lambda/a) \le (l_{FS}(t)/a)$; during $(t_{nr} < t < t_{eq})$, (l_{FS}/a) grows logarithmically; for $t > t_{eq}$, (l_{FS}/a) grows still more slowly and saturates at the value

$$\lambda_{FS} \equiv l_{FS}(t_0) = \left(\frac{a_0}{a_{nr}}\right)(2t_{nr})\left(\frac{5}{2} + \ln\frac{a_{eq}}{a_{nr}}\right) \tag{2.70}$$

Since this is the largest value of l_{FS}, all proper wavelengths $\lambda > \lambda_{FS}$ will survive the process of freestreaming.

A simple minded derivation of the above result for λ_{FS} is as follows: When the dark matter is relativistic, it travels with the speed of light and covers a proper distance of $(2t_{nr})$ by $t = t_{nr}$. This distance corresponds today to the length $(2t_{nr})$ (a_0/a_{nr}), which is identified as the freestreaming scale. Notice that this analysis gives the correct result upto a numerical factor.

To obtain numerical estimates, we need to identify the epoch t_{nr}. We may take this to be the time at which $T_{DM} \approx (m/3)$ where T_{DM} is the temperature of the DM-species and m is the mass of the dark matter particle. The temperature T_{DM}, in general, will not be the same as the radiation temperature $T_R \equiv T$ because the dark matter could have decoupled early. If n_{DM} is the number density of DM- particles then the quantity

$$\left(\frac{T_{DM}}{T}\right)^3 = \frac{n_{DM}}{n_\gamma} \tag{2.71}$$

is conserved during the expansion. Further,

$$\Omega_{DM} = \left(\frac{n_{DM}m}{\rho_c}\right)_{now} = \frac{mn_\gamma}{\rho_c}.\left(\frac{n_{DM}}{n_\gamma}\right) \simeq 30\left(\frac{m}{1\text{keV}}\right)\left(\frac{n_{DM}}{n_\gamma}\right)h^{-2} \tag{2.72}$$

giving

$$\left(\frac{T_{DM}}{T}\right)^3 = \left(\frac{n_{DM}}{n_\gamma}\right) \cong \left(\frac{\Omega_{DM} h^2}{30}\right) \left(\frac{m}{1\mathrm{keV}}\right)^{-1} \tag{2.73}$$

Using this relation and the numerical values (which can be easily derived from the expressions derived in Part 1).

$$\begin{aligned} \frac{a_{nr}}{a_0} &= 7 \times 10^{-7} \left(\frac{m}{1\mathrm{keV}}\right)^{-1} \left(\frac{T_{DM}}{T}\right) \\ t_{nr} &= 1.2 \times 10^7 \left(\frac{m}{1\mathrm{keV}}\right)^{-2} \left(\frac{T_{DM}}{T}\right)^2 \; sec. \\ \left(\frac{a_{eq}}{a_{nr}}\right) &= \left(\frac{m}{17\mathrm{keV}}\right) (\Omega h^2)^{-1} \left(\frac{T}{T_{DM}}\right) \end{aligned} \tag{2.74}$$

one finds that

$$\lambda_{FS} \simeq 40\mathrm{Mpc}\,(\Omega_{DM} h^2)^{-1} \left(\frac{T_{DM}}{T_R}\right)^4 = 0.5\mathrm{Mpc}\,(\Omega_{DM} h^2)^{1/3} \left(\frac{m}{1\mathrm{keV}}\right)^{-4/3} \tag{2.75}$$

This result is of extreme importance. It shows that the length scale below which perturbations will be wiped out, λ_{FS}, is essentially decided by the mass m and temperature T_{DM} of the dark matter. If, for example, the dark matter is made of neutrinos with $m \approx 30$ev, $(T_{DM}/T_R) \simeq 0.71$ and $\Omega_\nu h^2 \approx (m_\nu/91\mathrm{ev})$ then

$$\lambda_{FS} \simeq 28\mathrm{Mpc} \left(\frac{m_\nu}{30\mathrm{ev}}\right)^{-1} \tag{2.76}$$

This contains a mass of

$$M_{FS} \simeq 4 \times 10^{15} \left(\frac{m_\nu}{30\mathrm{ev}}\right)^{-2} M_\odot \tag{2.77}$$

Thus, in a universe with neutrino as dark matter, perturbations at all lower mass scales ($M < M_{FS}$) will be wiped out; there will be very little small scale power.

If, one the other hand, the DM-particle is much heavier with, say, $m \approx$ 1keV, $\Omega_{DM} h^2 \approx 1$, then

$$\lambda_{FS} \approx 0.5\mathrm{Mpc} \left(\frac{m}{1\mathrm{keV}}\right)^{-4/3} \tag{2.78}$$

containing only a mass of

$$M_{FS} \approx 6 \times 10^9 M_\odot. \tag{2.79}$$

In this case, power at scales above $10^9 M_\odot$ or so will survive dissipation due to free streaming. In general, a heavier dark matter candidate will let the power survive at smaller scales. The above discussion is summarised in fig. 2.5.

2.6. Collisional damping in the photon-baryon system

The damping of perturbations in the photon-baryon plasma occurs for a different - and somewhat simpler - reason. At $t \ll t_{dec}$, photons and baryons are very tightly

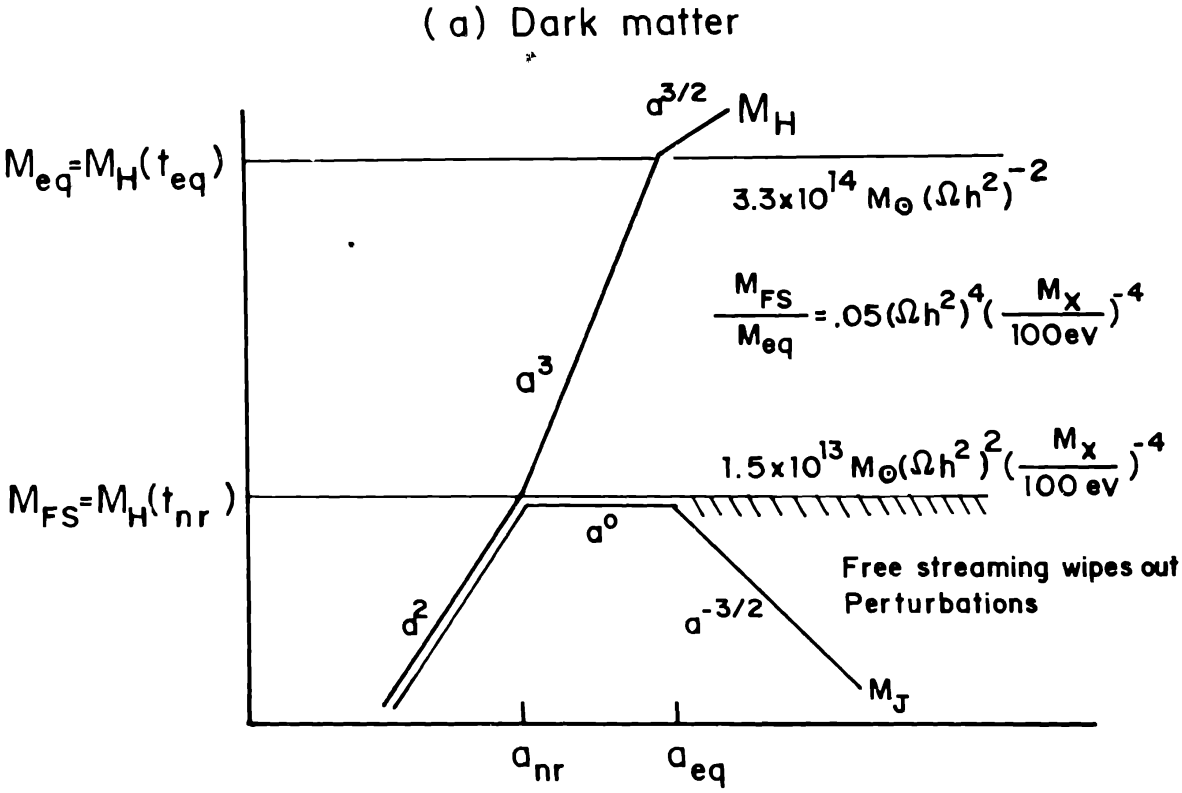
(a) Dark matter
$a^{3/2}$
M_H
$M_{eq}=M_H(t_{eq})$
$3.3\times10^{14}\,M_\odot(\Omega h^2)^{-2}$
$\frac{M_{FS}}{M_{eq}}=.05(\Omega h^2)^4(\frac{M_X}{100ev})^{-4}$
a^3
$1.5\times10^{13}\,M_\odot(\Omega h^2)^2(\frac{M_X}{100\,ev})^{-4}$
$M_{FS}=M_H(t_{nr})$
a^0
Free streaming wipes out Perturbations
a^2
$a^{-3/2}$
M_J
a_{nr}
a_{eq}

Figure 2.5. Free streaming of dark matter.

coupled due to Thomson scattering. The proper length corresponding to the photon mean free path at some time t is

$$l(t) = \frac{1}{X_e n_e \sigma} \simeq 1.3 \times 10^{29} cm X_e^{-1} (1+z)^{-3} (\Omega_B h^2)^{-1} \tag{2.80}$$

where X_e is the electron ionisation fraction. For wavelengths $\lambda \lesssim l$, the photon streaming will clearly damp any perturbation. But, actually, the damping effect is felt at even larger scales because of the following reason. (Silk 1968).

Consider a time interval Δt in which a photon suffers $N = (\Delta t / l(t))$ collisions. Between each collisions it travels a proper distance $l(t)$, or - equivalently - a coordinate distance $[l(t)/a(t)]$. Because of this random walk, it acquires a mean-square - coordinate displacement:

$$(\Delta x)^2 = N \left(\frac{l}{a}\right)^2 = \frac{\Delta t}{l(t)} \frac{l^2}{a^2} = \frac{\Delta t}{a^2} l(t) \tag{2.81}$$

The total mean-square coordinate displacement of the photon before decoupling is

$$x^2 \equiv \int_0^{t_{dec}} \frac{dt}{a^2(t)} l(t) = \frac{3}{5} \frac{t_{dec} l(t_{dec})}{a^2(t_{dec})} \tag{2.82}$$

which corresponds to the proper distance

$$l_s^2 = a^2(t_{dec}) x^2 = \frac{3}{5} t_{dec} l(t_{dec}) \simeq (3.5\mathrm{Mpc})^2 \left(\frac{\Omega}{\Omega_B}\right)^{1/2} (\Omega h^2)^{-\frac{3}{4}} \tag{2.83}$$

If we assume that baryons are tightly coupled to photons before t_{dec}, it follows that baryons will be dragged along with photons. Then all perturbations at wave lengths $\lambda < l_s$ will be wiped out. This wavelength corresponds to a mass

$$M_S \cong 6.2 \times 10^{12} M_{\odot} \left(\frac{\Omega}{\Omega_B}\right)^{3/2} (\Omega h^2)^{-5/4} \tag{2.84}$$

No baryonic perturbations carrying mass below M_S survives this damping process. (see fig. 2.6)

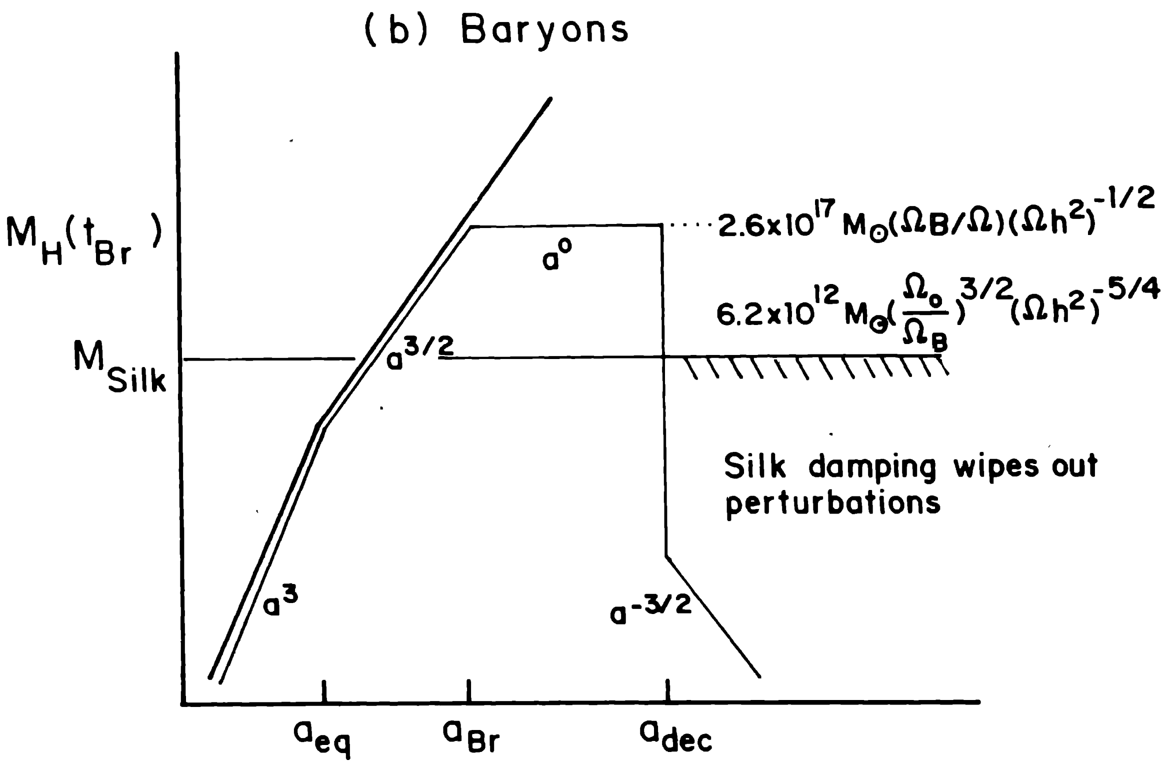
(b) Baryons
$M_H(t_{Br})$
M_{Silk}
a^0
$a^{3/2}$
a^3
$a^{-3/2}$
$2.6\times10^{17} M_{\odot}(\Omega B/\Omega)(\Omega h^2)^{-1/2}$
$6.2\times10^{12} M_{\odot}(\frac{\Omega_o}{\Omega_B})^{3/2}(\Omega h^2)^{-5/4}$
Silk damping wipes out perturbations
a_{eq}
a_{Br}
a_{dec}

Figure 2.6. Silk damping for baryons.

It is easy to determine the scale dependence of l_s and M_S. For $t < t_{eq}$, $l \sim a^3$, $t \sim a^2$ giving $l_S \sim a^{5/2}$ and $M_S \sim a^{9/2}$; for $t > t_{eq}$, $l \sim a^3$, $t \sim a^{3/2}$ giving $l_S \sim a^{9/4}$ and $M_S \sim a^{15/4}$.

Notice that this process occurs mainly around $t \simeq t_{dec}$. At $t \ll t_{dec}$, l_S is very small while for $t > t_{dec}$ baryons do not follow the photons. It should also be clear that $l_S \gg l(t_{dec})$ because $t_{dec} \gg l(t_{dec})$; thus the effect is felt at scales far larger than the mean free path.

The mean free path of the *electrons* $l_{elec} \simeq (n_\gamma \sigma)^{-1}$ is much smaller than the $l_{\text{photons}} \simeq (n_e \sigma)^{-1}$ because $n_\gamma \gg n_e$. Hence any damping due to the random walk of electrons will be subdominant to the effects considered above.

2.7. The processed final spectrum

We have now assembled all the ingredients to evolve an initial power spectrum at some $t = t_i \ll t_{eq}$ to a final value at $t > t_{dec}$. As usual, it is better to consider dark matter and baryons separately; we will discuss DM-spectrum in this section and the baryonic spectrum in the next.

Let $\delta_\lambda(t_i)$ denote the amplitude of DM-perturbation corresponding to some wavelength λ at the initial instant t_i. To each λ, we can associate a wavenumber k or mass M; accordingly, we may label the perturbation as $\delta_M(t)$ or $\delta_k(t)$,as well, with the scalings $M \sim \lambda^3$; $k \sim \lambda^{-1}$. We are interested in $\delta_\lambda(t)$ at some $t > t_{dec}$.

To begin with, free streaming will wipe out power at all scales smaller that λ_{FS} corresponding to a mass M_{FS}. So we have the first result:

$$\delta_M(t) \approx 0 \qquad (\text{for } M < M_{FS}; \lambda < \lambda_{FS}). \tag{2.85}$$

Consider next the range of wave lengths $\lambda_{FS} < \lambda < \lambda_{eq}$. These modes enter the horizon in the radiation dominated phase; however, they do not grow until $t = t_{eq}$. Therefore, for these wavelengths $\delta_\lambda(t_{eq}) = \delta_\lambda(t_{\text{enter}})$. After matter domination, they grow as the scale factor a. Therefore

$$\delta_M(t) = \delta_M(t_{\text{enter}}) \left(\frac{a}{a_{eq}}\right) \qquad (\text{for } M_{FS} < M < M_{eq}; a > a_{eq}) \tag{2.86}$$

Consider next the modes with $\lambda_{eq} < \lambda < \lambda_H$ where $\lambda_H \equiv H^{-1}(t)$ is the Hubble radius at the time t when we are studying the spectrum. These modes enter the Hubble radius in the matter dominated phase and grow proportional to a after entering the matter dominated phase. So

$$\delta_M(t) = \delta_M(t_{\text{enter}}). \left(\frac{a}{a_{\text{enter}}}\right) \qquad (\text{for } M_{eq} < M < M_H) \tag{2.87}$$

This may be rewritten as

$$\delta_M(t) = \delta_M(t_{\text{enter}}) \left(\frac{a_{eq}}{a_{\text{enter}}}\right) \left(\frac{a}{a_{eq}}\right) \tag{2.88}$$

But notice that, since $t_{\rm enter}$ is fixed by the condition $\lambda a_{\rm enter} \sim t_{\rm enter} \sim \lambda t_{\rm enter}^{2/3}$ it follows that $t_{\rm enter} \propto \lambda^3$. Further $(a_{eq}/a_{\rm enter}) = (t_{eq}/t_{\rm enter})^{2/3}$ giving

$$\left(\frac{a_{eq}}{a_{\rm enter}}\right) = \left(\frac{\lambda_{eq}}{\lambda}\right)^2 = \left(\frac{M_{eq}}{M}\right)^{2/3} \tag{2.89}$$

Substituting (2.89)in (2.88), we get

$$\begin{aligned} \delta_M(t) &= \delta_M(t_{\rm enter}) \left(\frac{\lambda_{eq}}{\lambda}\right)^2 \left(\frac{a}{a_{eq}}\right) \\ &= \delta_M(t_{\rm enter}) \left(\frac{M_{eq}}{M}\right)^{2/3} \left(\frac{a}{a_{eq}}\right) \end{aligned} \tag{2.90}$$

comparing (2.86)and (2.90)we see that, the mode which enters the Hubble radius after t_{eq} has its power decreased by a factor $M^{-2/3}$.

Finally, consider modes with $\lambda > \lambda_H$ which are still outside the Hubble radius at t and will enter the Hubble radius at sometime $t_{\rm enter} > t$. During the time $(t, t_{\rm enter})$, they will grow by a factor $(a_{\rm enter}/a)$. Thus

$$\delta_\lambda(t_{\rm enter}) = \delta_\lambda \left(\frac{a_{\rm enter}}{a}\right) \tag{2.91}$$

or

$$\delta_\lambda(t) = \delta_\lambda(t_{\rm enter}) \left(\frac{a}{a_{\rm enter}}\right) = \delta_M(t_{\rm enter}) \left(\frac{M_{eq}}{M}\right)^{2/3} \left(\frac{a}{a_{eq}}\right) (\lambda > \lambda_H) \tag{2.92}$$

[The last equality follows from the previous analysis]. Thus the behaviour of the modes is the same for all $\lambda > \lambda_{eq}$. We can state the final result as follows: (see fig. 2.7)

$$\delta_\lambda(t) = \begin{cases} 0 & \lambda < \lambda_{FS} \\ \delta_\lambda(t_{\rm enter}) \left(\frac{a}{a_{eq}}\right) & \lambda_{FS} < \lambda < \lambda_{eq} \\ \delta_\lambda(t_{\rm enter}) \left(\frac{a}{a_{eq}}\right) \left(\frac{\lambda_{eq}}{\lambda}\right)^2 & \lambda_{eq} < \lambda \end{cases}$$

or, equivalently

$$\delta_M(t) = \begin{cases} 0 & M < M_{FS} \\ \delta_M(t_{\rm enter}) \left(\frac{a}{a_{eq}}\right) & M_{FS} < M < M_{eq} \\ \delta_M(t_{\rm enter}) \left(\frac{a}{a_{eq}}\right) \left(\frac{M_{eq}}{M}\right)^{2/3} & M_{eq} < M \end{cases} \tag{2.93}$$

Thus the spectrum at late times is completely fixed by the amplitude of the spectrum when it enters the Hubble radius. Of course, we can relate $\delta(t_{\rm enter})$ to $\delta(t_i)$ for some $t_i < t_{\rm enter}$; but it is much more convenient to use $\delta_\lambda(t_{\rm enter})$ to characterise the fluctuations.

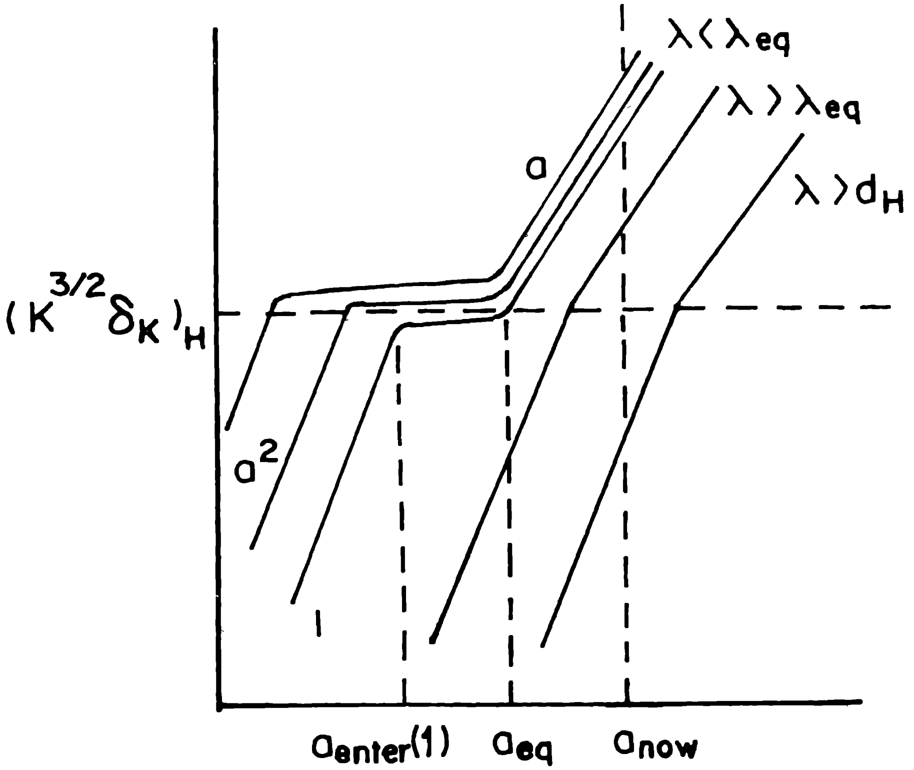

Figure 2.7. Growth of perturbation amplitudes for different modes.

2.8. Harrison-Zeldovich spectrum

The specification of $\delta_\lambda(t_{\rm enter})$ - or, equivalently, the specification of $\delta_\lambda(t_i)$ at some t_i- is a fundamental, unsolved, problem in cosmology. Any complete theory for structure formation must specify this function based on some physical considerations. In the absence of such a theory, we will have to make some reasonable assumption for this quantity, and compare the results with observations.

Notice that the symbol $\delta_\lambda(t_{\rm enter})$ actually stands for the function $\delta(\lambda, t)$ evaluated at $t = t_{\rm enter}\ (\lambda)$. Thus $\delta_\lambda(t_{\rm enter}) \equiv \delta(\lambda, t_{\rm enter}(\lambda)) = F(\lambda)$, some function of λ . The simplest choice for this function is a powerlaw which is examined in the literature in great detail. In this model, we take

$$\delta_\lambda(t_{\rm enter}) = A\lambda^\alpha \propto k^{-\alpha} \propto M^{\alpha/3} \tag{2.94}$$

It immediately follows that

$$\delta_\lambda(t) = \delta_M(t) \propto \begin{cases} 0 & (\lambda < \lambda_{FS}; M < M_{FS}) \\ \lambda^\alpha \propto M^{\alpha/3} & (\lambda_{FS} < \lambda < \lambda_{eq}; M_{FS} < M < M_{eq}) \\ \lambda^{\alpha-2} \propto M^{\alpha/3-\frac{3}{2}} & (\lambda_{eq} < \lambda; M_{eq} < M) \end{cases} \tag{2.95}$$

We will see in the later sections that the quantity which is physically relevant is the "power-per-octave" in the perturbations:

$$k^3|\delta_k(t)|^2 \propto M^{-1}|\delta_M(t)|^2 \propto \begin{cases} 0 & (M < M_{FS}) \\ M^{2\alpha/3-1} & (M_{FS} < M < M_{eq}) \\ M^{2\alpha/3-\frac{7}{3}} & (M_{eq} < M) \end{cases} \tag{2.96}$$

At the time of entering the Hubble radius, this quantity has the dependence (use (2.94))

$$(k^3|\delta_k|^2)_{t=t_{\rm enter}} \propto k^{3-2\alpha} \propto M^{2\alpha/3-1} \tag{2.97}$$

This expression shows that the value $\alpha = (3/2)$ is somewhat "special". If $\alpha > (3/2)$, then $k^3|\delta_k|^2$ will have more power at large M (at the time of entering the Hubble radius); if $\alpha < (3/2)$, then most of the power will be concentrated on small scales. For the special value of $\alpha = (3/2)$, neither small or large scales dominate. Such a spectrum (called scale-invariant spectrum or 'Harrison-Zeldovich spectrum') (Harrison 1970 ; Zeldovich 1972) is indeed predicted by inflationary models. It is, therefore, usual to take $\alpha = \frac{3}{2}$. In such a case

$$k^3|\delta_k(t)|^2 = \begin{cases} 0 & (M < M_{FS}) \\ constant & (M_{FS} < M < M_{eq}) \\ M^{-\frac{4}{3}} & (M_{eq} < M). \end{cases} \tag{2.98}$$

Most of the power is in the region $M_{FS} < M < M_{eq}$.

We have parameterised the spectrum by the form of $\delta(k,t)$ at $t = t_{enter}(k)$. One can also specify the same function by the power law $|\delta(k,t)|^2 \propto k^n$ *at a given time* t. It is easy to verify that the index n is related to α by $n = (4 - 2\alpha)$. The scale invariant (Harrison - Zeldovich) spectrum $\alpha = (3/2)$ corresponds to a value of $n = 1$.

The actual shape of the spectrum depends crucially on the ratio

$$\frac{M_{FS}}{M_{eq}} = 0.05(\Omega h^2)^4 \left(\frac{m}{100\text{ev}}\right)^{-4} \tag{2.99}$$

If neutrinos with $m \approx 30$ev constitute the dark matter then $(M_{FS}/M_{eq}) \approx 4(\Omega h^2)^4$ which is around the range of unity. Thus the spectrum will have a relatively sharp peak around M_{FS}. If, on the other hand, the dark matter particle is heavier (say 1 MeV or so) and makes $M_{FS} \ll M_{eq}$, then the spectrum will be relatively flat between M_{FS} and M_{eq}. (see fig. 2.8).

In this context, it is important to point out an extra complication: The flatness of the spectrum between M_{FS} and M_{eq} is a direct consequence of our assumption that the modes which enter Hubble radius before M_{eq} start growing only after t_{eq}. As we mentioned before, this result is not strictly true; there is a small growth in the interval $t_{\rm enter} < t < t_{eq}$. Because of this reason the spectrum will not be completely flat for $M_{FS} < M < M_{eq}$ but will be sloping gently downwards. Thus the CDM spectrum will have : (i) maximum power around M_{FS} (ii) a slow, gently sloping spectrum from M_{FS} to M_{eq} and (iii) a steep $M^{-\frac{2}{3}}$ fall after M_{eq}.

As time goes on, this spectrum just grows in proportion to $a(t)$. Once δ becomes of order unity at some wavelength, the linear perturbation theory fails around that

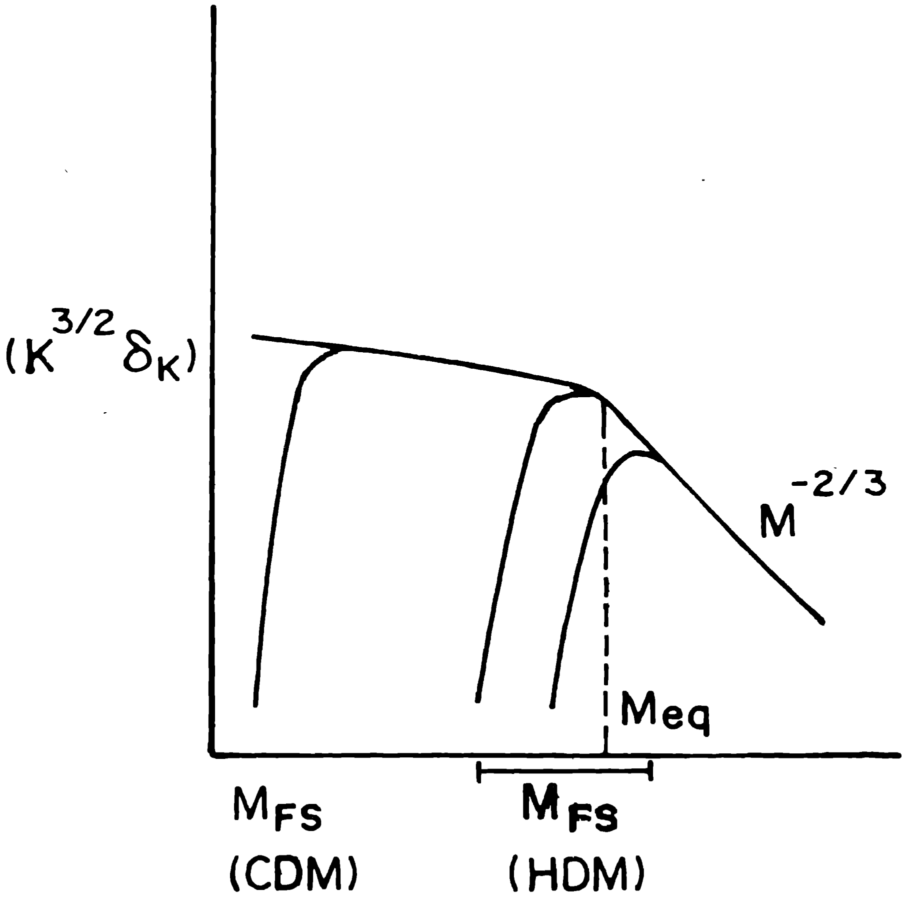

Figure 2.8. The processed spectrum.

wavelength. For HDM, the scale with mass $M \approx M_{FS}$ reaches nonlinearity first. Further evolution, involving pancaking, fragmentation etc will generate power at smaller scales. For CDM, the smallest scales $M \approx M_{FS}$ reach non- linearity first; however, since the spectrum is relatively flat for $M_{FS} < M < M_{eq}$, there is lot of "cross talk" between the various scales. In other words, larger and larger scales go nonlinear fairly quickly and "fall on top of" smaller scales. These non-linear proceses will be discussed in a later section.

The situation for baryons is somewhat analogous. To begin with, the collisional damping wipes out all power in the scales $M < M_S$, so that we only need to consider $M > M_S$. These perturbations too do not grow until t_{dec}; i.e.

$$\delta_M(t) = \delta_M(t_{\text{enter}}) \quad (\text{for } t < t_{dec}) \tag{2.100}$$

Just after decoupling, the baryonic perturbations start growing. At this epoch, they are driven by the dark matter perturbations which would have been growing since t_{eq}. Therefore

$$\delta_B(\lambda, t) \approx \delta_{DM}(\lambda, t) \quad (\text{for } t \gtrsim t_{dec}). \tag{2.101}$$

This analysis, of course, works for all λ at which $\delta_{DM}(\lambda, t)$ is significant. In CDM, this range is fairly wide - say, from $10^8 M_\odot$ to $10^{14} M_\odot$. Baryonic perturbations are now *generated* at all these scales even if they were originally absent; i.e., even at scales below $M_S \approx 10^{12} M_\odot$ in which Silk damping wipes out the power, this process regenerates the power. In HDM, the situation is somewhat different: The δ_{DM} (λ, t) is concentrated on a narrow band around $M_{FS} \approx 10^{14} M_\odot$ and the above process only generates baryonic power in that region. At $M < M_S \approx 10^{12} M_\odot$, all power is lost due to collisional damping. In the window ($10^{12} M_\odot$ to $10^{13} M_\odot$), there can be some small amount of baryonic power surviving from the early epochs; but they are not enhanced by the dark matter. Thus in HDM scenario, most of the baryonic and dark matter power is concentrated around $M \approx M_{FS} \approx 10^{14} M_\odot$ for $t \gg t_{dec}$.

Further evolution of baryons is complicated by the fact that they can radiate and cool. This will be discussed in detail in Part 4.

2.9. Gaussian random fields

The entire approach described so far is based on the separation $\rho(\mathbf{x}, t) = \overline{\rho}(t)$ $[1 + \delta(\mathbf{x}, t)]$ of the density into a mean density and the fluctuation. In this equation $\mathbf{x}$ stands for a particular location in space. It will be rather preposterous to expect that *any* theory will be able to predict the value of ρ - or, equivalently the value of δ - at some specific point $\mathbf{x}$ in the universe. Neither will observations determine $\rho(\mathbf{x}, t)$ at any single, *specific* location. The question, therefore, arises as to which essential properties of $\delta(\mathbf{x}, t)$ are to be fixed by the theory - and are directly relevant to observation. To attack this question it is more convenient to concentrate on the Fourier transform

$$\delta_k = \delta_k(t) = \int \delta(\mathbf{x}, t) e^{-i\mathbf{k}\cdot\mathbf{x}} d^3\mathbf{x} \tag{2.102}$$

[since the rest of the discussion will focus on a fixed time, we will omit t and wirte $\delta_k(t)$ as δ_k]. The complex number δ_k can be separated into an amplitude and phase:

$$\delta_k = \sqrt{|\delta_k|^2} \exp i\phi_k \tag{2.103}$$

We will show that the amplitude $|\delta_k|^2$ is most directly accessible to observation.

To see this consider a simple question regarding the distribution of matter in the universe: How much excess matter (over the average) do we expect in a typical region of radius R in the universe? This question is statistical in nature. If we divide the universe into spheres of radius R then we will expect to find each sphere to contain different amount of mass M. All that we are really interested in is the probability distribution $P(M, R)dM$ that one randomly chosen sphere will have mass in the range $(M, M + dM)$.

Such a probability distribution $P(M)$ can be equally well characterised by the moments of the distribution. In general, for an arbitrary $P(M)$, we have to specify an infinite number of moments; but if $P(M)$ is sufficiently simple, then the number of non-zero moments may be only a few. A particularly important statistic will be the rms value in M: $[< M^2 > - < M >^2]^{\frac{1}{2}}$, which can be directly related to the Fourier transform $\delta_k(t)$ of $\delta(\mathbf{x}, t)$. To do this we use the concept of "smearing", or window function. A window function is a non negative, dimensionless, function $W(r)$ with the property that $W(r) \approx 1$ for $r < R$ and $W(r) \approx 0$ for $r \gtrsim R$ [a Gaussiun with a standard deviation of R will be quite ideal]. Such a function defines an effective volume

$$V_W = \int_0^\infty r^2 W(r) dr \simeq \mathcal{O}(R^3) \tag{2.104}$$

Now consider the quantity

$$f_W(\mathbf{x}) \equiv \int f(\mathbf{x}+\mathbf{r}) W(\mathbf{r}) \frac{d^3\mathbf{r}}{V_W} \tag{2.105}$$

for any function $f(\mathbf{x})$. This defines an average value for f, at $\mathbf{x}$, with an averaging taken over a typical region of size R^3 centred at $\mathbf{x}$. In particular the excess mass in a sphere of radius R, centred at the point $\mathbf{x}$, will be

$$\delta M_R(\mathbf{x}) = \int \bar{\rho}\delta(\mathbf{x}+\mathbf{y}) W(\mathbf{y}) d^3\mathbf{y} = \bar{\rho} \int \delta_{\mathbf{k}} W_{\mathbf{k}}^* e^{i\mathbf{k}\cdot\mathbf{x}} \frac{d^3\mathbf{k}}{(2\pi)3} \tag{2.106}$$

From this we get

$$\left(\frac{\delta M}{M}\right)^2 \equiv \frac{[\delta M_R(\mathbf{x})]^2}{M_R^2} = \frac{1}{V_W^2} \int \delta(\mathbf{x}+\mathbf{y}) \delta(\mathbf{x}+\mathbf{z}) W(\mathbf{y}) W(\mathbf{z}) d^3\mathbf{y} d^3\mathbf{z} \tag{2.107}$$

To obtain the mean value of this quantity we average it over a large volume V; we get

$$\begin{aligned} < \left(\frac{\delta M}{M}\right)^2 >= \frac{< [\delta M_R(\mathbf{x})]^2 >}{M_R^2} &= \frac{1}{V_W^2} \int \frac{d^3\mathbf{x}}{V} \int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{d^3\mathbf{p}}{(2\pi)^3} \delta_k \delta_p^* W_k^* W_p e^{i(\mathbf{k}-\mathbf{p})\cdot\mathbf{x}} \\ &= \frac{1}{V_W^2} \int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{|\delta_k|^2 |W_k|^2}{V} \\ &= \int_0^\infty \frac{dk}{k} \Delta^2(k) \frac{|W_k|^2}{V_W^2} \end{aligned} \tag{2.108}$$

where

$$\Delta_k^2 \equiv \frac{k^3 |\delta_k|^2}{2\pi^2 V} \tag{2.109}$$

is the dimensionless power spectrum. If we take $W(r)$ to be a gaussian

$$W(r) = \exp\left(-\frac{r^2}{2R^2}\right) \tag{2.110}$$

and $|\delta_k|^2 = Ak^n$, the above integral can be evaluated to give

$$\left(\frac{\delta M}{M}\right)_R^2 = \frac{A}{2} \Gamma\left(\frac{n+3}{2}\right) \Delta^2(k = R^{-1}) \tag{2.111}$$

Using the relation $M \propto R^3 \propto k^{-3}$ we can also write this in a form which we use later in part 4

$$\left(\frac{\delta M}{M}\right)^2_R \propto M^{-(n+3)/3}. \tag{2.112}$$

Similar results will be obtained for other window functions. Thus the mean-square fluctuation in mass over regions of size R will be determined by the value of $\Delta^2_{\lambda=R}$. [This is why the quantity $k^3|\delta_k|^2$ was given special significance in the previous section]. It is also easy to see that

$$<\delta^2(\mathbf{x})>\equiv \left(\frac{\delta\rho}{\rho}\right)^2 = \frac{1}{V}\int_0^\infty \frac{dk}{k}\frac{k^3|\delta_k|^2}{2\pi^2} = \int_0^\infty \frac{dk}{k}\Delta_k^2 \tag{2.113}$$

Thus Δ_k^2 measures the contribution from the octave around wavenumber k to the fluctuation in the density contrast.

In these expressions, we have interpreted the symbol $<\cdots>$ as the volume average over a large size V in the universe. It is possible to provide an alternative interpretation which is extremely useful. Let us suppose that the phase ϕ_k of each mode δ_k is a random variable distributed uniformly in the interval $(0, 2\pi)$ thereby making δ_k itself a random variable. In that case we can define the mean values by averaging the phases over the interval $(0, 2\pi)$. Writing the sum over modes in a normalisation volume V as

$$\delta(\mathbf{x}) = \frac{1}{V}\sum_k \delta_k e^{i\mathbf{k}.\mathbf{x}} = \frac{1}{\sqrt{V}}\sum_k (a_k e^{i\phi_k})e^{i\mathbf{k}.\mathbf{x}}; \qquad |a_k|^2 \equiv \frac{1}{V}|\delta_k|^2 \tag{2.114}$$

it follows that

$$\begin{aligned} <\delta(\mathbf{x})> &= \frac{1}{\sqrt{V}}\sum_k a_k e^{i\mathbf{k}.\mathbf{x}} <e^{i\phi_k}>= 0 \\ <\delta^2(\mathbf{x})> &= \frac{1}{V}\sum_{kp} a_k a_p^* e^{i(\mathbf{k}-\mathbf{p}).\mathbf{x}} <e^{i(\phi_k-\phi_p)}>= \frac{1}{V}\sum_k |a_k|^2 \end{aligned} \tag{2.115}$$

Since the number of eigenmodes in the interval $(k, k+dk)$ is $(Vk^2/2\pi^2)dk$, we get

$$<\delta^2(\mathbf{x})>= \int_0^\infty dk\frac{k^2|a_k|^2}{2\pi^2} = \int_0^\infty \frac{dk}{k}\frac{k^3|a_k|^2}{2\pi^2} = \int_0^\infty \frac{dk}{k}\Delta_k^2 \tag{2.116}$$

which is precisely the result obtained before. In other words, we can either interpret the averages as spatial averages or as an average in an ensemble of δ_k's (with each realisation being given by the choice of phases ϕ_k for all k). The latter approach is often more convenient.

The manner in which we have defined the ensemble average (viz. uniform averaging over the phases ϕ_k, taken as independent random variables) makes δ_k a "Gaussian random variable". It can be easily verified that all moments of δ_k can be expressed in terms of the second moment.

$$<\delta_k\delta_p^*>\equiv \sigma_k^2 \tag{2.117}$$

For such a gaussian random variable, the probability that a particular set of values for the amplitudes will be realised is given by

$$P[\{\delta_k\}] = \prod_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp - \frac{|\delta_k|^2}{\sigma_k^2} \tag{2.118}$$

In the continuum limit, this is equivalent to a Gaussian law for the density function $\delta(\mathbf{x})$. The probability that the density contrast is described by a function $\delta(\mathbf{x})$ is given by

$$P[\delta(\mathbf{x})] = N \exp - \frac{1}{2} \int d^3\mathbf{x} d^3\mathbf{y} \delta(\mathbf{x}) G(\mathbf{x} - \mathbf{y}) \delta(\mathbf{y}) \tag{2.119}$$

where $G(\mathbf{x} - \mathbf{y})$ is a given function. This is completely equivalent to saying that each of the modes δ_k [obtained by Fourier transforming $\delta(\mathbf{x})$] acts as an independent random variable; the amplitude of δ_k is specified completely but the phase of δ_k is randomly and uniformly distributed. All the previous results can be obtained from such a description.

The rms fluctuations computed above constitute the second moment of the probability distribution. If the probability distributions were gaussian distributions, then the second moment, along with the mean provide a complete description of the distribution. If that is the case, then we expect the probability $P(M, R)$ that a randomly spaced sphere of radius R will contain a mass M to be

$$P(M) = \left(\frac{1}{2\pi s_M^2}\right)^{1/2} \exp - \frac{(M - \overline{M})^2}{2s_M^2} \tag{2.120}$$

with

$$\overline{M} = \frac{4\pi}{3} R^3 \overline{\rho}; \quad s_M^2 = \overline{M}^2 \left(\frac{\delta M}{M}\right)^2 \cong \overline{M}^2 \Delta^2(k = R^{-1}) \tag{2.121}$$

Similarly, the probability that the mean fractional excess density is δ at some point is given by the distribution

$$P(\delta) = \frac{1}{\sqrt{2\pi}\Delta} \exp\left(-\frac{\delta^2}{2\Delta^2}\right); \quad \Delta^2 = \int_0^\infty \frac{dk}{k} \Delta_k^2. \tag{2.122}$$

Whether these distributions are gaussian or not depends on the processes which generate the distribution. Inflationary models favour gaussian distributions but other processes can give raise to non-gaussian probabilities.

For a given R, the results of the averagings described above are pure numbers. They do not tell us anything about the spatial coherence of the density distribution. Such information is provided by the two-point-correlation function. This quantity is defined to be

$$\xi(\mathbf{r}) =< \delta(\mathbf{x} + \mathbf{r}) \delta(\mathbf{x}) >= \int \frac{d^3k}{(2\pi)^3 V} |\delta_k|^2 e^{-i\mathbf{k}.\mathbf{r}} \tag{2.123}$$

In other words, the correlation function is the Fourier transform of the power spectrum.

We conclude by deriving a simple relation between $(\delta M/M)^2$ and $\xi(r)$. Consider the integral

$$J_3(R) \equiv \int_0^R \xi(r)r^2dr \approx \frac{1}{4\pi}\int \xi(\mathbf{r})W(\mathbf{r})d^3\mathbf{r} \tag{2.124}$$

where $W(\mathbf{r})$ is a window function with range R. Then

$$\begin{aligned} J_3(R) &= \frac{V_w}{4\pi}\int \frac{d^3k}{(2\pi)^3V}|\delta_k|^2W_k \cong \frac{V_w}{4\pi}\left[\frac{|\delta_k|^2k^3}{2\pi^2V}\right]_{k=R^{-1}} \\ &= \frac{R^3}{3}\Delta^2(k=R^{-1}) \end{aligned} \tag{2.125}$$

On the other hand, we saw earlier that

$$\left(\frac{\delta M}{M}\right)^2_R \cong \Delta^2(k=R^{-1}) \tag{2.126}$$

Combining these two relations, we get

$$\left(\frac{\delta M}{M}\right)^2_R \simeq \frac{3J_3(R)}{R^3} = \frac{3}{4\pi R^3}\int_0^R \xi(r)r^2dr. \tag{2.127}$$

In other words, the average value of the correlation function in a region of size R is directly related to $(\delta M/M)^2$ at that size R and hence to the quantity Δ^2 at that scale.

3: Applications of the linear theory

3.1. Normalisation of the perturbation spectrum

The analysis in the previous sections used a density perturbation spectrum which is a power law, $\delta_k = Ak^{-\alpha}$ at the time of entering the horizon. Most inflationary models predict the value $\alpha = 3/2$; however, these models do not give a consistent value for the amplitude A. The value of A depends crucially on the model chosen for inflation [and the most natural models predict an absurdly high value for A].

Since we do not have a unique, acceptable prediction for A, it is necessary to consider it as a free parameter and fix its value from observation. This task - which goes under the name "normalising the spectrum" - turns out to be quite non-trivial.

To fix the value of A, we should work out some observable effect which depends on A and compare it with astronomical data. Two simple observations discussed in the last section can be used for this purpose.

The first one concerns the distribution of excess mass in random spheres of radius R, and uses the result:

$$\left(\frac{\delta M}{M}\right)_R \simeq A\Delta(k=R^{-1}) \tag{3.1}$$

Detailed analysis of Cf A - survey shows that $(\delta M/M) \simeq 1$ at $R = 8h^{-1}$Mpc. This can be used to determine A.

Alternatively, one can use the relation between $J_3(R)$ and $(\delta M/M)_R$ derived earlier:

$$\left(\frac{\delta M}{M}\right)_R^2 = \frac{3J_3(R)}{R^3} \tag{3.2}$$

The study of galaxy-galaxy correlation function allows one to determine the right-hand-side directly:

$$J_3(R) = \begin{cases} 270h^{-3}\text{Mpc}^3; & (at\, 10h^{-1}\text{Mpc}) \\ 600h^{-3}\text{Mpc}^3; & (at\, 30h^{-1}\text{Mpc}) \end{cases} \tag{3.3}$$

This gives

$$\left(\frac{\delta M}{M}\right)_R \simeq \begin{cases} 0.9 & (at\, 10h^{-1}\text{Mpc}) \\ 0.25 & (at\, 30h^{-1}\text{Mpc}) \end{cases} \tag{3.4}$$

suggesting that a scale of about $10h^{-1}$Mpc separates the linear regime from the one which has already gone non-linear. The values of $(\delta M/M)$ given above, of course, fixes the value of A. (Huchra *et al.* 1983; Davis & Peebles 1983).

There are two difficulties with the above approach which must be mentioned. Firstly, in order to normalise the spectrum reliably, we should use the value of R at which linear theory is valid. Since $\xi_{gg} \simeq 1$ at $5h^{-1}$Mpc, R has to be significantly greater than $5h^{-1}$Mpc. But at these large scales, survey estimates are not very reliable [for example, in the above equation, the value at $30h^{-1}$Mpc much less reliable compared to the $10h^{-1}$Mpc estimate.]

Secondly, our astronomical observations use galaxies as tracers of mass distribution. This assumption could be in error if - for example - galaxies formed preferentially at the high peaks of a random distribution. Such peaks would be clustered a lot more than the original mass distribution. The simplest - ad hoc - assumption regarding such a "biased galaxy formation" will be to take

$$(\delta\rho/\rho)_{\text{galaxies}} = b\,(\delta\rho/\rho)_{\text{mass}} \tag{3.5}$$

where b is some constant. Once we accept biasing, observations do not directly determine A but only the combination Ab.

It would be, therefore, nice if we could use some other theoretical consequence - which does not take galaxies as tracers of mass - to determine A. Two such consequences, viz. the induced anisotropies in CMBR temperature and the peculiar motion of galaxies provide one with a glimmer of hope. We will discuss these. effects in the next two sections.

3.2. Anisotropies in CMBR temperature

The observed temperature of the CMBR is expected to show fluctuations because of (at least!) five different reasons: (i) Since our galaxy has a peculiar velocity with respect to the cosmic frame, we expect the CMBR photons to show a dipole anisotropy with respect to our direction of motion. (Fortunately, such an anisotropy has been observed!) (ii) We receive CMBR photons from a redshift of about 10^3. If this 'last scattering surface (LSS)' had a gravitational potential varying in space, then the photons reaching us from the crests and troughs of the potential wells would have undergone different amounts of redshift - leading to a temperature fluctuation. (iii) Just as our galaxy has a peculiar velocity, the matter on the LSS can have different amount of peculiar velocity with respect to us in different locations of LSS. This will, again, lead to temperature fluctuations across the sky. (iv) The radiation field on LSS might have an intrinsic density fluctuation δ_R which will, of course, lead to a temperature anisotropy. (v) Lastly, the processes which take place between $z = 10^3$ and $z = 0$ can generate (or suppress) temperature fluctuations. These processes include gravitational lens effects, reionisation or inverse compton scattering by hot intergalactic gas.

Though all these processes produce temperature anisotropies - usually denoted by $(\Delta T/T)$ - they operate at different angular scales and have different magnitudes. This allows one to consider them separately.

To begin with, the peculiar motion of our galaxy produces an anisotropy

$$\frac{\Delta T}{T} \simeq v \cos\theta \tag{3.6}$$

where v is the peculiar velocity and θ is measured with respect to the direction of motion. Such an effect has been detected; our local group is found to be moving with a velocity of $600 \pm 50 \mathrm{kms}^{-1}$ towards the directions $\alpha = 11^h$, $\delta = -25^0$. In studying $T(\theta,\phi)$ in the sky, we will, hereafter, subtract this motion by going to the cosmic frame.

The rest of the processes can be conveniently separated into those which operate on large scales and those which produce $(\Delta T/T)$ at small angular scales. A useful separation of scales into large and small can be arrived at in the following manner. In the linear theory, a density perturbation at wavelength λ will induce a ΔT at the same scale λ. It can be easily shown that, in a Friedmann universe, a length scale λ at redshift z will subtend in the sky an angle $\theta(\lambda)$ where

$$\theta(\lambda) \simeq \left(\frac{a_0\lambda}{H_0^{-1}}\right)\frac{\Omega_0}{2} \simeq 34.4''(\Omega_0 h)\left(\frac{\lambda}{1\mathrm{Mpc}}\right) \quad (\text{for } z \gg 1) \tag{3.7}$$

Therefore, the Hubble radius $d_H(z_{dec})$ at decoupling subtends an angle

$$\theta(d_H) \simeq 0.87^0 \Omega_0^{1/2}\left(\frac{z_{dec}}{1100}\right)^{-\frac{1}{2}} \tag{3.8}$$

which is about one degree. Thus temperature fluctuations at scales larger than a degree are caused by density fluctuations at wavelengths bigger than the Hubble radius at decoupling. In other words (ΔT) at $\theta > 1^0$ ('large' angles) directly probes the superhorizon scale modes which are still in the primordial state, unaffected by "dirty" astrophysical processes.

At these scales, the second effect - usually called Sachs-Wolfe effect- dominates. If $\phi(\mathbf{x}, t_{dec})$ is the gravitational potential on LSS, then we expect $(\Delta T/T) \simeq \phi$; a rigorous, general relativistic, analysis gives the result (Sachs & Wolfe 1967).

$$\left(\frac{\Delta T}{T}\right) = \frac{1}{3}\phi(\mathbf{x}, t_{dec}) == -\frac{1}{3}Ga^2(t)\rho(t)\int d^3\mathbf{x}'\frac{\delta(\mathbf{x}',t)}{|\mathbf{x}-\mathbf{x}'|} \tag{3.9}$$

Notice that, as long as linear theory holds $\delta \propto a(t)$, $\rho \propto a^{-3}$ so that the last expression is independent of time. We can conveniently evaluate it at the present epoch; using $H^2 = (8\pi G\rho_0/3)$ and Fourier transforming $\delta(\mathbf{x}, t)$ we get

$$\left(\frac{\Delta T}{T}\right) = -\frac{a_0^2 H_0^2}{2(2\pi)^3}\int \frac{\delta_k}{k^2} e^{-i\mathbf{k}.\mathbf{x}} d^3\mathbf{k} \tag{3.10}$$

where $\mathbf{x}$ now points to the LSS and has a length $|\mathbf{x}| = 2H_0^{-1}$ which is the distance to the horizon. Since $|\mathbf{x}|$ is fixed, $(\Delta T/T)$ is essentially the function of the angles (θ, ϕ) on the sky. It is, therefore, best to expand $(\Delta T/T)$ as

$$\frac{\Delta T}{T} = \sum_{l=2}^{\infty}\sum_{m=-l}^{l} a_{lm} y_{lm}(\theta, \phi) \tag{3.11}$$

Using the previous expression for $(\Delta T/T)$ and the inversion formula for spherical harmonics, it is straightforward to show that

$$< |a_{lm}|^2 >= \frac{H_0^4}{2\pi V}\int_0^{\infty} \frac{dk}{k^2}|\delta_k|^2 |g_l(kx)|^2 \tag{3.12}$$

where g_l is the spherical Bessel function. Taking $|\delta_k|^2 = AVk^n$ we get

$$< |a_{lm}|^2 >= \frac{AH_0^{n+3}}{16}\frac{\Gamma(3-n)}{\Gamma^2[[(4-n)/2]}\frac{\Gamma[(2l+n-1)/2]}{\Gamma[(2l+5-n)/2]} \tag{3.13}$$

Notice that $\Delta^2 \propto (\Delta M/M)_R^2 \sim (k^3|\delta_k|^2)_{k=R^{-1}} \sim AR^{-(n+3)}$; so $< |a_{lm}|^2 >^{1/2} \propto (\Delta M/M)$ evaluated at $R = H_0^{-1}$. Thus, the CMBR anisotropy at large angles directly proble $(\Delta M/M)$ at horizon size today.

Taking $n = 1$ (which corresponds to $\alpha = 3/2$) and using a spherical window function to define $(\Delta M/M)$, we get

$$< |a_{2m}|^2 >= \frac{\pi}{3}(H_0 R)^4 \left(\frac{\delta M}{M}\right)_R^2 \tag{3.14}$$

[For $n = 1$, $(\delta M/M)^2 \propto R^{-(n+3)} \propto R^{-4}$; so the right hand side is indeed independent of R as it should be]. We may use the value $(\delta M/M) \approx 0.25$ at $R \approx 30h^{-1}$Mpc to obtain

$$< |a_{2m}|^2 >^{1/2} \simeq 2 \times 10^{-5} \tag{3.15}$$

Observations suggest that this value is less than 10^{-4} at 90% confidence. Theory and observations are consistent but do not provide us with any extra information.

Our original expression (3.9) also shows that, the r.m.s. fluctuation scales with the angle as:

$$< \left(\frac{\Delta T}{T}\right)^2 >^{1/2} \equiv \sigma_T \simeq k^{-2}(k^{3/2}\delta_k) \approx \begin{cases} \lambda^{\alpha+\frac{1}{2}} \propto \theta^{\alpha+\frac{1}{2}} & (\lambda < d_H) \\ \lambda^{\alpha-\frac{3}{2}} \propto \theta^{\alpha-\frac{3}{2}} & (\lambda > d_H) \end{cases} \tag{3.16}$$

For $\alpha = \frac{3}{2}$,

$$\sigma_T \propto \begin{cases} \theta^2 & (\text{for small } \theta; \theta < 1^0) \\ \text{constant} & (\text{for large } \theta; \theta > 1^0) \end{cases} \qquad (3.17)$$

This clearly shows why the Sachs-Wolfe effect dominates on large scales.

Let us now consider the contributions to $(\Delta T/T)$ at smaller scales. To begin with, one must notice that $(\Delta T/T)$ at very small scales are going to be wiped out by the finite thickness of LSS. Decoupling is not an instanteneous effect, but has a width - in redshift space - of about $\Delta z \approx 80$. (Jones & Wyse 1985) This width corresponds to the comoving scale of $\Delta l \approx 15$Mpc and a subtended angle of $\Delta\theta \approx 8'$. The photons we see could have originated anywhere within this thickness of LSS; in other words, we are only capable of observing quantities averaged over a size of about $8'$ in the sky. Any $(\Delta T/T)$ at smaller angular scales will be washed out.

Between the angular scales $8'$ and 1^0, the dominant effect is the intrinsic fluctuations in the radiation field itself. If the fluctuations are adiabatic, then we would expect

$$\left(\frac{\Delta T}{T}\right)_{\theta(\lambda)} = \begin{cases} \frac{1}{3}(\delta\rho_B/\rho_B)_\lambda & (\text{baryonic universe}) \\ \frac{1}{60}(\Omega h^2)^{-1}(\delta\rho_{DM}/\rho_{DM})_\lambda & (\text{with dark matter}) \end{cases} \qquad (3.18)$$

[Thus the density perturbation at wavelength λ translates into the angular scale $\theta(\lambda) \simeq 34.4''\ (\Omega h)(\lambda/1\text{Mpc})$]. Putting in the numbers, we get, for the baryonic universe

$$\left(\frac{\Delta T}{T}\right) \simeq \begin{cases} 10^{-3\ 5} & (\Omega = 1, h = 0{:}5) \\ 10^{-2.8} & (\Omega = 0.1, h = 0.5) \end{cases} \qquad (3.19)$$

and for the models with fermionic dark matter

$$\left(\frac{\Delta T}{T}\right) \simeq \begin{cases} 10^{-4\ 8} & (\Omega = 1, h = 0.75) \\ 10^{-4\ 0} & (\Omega = 0.2, h = 0.75) \end{cases} \qquad (3.20)$$

This is to be compared with observational bounds on $(\Delta T/T)$ which are all in the region of 10^{-5}. The RELICT experiment has set a constraint $(\Delta T/T) < 1.6 \times 10^{-5}$ at $\theta \gtrsim 3^0$; (Strutkov *et al.* 1987; Klypin *et al.* 1983) the OVRO measurements give $(\Delta T/T) < 1.5 \times 10^{-5}$ at $\theta = 7.15'$; (Readhead *et al.* 1989) the IRAM experiment gives $(\Delta T/T) < 2.6 \times 10^{-4}$ at $\theta \simeq 11''$ (Kreysa *et al.* 1989) and from VLA one has the bound of $(\Delta T/T) \leq 2 \times 10^{-4}$ for $16'' < \theta < 18''$. The most recent measurements, from COBE, gives $(\Delta T/T) < 4 \times 10^{-5}$ at $\theta = 7^0$. (Smoot *et al.* 1991) There has been several other attempts - some of which even claimed a detection initially - but these values seem to be the most reliable. We see that the baryonic model is completely ruled out while dark matter models are still consistent with observations.

The angular profile of small scale anisotropies is complementary to that of the Sachs-Wolfe effect. From (3.18) we see that

$$\sigma_T \simeq k^{3/2}|\delta_k| \simeq \begin{cases} \lambda^{\alpha-3/2} \simeq \theta^{\alpha-3/2} & (\lambda < d_H) \\ \lambda^{\alpha-7/2} \simeq \theta^{\alpha-7/2} & (\lambda > d_H) \end{cases} \qquad (3.21)$$

giving, for $\alpha = 3/2$, the dependence:

$$\sigma_T \simeq \begin{cases} \textit{constant} & (\theta < 1^0) \\ \theta^{-2} & (\theta > 1^0) \end{cases} \qquad (3.22)$$

The two remaining processes which affect $(\Delta T/T)$ at small angular scales are the peculiar velocities on the LSS and astrophysical effects which take place between $z = 10^3$ and $z = 0$. Of these, the effects due to peculiar velocities are usually subdominant to the intrinsic anisotropics discussed above [and can be calculated by methods to be discussed in the next section]. The astrophysical effects cannot be taken care of in a unified way but need separate, detailed, modelling.

One such effect which deserves special mention is the process called 'reionisation'. In the thermal history of the universe we have adopted, the material in the universe has become neutral for $z \lesssim 10^3$ and never .become a plasma again; that is why our LSS is at $z \simeq 10^3$. Suppose, however, that the medium got reionised - and became a plasma again - at some intermediate redshift, z_{ion}. If this is the case, then LSS comes much closer. Any primordial $(\Delta T/T)$ will get wiped out at angular scales which are smaller than the horizon size at z_{ion}. From (3.8), we can see that a value of $z_{ion} \approx 10$ will wipe out fluctuations below 9^0 or so.

This result emphasises the importance of anisotropy measurements at large angles $(\theta \gg 1^0)$. The absence of temperature fluctuations at small scales could be due to various effects (including reionisation) which cannot be entirely ruled out; but most of these effects cannot wipe out the large scale anisotropies.

3.3. Peculiar velocity field

In our study of density perturbations we concentrated on the evolution of $\delta_k(t)$. But such a non-zero δ_k must induce a peculiar velocity field $\mathbf{v}(\mathbf{x},t)$ to ensure local conservation of mass. From the Fourier transforms of continuity equation, Euler equation and the Poisson equation

$$\dot{\delta}_k = \frac{i\mathbf{k}}{a(t)}.\mathbf{v}_\mathbf{k}(t) \tag{3.23}$$

$$\frac{d}{dt}(av_k) - i\mathbf{k}c^2\delta_k - i\mathbf{k}\phi_k = 0 \tag{3.24}$$

$$\phi_k = -\frac{4\pi G\rho_0}{k^2}a^2\delta_k \tag{3.25}$$

it is fairly easy to show that, the velocity field $\mathbf{v}(t,\mathbf{x})$ arising from the component v which is parallel to $\mathbf{k}$, is given by

$$\mathbf{v}(t,\mathbf{x}) = \left(\frac{2}{3}\Omega_0^{-1}\frac{d\ln\delta}{d\ln a}\right)(H_0^{-1}\nabla\phi) \tag{3.26}$$

where $\nabla\phi$ is the peculiar acceleration field generated by the perturbed potential

$$\phi(t,\mathbf{x}) = -Ga^2\int\frac{\rho_b\delta(\mathbf{x}',t)}{|\mathbf{x}-\mathbf{x}'|}d^3\mathbf{x}' \tag{3.27}$$

[Naively speaking one would have expected $v \simeq gt_{univ}$; this is essentially what the above equation implies]. If we are in a $\Omega = 1$ universe, $\delta \propto a$ and the prefactor $(ad\delta/\delta da)$ in (3.26) becomes unity. However, for a general Ω, this factor is a slowly varying function of Ω. For all models with $\Omega \leq 1$, one can approximate this factor as (Peebles 1980):

$$\frac{d\ln\delta}{d\ln a} \simeq \Omega^{0.6} \tag{3.28}$$

Taking Fourier transforms, we see that

$$(v)_\lambda \simeq \frac{\lambda}{H_0^{-1}}\Omega_0^{0.6}\left(\frac{\delta\rho}{\rho}\right)_\lambda \tag{3.29}$$

If $(\delta\rho/\rho)_\lambda$ decreases more rapidly than λ^{-1} [or, equivalently, $(\delta\rho/\rho)_M$ falls more rapidly than $M^{-1/3}$, as it does in Harrison-Zeldovich spectrum], then the contribution from large scales will be negligible, justifying the Newtonian approach we have adopted.

The relation (3.26) is extremely significant. It shows that the peculiar velocity field of any class of test particles directly probes the underlying mass distribution. If the peculiar velocities of a large sample of galaxies can be measured - thereby providing $\mathbf{v}(\mathbf{x}, t_0)$ - then the divergence of this field will immediately give the underlying *mass* distribution. Given some independent data on mass distribution (say, from galaxy surveys) one can test the consistency of the theory as well as fix the value of $\Omega^{0.6}$ [or, rather, the value of $b\Omega^{0.6}$ if a biasing parameter is invoked].

From (3.26) one can also compute the rms value of peculiar velocity averaged over a region of size R:

$$\begin{aligned} v^2(R) &= \frac{1}{V_W^2} < \left(\int \mathbf{v}(\mathbf{r}+\mathbf{x})W(\mathbf{x})d^3\mathbf{x}\right)^2 > \\ &= \frac{H_0^2}{V}\left(\frac{d\ln\delta}{d\ln a}\right)^2 \int_0^\infty \frac{|\delta_k|^2|W_k|^2}{2\pi^2 V_W^2}dk \\ &\propto R^{-(n+1)} \end{aligned} \tag{3.30}$$

for a spectrum with $|\delta_k|^2 \propto k^n$. [The above expression shows that the relevant integrand here is $|\delta_k|^2 \propto (\Delta_k^2/k^3)$ rather than (Δ_k^2/k) which occurred in the rms values of density distribution. Thus peculiar velocities are more sensitive to larger scales than the density distributions]. For $n = 1$, $v^2(R) \propto R^{-1}$; putting in the numbers, and using the J_3 normalisation discussed in Sec. 3.1, we get

$$v(R) \simeq \begin{cases} 200\text{kms}^{-1}(R/25h^{-1}\text{Mpc})^{-1}h^{-5/3} & (HDM) \\ 160\text{kms}^{-1}(R/25h^{-1}\text{Mpc})^{-1}h^{-3/4} & (CDM) \end{cases} \tag{3.31}$$

Though the peculiar velocity field of the galaxies is an important diagnostic, it is not an easy quantity to measure. If the actual velocity of a galaxy, located at the position $\mathbf{r}$ is $\mathbf{v}$, then the 'peculiar' velocity is defined to be

$$\mathbf{v}_{pec} = \mathbf{v} - H_0\mathbf{r} \tag{3.32}$$

The only peculiar velocity which can be measured reliably is that of earth with respect to the isotropic MBR background. Correcting for known motions, it is estimated that our local group is moving with a velocity of 650 ± 30km s^{-1} in the direction of $(11.2 \pm 0.04hr, -7 \pm 0.05$ degrees).

The gravitational acceleration due to a distant mass falls as (M/r^2); the flux of radiation from that source also decreases as (L/r^2). Therefore, if (M/L) is a constant, we should be able to obtain the direction of peculiar acceleration from the distribution of sources. This study has been done for optical and infrared sources. The direction of acceleration obtained in this manner agrees reasonably with direction of motion determined by MBR anisotropy. This acceleration seems to originate within $40h^{-1}$Mpc.

To compute $\mathbf{v}_{pec}$ for distant galaxies, we need an estimate of $\mathbf{r}$, which is independent of redshift. Several such measures have been suggested and used in the literature for this purpose. Initial studies led to controversial claims in the literature. [The quoted values for $< v_{pec}^2 >^{1/2}$ ranged from 300 km s^{-1} to 10^4kms^{-1} (Rubin *et al.* 1976; Hart & Davis 1982; Collins *et al.* 1986; James *et al.* 1987)]. More recently a redshift-independent relation, $D_n \propto \sigma_0^{4/3}$, where D_n is a suitable defined isophotal diameter and σ_0 is the central velocity dispersion, was used to estimate distances. When applied to a sample of about 400 ellipticals with a redshift depth of 0.02 (i.e., for Hr $\lesssim$ 6000km s^{-1}), this method yields a large value $< v_p^2 >^{1/2} \simeq 600 \pm 100$km s^{-1} on a scale of about $50h^{-1}$Mpc. The direction of motion is consistent with previous observations. (Lynden-Bell *et al.* 1988)

More recently; the IRAS redshift survey was used to map the velocity field. (Rowan-Robinson *et al.* 1990) Two conclusions are suggested by these investigations: (1) It appears that, the peculiar velocities are bigger than what was predicted by simple models. (2) The peculiar velocity of our local group seems to be generated mostly from matter contained within a region of $100h^{-1}$Mpc or so.

It seems safe to claim that our models predict values which are smaller than the observed ones. This, in fact, is the most serious challenge faced by the theory at present.

Lastly, one may use (3.26), evaluated at $t = t_{dec}$ to compute the anisotropies in CMBR induced by peculiar velocities of matter on the LSS as $(\Delta T/T) \simeq (v/c)_{LSS}$. This gives values which are slightly subdominant to the intrinsic variations discussed earlier. The angular dependence of this effect, however, is quite different:

$$\begin{aligned} \sigma_T &\simeq (k^{3/2}\delta_k)k^{-1} \simeq \lambda^{\alpha-1/2} \simeq \theta^{\alpha-1/2} \quad (\text{for small } \theta) \\ &\simeq \theta \end{aligned} \tag{3.33}$$

if $\alpha = 3/2$. Most of the contribution comes from the region near $\theta \approx 1^0$.

4 : Non linear evolution

In the previous sections we have been considering the linear growth of density fluctuations and possible observational probes of the linear theory. However the real world of galaxies is highly non linear, in the sense that the density contrasts between galaxies and the background Universe are much larger than unity. How does one connect the linear theory with this highly non linear Universe ? One possible route which has generally been adopted is to do N - body simulations with a large number of particles interacting via gravity. Another way is to make simple but nevertheless non-trivial models of the non linear evolution. In most of this section we follow this latter route.

We begin in the next section with possibly the simplest nontrivial discription of nonlinear evolution, the spherical model. The power of the spherical model lies in the fact that at very little computational cost, one can estimate properties of 'nonlinear' bound objects which form as a result of gravitational instability. We then discuss an important class of theories of galaxy formation, the hierarchical clustering theories. We consider the scaling laws and the mass functions which obtain in such theories. Another route for examining non linear evolution, particularly in the context of top-down theories of galaxy formation, is via the elegant approximate solution proposed by Zeldovich. We examine the Zeldovich approximation in section 4.4 and its recent extension by Gurbatov, Saichev and Shandarin, the adhesion model, in the following section. Any discussion of nonlinear evolution would be incomplete without mention of the important results from N- Body simulations. In section 4.6 we outline the main results of such simulations, concentrating on two specific theories, the Cold DM and the Hot DM theories of structure formation. The rest of part 4, is concerned with trying to elucidate how some general properties of galaxies arise. What sets galactic mass scales ? How does the galactic angular momentum arise ? How do disk and elliptical galaxies form and what decides if a given protogalactic cloud becomes a disk galaxy or an elliptical ?

4.1. The spherical model

In the spherical model one assumes that the matter distribution and the geometry of the universe are spherically symmetric about a point, which we can define to be the origin of our co-ordinate system. It is also customary to assume that matter is described by a pressureless fluid. If one wants to study the evolution of density perturbations in a Friedmann- Robertson Walker (FRW) universe, one will also demand that far away from the origin the matter density and the geometry become uniform. It is remarkable that there exists an exact solution of the Einsteins equations describing such a spherically symmetric Universe (see below). For most purposes however, we do not need to use this solution. This is because we mostly deal with the situation where the density perturbation δ on a given physical scale l is very much smaller than unity, when the perturbation enters the horizon ($l = d_H$). And becomes comparable to unity, only long after by which time $l << d_H$. In this case the non linear evolution can be adequately studied in the Newtonian limit.

In this limit, as we described in Appendix 1, geometry can be described completely by using a 'Newtonian' potential ϕ (cf. equation A1.2), which satisfies the standard Poisson equation

$$\nabla^2\phi = 4\pi G\rho. \tag{4.1}$$

Also the evolution of matter is governed by the standard Newtonian continuity and Euler equations (A1.4 and A1.5). In particular for a pressureless fluid the euler equation reduces to

$$\frac{d^2\mathbf{r}}{dt^2} = -\nabla\phi, \tag{4.2}$$

where $\mathbf{r}$ is gives the position of the fluid particle.

To begin with consider the Newtonian limit of a FRW Universe which we discussed in Appendix 2.1. In this case the potential ϕ is given by equation (A1.3)

$$\phi = \phi_b(t,\mathbf{r}) = -(\frac{1}{2})(\frac{\ddot{a}}{a})|\mathbf{r}|^2. \tag{4.3}$$

The fluid equations integrate to give (see appendix 2.1)

$$\mathbf{r} = a(t)\mathbf{x}; \quad \mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{\dot{a}}{a}\mathbf{r} \equiv H(t)\mathbf{r}; \quad \rho(t) = \rho_b(t) \propto a^{-3} \tag{4.4}$$

From the Poisson equation, $\nabla^2\phi_b = -3\ddot{a}/a = 4\pi G\rho_b$. So one gets the equation of motion for a

$$\ddot{a} = -\frac{4\pi}{3}G\rho_b(t)a. \tag{4.5}$$

Substituting for ρ_b from (4.4) this integrates to

$$\frac{\dot{a}^2}{2} - \frac{4\pi G\rho_i a_i^3}{3a} = -\frac{k}{2} \tag{4.6}$$

or the standard equation for the background Universe

$$\frac{1}{a^2}((\frac{da}{dt})^2 + k) = \frac{8\pi G\rho_b(t)}{3}; \rho_b = \rho_i(\frac{a_i}{a})^3, \tag{4.7}$$

where ρ_i is the density when the scale factor is a_i. Also the equation of motion for a fluid particle can be rewritten using (4.5) as

$$\frac{d^2\mathbf{r}}{dt^2} = -(\frac{4\pi G\rho_b(t)}{3})\mathbf{r} \equiv -\frac{GM_b}{r^3}\mathbf{r}, \tag{4.8}$$

where we have defined M_b, the mass contained in a sphere of radius $r = |\mathbf{r}|$.

Now suppose the density around the origin differs from the background density ρ_b, at an initial time t_i, by a small spherically symmetric perturbation $\delta\rho(r,t_i)$, that is

$$\rho(r,t_i) = \rho_b(t_i) + \delta\rho(r,t_i) \equiv \rho_b(t_i)(1+\delta_i(r)). \tag{4.9}$$

Here δ_i is the initial fractional excess density contrast. We will also suppose that at an initial time t_i, the perturbed region moves a velocity $H_i r + v_i(r)$, where $H_i r$ is the hubble velocity of the region and $v_i(r)$ the peculiar velocity. In the Newtonian limit the effect of the perturbation is to add to ϕ_b a term $\delta\phi$ such that $d(\delta\phi)/dr = G\delta M/r^2$, where δM is the excess mass over and above M_b, contained within r. We can then use (4.2) to work out the evolution of such a perturbed region.

For this consider a spherical shell whose radius is r_i at the initial time t_i. The proper radius of the shell $r(t)$ obeys the equation of motion (4.2) , that is

$$d^2r/dt^2 = -GM/r^2, \tag{4.10}$$

where

$$M = \rho_b(\frac{4\pi}{3}r_i^3)(1+\bar{\delta}_i), \tag{4.11}$$

and

$$\bar{\delta}_i = (\frac{3}{4\pi r_i^3}) \int_0^{r_i} \delta_i(r) 4\pi r^2 dr \tag{4.12}$$

is the average value of δ within r_i. Let us assume that the perturbed density is such that shells with different initial radii dont cross as they evolve. This implies that M is constant. Then the first integral of equation (4.10) is

$$\frac{1}{2}(\frac{dr}{dt})^2 - \frac{GM}{r} = E \tag{4.13}$$

where E is a constant of integration. The sign of E decides if a given mass shell will expand for ever or eventually decouple from the expansion and collapse. If $E \geq 0$, then from (4.13) we can see that $\dot{r}$ will never become negative, and the shell will expand for ever. On the other hand for $E < 0$, as r increases $\dot{r}$ will become 0 and then negative, implying a collapse.

It is simple and instructive to derive the condition on the density perturbation for such a collapse to occur. For this consider (4.13) at the initial time t_i. Since the shell has a peculiar velocity v_i in addition to its Hubble velocity $H_i r_i$, we have $\dot{r}_i = H_i r_i + v_i(r_i)$, at time t_i. So

$$K_i = \frac{\dot{r}^2}{2}|_{t_i} = \frac{(H_i r_i + v_i)^2}{2}. \tag{4.14}$$

Further we have

$$(\frac{GM}{r})|_{t_i} = G\frac{4\pi}{3}\rho_b(t_i) r_i^2 (1 + \bar{\delta}_i) = \frac{1}{2} H_i^2 r_i^2 \Omega_i (1 + \bar{\delta}_i) \tag{4.15}$$

where Ω_i is the initial value of the density parameter Ω for the background smooth Universe. Then

$$E = \frac{H_i^2 r_i^2 \Omega_i}{2} [\Omega_i^{-1}(1 + (\frac{v_i}{H_i r_i}))^2 - (1 + \bar{\delta}_i)]. \tag{4.16}$$

We see that for the shell to collapse eventually one must have $(1 + \bar{\delta}_i) > \Omega_i^{-1}(1 + v_i/H_i r_i)^2$, or in otherwards satisfy the condition

$$\bar{\delta}_i > \Omega_i^{-1}(1 + \frac{v_i}{H_i r_i})^2 - 1. \tag{4.17}$$

This condition is easier to satisfy for a larger Ω_i and smaller v_i. If v_i were zero or negative then any overdense region with $\bar{\delta} > 0$ will eventually collapse in a closed or flat FRW universe, although at later times for smaller overdensities. On the other hand in an open universe with $\Omega_i < 1$, the overdensity has to be above a critical value for collapse to ensue. If $\delta(r_i)$ for example is centrally peaked, only shells within a critical initial radius r_{cr} such that $\bar{\delta}_i(r_{cr}) = \Omega_i^{-1}(1 + v_i/H_i r_i)^2 - 1$ will be able to collapse.

Using the fact that E is a constant of motion one can also derive the maximum radius which a bound shell attains. For this note that at the radius of turn around r_m, we have $\dot{r} = 0$ and so

$$E = -GM/r_m = -(\frac{r_i}{r_m})\frac{H_i^2 r_i^2}{2}\Omega_i(1 + \bar{\delta}_i). \tag{4.18}$$

Comparing this expression for E with the one given in (4.16) we get

$$\frac{r_m}{r_i} = \frac{(1+\bar{\delta}_i)}{[\bar{\delta}_i - [\Omega_i^{-1}(1+(v_i/H_i r_i))^2 - 1]]} \equiv \frac{(1+\bar{\delta}_i)}{D} \tag{4.19}$$

where we have defined the quantity D for later convenience.

Let us now consider the motion of the shell in more detail for the case $E < 0$, when the shell is bound. The solution to equation (4.13) for $E < 0$ is given in a parametric form by

$$r = A(1 - cos\theta), \quad t = B(\theta - \sin\theta) + \bar{t}; \quad A^3 = GMB^2 \tag{4.20}$$

where A and B are constants related according to the last equation in (4.20) and $\bar{t}$ is a constant of integration. The parameter θ increases with incresing t, while r increases to a maximum value before decreasing to zero. We see therefore that the shell enclosing M, initially expanding with the background Universe, slows down, reaches a maximum radius at $\theta = \pi$ before turning around and collapsing. The epoch of maximum radius is also referred to as the epoch of 'turnaround'. As we mentioned earlier, at turnaround, $dr/dt = 0$ and $r = r_m$.

The constants A, B and $\bar{t}$ can be fixed in a straightforward way by using (4.19) We have at $\theta = \pi$,

$$r(\pi) = r_m = 2A = r_i \frac{(1+\bar{\delta}_i)}{D} \tag{4.21}$$

Therefore

$$A = \frac{r_i}{2D}(1+\bar{\delta}_i). \tag{4.22}$$

Further using $A^3 = GMB^2$, given in (4.20) and the expression for M from (4.11) we have

$$B = \frac{1+\bar{\delta}_i}{2H_i\Omega_i^{1/2}D^{3/2}} \tag{4.23}$$

The constant of integration $\bar{t}$ can be related to the initial conditions as

$$\bar{t} = t_i - B(\theta_i - \sin\theta_i) \tag{4.24}$$

where θ_i is the initial value of θ given by

$$r_i = A(1 - \cos\theta_i). \tag{4.25}$$

Note that the expressions for A and B derived above differ from those given in Peebles (1980) beyond the leading order in δ_i. This is because Peebles (1980) derives A and B by matching with linear theory. Such a procedure will only give the constants correct to the leading order in δ_i, whereas in our derivation of A and B no such restriction need be placed. Of course since δ_i is generally assumed to be small compared to unity, this correction has little effect in practice.

The equations (4.20) with the constants A, B and $\bar{t}$ fixed by (4.22) , (4.23) and (4.24) give the complete information about how each perturbed mass shell evolves.

And we can use these equations to work out any interesting characteristic of a spherical perturbation. One such characteristic is the evolution of average density and the average fractional excess density contrast $\bar{\delta}(r,t)$ within each mass shell. Since M is constant for each mass shell the average density within it is simply

$$\bar{\rho}(t) = \frac{M}{(\frac{4\pi}{3}r^3)} = \frac{3M}{4\pi A^3(1-\cos\theta)^3} \tag{4.26}$$

In the special case when the initial density enhancement is homogeneous, the above average density is also the actual density. The density profile of such a constant density sphere is often referred to as the 'top hat' profile, for obvious pictorial reasons. To work out the time evolution of $\bar{\delta}(r,t)$, one also needs to know also how the background density evolves. For this one has to solve equation (4.7) describing the evolution of background Universe.

We consider here only the simplest case of a flat Universe with $k = 0$. The other cases of $k = 1$ and $k = -1$ can be examined in a similar fashion and the corresponding results can be found in Peebles (1980). One must note that theoretical prejudice about the early universe also favours the $k = 0$ model, so the simplest case may also be the most relevant. When $k = 0$ the solution to equation (4.7) is given by

$$a \propto t^{2/3}; \quad \rho_b(t) = \frac{1}{6\pi G t^2} \tag{4.27}$$

Dividing the average density $\bar{\rho}(r,t)$ in equation (4.26) by the background density in (4.27) we get for the average density contrast

$$\frac{\bar{\rho}(r,t)}{\rho_b(t)} = 1 + \bar{\delta}(r,t) = \frac{3M}{4\pi A^3} \times \frac{6\pi G B^2[(\theta - \sin\theta) + \bar{t}/B]^2}{(1-\cos\theta)^3}. \tag{4.28}$$

where we have used the relation between t and θ given in equation (4.20) . Since $A^3 = GMB^2$, from (4.20) , we then get

$$\bar{\delta} = \frac{9}{2}\frac{[(\theta - \sin\theta) + \bar{t}/B]^2}{(1-\cos\theta)^3} - 1. \tag{4.29}$$

It is interesting to examine the behaviour of the average density contrast in the limit of small $(t - \bar{t})$ or equivalently θ. We have from (4.29) and (4.20)

$$\bar{\delta} \approx \frac{3\theta^2}{20} + \frac{12\bar{t}}{B\theta^3}; \quad (t - \bar{t}) \approx \frac{B\theta^3}{6} \tag{4.30}$$

So for small θ

$$\bar{\delta} = \frac{3}{20}(\frac{6(t-\bar{t})}{B})^{2/3} + \frac{2\bar{t}}{(t-\bar{t})} \tag{4.31}$$

For a flat Universe we have $\Omega_i = 1$ and also $H_i = 2/(3t_i)$ from (4.27) . So to the leading order in the perturbed quantities (4.23) becomes

$$B = \frac{3}{4}\frac{t_i}{(\bar{\delta}_i - 3t_i v_i/r_i)^{3/2}} \tag{4.32}$$

and (4.24) reduces to

$$\bar{t} = t_i(\frac{\bar{\delta}_i}{5} + \frac{9v_it_i}{10r_i}) \tag{4.33}$$

Using these expressions for B and $\bar{t}$ in (4.31) we have

$$\bar{\delta} = \frac{3}{5}(\frac{t}{t_i})^{2/3}(\bar{\delta}_i - \frac{3t_iv_i}{r_i}) + \frac{2}{5}(\frac{t_i}{t})(\bar{\delta}_i + \frac{9t_iv_i}{2r_i}) \tag{4.34}$$

Here we have assumed that t is large enough so that $\bar{t}$ can be neglected compared to t in (4.31) , but at the same time small enough so that small θ approximation can be made. We see from the above that we have recovered the linear theory result for a general spherical density perturbation, including the decaying mode (compare for example with Peebles (1980)).

We, henceforth, assume with Peebles (1980) that $\bar{\delta}_i$ is so small compared to unity that it suffices to retain only the leading terms of $\bar{\delta}_i$ in A and B. We will also assume that initially the perturbation expands as the background universe, that is $v_i = 0$ and ignore the decaying mode by setting $\bar{t} = 0$. We then have

$$A = \frac{r_i}{2\bar{\delta}_i}; B = \frac{3t_i}{4\bar{\delta}_i^{3/2}} \tag{4.35}$$

These constants can also be conveniently expressed in terms of the comoving radius $x = a(t_0)r_i/a(t_i)$ and the average fractional density excess inside the shell measured at the current epoch, assuming linear growth $\bar{\delta}_0 = (a(t_0)/a(t_i))(3\bar{\delta}_i/5)$ (see (4.34)). Here t_0 denotes the present epoch. From (4.35) we then get

$$A = \frac{3x}{10\bar{\delta}_0}; B = (\frac{3}{5})^{3/2}\frac{3t_0}{4\bar{\delta}_0^{3/2}} \tag{4.36}$$

Collecting all our results together, the evolution of a spherical overdense region can be summarised by the following equations :

$$r(t) = \frac{r_i}{2\bar{\delta}_i}(1 - \cos\theta) = \frac{3x}{10\bar{\delta}_0}(1 - \cos\theta)$$

$$t = \frac{3t_i}{4\bar{\delta}_i^{3/2}}(\theta - \sin\theta) = (\frac{3}{5})^{3/2}\frac{3t_0}{4\bar{\delta}_0^{3/2}}(\theta - \sin\theta)$$

$$\bar{\rho}(t) = \rho_b(t)\frac{9(\theta - \sin\theta)^2}{2(1 - \cos\theta)^3} \tag{4.37}$$

where the last equation is just a rewritten form of (4.29) , with $\bar{t} = 0$.

The above analysis can be easily extended to work out how the density contrast profile evolves with time. From the conservation of mass we have

$$\rho(r(t))4\pi r^2 dr = \rho(r_i, t_i)4\pi r_i^2 dr_i \tag{4.38}$$

Dividing by the background density we have

$$\frac{\rho(r,t)}{\rho_b(t)} = \frac{\rho(r_i,t_i)}{\rho_b(t_i)}(\frac{t_i}{t})^{-2}(\frac{r_i}{r})^3(\frac{d\ln r_i}{d\ln r}) \tag{4.39}$$

Here we have used the fact that the background density $\rho_b \propto t^{-2}$ for a flat cosmological model. All the terms except the last one in (4.39) are easily evaluated from the equations for r and t above. From (4.37) we can write

$$\frac{d\ln r}{d\ln r_i} = 1 - \frac{d\ln \bar{\delta}_i}{d\ln r_i} + (\frac{\sin\theta}{1-\cos\theta})\frac{d\theta}{d\ln r_i} \tag{4.40}$$

The equation connecting t with θ gives

$$\frac{d\theta}{d\ln r_i} = \frac{3(\theta - \sin\theta)}{2(1-\cos\theta)}\frac{d\ln\bar{\delta}_i}{d\ln r_i} \tag{4.41}$$

and from the definition of $\bar{\delta}_i$

$$\frac{d\ln\bar{\delta}_i}{d\ln r_i} = 3\frac{(\delta_i - \bar{\delta}_i)}{\bar{\delta}_i} \tag{4.42}$$

Putting these results together we have for the evolution of the density profile

$$\frac{\rho(r,t)}{\rho_b(t)} = \frac{[9(\theta-\sin\theta)^2/2(1-\cos\theta)^3]}{1-3[(\delta_i/\bar{\delta}_i)-1][1-(3\sin\theta(\theta-\sin\theta)/2(1-\cos\theta)^2)]} \tag{4.43}$$

In this equation we must keep in mind that both θ and r_i are implicit functions of r and t through (4.20) .

Let us now derive some useful consequences of the spherical model. To make rough estimates it suffices to deal with average values of the relevant parameters. Ofcourse in the case of a top hat density profile, where the initial density perturbation is constant within a radius r_i and zero outside, one may replace the average excess density contrast in what follows by the actual value of δ. The first interesting set of parameters one can derive are the properties of the spherical perturbation at turn around. Putting $\theta = \pi$ in the equations (4.37) one gets the redshift z_m, the proper radius of the shell r_m and the average density contrast within the shell at 'turn around' $\bar{\delta}_m$, to be

$$(1+z_m) = \frac{\bar{\delta}_0}{1.062}$$

$$r_m = \frac{3x}{5\bar{\delta}_0}$$

$$(\frac{\bar{\rho}}{\rho_b})_m = 1 + \bar{\delta}_m = \frac{9\pi^2}{16} \approx 5.6 \tag{4.44}$$

After the spherical overdense region turns around the evolution equations (4.37) predict that it will continue to collapse until a time corresponding to $\theta = 2\pi$ when all the mass collapses to a point. However long before this happens our approximations, that matter is distributed in spherical shells, and that random velocities of the particles are small, are expected to break down. Shocks in the case of gas and 'violent relaxation' effects in the case of collisionless particles are expected to convert part of the potential energy at maximum expansion into 'random kinetic energy' until the virial theorem is satisfied. After virialisation of the spherical mass contained within a collapsed shell one expects the potential energy $P.E = -2(K.E)$, where $K.E$ is the kinetic energy. This implies that the energy $\bar{E} = P.E + K.E = -K.E$. At maximum expansion all the energy was in a potential form and so $\bar{E} \approx -3GM^2/5r_m$ where we have approximated $P.E$ by that relevant to a constant density sphere. One can then define a 'virial velocity' v and a characterestic 'virial radius' r_{vir} for the mass by

$$K.E = \frac{Mv^2}{2} = -\bar{E} = \frac{3GM^2}{5r_m} \qquad (4.45)$$

and

$$-P.E = \frac{3GM^2}{5r_{vir}} = 2(K.E) = Mv^2 \qquad (4.46)$$

This gives

$$v = (6GM/5r_m)^{1/2} \qquad (4.47)$$

$$r_{vir} = r_m/2. \qquad (4.48)$$

The time when a fluctuation collapses to a virial equilibrium is also of interest. A rough estimate of this collapse time t_{coll}, is given by the time corresponding to $\theta = 2\pi$. From equation (4.37) , t_{coll} and the redshift of collapse z_{coll} are given by

$$t_{coll} = \pi(r_m^3/2GM)^{1/2} \qquad (4.49)$$

$$(1 + z_{coll}) = \bar{\delta}_0/1.686 \qquad (4.50)$$

One can also estimate the mean density of the collapsed object to be

$$\rho' \approx (\frac{r_m}{r_{vir}})^3\rho_m = 8\rho_m \approx 44.8\rho_b(t_m) \approx 170\rho_b(t_{coll}) = 170\rho_0(1 + z_{coll})^3, \qquad (4.51)$$

where ρ_0 is the present cosmological density. Finally for any gaseous component one can estimate a 'virial' temperature T_{vir} from $3\rho_{gas}kT_{vir}/2\mu = \rho_{gas}v^2/2$, where ρ_{gas} is the gas density, μ is its mean molecular weight and k is the Boltzmann constant. This gives

$$T_{vir} = \mu v^2/3k \qquad (4.52).$$

It is useful to put in typical numbers for the various quantities derived above. Apart from the cosmological parameters, essentially two unknown parameters have to specified. We choose these to be the mass of the over dense region M and the present linearly extrapolated fractional average density contrast $\bar{\delta}_0$. We then get

$$\rho_0 = 4.55 \times 10^{-30}\Omega \mathrm{h}_{50}^2\mathrm{gcm}^{-3}$$

$$x = 1.52 M_{12}^{1/3}\Omega^{-1/3}\mathrm{h}_{50}^{-2/3}\mathrm{Mpc}$$

$$t_0 = 4/3 \times 10^{10}\mathrm{h}_{50}^{-1}\mathrm{yr}$$

$$r_m = 2r_{vir} = \frac{914}{\bar{\delta}_0} M_{12}^{1/3}\mathrm{h}_{50}^{-2/3}\mathrm{kpc}$$

$$t_{coll} = 2.19 t_0/\bar{\delta}_0^{3/2}$$

$$v = 76.4 M_{12}^{1/3}\bar{\delta}_0^{1/2}\mathrm{h}_{50}^{1/3}\mathrm{kms}^{-1}$$

$$T_{vir} = 1.4 \times 10^5 \bar{\delta}_0 M_{12}^{2/3}\mathrm{h}_{50}^{2/3} K$$

$$\bar{\delta}_0 = 1.062(1+z_m) = 1.686(1+z_{coll}) \tag{4.53}$$

Here Ω is the density parameter which is unity for the flat Universe, h_{50} the Hubble constant in units of $50\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$ and M_{12} is the mass in units of $10^{12}M_{\odot}$. Further all the equations except the first two in (4.53) assume the universe to be spatially flat with $k = 0$.

We can use the above to estimate the typical parameters of collapsed objects once we are given $\bar{\delta}_0$, say from assumptions about the initial density fluctuation and their subsequent evolution (Part 2 and see also below). Alternatively to get some rough idea one can specify the collapse redshift. For example if objects with $M = 10^{12}M_{\odot}$, typical of galaxies collapse at a redhift of say 2, then one gets $r_{vir} \approx 90$kpc, $t_{coll} \approx 2.5 \times 10^9$yr, $v \approx 132\mathrm{kms}^{-1}$, $T_{vir} \approx 4.1 \times 10^5 K$ and a present day density contrast of the galaxy $\approx 4.6 \times 10^3$.

We note the power of a simple model like the spherical model in being able to extrapolate linear theory to get some idea of the properties of 'non linear' bound objects which form.

We end this section with a discussion of the mesmerisingly simple general relativistic solution for a spherical density inhomogeneity in the FRW universe. The derivation of the metric, for such a spherically symmetric spacetime, with the matter in the form of a pressureless fluid, is given in Peebles (1980). It turns out that one can put the metric in the form

$$ds^2 = dt^2 - a^2(x,t)[\frac{dx^2}{1-k(x)x^2}(\frac{(ax)'}{a})^2 - x^2(d\theta^2 + \sin\theta d\phi^2)] \tag{4.54}$$

where $a(x,t)$ is a space dependent expansion factor, and $k(x)$ a space dependent curvature constant. It should be pointed out that the metric can be written as above only as long as mass shells at different values of x do not cross ; a condition which will be satisfied at least by density distributions where ρ decreases monotonically with x. We also note that the above metric reduces to the standard FRW metric if a and k are independent of x . The Einsteins equations give the time evolution of the expansion factor $a(x,t)$ and the matter density $\rho(x,t)$ as

$$\frac{\dot{a}^2 + k(x)}{a^2} = \frac{C}{a^3} = \frac{8\pi G\rho(x,t)}{3}\frac{(ax)'}{a} \tag{4.55}$$

where C is a constant. Here (4.55) is actually two equations, one giving the evolution of a and the other that of ρ. If ρ and hence a and k are independent of x these equations reduce to the standard equations for a FRW universe given by (4.7) .

This pleasantly simple generalisation of the homogeneous Universe model offers considerable insight in to the way a spherical overdense or underdense region behaves. Equation (4.55) tells us very simply that how a mass shell at a comoving radius x evolves is completely specified by the local value of the curvature constant k. If at some x, $k(x) \geq 0$, the corresponding mass shell will expand for ever, while if $k(x) < 0$ it will turn around at some stage and collapse.

One may wonder how the local value of k depends on the density distribution ? To elucidate this connection we first rewrite (4.55) as follows.

$$k(x)x^2 = \frac{Cx^3}{ax} - \dot{a}^2x^2 \tag{4.56}$$

This can be cast in a more familiar form by defining

$$r_e \equiv a(x,t)x$$

and

$$M_e \equiv \frac{Cx^3}{2G}. \tag{4.57}$$

From (4.54) the area of a sphere with constant x and t is given by $4\pi a^2x^2$. So $r_e = ax$ is a measure of the radius of such a sphere. Also we have using (4.57) and (4.55)

$$\frac{\partial M_e}{\partial r_e} = \frac{3Cx^2}{2G(ax)'} = 4\pi\rho r_e^2 \tag{4.58}$$

So $M_e(x)$ is a measure of the effective mass within a sphere at x. Using these quantities the equation for $k(x)$ becomes

$$k(x)x^2 = 2[\frac{GM_e}{r_e} - \frac{1}{2}\dot{r}_e^2] \tag{4.59}$$

We then see that the local value of k depends in a non local manner on the density $\rho(x,t)$, in the sense that it is the density integrated within a sphere that appears in the expression for $k(x)$.

We are now in a position to conceptually understand the evolution of all the different types of spherical density perturbations that may arise in the FRW Universe. We give below two examples ; other cases can be analysed in a similar way. Consider the case when $k(x)$ is positive for $x < x_0$, is zero at x_0 and goes to a constant negative value, say -1, far away from the origin. One way of realising such a situation, which can be seen from (4.59) is to embed a density hill centered around the origin, in an open FRW Universe and also start of the universe expanding uniformly. Ofcourse the density hill has to be high enough, compared to the background density in the open Universe, to make k positive near the origin. From the evolution equation (4.55) we can infer that, the region $x < x_0$ will eventually collapse, while the region $x \geq x_0$

will expand for ever. Here we see quite clearly that condensation in a local part of the universe does not alter the global topology of an open FRW universe. Similarly one can think of how to make expanding voids in a closed univese. In this case one demands that $k(x) < 0$ for $x < x_0$, say and positive elsewhere. This situation can be realised for example, if there is a deep enough density valley in a closed universe. The region within x_0 will keep on expanding, where as the region outside will initially expand slower and eventually recollapse. Infact at some stage an expanding shell may meet a recollapsing shell in a caustic. At this stage our metric will break down. Such expanding voids have been considered by many authors as an interesting way of generating structure.

4.2. Hierarchical clustering theories and scaling laws

It is often customary to assume that the fractional density contrast $\delta(x) = (\rho(x) - \rho_b)/\rho_b$ is a statistically homogeneous, isotropic, gaussian, random field. We discussed the properties of such fields in Section (2.9). We saw that all the information about such a field is contained in its power spectrum $P(k)$. We also worked out there the fractional excess of mass $\delta M/M$ in a sphere of radius r, containing on average a mass M. We showed that $\delta M/M$ is normally distributed with mean zero ($<\delta M/M>= 0$) and a standard deviation

$$\sigma(M) = (<(\delta M/M)^2>)^{1/2} = C \times M^{-(3+n)/6} \tag{4.60}$$

Here we have assumed a power spectrum in the form of a power law $P(k) \propto k^n$, and C is a constant to be fixed by matching the fractional mass excess on some scale to the fractional galaxy number excess, including any possible 'biasing' (see section 3.1).

From the definition of the fractional mass excess, one can easily recognise that it is the same as the $\bar{\delta}_0$ used in the spherical top hat model above; except that in the previous section $\bar{\delta}_0$ was thought of as a fixed number rather than a normally distributed random variable. Taking a particular sphere which has $\bar{\delta}_0 = \nu\sigma$, where ν is the number of standard deviations above the mean of the density peak, and substituting it into equations (4.53) we can express all physical quantities in (4.53) in terms of just the mass of the over dense region. This also gives various laws of how these quantities scale with the mass. We get

$$\begin{gathered}
t_{coll} \propto \nu^{-3/2} M^{(n+3)/4} \\
\rho \propto \nu^3 M^{-(n+3)/2} \\
r_{vir} \propto r_m \propto \nu^{-1} M^{(n+5)/6} \\
v \propto \nu^{1/2} M^{(1-n)/12} \\
T_{vir} \propto \nu M^{(1-n)/6}
\end{gathered} \tag{4.61}$$

In fact simpler scaling arguments using linear theory also give the same scaling laws and it is instructive to outline these. One first notes that δ grows as $t^{2/3}$ in a $k = 0$ model. So at any time $\delta \propto t^{2/3}\sigma(M) \propto t^{2/3} M^{-(3+n)/6}$. One asumes that the fluctuation turns around when $\delta \approx 1$, that is $t_{turn}^{2/3} M^{-(3+n)/6} \propto 1$. So

$t_{turn} \propto t_{coll} \propto (M^{(3+n)/6})^{3/2} = M^{(3+n)/4}$. The density $\rho \propto \rho_b(t_{turn}) \propto t_{turn}^{-2} \propto M^{-(3+n)/2}$, the radius $r_{vir} \propto r_m \propto (M/\rho)^{1/3}$, the virial velocity $v \propto (GM/r)^{1/2}$ and the virial temperature $T_{vir} \propto v^2$. One easily sees that the scaling laws in (4.61) are reproduced. However the ultimate justification for the assumptions made above and any quantitative estimate of the parameters has to come from a specific model like the spherical model.

For $n > -3$, σ decreases with increasing M and equation (4.61) then shows that on the average smaller masses turn around and collapse earlier than larger masses. Structure then grows by the gradual separation and recollapse of progressively larger units. As each unit condenses out it will in general be made up of a number of smaller condensations which have collapsed earlier because of their higher initial density contrast. A hierarchical pattern of clustering then builds up which has been compared and indeed identified with the hierarchical pattern seen in the distribution of galaxies (see Peebles 1980). White and Rees (1978) have argued that as a larger mass collapses its substructure is rapidly erased by the merging and tidal diruption of the smaller masses it contains. In this case the evolution of structure will be self-similar in time with a characteristic clump mass $M_c(t)$ which grows in time according the first equation in (4.61) , that is as

$$M_c(t) \propto t^{4/(n+3)} \propto a^{6/(n+3)}. \tag{4.62}$$

For much larger masses than $M_c(t)$, the fluctuations will still be in the linear regime ; on scales comparable to $M_c(t)$ structure will be turning around and collapsing and will show a hierarchical pattern ; while on mass scales much smaller than $M_c(t)$ structure will be smoothed out by non-linear relaxation effects (White 1982). Also as White (1982) notes the above picture will only be valid for $n < 1$. This is because for $n > 1$ the specific binding energy $\propto v^2$ in (4.61) decreases with M. In this case smaller masses have larger binding energy and cannot be dirupted as the larger mass collapses.

This hierarchical pattern of clustering will of course not hold for $n \leq -3$. This is because the spectrum would have to be cut-off above some wavelength λ_c so as to have a covergent $\sigma(M)$. And then the dominant effect would be due to density fluctuations of size of order λ_c, since these collapse on the average at the same time (for $n = -3$) or before (for $n < -3$) smaller scales. At the other extreme of large n, as we discuss in section 4.6, it may be that one can use the above scaling laws only for $n \leq 1$

Recall from Part 2 that in no theory is the post recombination power spectrum a pure power law. So the above scaling laws can only be applied piecewise, over mass scales where $P(k)$ can be described by a rough power law. For example in Cold Dark Matter theories, the effective $n \approx -2$ on galactic scales. In this case one sees from (4.61) that $M \propto v^4$, a relation which reproduces the observed Faber- Jackson and Tully-Fisher relations of Elliptical and Disk galaxies respectively, assuming reasonably that $M \propto L$, the luminosity. For a more precise comparison one can use the spherical top hat model in combination with the actual CDM power spectrum, and plot M vs v^2 for various ν. Such a plot is given in figure 4 of Blumenthal *et al.* (1984), where they also plot the observed points for different types of galaxies and clusters. It turns out that the CDM theory indeed fares very well on these mass scales.

4.3. Mass functions

Gravitationally bound objects in universe, like galaxies, span a large dynamic range in mass. Let $f(M)dM$ be the number density of bound objects in the mass range $(M, M+dM)$ [usually called the "mass function"] and let $F(M)$ be the number density of objects with masses *greater* than M.

Since the formation of gravitationally bound objects is an inherently non-linear process, it might seem that the linear theory is of no use to determine $F(M)$. This, however, is not entirely true. In any one realisation of the linear density field $\delta_R(\mathbf{x})$, (filtered using a window function of scale R), there will be regions with high density [i.e. regions with $\delta_R > \delta_c$ where δ_c is some critical value slightly greater than unity say]. It seems reasonable to assume that such regions will eventually condense out as a bound object. Though the dynamics of that region will be non-linear, the process of condensation is unlikely to change the mass contained in that region significantly. Therefore, if we can estimate the mean number of regions with $\delta_r > \delta_c$ in a gaussian random field, we will be able to determine $F(M)$.

An approximate way of achieving this is as follows (cf. Press & Schechter 1974) : Let us consider a density field $\delta_R(\mathbf{x})$ smoothed by a window function W_R of scale radius R. We have seen earlier that the probability that this field will have a value δ at any chosen point is

$$P(\delta, t) = \left[\frac{1}{2\pi\sigma^2(R,t)}\right]^{1/2} \exp\left(-\frac{\delta^2}{2\sigma^2(R,t)}\right) \tag{4.63}$$

where

$$\sigma^2(R,t) = \int \frac{d^3k}{(2\pi)^3} |\delta_k(t)|^2 W_k^2(R) \tag{4.64}$$

[To be precise we should always write $\sigma^2(R,t)$ since $\sigma^2 \propto t^{4/3}$ in the matter dominated phase; we will suppress this time-dependence when it is not likely to cause any confusion. We have also set $V = 1$]. As a first approximation, we may assume that the region with $\delta > \delta_c$ (t, t_i) (when smoothed on the scale R at time t_i) will form a gravitationally bound object with mass $M \propto \overline{\rho} R^3$ by the time t [The precise form of $M - R$ relation depends on the window function used; for a step function $M = (4\pi/3)$ $\overline{\rho} R^3$ while for a gaussian $M = (2\pi)^{3/2} \overline{\rho} R^3$]. Here $\delta_c(t, t_i)$ is the critical value needed at time t_i so that $\delta_c \simeq 1$ by the time t. For a flat universe, if we use linear theory to extrapolate the fractional excess density contrast, $\delta_c(t, t_i) \cong (t_i/t)^{2/3}$. Therefore, the fraction of bound objects with mass greater than M will be

$$\begin{aligned} F(M) &= \int_{\delta_c(t,t_i)}^{\infty} P(\delta, R, t_i) d\delta = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma(R, t_i)} \int_{\delta_c}^{\infty} \exp\left(-\frac{\delta^2}{2\sigma^2(R, t_i)}\right) d\delta \\ &= \frac{1}{2} erfc\left(\frac{\delta_c(t, t_i)}{\sqrt{2}\sigma(R, t_i)}\right) \end{aligned} \tag{4.65}$$

where $erfc(x)$ is the complementary error function. The mass function $f(M)$ is just $(\partial F/\partial M)$ and the (comoving) number density $N(M, t)$ can be found by dividing this expression by $(M/\overline{\rho})$. Carrying out these operations we get

$$N(M,t)dM = -\left(\frac{\overline{\rho}}{M}\right)\left(\frac{1}{2\pi}\right)^{1/2}\left(\frac{\delta_c}{\sigma}\right)\left(\frac{1}{\sigma}\frac{d\sigma}{dM}\right)\exp\left(-\frac{\delta_c^2}{2\sigma^2}\right)dM \tag{4.66}$$

Given the power spectrum $|\delta_k|^2$ and a window function W_R one can explicitly compute the right hand side of this expression.

There is, however, one fundamental difficulty with the equation (4.65). The integral of $f(M)$ over all M should give unity; but it is easy to see that, for the expression in (4.65),

$$\int_0^\infty f(M)dM = \int_0^\infty dF = \frac{1}{2} \tag{4.67}$$

This arises because we have not taken into account the underdense regions correctly.

To see this difficulty clearly, consider the interpretation of (4.65). If a point in space has $\delta > \delta_c$ when filtered at scale R, then that point should correspond to a system with mass greater than $M(R)$; this is taken care of correctly by equation (4.65). However, consider those points which have $\delta < \delta_c$ under this filtering. There is a *non-zero* probability that such a point will have $\delta > \delta_c$ when the density field is filtered with a radius $R_1 > R$. Therefore, to be consistent with the interpretation in (4.65), such points should *also* correspond to a region with mass greater than M. But (4.65) ignores these points completely and thus *underestimates* $F(M)$ [by a factor (1/2)].

To correct this mistake, we should replace (4.65) by the relation (Bond *et al.* 1991; Peacock & Heavens 1990)

$$F(M) = \int_{\delta_c}^\infty P(\delta, R)d\delta + \int_{-\infty}^{\delta_c} C(\delta_c, \delta)d\delta \tag{4.68}$$

where the second term represents the probability p_u that a point which has $\delta < \delta_c$ at the filter scale R has the density $\delta > \delta_c$ at a larger filter scale $R_1 > R$. For a sequence of filter scales $R_1, R_2, \cdots R_n$, we obtain a sequence of gaussian random fields parametrised by the dispersions $\Delta_1, \Delta_2 \cdots \Delta_n$. The probability that a point *remains underdense* (i.e. $\delta < \delta_c$) for all these filter scales is given by

$$p_{\text{survive}} \equiv p_s = \int_{-\infty}^{\delta_c} d\delta_1 \int_{-\infty}^{\delta_c} d\delta_2 \cdots \int_{-\infty}^{\delta_c} d\delta_n p_J(\delta_1, \delta_2 \cdots \delta_n) \tag{4.69}$$

where $p_J[\delta_i]$ is the joint probability distribution that the gaussian variables δ_i take the set of values simultaneously. Obviously, $(1 - p_s)$ gives the probability that a point becomes overdense somewhere along the sequence of filterings $(R_1, \cdots R_n)$.

The gaussian variables obtained by different filtering scales, unfortunately, are not independent. We can see that

$$\begin{aligned} < \delta_a\delta_b > &= \int \frac{d^3k}{(2\pi)^3}\frac{d^3p}{(2\pi)^3} W_k(R_a)W_p^\star(R_b) < \delta_k\delta_p^\star > e^{i(\mathbf{k}-\mathbf{p})\cdot\mathbf{x}} \\ &= \int \frac{d^3\mathbf{k}}{(2\pi)^3} W_k(R_a)W_k^\star(R_b)\sigma_k^2 \end{aligned} \tag{4.70}$$

is, in general, non-zero. Hence, calculating (4.69) is a non-trivial task.

We can look upon this process in a different, but equivalent, manner. Consider any one fixed location in space. When the filtering scale is some large value R_1 (with a dispersion Δ_1), let us assume that this point had a density contrast δ_1. When we reduce the scale to R_2, we will have a *new* probability distribution for δ; let the value of density contrast at our chosen point is now be δ_2. As we go through a sequence of filtering scales, R_1, $R_2, \cdots R_n$ (in the *decreasing* order), the density contrast performs a random walk through the points $(\delta_1, \delta_2 \cdots \delta_n)$. Suppose the *first* instance when δ crosses the value δ_c occurs at the k-th step. Then we will attribute chosen point to a mass $M_k \propto \overline{\rho} R_k^3$. Notice that, since $\delta < \delta_c$ for all the *higher* filtering scales - i.e. for all $(R_1, R_2 \cdots R_{k-1})$ - this point *does not* belong to any higher mass. [This takes care of the original difficulty in (4.65)] The random walk concept merely translates into a pictorial form the content of (4.69). This random walk problem is equally difficult to solve because the steps are not independent. In fact the answer will clearly depend on the correlation between the steps; and from (4.70) it follows that the answer will critically depend on the form of the window function.

Since no result which is independent of the form of the window function is possible, one might consider window functions for which the analysis is the simplest. This happens for window functions which are sharply truncated in k-space; that is for $W_k(R) = \theta(R^{-1} - k)$ which acts as a low-pass-filter in k-space (see Bond *et al.* 1991). From (4.70) it follows that for this window function,

$$< \delta_a \delta_{a+1} >= \sigma_a^2; \quad < \delta_a \delta_b >= \sigma_a^2 \quad (\text{for } a \leq b) \tag{4.71}$$

The step lengths of the random walks are $l_1 \equiv (\delta_2 - \delta_1)$, $l_2 = (\delta_3 - \delta_2)$, $\cdots l_a = (\delta_{a+1} - \delta_a)$ etc. Each of these is a gaussian variable with the dispersion

$$< l_a^2 >=< (\delta_{a+1} - \delta_a)^2 >= \sigma_{a+1}^2 + \sigma_a^2 - 2 < \delta_{a+1}\delta_a >= \sigma_{a+1}^2 - \sigma_a^2. \tag{4.72}$$

and *zero* cross correlation:

$$\begin{aligned} < l_a l_b > &=< (\delta_{a+1} - \delta_a)(\delta_{b+1} - \delta_b) > \\ &=< \delta_{a+1}\delta_{b+1} > - < \delta_{a+1}\delta_b > - < \delta_a b_{b+1} > + < \delta_a \delta_b > \\ &= \sigma_{a+1}^2 - \sigma_{a+1}^2 - \sigma_a^2 + \sigma_a^2 = 0 \end{aligned} \tag{4.73}$$

for $(a+1) < b$; other cases can be considered in a similar manner and can be shown to vanish. In other words, sharp filter in k-space produces a random walk in which each step l_a is independent and is drawn from a gaussian variable with dispersion $(\sigma_{a+1}^2 - \sigma_a^2)$. In the continuum limit, this random walk is described by a diffusion equation. The probability $P(\delta, \sigma^2)$ that the particle is at $(\delta, \delta + d\delta)$ when the dispersion is σ^2 obeys the diffusion equation

$$\frac{\partial P}{\partial \sigma^2} = \frac{1}{2}\frac{\partial^2 P}{\partial \delta^2} \tag{4.74}$$

We are interested in the probability that the trajectory reaches (δ, σ) without exceeding δ_c earlier, i.e. at smaller σ. This is equivalent to solving (4.74) with the boundary condition that there exists an absorbing barrier at $\delta = \delta_c$. This is straightforward and the answer is

$$P(\delta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\left[\exp\left(-\frac{\delta^2}{2\sigma^2}\right) - \exp\left(-\frac{(\delta - 2\delta_c)^2}{2\sigma^2}\right)\right] \tag{4.75}$$

Integrating this expression from δ_c to ∞ and differentiating with respect to M, we get

$$dF(M) = \sqrt{\frac{2}{\pi}} \cdot \frac{\delta_c}{\sigma^2} \cdot \left(-\frac{\partial\sigma}{\partial M}\right) \exp\left(-\frac{\delta_c^2}{2\sigma^2}\right) dM \tag{4.76}$$

or

$$N(M)dM = -\frac{\overline{\rho}}{M}\left(\frac{2}{\pi}\right)^{1/2} \frac{\delta_c}{\sigma^2}\left(-\frac{\partial\sigma}{\partial M}\right) \exp\left(-\frac{\delta_c^2}{2\sigma^2}\right) dM \tag{4.77}$$

which is precisely *twice* the value obtained by using (4.65). Of course, the normalisation problem is solved automatically.

This result can be expressed in an explicit form for power law spectra with $|\delta_k|^2 \propto k^n$. In that case $\sigma^2 = c^2 M^{-(3+n)/3} t^{4/3}$ where c is some constant. Since the limit of linear theory occurs when $\sigma \simeq 1$, the characteristic mass scale M_{nl} which goes nonlinear at time t obeys the scaling relation $M_{nl} \propto t^{4/(3+n)}$. Therefore, $c^2 = M_{nl}^{(3+n)/3}(t_0)t_0^{-4/3}$ where t_0 is the epoch at which we need the $N(M)$; say, the present epoch. Since $\delta_c(t_0, t_i) \cong (t_i/t_0)^{2/3}$,

$$\begin{aligned}\frac{\delta_c(t_0, t_i)}{\sigma(M, t_i)} &= \left(\frac{t_i}{t_0}\right)^{2/3} \cdot \frac{M^{(3+n)/6}}{t_i^{2/3}} \cdot \frac{t_0^{2/3}}{M_{nl}^{(3+n)/6}(t_0)} \\ &= \left[\frac{M}{M_{nl}(t_0)}\right]^{(3+n)/6}\end{aligned} \tag{4.78}$$

where $M_{nl}(t_0)$ is the scale which is going nonlinear today. The expression for $N(M, t_0)$ now becomes

$$N(M, t_0)dM = \frac{\overline{\rho}}{\sqrt{\pi}}\left(1 + \frac{n}{3}\right)\left(\frac{M}{M_{nl}}\right)^{(3+n)/6} \exp\left[-\frac{1}{2}\left(\frac{M}{M_{nl}}\right)^{(3+n)/3}\right]\frac{dM}{M^2}. \tag{4.79}$$

A different choice of the window function, in general, will give a different result especially in the low mass limit (Peacock & Heavens 1990).

4.4. Zeldovich approximation

An elegant approximate solution for the nonlinear evolution of density perturbations was proposed by Zeldovich (1970). The starting point of the Zeldovich approximation is the linear theory result for the growth of small perturbations, expressed as a relation between the Eulerian and Lagrangian co-ordinates of fluid particles. We also restrict ourselves to scales which are much smaller than the horizon scale d_H, so that the newtonian approximation applies. Consider first the smooth universe with uniform density $\rho_b(t)$. The actual position of any particle $\mathbf{r}(t)$, is related to its initial (Lagrangian) location $\mathbf{q}$ by

$$\mathbf{r}(t) = a(t)\mathbf{q} \tag{4.80}$$

where $a(t)$ takes account of the expansion of the Universe.

Now consider how this is altered in the presence of growing density perturbations. We know that at least in the linear regime, one can separate out the time dependence of a perturbation from its spatial dependence. And we would like to match with the linear theory for small density contrasts. So to take into account the perturbation it may suffice to add to (4.80) a separable function of t and $\mathbf{q}$ of the form say, $a(t)b(t)\mathbf{p}(\mathbf{q})$. That is we write for the motion of particles in the perturbed universe the equation

$$\mathbf{r}(t) = a(t)(\mathbf{q} + b(t)\mathbf{p}(\mathbf{q})) \equiv a(t)\mathbf{x}(t) \tag{4.81}$$

where we have defined $\mathbf{x}(t)$, the comoving Eulerian coördinate of a particle which has a Lagrangian position $\mathbf{q}$. Although (4.81) gives the evolution of fluid particles which started at various $\mathbf{q}$, it can also be thought as giving a co-ordinate transformation between the $\mathbf{q}$ space and the $\mathbf{r}$ or the $\mathbf{x}$ space ; provided the transformation in invertible. We shall often use (4.81) in this sense below.

In order to find the functions $b(t)$ and $\mathbf{p}(\mathbf{q})$ and also see more clearly how (4.81) describes the evolution of perturbations one has to calculate how the perturbed density evolves when fluid particles move according to (4.81) . Assume that the initial unperturbed density $\bar{\rho}$, is independent of $\mathbf{q}$. Since mass is conserved we have for the perturbed density

$$\rho(\mathbf{r},t)d^3\mathbf{r} = \bar{\rho}d^3\mathbf{q} \tag{4.82}$$

So

$$\rho(\mathbf{r},t) = \bar{\rho}det(\partial q_i/\partial r_j) = \frac{\bar{\rho}/a^3}{det(\partial x_j/\partial q_i)} = \frac{\rho_b(t)}{det(\delta_{ij} + b(t)(\partial p_j/\partial q_i))} \tag{4.83}$$

Here we have used the fact that the background density of the smooth FRW universe $\rho_b(t) = (\bar{\rho}/a^3(t))$. Expanding the Jacobian to first order in the perturbation $b(t)\mathbf{p}(\mathbf{q})$, we get

$$\frac{\delta\rho}{\rho} = \frac{(\rho - \rho_b)}{\rho_b} = -b(t)\nabla_{\mathbf{q}}\mathbf{p} \tag{4.84}$$

On the other hand from linear theory, we found that

$$\frac{\delta\rho}{\rho}(\mathbf{x},t) = g(t)\delta_i(\mathbf{x}) \tag{4.85}$$

where $\delta_i(\mathbf{x})$ is the fractional excess density contrast at an initial time t_i and $g(t)$ is the function describing the time evolution of the growing mode of δ. When $k = 0$, for example, we have $g(t) = \frac{3}{5}(\frac{t}{t_i})^{2/3}$. So we have from matching (4.84) and (4.85) the condition

$$g(t)\delta_i(\mathbf{x}) = g(t)\Sigma A_{\mathbf{k}}exp(i\mathbf{k}.(\mathbf{q} + b(t_i)\mathbf{p}(\mathbf{q}))) = -b(t)\nabla_{\mathbf{q}}(\mathbf{p}) \tag{4.86}$$

Here $A_{\mathbf{k}}$ is the fourier transform of $\delta_i(\mathbf{x})$. Let us we chose t_i such that the term $b(t_i)\mathbf{p}$ in (4.86) can be neglected compared to $\mathbf{q}$. Then we can satisfy (4.86) by identifying $b(t)$ with $g(t)$ and taking

$$\mathbf{p}(\mathbf{q}) = \Sigma\frac{i\mathbf{k}}{k^2}A_{\mathbf{k}}exp(i\mathbf{k}.\mathbf{q}). \tag{4.87}$$

If we do this then (4.81) does indeed reproduce the linear theory result for growth of small density perturbations. It was Zeldovich's remarkable insight to suggest that while (4.81) is in accord with linear theory, it may also provide a good approximate description of the evolution of density perturbations into the ***non linear regime*** where $\delta\rho/\rho$ greatly exceeds unity. The approximation (4.81) is aptly called the 'Zeldovich approximation'.

From the definition of $\mathbf{p}(\mathbf{q})$ given in (4.87) , one also has

$$\mathbf{p}(\mathbf{q}) = \nabla_{\mathbf{q}}\Phi_0(\mathbf{q}) \tag{4.88}$$

where

$$\Phi_0(\mathbf{q}) = \Sigma\frac{A_{\mathbf{k}}exp(i\mathbf{k}.\mathbf{q})}{k^2}. \tag{4.89}$$

From (4.84) we then have

$$\nabla_{\mathbf{q}}.\mathbf{p} = \nabla^2\Phi_0 = -\frac{(\rho-\rho_b)}{b\rho_b} \tag{4.90}$$

Using the Einstein equation $\ddot{a} = -(4\pi G\rho_b a)/3$, we can also write this equation as

$$\nabla^2_{\mathbf{q}}\Phi_0 = \frac{4\pi Ga^2(\rho-\rho_b)}{(3ab\ddot{a})} \tag{4.91}$$

Suppose we compare (4.91) with the equation for the fluctuation in the gravitational potential ϕ in an expanding universe

$$\nabla^2_{\mathbf{x}}\phi = 4\pi Ga^2(\rho-\rho_b) \tag{4.92}$$

at an early enough time such that $\mathbf{x}$ is very nearly equal to $\mathbf{q}$. Then we get

$$\phi = 3ab\ddot{a}\Phi_0 \tag{4.93}$$

So Φ_0 is proportional to the fluctuation of the gravitational potential.

Since $\mathbf{p}(\mathbf{q})$ is a gradient of a scalar function, the Jacobian in (4.83) is a real symmetric matrix and can be diagonalised at every point $\mathbf{q}$, to yield a set of eigenvalues and principal axes as a function of $\mathbf{q}$. If the eigenavlues of $\partial p_j/\partial q_i$ are $-\lambda_1(\mathbf{q})$, $-\lambda_2(\mathbf{q})$ and $-\lambda_3(\mathbf{q})$ then the perturbed density is given by

$$\rho(\mathbf{r},t) = \frac{\rho_b(t)}{(1-b(t)\lambda_1(\mathbf{q}))(1-b(t)\lambda_2(\mathbf{q}))(1-b(t)\lambda_3(\mathbf{q}))} \tag{4.94}$$

Note that in the above equation $\mathbf{q}$ as a function of $\mathbf{r}$ is given by inverting (4.81) , which can be done as long as the Jacobian matrix $(\partial x_j/\partial q_i)$ is non singular.

In an overdense region, equation (4.94) shows that the density will become infinite if one of the terms in brackets in the denominator of (4.94) becomes zero. Zeldovich argued that the eigenvalues will generically be different from each other and in any region of $\mathbf{q}$ space one of them, say λ_1, will be maximum. Then the density goes to infinity first when

$$(1-b(t)\lambda_1(\mathbf{q})) = 0, \tag{4.95}$$

that is when matter in a cube in this region of **q** space gets compressed to a sheet in the **r** space, along the principal axis corresponding to λ_1. At any time t the solution to equation (4.95) , if it exists at all, will define a surface in **q** space and hence generically a surface in **r** space also. Zeldovich therefore argued that sheetlike structures, or 'pancakes', will be the first non-linear structures to form when gravitational instability acts on small density perturbations.

Several interesting features of the Zeldovich approximation are worth commenting upon. Firstly this approximation differs from linear theory in that it predicts the formation of the first nonlinear objects from the high peaks of $\lambda_1(\mathbf{q})$, instead of the peaks of $\delta(\mathbf{q}) \propto (\lambda_1 + \lambda_2 + \lambda_3)$, as in linear theory. Also worth noting is that in spite of the formal similarity of (4.81) to inertial motion of the co-moving co-ordinate of fluid particles, it actually describes gravitational instability in an expanding Universe. If there were no gravity and only expansion, $b(t)$ would infact decrease with time.

Numerical simulations have been employed to test how well Zeldovich approximation works. It is found that at the beginning of the nonlinear stage it gives the general density distribution very well, also reproducing in an excellent fashion the formation, appearence and location of the pancakes (Doroshkevich *et al.* 1980, Efstathiou & Silk 1983). Infact (4.81) is even used by some simulators to set the initial conditions for numerical simulations of large scale structure formation where one wants to start from a rather late stage when δ is not small (say $\approx 0.1 - 0.5$), to save computational time (see Shandarin & Zeldovich 1989). At later times, however, it is found that while Zeldovich approximation predicts the caustics to increasingly blur out and pancakes to thicken, N - body simulations show that pancakes remain relatively thin. It turns out that the pancake thickness quickly stabilizes even in a collisionless medium due to the action of gravity. Particles falling into pancakes oscillate about the middle rather than moving out progressively along the direction of their initial velocity as predicted by (4.81) . Also the N-body simulations show that particles flow along pancakes to form filaments at the intersection of pancakes and finally clumps at the intersection of filaments, whose sharpness is not well reproduced.

In order to overcome some of these problems a number of authors have recently suggested a possible extension to the Zeldovich approximation which we consider in the next section. But before doing this we end the present one by asking why in the first place does the Zeldovich approximation work at all? At least until the formation of caustics, we have described above that (4.81) describes the evolution of density perturbations very well , even when the density contrasts are highly nonlinear. Why this should be so, for after all (4.81) is just an extrapolation of linear theory ? There are several probably equivalent ways of answering this question. First note that even when the term $b\mathbf{p}$ is small compared to **q**, the density perturbations need not be small since $\delta\rho/\rho$ depends on the derivatives of $b\mathbf{p}$ and not on its magnitude. So even if particle positions are not significantly perturbed, the density contrasts can be nonlinear. The Zeldovich approximation exploits this feature by describing the perturbations in terms of the perturbed tragectories of particles extrapolated from linear theory, instead of trying to extrpolate the evolution of density contrasts. Another way of looking at this question is to realise that that the basic quantity which is used in (4.81) is the perturbed newtonian potential, extrapolated from linear theory. The actual potential may not deviate from linear theory much even when the density contrasts become highly nonlinear.

4.5. The adhesion model

In the adhesion model every particle moves in accordance with (4.81) until it runs into another particle. Then they 'stick' and move together with a velocity which conserves their momentum. Such a model can roughly describe the formation of pancakes, filaments and compact clumps. Also by fiat it keeps the pancakes thin while allowing particles to move along the pancakes. Infact, the model allows the pancakes , filaments and clumps themselves to move as a whole and merge with each other. The one disadvantage is that the internal structure of pancakes and other objects which form cannot be inferred from this model.

The mathematical expression of these ideas is worked out in a number of papers by Gurbatov, Saichev & Shandarin (1983, 1985, 1989). An excellent review of these ideas can also be found in Shandarin & Zeldovich (1989). The starting point of this model is Zeldovich approximation (4.81) expressed in a slightly different form.

Let us consider the peculiar velocity of a particle $\mathbf{V}(t) \equiv a(t)d\mathbf{x}/dt$, in the Zeldovich approximation. We have using (4.81)

$$\mathbf{V} = a(t)\dot{b}\mathbf{p}(\mathbf{q}). \tag{4.96}$$

We can also look at the above equation as defining a peculiar velocity field $\mathbf{V}(\mathbf{q}(\mathbf{x}))$ in the $\mathbf{x}$ space as long as the transformation (4.81) is invertible, that is before caustics form. Suppose we define a new velocity variable $\mathbf{v} = \mathbf{V}/(a\dot{b})$, and use a new time variable $b(t)$ instead of t. Then from (4.96) we have

$$\mathbf{v} = \mathbf{V}/(a\dot{b}) = \mathbf{p}(\mathbf{q}) \tag{4.97}$$

Also from (4.81)

$$\mathbf{v} = \frac{1}{a\dot{b}}a\frac{d\mathbf{x}}{dt} = \frac{1}{\dot{b}}\frac{d\mathbf{x}}{dt} = \frac{d\mathbf{x}}{db} \tag{4.98}$$

Since $\mathbf{v}(\mathbf{x}) = d\mathbf{x}/db = \mathbf{p}(\mathbf{q})$, and $\mathbf{p}(\mathbf{q})$ is constant in time t and in 'b' time, the derivative of $d\mathbf{v}/db$ vanishes. So we have

$$\frac{d\mathbf{v}}{db} = (\frac{\partial\mathbf{v}}{\partial b})_{\mathbf{x}} + \mathbf{v}.\nabla_{\mathbf{x}}\mathbf{v} = 0. \tag{4.99}$$

One must keep in mind that this equation is valid only as long as the transformation between $\mathbf{q}$ and $\mathbf{x}$ is nonsingular, for only then can one define a single valued $\mathbf{v}(\mathbf{x})$.

The equation of continuity in the expanding Universe (Peebles 1980)

$$\frac{\partial\rho}{\partial t} + \frac{3}{a}\frac{da}{dt} + \frac{1}{a}\nabla_{\mathbf{x}}.(\rho\mathbf{V}) = 0, \tag{4.100}$$

can also be recast interms of the new variables by defining a new density variable $\eta = a^3\rho(\mathbf{x}, .t)$. We get

$$\frac{\partial\eta}{\partial b} + \nabla_{\mathbf{x}}.(\eta\mathbf{v}) = 0 \tag{4.101}$$

These equations describe the evolution of a force free fluid in 'b' time. Assuming that at small b the density is homogeneous, one can write down its solution in Lagrangian form

$$\mathbf{x} = \mathbf{q} + b\mathbf{p}(\mathbf{q}); \eta = \frac{\eta_0}{Det(\partial x^j/\partial q^i)} \tag{4.102}$$

As expected this is nothing but the Zeldovich solution (4.81) written in terms of comoving co-ordinates, if we choose the initial scaled velocity field $\mathbf{p}$ to be that given in (4.87) .

The equations (4.99) and (4.101) break down at points where many $\mathbf{q}$'s correspond to the same $\mathbf{x}$, that is after caustic formation. Gurbatov, Saichev and Shandarin's suggestion to rectify this problem was to modify (4.99) by adding a viscous term to its right hand side. Such a viscosity term will prevent $\mathbf{v}$ from becoming multivalued and will enable one to follow the evolution beyond caustic formation. Also by fiat it will keep pancakes thin once they form. This model is not intended to describe the evolution of the internal structure of pancakes or other clumps. So the form chosen for the additional viscous term is irrelevant. A particularly simple form for this term can then be chosen, one that makes the resulting equation analogous to a well known equation called Burgers' equation (Burgers 1940, 1974). Thus (4.99) is replaced by the equation

$$\frac{\partial \mathbf{v}}{\partial b} + \mathbf{v}.\nabla_{\mathbf{x}}\mathbf{v} = \nu\nabla^2_{\mathbf{x}}\mathbf{v} \tag{4.103}$$

Note that as $\nu \longrightarrow 0$ the viscosity term does not influence the motion, except in regions where there are rapid variations in velocity, that is at caustics. So in this limit the evolution outside caustics still follows the Zeldovich approximation, while at the same time preventing multistream flows at caustics. We shall see the implications of (4.103) better by solving it. Indeed, the main advantage of using Burgers' equation is that, remarkably, it has an analytic solution for 'potential' motion.

Suppose the velocity field can be expressed as the gradient of a velocity potential Φ, that is $\mathbf{v} = \nabla\Phi$. Note that this can indeed be done at the initial stages as we saw in the last section for the purely growing mode. Then from (4.103) Φ obeys the equation

$$\frac{\partial \Phi}{\partial b} + \frac{1}{2}(\nabla\Phi)^2 = \nu\nabla^2\Phi \tag{4.104}$$

which by means of the substitution

$$\Phi = -2\nu \ln U \tag{4.105}$$

is transformed to the linear diffusion equation

$$\frac{\partial U}{\partial b} = \nu\nabla^2 U \qquad (4.106).$$

The solution to this equation is

$$U(\mathbf{x}, b) = (\frac{1}{4\pi\nu b})^{3/2}\int exp(-\frac{(\mathbf{x}-\mathbf{q})^2}{4\nu b})U(\mathbf{q}, 0)d^3\mathbf{q}, \tag{4.107}$$

From (4.105) $U(\mathbf{q},0) = exp(-\Phi_0(\mathbf{q})/2\nu)$, where $\Phi_0(\mathbf{q})$ gives the initial value of the velocity potential and is the same function defined previously in Section 4.4. Expressing U in terms of Φ, and taking the gradient we finally get

$$\mathbf{v}(\mathbf{x},b) = \frac{\int \frac{(\mathbf{x}-\mathbf{q})}{b} exp(-G(\mathbf{x},\mathbf{q},b)/2\nu)d^3\mathbf{q}}{\int exp(-G(\mathbf{x},\mathbf{q},b)/2\nu)d^3\mathbf{q}} \tag{4.108}$$

where

$$G(\mathbf{x},\mathbf{q},b) = \Phi_0(\mathbf{q}) + \frac{(\mathbf{x}-\mathbf{q})^2}{2b}. \tag{4.109}$$

The solution (4.108) in the limit $\nu \rightarrow 0$ is of particular interest. As ν tends to zero the main contribution to the integral in (4.108) comes from the vicinity of the ***absolute minimum*** of $G(\mathbf{x},\mathbf{q},b)$ treated as a function of $\mathbf{q}$. From (4.109) this point $\mathbf{q}$ is therefore the solution of the equation

$$\frac{(\mathbf{q}-\mathbf{x})}{b} + \nabla\Phi_0 = 0 \tag{4.110}$$

which is nothing but the solution of Zeldovich in comoving coordinates. If $\mathbf{q}_{min}$ is a solution of (4.110) for which G is also an absolute minimum then the 'velocity' $\mathbf{v}(\mathbf{x},b)$ is given from the steepest decent approximation to the integral in (4.108) by

$$\mathbf{v}(\mathbf{x},b) = \frac{\mathbf{x}-\mathbf{q}_{min}(\mathbf{x},b)}{b} \tag{4.111}.$$

At early times, when density contrasts are still small, one expects a unique solution, say $\mathbf{q}_1$, to (4.110) : at every eulerian point $\mathbf{x}$ there is only one particle which has come from $\mathbf{q}_1$. But later on for some $\mathbf{x}$, there may be several roots $\mathbf{q}_1,\mathbf{q}_2,\mathbf{q}_3$, say, which all satisfy (4.110) . This means that several particles from diffrent $\mathbf{q}$ would have all come to the same $\mathbf{x}$ under the 'old' Zeldovich approximation. But for solution to (4.108) , one still gets a unique $\mathbf{q}_{min}$ and hence $\mathbf{v}$, that for which G is an absolute minimum. This is because the other particles which could reach $\mathbf{x}$, if the medium were collisionless, got stuck in pancakes earlier. Ofcourse there will be $\mathbf{x}$'s for which G will have an absolute minimum simultaneously at several points $\mathbf{q}$. This can happen if fluid elements at these $\mathbf{q}$'s have just met at this eulerian point $\mathbf{x}$. The set of all such $\mathbf{x}$'s will then trace out the caustics in eulerian space in the limit $\nu \rightarrow 0$.

An elegant geometrical method then suggests itself to study the solution (4.108) in the limit $\nu \rightarrow 0$. Given an eulerian co-ordinate $\mathbf{x}$ and a 'time' 'b', we construct the paraboloid

$$P(\mathbf{x},\mathbf{q},b) = -\frac{(\mathbf{x}-\mathbf{q})^2}{2b} + h. \tag{4.112}$$

From the discussion above, the coordinate of the absolute minimum of $G(\mathbf{x},\mathbf{q},b)$ is then given by the point $\mathbf{q}$ where this paraboloid is tangential to the hypersurface $\Phi_0(\mathbf{q})$ for the first time as one increases h from $-\infty$. The fact that the paraboloid is tangential to Φ_0, is equivalent to (4.110) and the property that the $\mathbf{q}$ is an absolute minimum is guaranteed by demanding that the paraboloid is tangential for the first time as one increases h from $-\infty$. The Eulerian coordinate $\mathbf{x}$ of the particle in question (with Lagrangian coordinate q) is by construction the coordinate of the top of the paraboloid. At early times when b is small the paraboloid has large curvature ($\propto b^{-1}$), is very narrow and is tangential to the hypersurface at only one point. But as time goes on b increases and the paraboloid becomes shallower. It may then be tangential to Φ_0 at two points for the first time as one keeps increasing h. This is illustrated schematically in figure 4.1. In this case these two points have just run into a pancake. The case of such double 'touching' is degenerate. So the points on the paraboloid tops, when double touching takes place form a set of lower dimensionality, and form sheets in Eulerian space. At later stages the paroboloids may touch the suface Φ_0 at three or even four points. The apices of these paroboloids then indicate the positions of filaments and clumps formed.

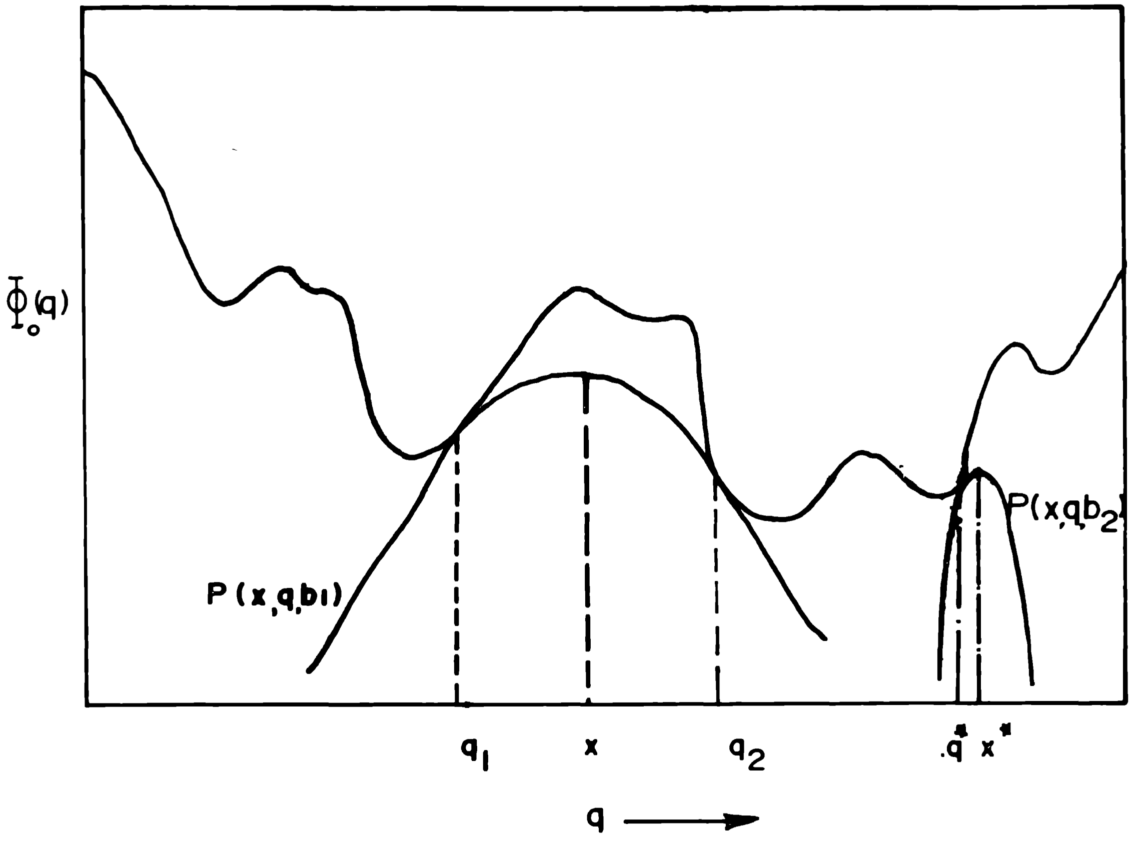

Figure 4.1. Graphical solution to the adhesion model.

Simulations of the evolution of large scale structure have been carried out using the adhesion model, and compared with N -body simulations. Both numerical solution of (4.108) (Weinberg & Gunn 1990 a, b) and simulations using the geometrical technique (Kofman *et al.* 1990; Sahni 1990) have been used to evolve the initial density field. We show for example in figure 4.2 a comparison of the results of a simulation by Sahni (1990) using the adhesion model (Fig. 4.2c and 4.2d) with direct N

- body simulations of Mellot & Shandarin (1989)(Fig. 4.2a), starting from the same initial conditions. We have also shown the result of using the Zeldovich approximation (Fig 4.2b) (see Sahni 1991 for details). We can see quite clearly that the adhesion model reproduces the results of fully nonlinear calculations very well. Infact, because (4.108) has an exact solution, the density field at some arbitrary time can be found without following the evolution for intermediate times, unlike in a N-body simulation. It turns out that efficient algorithms using much shorter computing time, compared to N-body simulations, can be developed to implement the the adhesion model. So one can simulate larger volumes of space provided an approximate treatment of small scale structure is acceptable.

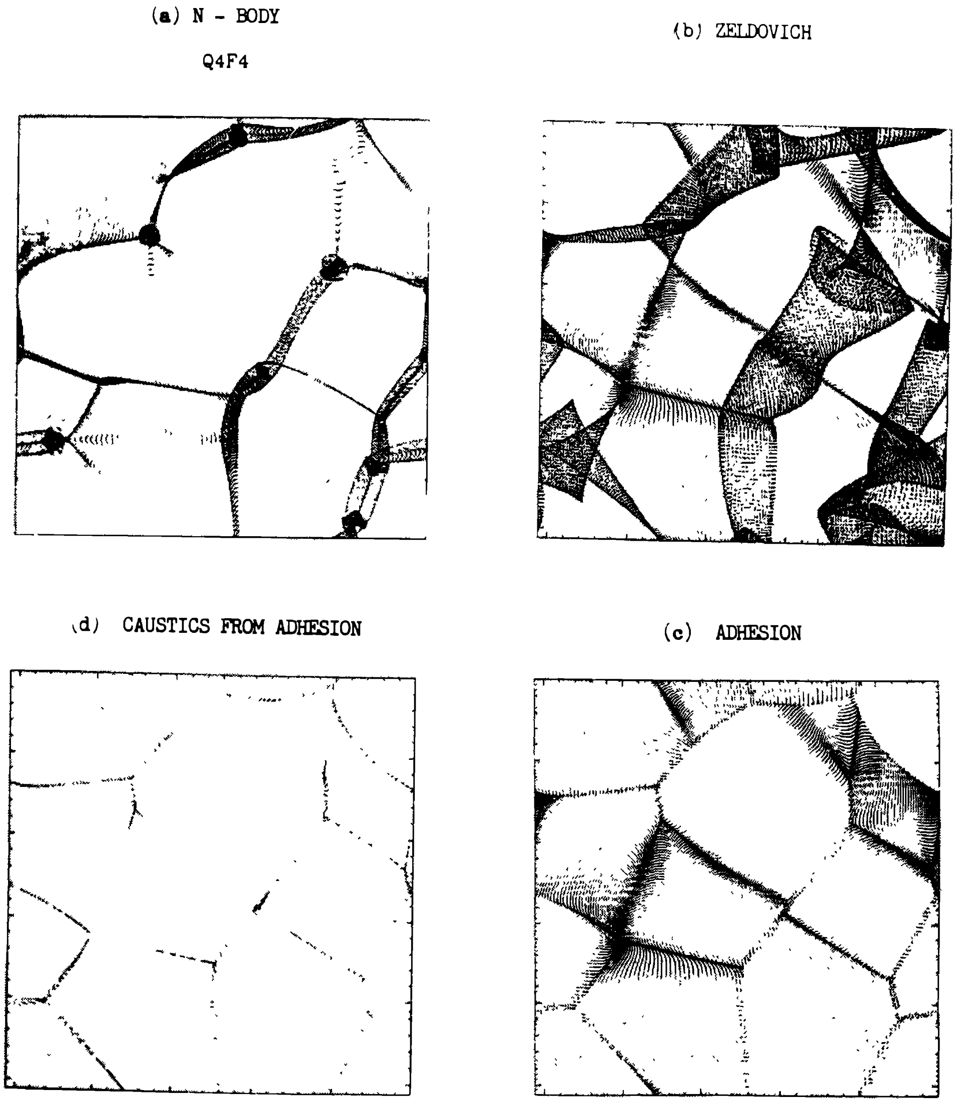

Figure 4.2. Comparison between N - body simulation (a), Zeldovich approximation (b) and the adhesion model (c) and (d), adapted from Sahni (1990, 1991).

Another aspect of this model is worth pointing out. In the adhesion model, as emphasised by Gurbatov, Saichev & Shandarin (1989), the evolution of the density field is completely given once the initial gravitational potential pertubation ϕ ($\propto \Phi_0$) is specified. For gaussian random fluctuations the power spectrum of ϕ is $\propto P(k) \times k^{-4}$, where $P(k)$ is the power spectrum of the density fluctuation. So the ϕ field has more large wavelength (small k) power than the density fluctuation field. This may have important consequences for the way the charecteristic mass M_c, of collapsed objects grows with time. In section 4.2 we saw that for $-3 \leq n \leq 1$ we have the scaling law

$$M_c \propto t^{4/(n+3)} \propto a^{6/(n+3)} \tag{4.113}$$

It was widely believed that the same scaling law would hold even for $1 \leq n \leq 4$. For spectra steeper than $n = 4$, it had been argued that nonlinear generation of the long wavelength part of the spectrum due to mode - coupling terms would dominate any intrinsic large scale power producing an effective $n = 4$ power spectrum even if initially $n \geq 4$. So a limiting growth law for M_c had been proposed for $n \geq 4$ of the form (cf. Peebles 1980)

$$M_c \propto t^{4/7} \propto a^{6/7}. \tag{4.114}$$

Gurbatov, Saichev and Shandarin found on the other hand, a very different scaling law for spectra with $n > 1$, using the adhesion model. Note that at large times $b(t)$ becomes large and the top of the paraboloid becomes very flat compared to the shapes of the peaks and troughs of the initial potential Φ_0. In this limit for any $\mathbf{x}$ the paraboloid is tangential to the hypersurface $\Phi_0(\mathrm{q})$ practically at its local minima and finally at the deepest minima. Thus the asymptotic behaviour of how structure evolves is governed by the statistics and spatial distribution of the deepest minima of the initial potential Φ_0. As b increases the characteristic scale where an absolute minimum can exist, and hence the mass of collapsed structures, will increse depending on the statistics of the minima. In particular, it turns out that (4.113) holds only as long as the variance of the *gravitational potential* diverges, as k tends to zero (see Gurbatov, Saichev & Shandarin 1989). For a pure power law spectrum with a small wavelength cut-off k_{max} the variance in the potential is given by

$$\sigma_\phi^2 \sim \int_0^{k_{max}} k^{(n-4)} k^2 dk \tag{4.115}$$

which diverges at small k only for $n \leq 1$. For $n > 1$, when σ_ϕ is finite, these authors found that M_c grows according to the limiting form

$$M_c \propto a^{3/2}, \tag{4.116}$$

corresponding to the growth law for $n = 1$ in (4.113) .

So the adhesion model clearly predicts a much more rapid growth of structure for steep power spectra compared to an extrapolation of linear theory or simple minded accounting of the effects of mode - coupling. There is some supportive evidence, from one dimensional numerical simulations, that the adhesion model indeed predicts the correct limiting growth law of M_c, that is the validity of the form (4.116) rather than (4.113) for steep spectra (Kotok & Shandarin 1989, Williams *et al.* 1991). Whether this result has any implication for structure formation theories is yet to be explored.

4.6. N - body simulations : the CDM and HDM models

One of the most fruitful and direct method of following the non linear evolution of density perturbations, is via cosmological N - body simulations. This is particularly the case if the universe is dominated by collisionless particles, since the evolution of large scale structure is then driven, to a good approximation, by gravity alone. In this section we discuss some of the results obtained by such simulations, concentrating specifically on two of the well explored theories of galaxy formation, the hot and the cold dark matter models.

In an N - body simulation one approximates the matter distribution as a collection of N particles interacting via gravity. The state of the system at any time t is given by the position and the velocities of the particles. These are evolved over a sequence of small timesteps. For this the force on a particle due to all other $(N-1)$ particles is first calculated. This is used to update the velocities which can then be used to update particle positions. In order to get a good representation of a smooth density distribution one has to make N as large as possible. The main limitation on this comes from the computer time required to calculate the forces. A detailed discussion of the various schemes employed for the force computation is beyond the scope of this review. An excellent book (Hockeny & Eastwood 1981) and several review articles exist (cf. Aarseth 1984, Efstathiou *et al.* 1985). We shall be content with just a brief qualitative discussion of these schemes.

Three different schemes for the computation of forces have been extensively explored so far. Possibly the simplest in concept is the direct summation method, (also known as the 'Particle - Particle' or PP scheme) where the force on a particle is calculated by directly summing the inverse square law forces due to all other particles. In general the optimum strategy is not to use the strict $1/r^2$ law all the way to zero r ; but rather to cut off the force exerted by a particle at a minimum separation. Although such softening of the force results in some loss of spatial resolution, it proves useful in terms of computer time. This is because under a strictly $1/r^2$ force law, the velocities of the particles involved change very rapidly at very small separations. One has to then use correspondingly small timesteps to follow their trajectories accurately, which is costly in computer time. The PP method gives fairly accurate forces but is not suitable for very large number of particles. About N^2 operations are required to evaluate the forces on N particles due to the other $N-1$ particles, and the computer time required then increases very rapidly with N. Clever algorithms using individual particle time steps and a temporal hierarchy of force evaluations can reduce the operation count to order $N^{1.6}$ but still, in practice, it turns out that the PP method is limited to $N \lesssim 10^4$.

A significant gain in N can be obtained if one calculates the forces on particles, not directly, but from a potential got from solving the Poisson equation on a mesh. The Particle - Mesh or the PM scheme adopts such an approach. All field variables like the density or the potential, which are functions over space, are approximated by their values on a regular array of mesh points. Differential operators are replaced by finite difference approximations on the mesh. The potentials and forces at the location of particles are derived by interpolation of the values defined on the mesh. At the same time the densities at the mesh points are got by the opposite process of smearing in a well defined manner the particle mass to a number of nearby mesh points.

Once the densities are specified on the mesh, say as $\rho(l,m,n)$ where (l,m,n) are integers giving the position of a mesh point, one can calculate the potential on the mesh $\phi(l,m,n)$ as the sum

$$\phi(l,m,n) = \Sigma_{\bar{l},\bar{m},\bar{n}} G(l-\bar{l}, m-\bar{m}, n-\bar{n})\rho(\bar{l},\bar{m},\bar{n}). \qquad (4.117)$$

Here $G(l,m,n)$ is the Greens function of the laplacian operator defined on the mesh. Using the convolution theorem, one can .calculate ϕ by first multiplying the finite fourier transforms of G and ρ, and then taking the fourier transform of the product. The main advantage of the PM method lies in the fact that an efficient algorithm, the fast Fourier transform (FFT), can then be employed to work out the potential. Also the Greens function has to be worked out only once. The assigning of the mass of the particles to the mesh points, computing the forces on the particles, updating the velocities and positions all require operations of order $10N$, The operation count for finding the potential on the mesh is found to be of order $5M^3 log_2 M^3$ for a $(M \times M \times M)$ mesh (Hockeny & Eastwood 1988). The total operation count is then of order $10N + 5M^3 log_2 M^3$. So one can see that for a fixed M the PM method wins out in the amount of computer time required compared to the PP method for a large enough N. For example for typical values of $M = 32$ and $N = 10^5$ the number of operations required in a timestep of the PM simulation is of order 4×10^6, compared to a number of order 10^{11} ($\sim 10N^2$) using the PP method. The enormous gain in speed using the PM method is at the cost of a loss in resolution in the evaluation of the force field. This is acceptable when one wants to study problems where the scales over which the potential varies is atleast as large as several mesh lengths. Such a situation can be arranged when studying the nonlinear evolution of a HDM universe, basically because the density fluctuations in such models have no power on small scales due to free streaming (see part 2). Indeed the PM method has been extensively used to simulate the evolution of HDM models.

On the other hand when one has a problem where structure is becoming non linear at about the same time on a wide range of length scales, like in a CDM universe, the PM method has to be modified. This brings us to the third scheme for evaluating the force, the P^3M technique. The P^3M scheme tries to combine the advantages of both the PP and PM methods. The trick is to split up the inter particle forces into two parts ; a short range rapidly varying part due to nearby particles and a slowly varying part due to more widely separated particles. The PP method is used to find the total short range part of the force on each particle and the PM method for the slowly varying force contribution. This is the reason for the name given to this method ; particle - particle / particle - mesh or P^3M scheme. The P^3M sheme has the advantage that it can evaluate the short range force accurately and the long range force rapidly. Its main disadvantage compared to the PM method is the extra time required to calculate short range force contribution by direct summation. This adds a number $\propto N_n N$ to the operation count of the corresponding PM calculation, where N_n is the typical number of neighbours which contribute to the short range force. The P^3M method has been extensively used in simulations of the CDM models. This completes our brief survey of the different N - body schemes which have been used in simulating the non linear evolution of structures. We now turn to discuss the results obtained for two particular theories, the CDM and HDM models.

In the cold dark matter model, recall that one assumes the dark matter to be composed of cold collisionless particles, with negligible random velocities. One also assumes the primordial density fluctuation field to be a random gaussian field with a power spectrum of the Harrison - Zeldovich form. Further there is a theoretical preference for a flat universe, especially if one believes in inflation. The post recombination power spectrum of density fluctuations is ofcourse different from the primordial $n = 1$ power law due to the processes discussed in part 2. It bends gently from the $n = -3$ power law on the subgalactic scales to a $n = 1$ index on scales larger than the horizon scale at the epoch of matter - radiation equality. Structure forms in this model by hierarchical clustering as outlined in section 4.2. However since $\delta M/M$ does not vary very much with the scale length right from subgalactic to cluster scales, there is substantial cross- talk between different scales.

N - body simulations to examine the non linear evolution of the CDM universe have been mainly carried out by Davis, Frenk, Efstathiou and White , and their results can be found in a series of papers (Davis *et al.* 1985; Frenk *et al.* 1985; Frenk *et al.* 1988; White *et al.* 1987a,b). In the first paper of this series, Davis *et al.* (1985) followed the evolution of 32, 768 particles using a P^3M code with periodic boundary conditions. The Zeldovich approximation was used to set up the initial positions and velocities of the particles. They considered an ensemble of models and looked at both the case of a flat universe and the case when $\Omega < 1$. The most important result of this study arose from an examination of the two point correlation function ξ of the mass distribution. Recall that ξ for galaxies is well approximated by a power law of the form $\xi(r) = (r/r_0)^{-1\ 8}$, where $r_0 = 5\mathrm{h}^{-1}\mathrm{Mpc}$. Any model of galaxy formation should be able to reproduce this behaviour. In the simulations with the CDM power spectrum it was found that the non linear evolution leads to a progressive steepening of the two point correlation function. As a result there is only one time when the mass auto correlation function has approxiamtely the same slope as that observed for galaxies. However, at this time, which occurs very rapidly in the simulations (after an expansion factor of only 1.8), the amplitude of ξ is smaller than that observed unless $h \lesssim 0.22$ in a flat universe or the universe is open with $\Omega < 1$. The first possibility is ruled out since the Hubble constant is limited by $0.5 < h < 1$. At the same time many theorists are loath to give up the idea that the universe has $\Omega = 1$; partly from the view point of the inflationary universe model to be discussed in part 6. One way of retaining a flat cosmology in the CDM picture is to invoke a very smooth contribution $\Omega_{smooth} = 1 - \Omega_{CDM} \approx 0.8$ due to some relativistic particle (Turner *et al.* 1984; Olive *et al.* 1985; Padmanabhan & Vasanthi 1987) or – more brazenly – by postulating a cosmological constant of the required magnitude.

A much more popular alternative was first pointed out by by Davis et al. The basic idea was to relax the assumption that galaxies trace the mass distribution. It was suggested that perhaps galaxies could form more readily in regions of higher density, that is around high peaks in a suitably smoothed version of the linear density fluctuation distribution. Such high peaks have an enhanced amplitude of clustering compared to the underlying mass distribution (Kaiser 1985). If one postulates for example, that galaxies form only in peaks with excess fractional density contrasts above 2.5σ (where σ is the rms mass fluctuation), then one can account for both the amplitude and slope of the observed ξ, with $h \sim 0.5$ and $\Omega = 1$. In this model the galaxies form a biased subset of the mass distribution. Considerable effort subsequent to this paper has been spent on how to realise in practice such 'biased galaxy formation'.

The first works in this direction concentrated on radiative and hydrodynamic processes which could supress galaxy formation in halos which collapse late, from low peaks in the density field. (Rees 1985; Silk 1985; see also the review by Dekel & Rees 1987). Later on it has been pointed out that purely gravitational processes may themselves lead to biasing. For example White *et al.* (1987a) found from their N - body simulations that the strength of galaxy clustering was larger for galaxies with a deeper potential well. This enhancement ranges from a factor of order 1.8 for galaxies with circular velocities $v_c > 100\text{kms}^{-1}$ to about 5 for $v_c > 250\text{kms}^{-1}$. This bias basically arises because clumps in higher density regions collapse earlier and accrete faster than similar objects in low density regions. So the typical velocity dispersion and mass of clumps is greater in protoclusters compared to protovoids. White *et al.* speculated that this enhanced clustering of bright galaxies compared to the mass may be sufficient to reconcile a flat CDM universe with observed galaxy correlation function. A somewhat different approach has also been pointed out by Carlberg, Couchman & Thomas (1990), which is termed velocity bias (see also Evrard 1986; West & Richstone 1988). In this approach a bias in the velocities of galaxies compared to the DM mass arises due to the dynamical friction of the galaxies moving through the DM. Such dynamical friction leads to a reduction in galaxy velocities compared to that of the mass on scales up to cluster scales. This reduction implies that any mass estimate on cluster scales using galaxy velocities is likely to be an underestimate. Also the resulting concentration of the galaxies with respect to the mass results in an enhanced amplitude of the galaxy autocorrelation function. One may then be able to reconcile a flat CDM universe with observations of galaxy clustering (Carlberg, Couchman & Thomas 1990).

Adopting the high peak model of biasing N - body simulations have been used to examine how well the CDM model explains a number of other features of galaxies. On galactic scales it turns out that the CDM model does very well ; dark galactic halos are predicted to have flat rotation curves as observed and the correct abundance of dark halos as a function of their potential well depth are obtained (Frenk *et al.* 1988). However it has been somewhat of a controversial issue whether the CDM model predicts equally well the observed large scale structure ; the sheets, filaments and voids seen in the CFA survey (see part 2), the abundance of rich clusters (cf. Teague, Carter & Gray 1990) and their correlation function (Bachall & Soneira 1983, Sutherland 1988), the bulk flows of order 600kms^{-1} on scales of about 50Mpc (Lynden Bell *et al.* 1988). The sympathisers of the CDM model have maintained that filaments, large sheets and voids do arise in the simulations of the CDM model (White *et al.* 1987b) ; that the abundance of rich clusters can be understood if one takes proper account of projection effects (Frenk *et al.* 1990 but see Peebles 1991) ; that bulk velocities may also be explained (Kaiser & Lahav 1989 but see Ostriker & Suto 1990).

More recently more substantive pressure on the CDM model has come from the results of several detailed surveys of galaxies. The APM galaxy survey of Maddox *et al.* (1990) has revealed that the angular correlation function of galaxies has substantially larger amplitude on scales above 20h^{-1}Mpc than predicted in the standard biased CDM model. Further a new redshift survey of IRAS galaxies, known as the QDOT survey has also revealed an amplitude of the galaxy correlation function larger than that expected in the standard biased CDM model (Efstathiou *et al.* 1990; Saunders *et al.* 1991). For example Saunders *et al.* (1990) find the fractional density excess in galaxies on scales of 20h^{-1}Mpc, $\delta\rho/\rho = 0.0669 \pm 0.019$, compared to $\delta\rho/\rho = 0.0192 \pm 0.0013$, which they expect in standard biased CDM model. So these latest detailed surveys cast serious doubts on the CDM model although clearly it is premature to abandon all aspects of the theory altogether.

Let us now turn to the other widely discussed theory of structure formation, the hot dark matter model. The original popularity of the HDM model was due to the fact that there is a natural candidate for HDM, the neutrino. The cosmological relevance of massive neutrinos was pointed out some two decades back by Cowsik & McClelland (1972) and Marx & Szalay (1972). But it was with the report of a mass measurement for the neutrino near the value needed to close the universe that structure formation scenarios with neutrinos as the DM, came into prominence. As we described in part 2, massive neutrinos with mass around 30*ev* are still relativistic when scales upto cluster scales enter the horizon. As a result free streaming of the neutrinos wipes out any primordial fluctuations on scales less than $\lambda_{FS} = 28(m_\nu/30\text{eV})^{-1}$Mpc (see eq. (2.76)). Due to this cutoff the first structure to form in the neutrino dominated universe are cluster mass objects. Also because of the cutoff of small scale power, the initial density and potential field are smooth on scales of order λ_c. Then as we described in section 4.5 the nonlinear evolution proceeds initially according to the Zeldovich approximation via the formation of pancakes, filaments and finally massive clumps. N - body simulations of this evolution have been done by a number of authors using PM codes (cf. Centrella & Mellot 1982; White, Frenk & Davis 1983). The simulations of White, Frenk & Davis (1983) brought out several potential problems with the HDM model, which led to a decline in its popularity.

The basic problem is due to the large value of λ_{FS}. In the HDM model galaxies can only form after the collapse of cluster sized pancakes, by say fragmentation. Suppose one wants galaxies to start forming sufficiently early, say at redshifts z_{form} comparablе to that of the farthest quasars. Then some fraction of the matter must have gone through pancakes by this redshift. For example White, Frenk and Davis defined the onset of galaxy formation as the epoch when one percent of the particles in their simulation had passed through caustics. It turns out that after the formation of the first pancakes, the clustering in an HDM universe proceeds rapidly until most of the mass is in very massive clumps. And unless $z_{form} \lesssim 0.5$, the autocorrelation function of particles which have gone through a caustic (and which are identified as potential galaxies) has a much larger amplitude than the observed ξ of galaxies. The clustering of real galaxies is also significantly weaker than that of the total mass distribution (neutrinos) for any acceptable redshift of galaxy formation.

White, Frenk and Davis also pointed out another difficulty with the HDM scenario somewhat independent of the problem associated with matching the observed galaxy clustering. This was to do with the very massive clusters that resulted in their simulation for what were considered to be reasonable values of $z_{form} \sim 2 - 3$. They argued that such clusters were unlike any known object in the universe ; and the accretion of gas in to their potential wells would produce very large, $\sim 10^{46}$ergs^{-1}, X - ray luminosities and large $\sim$ 45KeV gas temperatures. The existing X - ray observations do not reveal any such source although a large number of them should have been detected (see also White 1986).

One possible weakness in these arguments against the HDM model is the considerable uncertainity in deciding how galaxies are related to the mass distribution. If for example galaxies avoid forming in the dense filaments and clumps then the galxy distribution may be much smoother than that of the DM. Such antibiasing offers one way of reconciling the HDM model with the observed galaxy clustering (Braun, Dekel & Shapiro 1988). One would also have to argue in this picture that the gas was also somehow prevented from falling in to the deep potential wells of the massive clusters. Another possibility within the framework of the galaxies tracing the neutrino mass distribution has been explored by Centrella *et al.*(1988). They identified the present epoch as the time when the two particle correlation function in their simulation matched the observed ξ. This choice means that their model is less evolved compared to White *et al.* model ; the mass is not completely concentrated in gaint clusters at the present epoch. They argue that to form quasars at high redshift it is not essential for one percent of the mass to go through caustics, like White *et al.* assumed. Rather it is sufficient if a small fraction of the mass has become nonlinear with $\delta\rho/\rho \gtrsim 1$. Their model has the first nonlinear structure developing at a redshift of about 7 with most galaxies forming between $z = 7$ and $z = 1$. Also in their way of normalising the HDM spectrum, very massive clusters dont form and the problems with X - ray observations may be avoided. These authors conclude that the present state of ignorance about how galaxies form warrants caution in rejecting the HDM model. It is not clear ofcourse, whether simple nonlinearity in the density contrast is enough to form quasars and galaxies.

We see from the above discussions that, in their present form, neither the CDM nor the HDM model is able to accomodate all the observed features of galaxies. And N - body simulations have played a crucial role in arriving at this conclusion. This was the raison d'etre of this section ! We now turn to the consideration of how some of the basic properties of galaxies may arise, which is not tied down to either of these scenarios.

4.7. The origin of the characteristic mass scales of galaxies

Galaxies have typical masses $\approx 10^{11} M_{\odot}$. Galaxy formation theories based purely on the gravitational instability of density fluctuations in an expanding universe do not appear to provide any natural explanation for this charecteristic mass. Is there any physics which naturally selects galaxy mass scales ? We address this question in the present section. As we will see the cooling of gas may provide a possible answer.

The importance of gas cooling in setting galaxy scales was first hinted at by Hoyle (1953) and analysed further in the now classic papers of Rees & Ostriker (1977); Silk (1977) and Binney (1977). Suppose we consider a gas cloud of mass M and radius R, which is supported against gravitational collapse by gas pressure. The gas then has a typical temperature T given by $kT \approx GM\mu/R$. Such a cloud is on the verge of gravitational collapse, its further evolution being governed by the relation of its cooling timescale

$$t_{cool} \approx \frac{3\rho kT}{2\mu\Lambda(T)} \tag{4.118}$$

to its dynamical or freefall timescale

$$t_{dyn} \approx \frac{\pi}{2}(\frac{2GM}{R^3})^{-1/2} \tag{4.119}.$$

Here ρ is the average cloud density and $\Lambda(T)$ gives the cooling rate in units of say ergcm^{-3}s^{-1}. Note that we have taken t_{dyn} to be the freefall time of a uniform density sphere (any other definition will give similar results for the rough arguments which follow).

There are three possibilities. Firstly if t_{cool} is greater than the hubble time,t_{hubble}, that is the cosmic time elapsed since the redshift of cloud formation, the cloud will not have evolved much. On the other hand if $t_{hubble} > t_{cool} > t_{dyn}$, the gas can cool but as it cools the cloud can retain pressure support by adjusting its pressure distribution on a sound crossing timescale $\approx t_{dyn}$ which is less than t_{cool}. In this case the collapse of the cloud will be quasi - static on a timescale of order t_{cool}. Finally there is the possibility that $t_{cool} < t_{dyn}$. In this case the cloud can cool 'rapidly' (compared to its dynamical timescale) to a minimum T. It will then lose pressure support and undergo an almost freefall collapse. Also fragmentation into smaller units can occur, because as the collapse proceeds isothermally, smaller and smaller mass scales become Jeans unstable.

Rees & Ostriker (1977) and Silk (1977) suggested that it is the criterion $t_{cool} < t_{dyn}$, which sets the mass scale of galaxies. For only then will gravitating gas clouds collapse appreciably and also possibly fragment into stars. Further in any hierarchical theory of galaxy formation, unless a gas cloud cools on a dynamical timescale and becomes appreciably bound, collapse on a larger scale will disrupt the object (White & Rees 1978). In such theories galaxies are the largest masses which have resisted such disruption by being able to satisfy the above criterion. White & Rees (1978) further extended this argument to the case when some form of collisionless dark matter provides the dominant mass component of a galaxy. Before we come to this, it is instructive to examine the original case of evolution without DM in greater detail.

The cooling of primordial gas is mainly due to bremsstrahlung, hydrogen and helium recombinations, and Compton scattering of hot electrons by the colder cosmic background photons. Compton cooling turns out to be important only at high redshifts > 10 (see below). For the present we consider only the case when galaxy scales become nonlinear at redshifts below $\approx$ 10, coming back later on to discuss the case when Compton cooling dominates. The cooling rate can then be written as

$$\Lambda(T) = (A_B T^{1/2} + A_R T^{-1/2})\rho^2 \text{ergcm}^{-3}\text{s}^{-1}, \tag{4.120}$$

where the A_B term represents the bremsstrahlung contribution and the A_R term that due to recombination line cooling. Note that the (4.120) is only valid for temperatures above $\sim 10^4 K$: for lower temperatures the cooling rate drops drastically since hydrogen can no longer be significantly ionised by collisions. The cooling rate for a plasma in thermal equilibrium has been calculated by Raymond, Cox & Smith (1976) and one can use this in (4.118) to calculate t_{cool}. For a hydrogen plus helium plasma with a helium abundance $Y = 0.25$ and some admixture of metals, a very usefull approximate expression for the resulting t_{cool} has been given by Peacock & Heavens (1990). In a slightly modified form we have for a $T > 10^4 K$,

$$t_{cool} = \frac{8 \times 10^6}{n_0(T_6^{-1/2} + 1.5 f_m T_6^{-3/2})} \mathrm{yr}. \tag{4.121}$$

Here n_0 is the gas number density in units of $1\mathrm{cm}^{-3}$, T_6 the temperature in units of $10^6 K$ and f_m takes into account the possibility that the gas may be enriched with metals : $f_m \approx 1$ for no metals and $f_m \approx 30$ for solar abundance (Peacock & Heavens 1990). The first term in the denominator represents the effect of bremsstraulung while the second takes account of line cooling. For gas with primordial abundance ($f_m \approx 1$), one can see from (4.121) that there is a transition temperature $T^* \approx 10^6 K$. For temperatures above T^* bremsstrahlung dominates the cooling while line cooling dominates below T^*.

Now consider the ratio $\tau = t_{cool}/t_{dyn}$. This ratio is fixed once any two parameters of a cloud is given, say ρ and T. The condition $\tau = 1$ defines a curve on the $\rho - T$ plane, which demarcates the region of parameter space for which cooling occurs rapidly within a dynamical time from the region of slow cooling (see Figure 4.3 below). For $T < T^*$ when line cooling is dominant, we have $t_{cool} \propto T^{3/2}/\rho$ and $t_{dyn} \propto \rho^{-1/2}$. So $\tau \propto T^{3/2}/\rho^{1/2} \propto M_J$, the Jeans mass for this temperature and density. (We define the Jeans mass to be $M_J = (4\pi/3)\rho(\lambda_J/2)^3$, where the Jeans wavelength $\lambda_J = (\pi kT/\mu G\rho)^{1/2}$). The $\tau = 1$ curve will then be parallel to lines of constant Jeans mass in the $\rho - T$ plane, for $T < T^*$. Putting in numbers and using the cooling time from (4.121) we get

$$\tau = \frac{t_{cool}}{t_{dyn}} \approx \frac{M_J}{10^{12} M_\odot}; \quad T < T^* \tag{4.122}$$

We see that the Rees- Ostriker - Silk criterion for efficient cooling can be satisfied for masses below a critical mass $\approx 10^{12} M_\odot$, if $T < 10^6 K$.

On the other hand for $T > T^*$, when bremsstraulung dominates the cooling, $t_{cool} \propto T^{1/2}/\rho$ and $t_{dyn} \propto \rho^{-1/2}$. So $\tau \propto T^{1/2}/\rho^{1/2} \propto \lambda_J$, the Jeans wavelength. The curve $\tau = 1$ for temperatures much higher than T^* will then be parallel to lines of constant Jeans wavelength in the $\rho - T$ plane. In fact, using the t_{cool} in (4.121) , and defining the radius associated with the cloud to be $R_J = \lambda_J/2$, we get

$$\tau = \frac{t_{cool}}{t_{dyn}} \approx \frac{R_J}{70\mathrm{kpc}}; \quad T > T^* \tag{4.123}$$

Therefore high temperature clouds have to shrink below a critical radius of about 70kpc before being able to cool efficiently to form galaxies.

These features are illustrated schematically in figure 4.3, which we will refer to as a " cooling diagram ". (In the figure we have used n instead of ρ). Such diagrams have been given by Rees & Ostriker (1977) and Silk (1977) and they are useful in visualising much of the physics of cooling gas clouds. We have indicated there three regimes A,B and C. A gas cloud with constant mass evolves roughly along lines of constant M_J, with $T \propto \rho^{1/3} \propto n^{1/3}$, if pressure supported. Gas clouds in region A have $t_{cool} > t_{hubble}$ and never cool much. Those in B cool slowly and undergo quasistatic pressure supported collapse until they enter region C where $\tau < 1$. Gas clouds in C can cool efficiently to form galaxies. We saw above that to lie in this region they have to have a mass below $\approx 10^{12} M_\odot$ or shrink to a radius below $\approx$ 70kpc. Rees & Ostriker (1977) and Silk (1977) made the interesting point that these masses and radii compare rather well with the scales characteristic of galaxies. Over the decade since these arguments were given, theories of galaxy formation have varied greatly according to the current fashion. But some essence of the above ideas remains in most theories as a hint to explain galaxy masses.

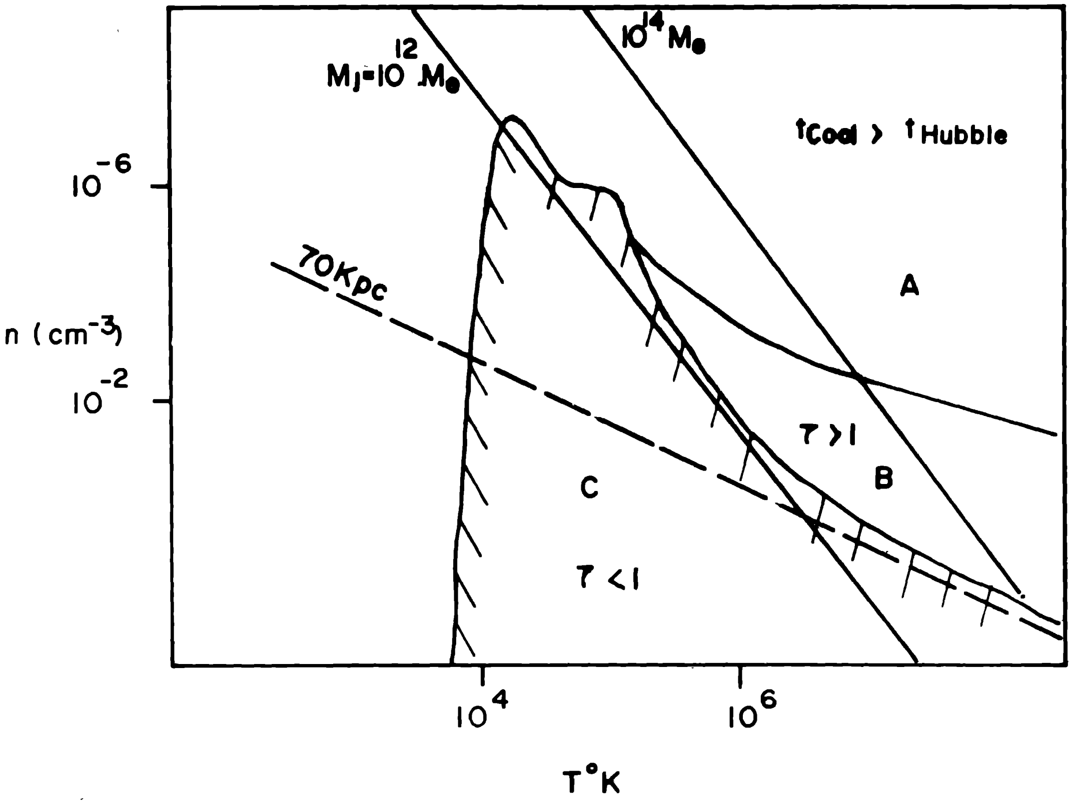

Figure 4.3. A schematic cooling diagram.

We now consider, following White & Rees (1978), the effects of incorporating a gravitationally dominant DM component and that of the expansion of the Universe. In the cosmological setting one starts by considering a cloud of DM and gas which is initially expanding, until gravitational instability leads to a 'turn around' and collapse. The dynamical time scale is determined by the total, DM + gas density, whereas the cooling time still depends only on the density of the gas. Also, in this case the gas will not initially be at the virial temperature. It is only during collapse that the gas gets heated up by shocks formed when different bits of gas run into each other. If the cooling timescale of the shocked gas is larger than the dynamical timescale in which the cloud settles down to an equilibrium, then the gas will eventually get heated up to the virial temperature. On the other hand were the cooling time shorter, the gas may never reach such a pressure supported equilibrium. Efficient cooling would result in the gas sinking to the centre of the forming DM potential well, until it is halted by rotation or fragmentation into stars. We see that once again it is the ratio of the cooling to the dynamical time of the object which governs its evolution. Further, in hierarchical clustering theories, we pointed out earlier that smaller mass clumps are disrupted as larger masses turn around and collapse. However if the gas component can cool efficiently enough, it may shrink sufficiently to the centre of the DM potential well to resist disruption and so break the hierarchy. In these theories, galaxies are thought to be the largest masses that have survived hierarchical clustering.

The spherical top hat model can be used to estimate the relevant dynamical timescale. We assume t_{dyn} to be roughly comparable to $t_{coll}/2$, the time taken for a spherical top hat fluctuation to collapse after turn around (section 4.1). Infact this is the same expression as the t_{dyn} given in (4.119) above, if we identify R there with the radius of turn around r_m. We have

$$t_{dyn} \approx \frac{t_{coll}}{2} \approx 1.5 \times 10^9 (\frac{M}{10^{12} M_\odot})^{-1/2} (\frac{r_m}{200\mathrm{kpc}})^{3/2} \mathrm{yr}. \tag{4.124}$$

For estimating the cooling timescale, we use (4.118), and assume that the gas makes up a fraction F of the total mass and is uniformly distributed within a radius $r_m/2$. The gas temperature is taken to be of order the virial temperature obtained in the spherical model, that is $T_{vir} \sim (\mu v^2/3k)$, where $v^2 \sim 6GM/(5r_m)$. This corresponds roughly to heating by shocks with a velocity of order the collapse or the virial velocity. We then have

$$t_{cool} \approx 2.6 \times 10^9 f_m^{-1} (\frac{F}{0.1})^{-1} (\frac{M}{10^{12} M_\odot})^{1/2} (\frac{r_m}{200\mathrm{kpc}})^{3/2} \mathrm{yr}. \tag{4.125}$$

Here we have further assumed that the line cooling dominates for T_{vir} relevant to galaxies and adopted a typical value of $F \sim 0.1$. Note that the collapse in general is likely to be highly inhomogeneous, and the above estimates are only to get a rough idea of the numbers involved. From (4.124) and (4.125) we can once again estimate that

$$\tau = t_{cool}/t_{dyn} \approx 1.7 f_m^{-1} (\frac{F}{0.1})^{-1} (\frac{M}{10^{12} M_\odot}) \tag{4.126}$$

So efficient cooling with $\tau < 1$ implies the condition

$$M < M_{crit} \approx 5.7 \times 10^{11} M_\odot f_m (\frac{F}{0.1}) \tag{4.127}$$

It is pleasing that masses of order galactic masses are once again preferentially picked out when one includes DM and takes into account the cosmological setting.

We now have the machinery to be able to apply the above ideas to any particular theory of structure formation, especially those involving hierarchical clustering. The starting point to understand much of the galaxy formation physics is to plot the cooling diagram. Firstly one plots the curve in the $n_b - T$ plane on which $\tau = 1$. Here n_b is the baryon density assumed to be a fraction , say 0.1 of the total density. Furthermore, given the power spectrum of density fluctuations, one can work out $\bar{\delta}_0 = \nu\sigma(M)$ (see section 4.2). Then the various properties, like n_b and T, of the collapsed objects which form, can be estimated using for example the spherical top hat model. We saw in section 4.1 that these properties depend only on one parameter M, once the density contrast δ_0 is fixed. So for each ν one gets a curve on the $n_b - T$ plane, giving the properties of collapsed objects. These curves assume that the proto condensations have virialised, but that the gas has not cooled and condensed. Cooling moves points on these curves to higher densities. In the same diagram one can also plot for comparison the observed positions of galaxies , groups and clusters of galaxies.

Such a cooling diagram for the once popular cold dark matter theory, mentioned in Section 2, is given schematically in figure 4.4. The figure is an adaptation of the cooling diagram given by Blumenthal *et al.* (1984). It indicates that while galaxies indeed show evidence of having cooled and condensed within their their dark halos, dwarf spheroidals are only marginally able to cool and groups and clusters of galaxies have too long a cooling time to have dissipated much of their energy. From the diagram one can also see that roughly for mass scales with $10^8 M_\odot < M < 10^{12} M_\odot$ gas of primordial composition can cool within the dynamical timescale. We have already discussed the upper limit. The lower limit comes from the fact (mentioned earlier) that the cooling rate drops drastically below about $10^4 K$, when hydrogen can no longer be significantly ionised by collisions.

Some complicating features which affect the above simple ideas, deserve mention. Firstly, we have ignored star formation and its feedback effects on the gas. If star formation is very efficient, the supernovae from the massive stars may provide an important heat input. It may then drive out the gas if the potential well is shallow enough (Dekel & Silk 1986). Infact such effects may be crucial in preventing the baryons from being all locked up in small objects, before typical galaxies form (White & Frenk 1991). Also, we see from (4.126) that if the gas were enriched with metals, much larger masses can cool within a dynamical time because of the increased cooling rate. So the chemical history of the gas could also be important.

One may wonder at this stage whether cooling is relevant in setting galactic scales in theories like the Pancake or Explosion pictures of structure formation ; where large masses form first and then fragment into smaller masses. Silk & Norman (1981) have argued that such considerations are still relavent, if the fragments are subgalactic, as indeed seems to be the case in pancake theories. The coagulation of subgalactic clouds to form galaxies leads to a picture not very different from inhomogeneous protogalactic collapse. The condition of efficient cooling of the shocks formed when two clouds collide then leads to a condition equivalent to $\tau < 1$. (Silk 1983; Efstathiou & Silk 1983).

We turn now to a consideration of the effect of Compton cooling, which we have ignored so far. The cooling rate of a gas with electron density n_e and temperature T in a blackbody radiation field of density ρ_r and temperature T_r is given by

$$\Lambda_{comp} = \frac{4\sigma_T n_e ck(T - T_r)\rho_r}{m_e c^2} erg\mathrm{cm}^{-3}\mathrm{s}^{-1}. \qquad (4.128)$$

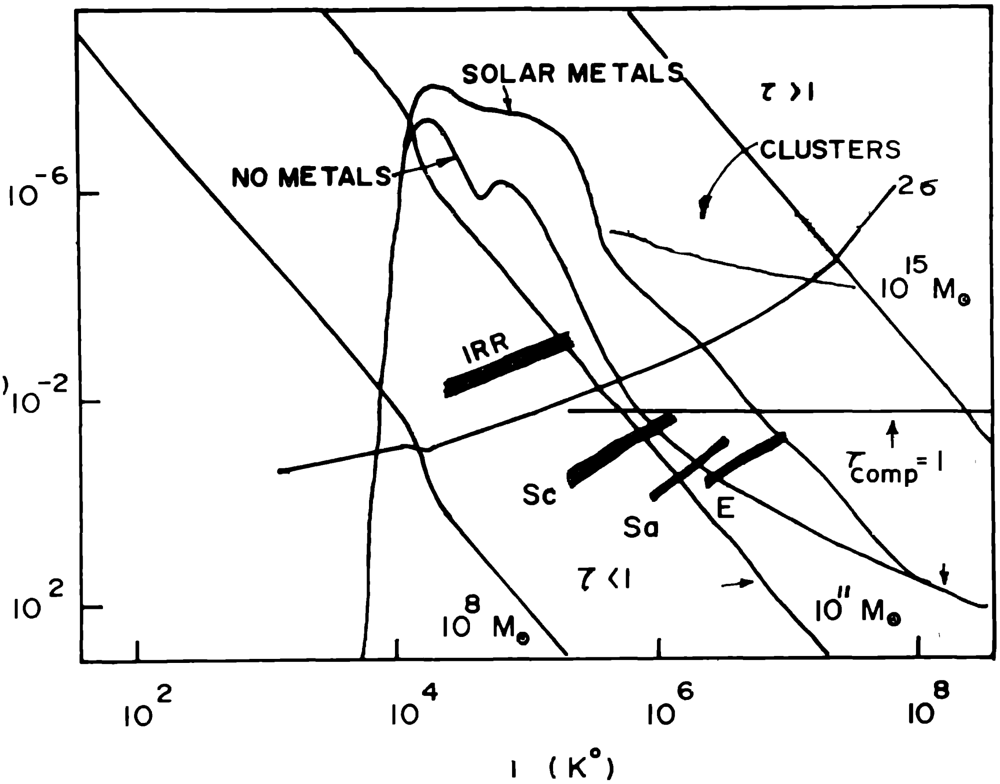

Figure 4.4. Schematic cooling diagram for the CDM model adapted from Blumenthal *et al.* (1984).

Here σ_T is the Thompson cross section, m_e the mass of the electron and c the velocity of light. The cooling time due to inverse compton scattering off the cosmic background photons, using (4.118), is therefore

$$t_{comp} = \frac{3m_p m_e c(1+z)^{-4}}{8\mu\sigma_T\rho_{r0}} \approx 2.1 \times 10^{12}(1+z)^{-4}\text{yr}. \tag{4.129}$$

Here we have assumed $T >> T_r$ and used $\rho_r = \rho_{r0}(1+z)^4$ to take into account the expansion of the Universe.

Suppose we compare this time with the dynamical time in (4.124) using the expression for t_{coll} in (4.21). We then get

$$\tau_{comp} = \frac{t_{comp}}{t_{dyn}} \approx 240(1+z_{coll})^{-2.5}, \tag{4.130}$$

which is less than unity for a $z_{coll} > 8$, independent of the mass of the object. So compton cooling can efficiently cool an object if it collapses before a redshift ~ 10, independent of its mass. It is not clear whether galaxies can collapse that early, but if they do then galaxy scales can not be preferentially picked out because of the cooling processes outlined above. However an interesting feature emerges if one plots the the $\tau_{comp} = 1$ line in the cooling diagram. Note that this line is parallel to the T axis since τ_{comp} does not depend on T. It turns out that galaxies and clusters are neatly separated by the Compton cooling line, suggesting that galaxy formation ceased when Compton cooling became inefficient (Gunn quoted in Blumenthal *et al.* 1984 and see also Gunn 1982). If galaxies could form this early, this would be provide a qualitatively different reason, from that of Rees & Ostriker (1977) and Silk (1977), for their charecteristic masses.

4.8. The origin of the angular momentum of galaxies

Another key parameter of galaxies apart from their masses is their angular momentum content. One important class of galaxies, the disks, owe their equilibrium to rotational support. The origin of galactic angular momentum is therefore a very important question which any theory of galaxy formation must address. At present the most popular idea is that galxies aquire their angular momentum due to the tidal toques of their neighbours (Hoyle 1949; Peebles 1969; Doroshkevich 1970; White 1984). Most of this section is devoted to examining this possibility. At the end we also consider another possible mechanism by which galaxies can aquire angular momentum, one which is particularly relevant for 'top - down ' theories, like the pancake picture of galaxy formation.

The growth of spin in a protogalaxy, due to tidal torques, is most easily analysed, in the linear stages, using the Zeldovich approximation of section 4.4. Ofcourse, in the case when the density field has lot of power on scales much smaller than galaxies, these scales become nonlinear much before galaxies form and Zeldovich approximation will cease to describe the evolution of galaxies. To avoid this formal problem one must apply Zeldovich's equation (4.81) not to the actual density field but to a field smoothed on protogalactic scales. One is assuming here that small scale nonlinear structures have negligible influence on the gravitational evolution of larger quasilinear scales, an assumption which seems reasonable but needs further justification.

The spin angular momentum of the mass which will form a galaxy is given by

$$\mathbf{L}(t) = \int_V (\mathbf{r}(\mathbf{q},t) - \bar{\mathbf{r}}(t)) \times (\mathbf{V}(\mathbf{q},t) - \bar{\mathbf{V}})\rho(\mathbf{r},t)d^3\mathbf{r}, \qquad (4.131)$$

where $\mathbf{r}$ and $\mathbf{V}$ describe the proper position and the peculiar velocity of a mass element as before (see section 4.4). The integral is over the region which will eventually end up forming the galaxy, $\bar{\mathbf{r}}$ is the centre of mass and $\bar{\mathbf{V}}$ the centre of mass peculiar velocity of the system at time t. Using equation (4.26) for conservation of mass, one can convert the integral to one over the lagrangian volume, say V_L, which initially contained the galaxy mass. Substituting the Zeldovich solution (4.25) for $\mathbf{r}$ we then get

$$\mathbf{L}(t) = a^2 \int_{V_L} [([\mathbf{q} - \bar{\mathbf{q}}] + b(t)[\mathbf{p}(\mathbf{q}) - \mathbf{p}(\bar{\mathbf{q}})]). \times \dot{b}(\mathbf{p}(\mathbf{q}) - \mathbf{p}(\bar{\mathbf{q}}))] (\bar{\rho} d^3\mathbf{q}). \qquad (4.132)$$

The cross product of the b and the $\dot{b}$ terms vanish because they are parallel. Since $\mathbf{p}(\mathbf{q}) = \nabla\Phi_0(\mathbf{q})$, we get

$$\mathbf{L}(t) = \bar{\rho} a^2 \dot{b} \int_{V_L} (\mathbf{q} - \bar{\mathbf{q}}) \times (\nabla\Phi_0(\mathbf{q}) - \nabla\Phi_0(\bar{\mathbf{q}}))d^3\mathbf{q}. \qquad (4.133)$$

We see that the angular momentum aquired is first order in the perturbations. Also its time evolution is determined by the behaviour of $a^2\dot{b}$, since the integral and $\bar{\rho}$ are constant with time. For a spatially flat Universe with $b(t) \propto a(t) \propto t^{2/3}$, the $a^2\dot{b} \propto t$, and so the angular momentum grows linearly with time.

Further insight into the meaning of (4.133) can be got if we assume that Φ_0 within V_L, can be approximated by the first three terms of its Taylor series about $\bar{\mathbf{q}}$

$$\Phi_0(\mathbf{q}) = \Phi_0(\bar{\mathbf{q}}) + (q_i - \bar{q}_i)\frac{\partial\Phi_0}{\partial q_i}|_{\mathbf{q}} + \frac{1}{2}(q_i - \bar{q}_i)\frac{\partial^2\Phi_0}{\partial q_i \partial q_j}|_{\mathbf{q}}(q_j - \bar{q}_j). \qquad (4.134)$$

Putting this in equation (4.133)we have

$$L_i(t) = a^2\dot{b}\epsilon_{ijk}I_{jl}T_{lk}, \qquad (4.135)$$

where

$$I_{jl} = \int_{V_L} (q_j - \bar{q}_j)(q_l - \bar{q}_l)\bar{\rho}d^3\mathbf{q}$$

is the moment of inertia tensor of the mass in V_L, and

$$T_{lk} = \frac{\partial^2\Phi_0}{\partial q_l \partial q_k}$$

is proportional to the tidal gravitational field at $\bar{\mathbf{q}}$.

The tensor product in (4.135) is of the usual form used to calculate the torque on an extended body in a tidal field. **L** vanishes if and only if the tensors T_{ij} and I_{ij} have the same principal axes. Since the inertia tensor depends only on the shape of the protogalaxy, while the tensor T_{ij} depends in addition on the distribution of neighbouring protoclumps, this is not expected to happen in general. Equation (4.135) then shows that the angular momentum of a galaxy arises in first order due to the coupling of the first order tidal field with the zeroth order quadrapole moment of the irregular boundary of the protogalaxy (White 1984). This result was derived two decades back by Doroshkevich (1970), but it was only relatively recently that it has been clearly elucidated in the western literature by White (1984).

Although the above calculation to derive (4.135) provides some insight into how **L** originates due to tidal torquing, it cannot be used in a staightforward way to get good numerical estimates. The problem is basically twofold : Firstly, it is difficult to decide how V_L should be specified so that it encompasses all the matter that ends up in a collapsing protogalaxy. Secondly, one has to take account of the growth and exchanges of **L** during the fully nonlinear stages of protogalactic evolution. At present the best way of addressing both these problems seems to be via N - body simulations (Barnes & Efsthathiou 1987), though Heavens & Peacock (1988) make a brave, fully analytical attempt starting from (4.135). Before discussing the results of these papers, it is instructive to look at rough estimates of **L** using (4.135), and dimensional arguments.

Consider a galaxy scale fluctuation of mass M, which attains a maximum size $\sim R$ before turning around and collapsing. From (4.135) the angular momentum aquired by this object initially grows as t and therefore it experiences a constant tidal torque ($d\mathbf{L}/dt$), at least during the quasilinear stages of growth. After turn around the torque on the protogalaxy is expected to decrease. This is because collapse will lead to a decrease in its moment of inertia and also cosmological expansion will result in the protogalaxy moving away from its neighbours, and hence a decrease in the tidal force. It then seems reasonable to assume that the maximum torque on the protogalaxy arises when it is near about the time of turn around. We can then derive a rough measure of the angular momentum aquired by extrapolating the constant torque got from differentiating (4.135), to the the turn around time and multipling it by the time elapsed till turn around. We get

$$L \sim (\frac{GM}{R^3})^{-1/2} \times (\frac{GM}{R^3}) \times (MR^2) \sim MR^2(\frac{GM}{R^3})^{1/2} \tag{4.136}$$

where first term in (4.136) represents the time at turn around, the second the tidal potential while the third term describes the moment of inertia. (Recall from section 4.4 that the gravitational potential $\phi = 3\ddot{a}ab\Phi_0$, atleast in the quasi linear stages.) We can compare the rotational frequency aquired by a mass element

$$\omega \sim L/MR^2 \tag{4.137}$$

with that required for rotational support against gravity, say ω_0, got from the relation

$$\omega_0^2 R \sim GM/R^2. \tag{4.138}$$

From (4.136), (4.137)and (4.138)we have

$$\frac{\omega}{\omega_0} \sim 1 \tag{4.139}$$

So tidal torques could in principle give a L which is dynamically important. Ofcourse in the above argument we reduced all mass scales to a single M and length scales to just the radius at turnaround. So it should not be surprising that ω/ω_0 is of order unity. In general this would not be the case since the mass scale and length scale involved in the expression for the torque is different from that of the galaxy under consideration, and more detailed calculations would be needed to derive ω/ω_0 and acess the significance of tidal torques.

Such a detailed study using N - body simulations has been carried out most recently by Barnes & Efstathiou (1987). A somewhat complementary, analytical approach has also been explored by Heavens & Peacock (1988). In all such studies, the angular momentum is usually given in terms of a convenient parameter,

$$\lambda \equiv \frac{LE^{1/2}}{GM^{5/2}}, \tag{4.140}$$

whose significance we briefly discuss (Narlikar 1983).

Consider a self gravitating system of mass M, energy E and angular momentum L. Let it have a charecteristic radius R. We can once again define a characteristic angular frequency of the system ω by the relation (4.137) and also a hypothetical angular frequency ω_0, that is needed to support the system purely by rotation using (4.138). We have for the ratio

$$\frac{\omega}{\omega_0} \sim \frac{(L/MR^2)}{(GM/R^3)^{1/2}} \sim \frac{L}{M^{3/2}G^{1/2}R^{1/2}}. \tag{4.141}$$

Since the energy of the of the system $E \sim \frac{GM^2}{R}$, we can eliminate R in (4.141), in terms of E and M. We get

$$\frac{\omega}{\omega_0} \sim \frac{L}{M^{3/2}G^{1/2}}(\frac{E}{GM^2})^{1/2} \equiv \lambda. \tag{4.142}$$

So a self gravitating system with appreciable rotational support has a λ comparable to unity.

The N - body simulations of Barnes & Efstathiou (1987) and the analytical work of Heavens & Peacock (1988) both indicate a broad distribution of λ values for collapsed objects identified with the DM halos of galaxies, ($\sim 10^{-1} - 10^{-2}$) with a median $\lambda \approx 0.05$. In their simulations Barnes and Efstathiou looked at both white noise initial conditions ($n = 0$) and initial conditions which have more power on large scales, as predicted by models of galaxy formation with cold DM. They verifed that $L(t) \propto t$, when density contrasts are small compared to unity, in accord with the predictions of linear theory. As clumps become nonlinear, their L grows more slowly and can even decrease at later times, depending upon their substructure. For example, if an object has significant substructure, it initially aquires more angular momentum because of a stronger coupling to the tidal field. But subsequently as the subclumps in the object sink to the centre and merge, they lose their orbital angular momentum to the outer parts, resulting in a high density, low L core.

Barnes and Efstathiou also found that the median λ for the collapsed objects is quite insensitive to the shape of the initial power spectrum of density fluctuation, a result corraborated by Heavens and Peacock from their analytical estimates. Another important question (which we come back to in section 4.9), investigated by these authors was whether the angular momentum aquired by an object is systematically dependent on the magnitude of its initial overdensity ? The answer was negative. Heavens and Peacock explained the results in a nice way : Higher peaks have a shorter collapse time. So one may have thought that they will have less time to get spun up. But, it turns out that higher peaks are also more clustered and experience a stronger torque. Which effect is dominant depends on the power spectrum. For Cold DM spectra these two effects nearly cancel. Also for all power spectra the large spread in the angular momenta aquired washes out any systematic dependence on peak height.

The net result can be summarised as follows. Tidal torques can give angular momentum to a protogalaxy. But both analytical estimates and N - Body simulations indicate that the resulting λ distribution has a large spread with a median value ~ 0.05, more or less independent of the shape of the initial spectrum of density fluctuations, or the peak height. Tidal torques are therefore able to give at best 5 to 10%, of the angular momentum needed for rotational support. How then can we explain disk galaxies ? We will come to this question in the next subsection.

Before this it is interesting to discuss briefly another, completely different way by which galaxies may aquire their angular momentum. This was also analysed first by Doroshkevich (1973) who proposed that in pancake theories the galactic angular momentum can arise due to vorticity generation in shock waves. Note that in theories of structure formation which invoke gravitational instability one generaly assumes that the initial perturbations were irrotational. (Any vorticity component of the perturbation decays with time. So to get significant vorticities at the present requires a much larger vorticity at recombination which is in general incompatible with the smoothness of the cosmic background radiation). It is well known that as long as viscous forces are negligible, vorticity is conserved by fluid motion (Kelvin's theorem). However this breaks down in the presence of shocks. In theories like the HDM theory, the first structure to form are pancake like caustics. Shock waves will go through the gaseous component as the pancake develops. In general the pancake will be curved and the velocity vector of the gas will be oblique to the normal vector to the shock surface. In this case one knows that the normal component of the velocity of the gas will be much reduced, while the tangential component remains unchanged.

Suppose at some point on the shock surface we fix our axes such that the $x-axis$ is normal to the shock and the velocity vector lay on the $x - y$ plane. Also let the vorticity be zero for the gas, before it enters the shock wave. Then initially

$$\partial v_x/\partial y|_i = \partial v_y/\partial x|_i \tag{4.143}$$

After passage through a strong shock we have

$$v_x \approx 0, \partial v_x/\partial y \approx 0; \quad \partial v_y/\partial x|_f \approx 4\partial v_y/\partial x|_i. \tag{4.144}$$

The factor of 4 takes account of the gas compression in a strong shock. So the z component of the vorticity is about $4\partial v_y/\partial x|_i$. Doroshkevich pointed out that when galaxies fragment from the shocked pancake of gas this vorticity would lead to a net angular momentum of the galaxy. His order of magnitude estimates of the resulting specific angular momentum gave a number of the same order as seen in disk galaxies, about $10^{30}\mathrm{cm}^2\mathrm{s}^{-1}$. As he himself noted more detailed calculations are neccessary to establish the validity of this mechanism.

4.9. Formation of disk galaxies

The previous two sections have been devoted to general ideas on whether it is possible to account for two of the basic properties of galaxies, their characteristic mass and angular momentum. In this and the subsequent section we consider more specific details of the formation of different types of galaxies. Galaxies are divided generally using the Hubble classification system (Binney & Tremaine 1987), into four broad categories : ellipticals, lenticulars, disks and irregulars. Further the first three types of galaxies are subdivided to form a sequence known as the Hubble sequence, from early type, round elliticals to late type disks. However when one takes a broad look, one comes to the conclusion, that at the zeroth order, there are two basic types of galactic systems whose origin one should try to understand, the disks and the ellipticals. This section is devoted to the disk galaxies and the next to the formation of ellipticals.

The first basic type, the spiral or disk galaxies, are basically galaxies whose mass lies in a thin disk. They are supported against gravity by rotation, having a value of the dimensionless parameter $\lambda \sim 0.4 - 0.5$. The rotation velocity of the disk is nearly constant in radius exept near the centre where it goes to zero. Typical rotation velocities range from $200 - 300\text{kms}^{-1}$. A remarkable feature of many disk galaxies is the fact that their rotation speed curves remain flat even at radii well beyond the visible galaxy. This implies that these galaxies may be embedded in large halos of dark matter, which dominates gravitationally outside the visible disk. The surface brightness of the disk, and hence its surface density, displays a typical exponential profile of the form

$$I(r) = I_0 exp(-r/r_c)$$

, where I_0 is the central surface brightness, typically about $170 L_{\odot} pc^{-2}$ and r_c the disk scale length, has a typical value of $\sim 3\text{h}^{-1}\text{kpc}$. Ofcourse the most visible feature of the disks are their spiral arms, which is what gives them their name. The abundance of disk galaxies is sensitively dependent on environment. It is observed that in low density regions of the Universe, almost 80% of the galaxies are disks, while this fraction drops to about $\sim$ 10%, in dense regions such as the cores of rich clusters of galaxies. How does one account for all these properties in any theory of galaxy formation ?

Most of the work on accounting for the detailed properties of disks has been done in the context of hierarchical theories of galaxy formation, although some ideas may also apply more generally. We shall have to keep this in mind in what follows. Firstly in the hierarchical picture, when one incorporates a dominant DM component, it is relatively straightforward to understand the luminous core - extended dark halo structure of galaxies. The dark halo is supposed to form as a galactic scale fluctuation in the DM component grows, turns around, collapses, virialises and settles into an equilibrium. Meanwhile the gas in the halo also collapses with the halo, and can attain a pressure supported equilibrium if heated up by shocks, to the virial temperature of the system. However gas can cool radiatively and so collapse further into the core of the dark halo. Initially since the gas is supposed to make up only a small fraction ~ 0.1, of the total mass, the gravity of the halo dominates that of the gas. At this stage, star formation is expected to be suppressed by the tidal forces of the halo, except in regions of large gas density (White & Rees 1978; Faber 1982). But as the gas sinks to the halo core, it comes to be dominated by its own self gravity and can fragment to form stars, resulting in a luminous galaxy in the core of an extended dark halo.

Dissipative collapse in a dark halo also seems to be crucial to understand the rotation of disk galaxies. Recall from the last section that tidal torques give only 5 – 10% of the angular momentum required for rotational support, with a spin parameter $\lambda \sim 0.05$. This is far below the observed value of the spin parameter $\lambda_d \sim 0.4 - 0.5$. How then can one reconcile the tidal torque theory with observations ? One possibility is for the gas to collapse due to cooling, increasing its binding energy, at the same time conserving its mass and angular momentum. Since the spin parameter $\lambda \propto |E|^{1/2}$, it will increase as the binding energy increases. As we show below this does not work unless we invoke massive dark halos.

For this, suppose we were to ignore the presence of massive dark halos around galaxies, and assume that a protogalaxy was just a self gravitating cloud of gas without any DM. The binding energy of the protogalaxy will be $|E| \sim GM^2/R$, where R is its charecteristic radius. Since M is constant during collapse, $|E| \propto R^{-1}$ and so $\lambda \propto R^{-1/2}$. The gas cloud then has to collapse by a factor $\sim (\lambda_d/\lambda_i)^2 \approx (0.5/0.05)^2 \sim 100$, before it can spin up enough to be rotationally supported, where λ_i is the initial value of λ given by tidal torques. Therefore to form a typical rotationally supported galactic disk of mass $\sim 10^{11} M_{\odot}$ and radius $\sim$ 10kpc, matter needs to collapse from an initial radius $\sim$ 1Mpc to 10kpc. From the expression for the collapse time given in (4.124) , this would take an inordinately long time $t_{coll} \sim 10^{11}$yr, much longer than the age of the Universe. Note that even the matter within a scale length $r_c \sim$ 3kpc, would need to collapse from $\sim$ 300kpc and would still take a time $\sim 2 \times 10^{10}$yr.

This timing difficulty is easily avoided if one invokes massive halos. In the presence of a massive dark halo, the spin parameter of the system, before collapse of the gas, can be written as

$$\lambda_i = \frac{L|E|^{1/2}}{GM^{5/2}}. \tag{4.145}$$

Here the the various quantities, L, E and M refer to the combined DM - gas system, although the gas contribution is negligible compared to that of the DM halo. After collapse the gas becomes self gravitating and one gets for the spin parameter of the resulting disk galaxy

$$\lambda_d = \frac{L_d|E_d|^{1/2}}{GM_d^{5/2}}, \tag{4.146}$$

where the parameters now refer to the disk. So

$$\frac{\lambda_d}{\lambda_i} = (\frac{L_d}{L})(\frac{|E_d|}{|E|})^{1/2}(\frac{M_d}{M})^{-5/2} \tag{4.147}$$

The energy of the virialised DM - gas system, assuming the gas has not yet collapsed can be written as

$$|E| = k_1\frac{GM^2}{R_c}, \tag{4.148}$$

while that of the disk is given by

$$|E_d| = k_2\frac{GM_d^2}{r_c}. \tag{4.149}$$

Here R_c and r_c are the characteristic radii associated with the DM -gas and the disk systems respectively, while k_1, k_2 are constants of order unity which depend on the precise density profile and geometry of the two systems. The ratio of the binding energy of the collapsed disk to that of the DM - gas system is then

$$\frac{|E_d|}{|E|} = \frac{k_2}{k_1}(\frac{M_d}{M})^2(\frac{r_c}{R_c})^{-1}. \tag{4.150}$$

Also one expects that the total angular momentum per unit mass aquired by the gas destined to form the disk should be the same as that of the DM. This is because all the material in the system generally experience the same external torques before the cooling of gas separates it into two distinct components. Assuming that the gas roughly conserved its angular momentum during the collapse to form the disk we then have

$$\frac{L_d}{M_d} = \frac{L}{M} \tag{4.151}$$

So the ratio

$$\frac{\lambda_d}{\lambda_i} = (\frac{M_d}{M})(\frac{k_2}{k_1})^{1/2}(\frac{M_d}{M})(\frac{R_c}{r_c})^{1/2}(\frac{M}{M_d})^{5/2} = (\frac{k_2}{k_1})^{1/2}(\frac{R_c}{r_c})^{1/2}(\frac{M}{M_d})^{1/2} \tag{4.152}$$

where we have used (4.150) and (4.151) to simplify (4.147). The gas originally occupied the halo before collapsing, so had a precollapse radius of $\sim R_c$. So the collapse factor of the gas is

$$\frac{R_c}{r_c} = (\frac{k_1}{k_2})(\frac{M_d}{M})(\frac{\lambda_d}{\lambda_i})^2. \tag{4.153}$$

We see that the required collapse factor for the gas to attain rotational support has been reduced by a factor $\sim M_d/M$, from that needed without a dominant dark halo. For a typical galaxy with a halo mass ~ 10 times the disk mass, one needs a collapse only by a factor of ~ 10 before the gas can spin up sufficiently to attain rotational support.

The above dimensional arguments can be somewhat improved if one assumes realistic models of the halo and the disk. For example suppose we take an exponential disk with a surface density

$$\Sigma(r) = \frac{M_d}{2\pi r_c^2} exp(-\frac{r}{r_c}). \tag{4.154}$$

Let us also assume that its rotational velocity v_m is constant with radius. Then $L_d = 2r_c v_m M_d$ and from the defenition of λ_d,

$$\lambda_d = k_2^{1/2}(\frac{r_c}{GM_d})^{1/2} v_m \tag{4.155}$$

Putting this in (4.153) and simplifying, we have

$$\frac{R_c}{r_c} = \frac{1}{\lambda_i} 2k_1 v_m (\frac{M}{|E|})^{1/2} = \frac{1}{\lambda_i}\frac{2\sqrt{2}k_1 v_m}{v}, \tag{4.156}$$

where v is the virial velocity v of the DM - gas system defined in section 4.1. Note that the constant k_2 has dropped out of (4.156). So far we have no assumptions about the halo properties. Suppose we take the halo to originate initially from a spherical top hat fluctuation, we can identify R_c with r_{vir} in section 4.1 and take $k_1 = 3/10$. To relate v_m and v we can assume that far out in the halo the rotation curve is detemined only by the gravity of the halo, implying $v_m = \sqrt{\frac{2}{3}}v$. In this case we get for the collapse factor

$$\frac{r_{vir}}{r_c} = \frac{1}{\lambda_i}(\frac{2\sqrt{3}}{5}). \tag{4.157}$$

We see that collapse factors of order 10 are once again implied.

Detailed computations, using a more realistic rotation curve and including the self gravity of the disk, have been made by Fall & Efstathiou (1980). They also use a stronger conservation law than the conservation of total L_d of the gas, that each element of the gas conserves its angular momentum from the time it begins to collapse. They can then use the disk properties to deduce the halo mass and the gas collapse factor. They deduce that collapse factors $\sim$ 10 and halo to disk mass ratios of five or more are required to explain the rotation of disk galaxies, agreeing substantially with the above more rough estimates.

The above picture of dissipative collapse of gas in the potential of a dominant DM halo, has been further elaborated upon by Gunn (1982), in an attempt to explain the systematic properties of disk galaxies. An added ingredient in this work is the hypothesis that the collapse to form the disk occurs over a prolonged period of time, and may even be continuing till the present time. Gunn pointed out that density peaks inevitably have 'tails', which collapse much later than the material around the peak. As a mass shell turns around and collapses conserving its angular momentum, it will settle into a disk at a radius when rotation can balance gravity. If the angular momentum per unit mass increases with radius, as successive shells around a density peak collape, the disk will grow secularly inside out.

Gunn suggested a nice explanation for why disks are exponential which deserves discussion here. Mestel (1963) had noted a long time back that disks of spirals have an angular momentum distribution which are not very different from that of a uniformly rotating constant density sphere

$$m(h) = M[1 - (1 - h/H)^{3/2}]. \tag{4.158}$$

Here $m(h)$ is the mass with specific angular momentum h or smaller and H is the maximum value of h. Mestel had examined what kind of self gravitating disks have such an angular momentum distribution. He found that there are at least two solutions. A uniformly rotating disk with a very flat density profile, and another which is quite centrally concentrated and which was strongly differentially rotating (called popularly as Mestels disk). Gunn considered the effect of embedding the disk in an isothermal dark halo. He asked what would be the suface density distribution of a disk, which obeys (4.158), has a flat rotation curve and is supported by rotation against both its own self gravity and the gravity of the halo ? He showed that this surface density distribution could be fitted very well by an exponential profile over 3 to 4 scalelengths !

This led Gunn to suggest that if the original angular momentum distribution resembled that of a uniformly rotating sphere and if it was conserved approximately during collapse, then the disk formed would have a characteristic exponential form. Also as succesive shells of mass fell in, this form would be preserved except for an increase in the scale length of the exponential. Gunn also showed how many of the systematic properties of disk galaxies like the kind of spiral structure they display, can be explained as a consequence of their differing accretion rates. He also pointed out that in dense clusters the infall may be totally cut off, since the formation of the cluster would itself have heated up all the intracluster gas to the virial temperature of the group. This could be one reason for the deficiency of spirals in dense regions.

To summarise we see that the idea of disk galaxies forming by dissipative collapse in the potential of a massive dark halo has the potential of explaining many of their observd features. The situation is not at all so clear for the ellipticals as we shall see in the next section.

4.10. The enigma of elliptical galaxy formation

We now turn to consider the formation of the second basic type of galaxy, the ellipticals. Elliptical galaxies are systems of stars which are basically supported against gravity by stellar random motions. Their projected luminosity profiles are smooth and are well approximated by de Vaucouleurs' $r^{1/4}$ law

$$I(r) = I(0)exp(-kr^{1/4}) \equiv I_e exp(-7.67[(\frac{r}{r_e})^{1/4} - 1]), \qquad (4.159)$$

where r_e is the effective radius, which encloses half the total luminosity and I_e is the surface brightness at r_e. The effective radius is typically $\sim$ 3h^{-1}kpc for bright ellipticals and is smaller for fainter galaxies. Another successful fitting formula is what is known as the Hubble - Reynolds law $I(r) = (I_0 r_H^2)/(r + r_H)^2$, where the radius r_H is typically $0.1 r_e$. Ellipticals owe their name to the fact that their isophotes are in general elliptical, and they are classified according to the degree of ellipticity of their isophotes. It was earlier thought that ellipticals were oblate spheroids and that their flattening was due to rotation. But ever since the now classic work of Binney (1978), it has been realised that ellipticals rotate too slowly for this explanation to work. Rather it is now believed that ellipticals are triaxial and owe their shape to anisotropic velocity dispersions. Interms of the spin parameter introduced in section 4.7 the slow rotation of ellipticals implies a $\lambda \sim 0.05$. The abundance of elliptical galaxies is also sensitive to the environment and infact the sense of this dependence is opposite to that of spiral galaxies. In low density regions elliptical galaxies form only $\sim$ 10% of the population while in rich clusters their abundance rises to $\sim$ 40%. How does one account for these properties of ellipticals in any theory of galaxy formation ? And what decides if a given protogalctic fluctuation becomes an elliptical or a spiral galaxy ?

A concept which has been at the foundation of many models to explain the density profiles and the relaxed appearence of elliptical galaxies is that of 'violent relaxation', and it is worth mentioning it atleast briefly. As described above elliptical galaxies have remarkably similar density profiles, which points to some common relaxation process. The relaxation time t_R for two body encounters between stars to relax a galaxy of N stars is about $N(\ln N)^{-1}$ times larger than the dynamical timescale of the system. For a galaxy with $\sim 10^{11}$ stars this time is very much larger than the age of the Universe. So some other process is involved which leads to a relaxed elliptical galaxy. Lynden-Bell (1967) suggested a very interesting mechanism to solve this problem. He argued that during the collapse of a protogalaxy, one would expect large fluctuations in the gravitational potential, in a time of the order of the collapse time. Since the potential is changing, individual stars do not follow energy conserving orbits. Clearly the change in the energy of a star depends in a complex way on its initial position and velocity, but the net effect is to widen the range of energies of stars in a time scale of the order the collapse time, which is much shorter than t_R. In this way a time varying potential provides a relaxation mechanism which has been termed violent relaxation, since it operates on the relatively short dynamical or collapse timescale of the system. Violent relaxation and another process known as phase mixing (see for example Binney & Tremaine (1987)) are thought to be the key ingredients in many models of dissipationless collapse for the formation of ellipticals.

Historically, these ideas about violent relaxation encouraged Gott (see for example the review by Gott (1977)) to suggest that ellipticals and spiral bulges were formed by dissipationless collapse of stars at high redshift, while spirals formed later on with considerable dissipation. Gott (1977) promoted a picture where stars formed rapidly in dense protogalaxies even before maximum expansion, and then violently relax to form ellipticals and spiral bulges. He needed secondary infall to reproduce the observed de vaucouleur profiles. He argued that disk galaxies result when star formation is inefficient and the left over gas gathers in a disk, perpendicular to the angular momentum vector. At about the same time models of ellipticals were also made, by Larson (1975), where dissipation also played a crucial role. In these models the time scale of dissipation, collapse and star formation were all assumed to be comparable. Larson also reproduced the observed light profile and the debate was whether ellipticals formed by dissipationless or dissipative collapse. However both these sets of models suffered from a crucial flaw, on hindsight. The problem was that in both sets of models the ellipticity of the galaxy was explained as due to flattening by rotation. However, as we discussed above, this is not correct and these models are no longer fashionable as they stand.

The works of Gott and Larson however brought to the forefront the idea that the rate of star formation is likely to be a crucial parameter in any scheme to explain the appearence of the different types of galaxies. It seems that to form a thin disk, it is important that star formation does not exaust the gas supply. On the other hand in ellipticals star formation must have been efficient enough, so that the stars formed before the gas could settle into a disk. However the question which has to be addressed is what decides the star formation rate ? Further, there may be other parameters which are also crucial in understanding the origin of ellipticals and spirals. One such crucial parameter is likely to be angular momentum, to which we now turn.

We saw in section 4.7 that tidal torques lead to a median value of the spin parameter $\lambda \sim 0.05$. This value is ideal to explain the observed rotation of elliptical galaxies. One needs to assume however that galactic scale fluctuations collapse without much dissipation. To explain both the typical mass and the size of ellipticals it may be neccessary that this collapse occured rather early. For example, from the spherical top hat model to get $r_{vir} \sim 10$kpc and $M = 10^{11} M_{\odot}$, one needs $\delta_0 \sim 21.2$ and the collapse to occur at redhifts of ~ 11.6. It is not clear whether such early galaxy formation can be reconciled with MBR isotropies. Also since ellipticals are not thought to be dominated by DM atleast within their luminous radii, such a theory of galaxy formation will not be able to accomodate DM clustered on galactic scales. Finally how can one then understand disk like systems which have λ comparable to unity ?

On the other hand if one argues as we did in the last section that galaxies form by dissipative collapse in the potential well of a massive dark halo, then λ will increase to a value of order unity. Although this is what we want for explaining disks, we do not want λ to increase for the case of ellipticals. So in this picture, which explains so nicely many facets of disk galaxies, it is a mystery why rotation is not dynamically important for ellipticals also ? It is not obvious as yet how to understand both ellipticals and disks in a unified manner. We discuss below some of the ideas which have been put forward to solve this problem.

One of the most widely discussed idea is that all galaxies first formed as disk galaxies and ellipticals then form from mergers between disk systems. This idea popularised originally by Toomre (1977), has been a source of considerable controversy. There are some obvious positive features of the merger hypothesis. Firstly galaxy mergers are seen to occur in the local universe and probably were more frequent in the past. Secondly a number of ellipticals which have smooth de vaucouler type light profile, do show signs that they have experienced mergers. (cf. Schweizer 1982, 1986). These telltale signs include tidal tails, shells and in some cases gas disks inclined to the principal planes. Furthermore numerical simulations of mergers between galaxies show that the resultant starpiles generally resemble ellipticals. Their density profiles closely match the density profile implied by the Hubble - Reynolds law, $\rho \propto r^{-3}$ (White 1979; Negroponte & White 1983). Also the disks which are merging will in general have their spins randomly oriented with respect to each other. Then the spin angular momentum of the merger remanent can be considerably smaller than the progenitor disks ; especially if several disk galaxies are involved in the merger and the orbital angular momentum is negligible Fall (1979). So over all it does seem eminently reasonable that some ellipticals can result from mergers.

The merger hypothesis also has several potential problems.

1. Ellipticals are more abundant in rich clusters. However in such clusters the galaxies have velocities $\sim 1000 \mathrm{kms}^{-1}$, hardly conducive to mergers.
2. Special orbits may be needed for disks galaxies firstly to merge and also for their orbital angular momentum not to lead to an excessive rotation for the merger remnant.
3. In a dissipationless merger, if energy and mass are nearly conserved, the energy per unit mass or velocity dispersion or equivalently the depth of the potential well of the remnant will be similar to that of the disks. However ellipticals have much deeper potential wells than typical disks. How can mergers lead to this increase in the binding energy ?
4. Some ellipticals have higher core phase space densities than found in any disk galaxy. How can a dissipationless merger, increase the phase space density ?
5. Ellipticals display metallicity luminosity correlations and metallicity gradients, which are difficult to obtain from mergers of purely stellar disks.
6. Ellipticals have systematically more globular clusters per unit luminosity (or mass) than spiral disks. Van den Berg defines a parameter S which is the number of globular clusters in a galaxy per $M_V = -15$ of the parent galaxy luminosity. He finds that S is an order of magnitude smaller in disk systems than it is in ellipticals (Van den Bergh 1990).

Some of these problems may be understood by refining and modifying the original merger hypothesis, to include the effects of hierarchical clustering, dark matter and dissipation.

The first problem, for example, can be resolved in theories of galaxy formation involving hierarchical clustering. We saw in section 4.2 that in such theories the velocity dispersion increases with mass as long as the power spectrum of density fluctuations is shallower than a $n = 1$ power law over the relevant scales. So in small mass sub clusters the random bulk velocities of galaxies will be smaller than in a rich cluster. In such an environment merging can take place more easily to form ellipticals. Since these subclusters merge to form rich clusters, the rich clusters will also have ellipticals resolving the first objection (White 1982).

Some of the other potential problems (items 2 and 3 above) of forming ellipticals by mergers may also get resolved if one takes into account the prsence of extended dark halos around disk galaxies. The latest simulations of the merging of disks with dark halos to form ellipticals are by Barnes (1989). In these simulations Barnes follws the evolutions of a small compact group of disk galaxies with dark halos. The presence of the dark halos lead to significant dynamical friction on the galaxies as they move through each others halo, leading to the galaxies merging even if they were initially on a parabolic orbit. It turns out that the compact group evolves through a sequence of mergers on a timescale of only a few crossing times. The dark halo also acts as a sink for energy. As the disks merge they can become more strongly bound by transfering energy to the halo. Also because of the dynamical friction the galaxies can lose their orbital angular momentum before merging leading to a slowly rotating merger remanent. Barnes finds that the merger remnants are typically slowly rotating triaxial systems with de Vaucoulers' law luminosity profile and structural parameters generally consistent with bright ellipticals. He infact goes on to ask if most ellipticals are produced from mergers in such compact groups ?

Several of the other potential problems (like items 4 and 5) may require the presence of gas in the progenitor galaxies. During the merger this gas can dissipate energy and sink to the centre leading to cores with high phase space densities. Kormendy (1989) points out that this may be essential at least for the low luminosity ellipticals. For example, Kormendy estimates a phase space density for M32 $\sim .08 M_{\odot} pc^{-3} (\text{kms}^{-1})^{-3}$ a factor $\sim 10^7$ higher than in typical disks ! He concludes that the formation of the low luminosity elliptical galaxy M32 atleast must have involved considerable amounts of dissipation. Modest amount of gaseous dissipation and subsequent sinking to the core regions may also help in setting up metalicity gradients. But it is not at all clear whether the excess of globular clusters can be explained even by the presence of gas in the progenitor galaxies which are merging to form the elliptical (see Van den Bergh 1989, and Schweizer 1986 for opposite views). As a hypothesis the idea of ellipticals forming by mergers of disks has led to a lot of fruitfull work. In the process it has evolved a great deal and a suitably modified version seems to have some amount of suceess. Nevertheless it is perhaps wise to keep an open mind to other possibilities, a few of which we shall now discuss.

It has always been a fond hope that one can find a small set of parameters of a collapsing protogalaxy, which once specified, will decide the type of galaxy formed. We have discussed the role of star formation rate and angular momentum. Are there any other such crucial parameters ? In the early stages of the development of the Cold DM theory of structure formation, it was hypothesised that the height of a density peak on the galactic scale was one such parameter (Blumenthal *et al.* 1984). It was argued that the higher peaks collapsed earlier, had less time to be torqued up and so formed ellipticals with low λ. However we mentioned in section 4.7 that the numerical simulations which followed the formation of halos did not bear out this picture. These simulations by Barnes & Efstathiou (1987) and earlier work by Frenk *et al.* (1988) however brought to the fore another possible crucial parameter, the amount of substructure in a protogalaxy.

It was noticed that when a prot‹ ȷalaxy underwent a clumpy collapse significant amounts of angular momentum can be transferred outwards as the clumps sank to the centre and merged (see also section 4.7), leading to slowly rotating cores embedded in extended halos. If stars could form in these clumps before the collapse of the protogalaxy then the resulting stellar system in the core will also be slowly rotating, and be an ideal candidate for an elliptical. On the other hand it was found that quiescent protogalactic collapses, did not lead to such a transfer of angular momentum. These halos could then be ideal sites for disk formation (Frenk et.al. 1988). This picture elliptical galaxy formation is very similar to the merger picture except that the progenitors are sub galactic clumps and not full blown disks. Of course it has not yet been clarified whether this idea can work in detail, but it looks promising, even if the Cold DM theory goes out of fashion.

Another possible candidate crucial parameter which has been suggested is the shape of the dark halo in which the galaxy forms (Subramanian 1988; Katz & Gunn 1991).. Dark halos are unlikely to be symmetric, more generally they will be triaxial. In fact nearly 25% of the halos in the N - body simulation of Frenk et al (1988) have a short to long axis $< 1 : 4$. As gas collapses in such a halo potential it may lose its angular momentum to the halo. The magnitude of this effect was worked out in detail in an idealised picture of protogalactic collapse by Subramanian (1988). The angular momentum loss of the collapsing gas was found to be significant when the initial protogalaxy rotates about the halo middle axis and the halo had a short to long axis ratio less than $\sim 1 : 4$. It was suggested that ellipticals were those protogalaxies where the angular momentum has been transferred from the gas to the dark halo. It was also suggested that in regions of larger galaxy density halos would be more misshapen due to the tidal forces. Whether this will work in practice is not yet clear.

One can see from the plethora of ideas expressed in this section that the problem of explaining how elliptical galaxies arise is still quite open. Certainly the star formation rate, the angular momentum content of the protogalaxy, the complex ways in which it gets re distributed during collapse, the subsequent mergers are all likely to have some relavance. The difficulty lies in not yet having a clean grand picture, which seems patently obvious to everyone !

This completes our grand tour of the possible processes which operate during the non linear evolution of structure. We now turn to consider whether the observed universe at high redshifts has any clues to offer about when and how galaxies formed.

5: High redshift objects and Galaxy formation

5.1 Introduction

So far we have analysed the idea that gravitational instability amplified small density fluctuations in to galaxies and other structures. We also discussed some of the constraints on the theory arising from the CMBR and large scale velocities, which largely probe density fluctuations when they were still in the linear regime. How does one learn about the nonlinear phases of galaxy formation ? Presumably, the nonlinear phase of structure formation occurs much after decoupling but much earlier than the present. So a study of objects in the universe at larger and larger redshifts will be fruitful in constraining different models of galaxy formation. We therefore examine in what follows the possible hints which the high redshift universe offers as to how and when gàlaxies formed.

The highest redshift objects which have been discovered so far are of course the Quasi stellar objects (QSO 's or quasars). The QSO PC1247 + 3406, with $z = 4.897$, holds the present record for the highest redshift (Schneider, Schmidt & Gunn (1991)). In recent years there has also been a spurt in the discovery of galaxies at high redshift using their radio properties, the highest redshift being 3.8 for the galaxy 4C 41.17, discovered by Chambers, Miley & van Breugel (1990). The very existence of such objects puts severe constraints on theories which predict a late epoch of galaxy formation. The high redshift radio galaxies also have several important differences from their low z counterparts, which are indicative of vastly different physical conditions at these redshifts. Furthermore, the QSO's and radio galaxies show an evolution in their comoving number density, with a possible peak at redshifts of around 2. The relation of this epoch to the epoch of galaxy formation is intriguing. These and other important constraints imposed by the high z quasars and radio galaxies on galaxy formation are studied in sections 5.2 and 5.3.

The high z QSOs are not only interesting in themselves but also very useful because they reveal the presence of interesting types of intervening objects, which cause absorption lines in the QSO 's spectra. A study of the absorption line systems can be used to probe the content of the universe at redshifts below the QSO redshift. Also since an absorption line producing object comes along the line of sight to the QSO by chance, they are more representative of their class. This contrasts with the case of QSO's or radio galaxies whose properties may not be representative of the general class of galaxies, since it may require special conditions to make a galaxy active. We shall highlight some key results of absorption line studies in section 5.4, which have particular relevance for structure formation theories.

One of the most exciting prospects of looking at objects at high redshifts is to actually detect a forming galaxy. In fact it is not even clear at present how we will recognise such an object. We therefore end this part with a consideration of primeval galaxies, what are they ? when did they form ? and what should they look like ? Much of the material to be discussed in the various sections of this part is still a subject of active research and debate ; the conclusions we draw are therefore only preliminary.

5.2 Quasars and galaxy formation

Ever since their discovery, quasars have been recognised as potentially valuable probes of the high z universe. Their high intrinsic luminosity makes it possible to see quasars to high redshifts. Also if the QSO phase lasts only for a short time compared to the Hubble time at the QSO redshift, then the number of galaxies which go through a QSO like phase may be much more than the number of QSOs seen at that z. Quasars may therefore be tracers of the galaxy population at high z. Systematic observations of quasars has recently produced several potentially important results.

These include a characterisation of how the QSO luminosity function evolves with redshift, and more importantly the discovery more and more quasars at higher and higher redshifts (about 20 are known now with $z > 4$). We discuss the relevance of each of these to galaxy formation, in turn.

The quasar luminosity function is usually defined as the number of quasars per unit comoving volume per unit luminosity. One can also separate the data into various redshift bins and examine how this luminosity function evolves with redshift. The most interesting information from the viewpoint of galaxy formation is the redshift dependence of the number density of quasars above a fixed luminosity. This is got by integrating the luminosity function at a redshift z over luminosity. There is considerable evidence that the integrated comoving number density of quasars increases with redshift, upto a redshift $z \sim 2$. This rise holds for both radio quiet QSOs (Boyle *et al.* 1987) and for radio quasars (Dunlop & Peacock 1990). What happens after this redshift is only beginning to be elucidated. A particularly interesting question is whether the quasar number density cuts off drastically after some maximum redshift. Such an epoch of peak quasar density could indicate that this epoch, say $z \sim 2$, is in someway special, perhaps because it was the epoch when galaxies which could host quasars formed in abundance.

As we summarise below, with the exception of the brightest quasars, the number density of quasars and radio galaxies does indeed show a decline between redshifts of 2 and 4. However this decline seems to be gradual, not a drastic redshift cut -off in the AGN number densities. In the case of radio quasars Dunlop & Peacock (1990) find a decrease in comoving number density by a factor ~ 5, betweeen $z = 2$ and $z = 4$ and also tentative evidence for a similar decrease in the number density of steep spectrum radio sources (mostly radio galaxies). For optically selected QSOs the situation is not so simple. For QSOs with absolute magnitude $M_B < -26$, there is evidence for a sharper decline, (cf. the review by Green (1989)) with the number density decreasing by about an order of magnitude between redshifts of 2 and 4. However, brighter QSOs with with $M_B < -28$ do not show a decline in their space density between these redshifts (Boyle 1990). One should keep in mind that these are difficult observational questions to settle, because of selection effects and the fact that the brightest objects are more easily seen. We should therefore not overinterpret the decline in AGN number density until the observational situation clarifies. On the theoretical side it is not yet clear what turns on or turns off the QSO phenomenan.

More important than the detailed evolution of the quasar luminosity function, it is just the very existence of quasars at high redshifts, say $z > 4$ that provides crucial constraints on galaxy formation (Efstathiou & Rees 1988; Turner 1991; Kashlinsky & Jones 1991). Quasars are thought to be powered by accretion onto massive black holes, at the centres of galaxies. So before any quasar can turn on, some galaxies must evolve at least to the stage of developing a compact and massive enough nuclei. These galaxies have to then collapse and settle down at redshifts higher than that of the quasar. However in some theories like the standard CDM theory or the HDM pancake theory galaxies form at a relatively late epoch at $z \sim 2$, say. In this case finding many quasars at high $z \gtrsim 4$ could be embarassing. This is the dichotomy which we elaborate more quantitatively below.

Quasars powered by accretion onto massive black holes have associated with them a characteristic limiting luminosity known as the Eddington luminosity. This is the luminosity, say L_E above which the radiation pressure on accreted plasma exceeds the gravitational attraction of the black hole, thus preventing accretion. For an accreting black hole of mass M_{BH},

$$L_E = \frac{4\pi G m_p c M_{BH}}{\sigma_T} \approx 1.3 \times 10^{47} (\frac{M_{BH}}{10^9 M_\odot}) \mathrm{ergs}^{-1} \tag{5.1}$$

Super Eddington luminosities are possible, but need special models, like for example the electromagnetic extraction of the rotational energy of a spinning black hole (cf. Rees 1984) Unless such special models are involved one can then infer a characteristic black hole mass from (5.1) using the observed luminosity of the quasar. The quasars with $z > 4$ have typical luminosities $\sim 10^{47} \mathrm{ergs}^{-1}$, assuming a cosmological model with $q_0 = 1/2$ and $h = 1/2$ (cf. Turner 1991, Schneider *et al.* 1989 a,b). This implies that they involve black holes of mass $M_{BH} \sim 10^9 M_\odot$.

From the luminosity one can also estimate the amount of fuel that must be present to power the quasar for a lifetime t_Q. If ϵ is the efficiency with which the rest mass energy of the fuel is converted into radiation, then the fuel mass is

$$M_f = \frac{L t_Q}{\epsilon c^2} \approx 2 \times 10^9 M_\odot L_{47} t_{Q8} \epsilon_{0.1}^{-1}. \tag{5.2}$$

Here L_{47} is the luminosity in units of $10^{47} \mathrm{ergs}^{-1}$, t_{Q8} is time in units of 10^8yr and $\epsilon_{0.1} = \epsilon/(0.1)$. So for a lifetime of around 10^8yr, and reasonable efficiencies the required fuel mass is comparable to the mass of the central black hole inferred on the basis of (5.1) .

The above masses only refer to that involved in the central engine of the quasar. This will in general be a small fraction say F of the mass of the host galaxy. Efstathiou & Rees (1988) write F as a product of three factors. Firstly only a fraction f_b of matter in the universe may be in baryonic form. Then when the galaxy forms only some fraction f_{ret} of the baryons originally associated with the galaxy be retained rather than expelled via a supernova - driven wind. Finally only a fraction f_{hole} of the baryons retained be able to participate in the collapse to form the compact central object. So $F = f_b f_{ret} f_{hole}$ and we get for the mass of the host galaxy

$$M_G \approx 2c_1 \times 10^{11} M_\odot \tag{5.3}$$

where $c_1 = L_{47} t_{Q8} \epsilon_{0.1}^{-1} F_{0.01}^{-1}$ and $F_{0.01} = F/0.01$. For $F_{0.01} \sim 0.1 - 1$, with the earlier value being more likely, (5.3) implies a mass for the host galaxy $\sim 10^{11} - 10^{12} M_\odot$ assuming that the other dimensionless parameters in (5.2) are of order unity. So the existence of quasars at $z > 4$ implies that atleast some objects with galactic masses should have formed by these redshifts.

It is relatively easy to estimate the typical mass of collapsed objects at any z in hierarchical models. Recall from Part 3 and Part 4 that the fractional density excess, $\bar{\delta}_0$, in a sphere of radius R containing on the average a mass M is given by

$$\bar{\delta}_0(M) = \frac{\nu}{b} 0.9 (\frac{M}{2.3 \times 10^{15} \Omega H_{50}^{-1} M_\odot})^{-(3+n)/6} \tag{5.4}$$

Here we have used the J_3 normalisation given in (3.4), ν is the peak height in units of the standard deviation and b is the bias factor as before. One can also relate $\bar{\delta}_0$ to the collapse redshift of an object using for instance the spherical top hat model. For a flat universe we have from (4.50),

$$(1 + z_{coll}) = \frac{\bar{\delta}_0}{1.686} \tag{5.5}$$

From (5.4) and (5.5) the characteristic mass (with $\nu = 1$) which collapses at a redshift z is given by

$$M_c(z) = 2.3 \times 10^{15} (\frac{1.686b(1+z)}{0.9})^{-(6/(n+3))} H_{50} M_{\odot} . \tag{5.6}$$

From (5.6) we see that M_c drops steeply with redshift for for negative n. For example $M_c \propto (1 + z)^{-6}$, for $n = -2$, relavent to the CDM power spectrum at galactic scales. Also if $b > 1$, M_c is decreased correspondingly by a factor $b^{6/(n+3)}$. In table 5.1 we have given M_c for different values of n, the slope of the power spectrum of density fluctuations, and for several values of z. These values of M_C assume no biasing. We see from the table that, for $b = 1$, the characteristic mass is comparable to galactic masses for $z \sim 4 - 5$ for $n = -1$, while for $n = -2$, this happens only at redshifts $\sim 1 - 2$. So in theories with $n = -2$, galactic mass objects at $z > 4$, will be rare. For theories like the Cold DM theories, the effective value of n varies with mass scale, from $n \gtrsim -3$ at small masses through $n \sim -2$ at galactic scales to $n \sim -1$ at cluster mass scales. Also the standard Cold DM model needs a biasing factor $b \sim 2$, and consequently a reduction in M_c at a fixed redshift. It turns out that even in these models galactic scale objects are rare at redshifts ~ 4 ; infact galaxies form in abundance only at a $z \lesssim 2$ (cf. Frenk *et al.* 1988; Frenk 1989).

Table 5.1
M_c in units of $M_{\odot}$ for different values of n and for different redshifts

	$b^6 M_c$	$b^3 M_c$	$b^2 M_c$	$b^{1.5} M_c$
	$n = -2$	$n = -1$	$n = 0$	$n = 1$
$z = 0$	5.4×10^{13}	3.5×10^{14}	6.6×10^{14}	9.0×10^{14}
$z = 2$	7.4×10^{10}	1.3×10^{13}	7.3×10^{13}	1.7×10^{14}
$z = 4$	3.4×10^{9}	2.8×10^{12}	2.6×10^{13}	8.1×10^{13}
$z = 5$	1.1×10^{9}	1.6×10^{12}	1.8×10^{13}	6.2×10^{13}

We note in passing that how one normalises the spectrum may play an important role in the above considerations. Suppose we know that the fractional excess density contrast in galaxies $\delta N/N$ is unity at some mass scale M_0. Then naively one may have written $\sigma(M_0) = 1/b$. However this ignores the fact that for excess density contrasts comparable to unity, the effects of nonlinear evolution will be important. In particular, to get a $\delta\rho/\rho \sim 1$ at some time, one would need a smaller initial density contrast, when one takes into account the effect of nonlinear evolution, than if one just extrapolates linear theory. For example suppose we use the spherical model to follow the nonlinear evolution. Then $\bar{\delta} = \rho/\rho_b - 1 = 1$ when $\theta \approx 2\pi/3$. For this θ we then get from (4.37), the fractional excess density contrast linearly extrapolated to the present epoch $\bar{\delta}_0 = 0.57$. So we should correctly normalise our spectrum by demanding $\sigma(M_0) = 0.57/b$. This is infact how Kashlinsky & Jones (1991) normalise the spectrum. Under this normalisation the characteristic mass which collapses at any z becomes

$$\bar{M}_c(z) = 1.2 \times 10^{15}(\frac{1.686b(1+z)}{0.57})^{-(6/(n+3))} H_{50} M_{\odot}. \qquad (5.7)$$

where we have used the fact that $\delta N/N = 1$ on a scale of $8h^{-1}$Mpc. We see that the effect of this different normalisation is roughly equivalent to introducing a bias $b \sim 2$ in the expression for M_c in (5.6) . So $\bar{M}_c$ is in general smaller than M_c. For example for $n = -1$, we get $\bar{M}_c = 3.7 \times 10^{11} M_{\odot}$ at a redshift of 4 compared to $M_c = 2.8 \times 10^{12} M_{\odot}$ in table 5.1. So using this normalisation instead of the J_3 normalisation implies an even smaller abundance of quasars at high z.

How rare the possible host galaxies of quasars are in the CDM theory has been quantified by Efstathiou & Rees (1988), using the Press - Schechter theory outlined in section 4.3. Recall that we can get, $f(M,z)dM$ the comoving number density of collapsed objects in a mass range dM as a function of redshift, using this formalism. Assuming that every halo of mass M_G or greater forms a quasar with lifetime t_Q, the expected number density of quasars is (cf. Efstathiou & Rees 1988; Kashlinsky & Jones 1991)

$$N_Q(> L_{47}, z) \approx f_Q \int_{M_G}^{\infty} f(M,z)dM \qquad (5.8)$$

Here $f_Q = min(1, t_Q/t(z))$, takes into account that only a fraction of order $t_Q/t(z)$ of halos will display quasar activity if the quasar lifetime is smaller than than the age of the universe $t(z)$ at redshift z. Efstathiou and Rees estimate

$$N_Q(> L_{47}, z) \approx 10^{-3} t_{Q8}(1+z)^{5/2} c_1^{-0.866} exp[-0.21(c_1)^{0.266}(1+z)^2]\mathrm{Mpc}^{-3} \qquad (5.9)$$

At large z the exponential in (5.9) dominates and leads to a sharp decline in the quasar density.

The comoving number density of luminous quasars with $L_{47} \gtrsim 1$ at $z \sim 2$ is $\sim 1.2 \times 10^{-7}\mathrm{Mpc}^{-3}$(Efstathiou & Rees 1988). One can estimate a critical redshift z_{crit} at which the quasar density in (5.9) will drop below this value. We mentioned earlier that the comoving number density of brightest quasars does not show any decrease between $z \sim 2$ to a $z \sim 4$. So it may be desirable for z_{crit} to be atleast greater than ~ 4. This implies from (5.9) the condition

$$\frac{\epsilon_{0.1} F_{0.01}}{t_{Q8}} \gtrsim 0.1 \qquad (5.10)$$

The value of $F_{0.01}$ depends on uncertain astrophysics ; Efstathiou and Rees consider that a value ~ 0.1 (corresponding to 1% of the baryons collapsing to form the central object), may be a realistic upper limit. The radiative efficiency $\epsilon_{0\ 1}$ cannot substantially exceed 1 (cf. Phinney 1983). So the product $\epsilon_{0.1}F_{0.01}$ is unlikely to larger than 0.1. The inequality (5.10) then implies that high $z \sim 4$ quasars can only exist in sufficient number if the quasar lifetime t_Q is shorter than $\sim 10^8$yr, or from (5.3) a smaller mass for the host galaxy. However for lifetimes so small that $M_f < M_{BH}$, we should not use the fuel mass to estimate the mass of the host galaxy. Rather we should in this case use the black hole mass M_{BH} and write $M_G \approx FM_{BH}$. Assuming that quasar luminosities are limited by L_E, one then gets a lower limit to what we can take for t_Q in equation (5.10) given by equating M_f and M_{BH}. We get

$$t_Q > t_S = \epsilon t_E = \frac{\epsilon \sigma_T c}{(4\pi G m_p)} \approx 4 \times 10^7 \epsilon_{0.1}\text{yr}. \tag{5.11}$$

Interestingly this is also the time scale (Salpeter 1964) in which the mass of a black hole accreting at the Eddington limit doubles due to the accretion. Keeping in mind this lower limit to $t_Q \sim t_S$ we see that (5.10) can be satisfied but not by a wide margin. Further if $z_{crit} \sim 4$, the abundance of high luminosity quasars must decline exponentially for higher redshifts, which does not seem to be supported by recent observations (Irwin *et al.* 1991). Finally we should note that the above analysis has assumed that quasars can turn on immediately after their hosts collapse. More realistically it will take some time for hosts to develop compact nuclei to power quasars. In this case one has to account for higher collapse redshifts which is a greater problem for a theory like the standard Cold DM model.

We have discussed the constraints implied by high z quasars on the CDM theory in such detail, since this theory has recieved considerable attention in the recent past. It must be emphasised that the existence of high redshift quasars is a potential problem for any theory where galaxies form late. For example in some current top- down theories, like the HDM picture super clusters collapse into pancake like structures first and subsequently fragment to form galaxies. As discussed in section 4.6 studies of non linear clustering on scales $\lesssim$ 10Mpc show that supercluster collapse must have occured quite recently, at $z \lesssim 2$, so as to avoid exessive clustering. (Frenk, White and Davis 1983). The existence of high $z \gtrsim 4$ quasars is then hard to understand in such theories also.

Until recently quasars were the only class of objects which could be seen to high redshifts. This monopoly has now been broken and a new class of intriguing high z objects have been discovered, the high redshift radio galaxies, which we now consider.

5.3 High redshift radio galaxies

The explosion in the discovery of high redshift radio galaxies is a fairly recent phenomena which has come about basically from the optical study of large samples of strong, steep spectrum radio sources. It is still somewhat controversial as to why and whether this criterion, that the radio source has ultra steep spectrum, is important for discovering high z objects. Nevertheless most of the galaxies discovered to date with $z > 1$, including 4C 41.17, the galaxy with the largest redshift of 3.8, have come from the application of this method. At present more than ~ 20 galaxies are known with $z > 2$, and out of these several have a redshift greater than 3. These numbers are expected to increase in the next few years.

The high z radio galaxies are relevant to galaxy formation in several ways. Firstly as we discuss below, their optical properties seem to be very different from their lower z counterparts. So their study may help one to understand the physical conditions and the evolution of at least those, perhaps exceptional, high redshift galaxies, which host radio sources. Secondly the very existence of the highest z radio galaxies will constrain theories of late galaxy formation, just as in the case of quasars discussed in the previous section. In addition, in the case of these galaxies, one may be able to independently estimate their age and minimum mass, from the flux they emit at various wavebands (see below). This will provide additional constraints on the epoch of structure formation.

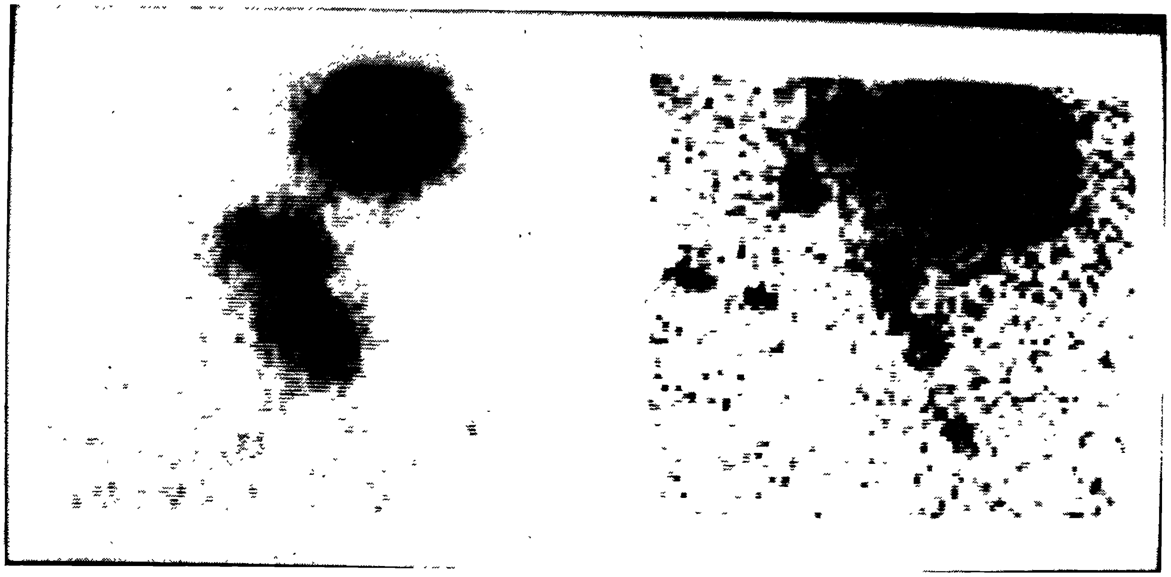

Figure 5.1a,b. The high redshift radio galaxy 2104 – 242 at $z = 2.491$ showing the alignment effect. The Lyman α image on the left (5.1a) and the r- band image on the right (5.1b) are both elongated in the same direction as the radio source (Figure 5.1c) appearing below. The images are 25" on a side and the object to the upper right in (5.1a) and (5.1b) is a star.

Radio galaxies at high z differ from their lower redshift counterparts in basically the following ways : Firstly, they have associated with them optical emission - both lines and continuum - extended over large regions, from several tens of kpc to over 200 kpc (cf. McCarthy *et al.* 1987a). The emission line gas in several objects also displays large velocity gradients up to $\sim 2000\mathrm{kms}^{-1}$ (McCarthy *et al.* 1987a). And most intriguing is the fact that the optical emitting region is elongated, its major axis being preferentially aligned with the axis defined by the radio source (McCarthy *et al.* 1987b; Chambers, Miley & van Breugel 1987; McCarthy 1989). This is in complete contrast to what is seen in the case of radio galaxies at low redshift, where if at all there is alignment it is between the radio and the minor axis of the optical galaxy (

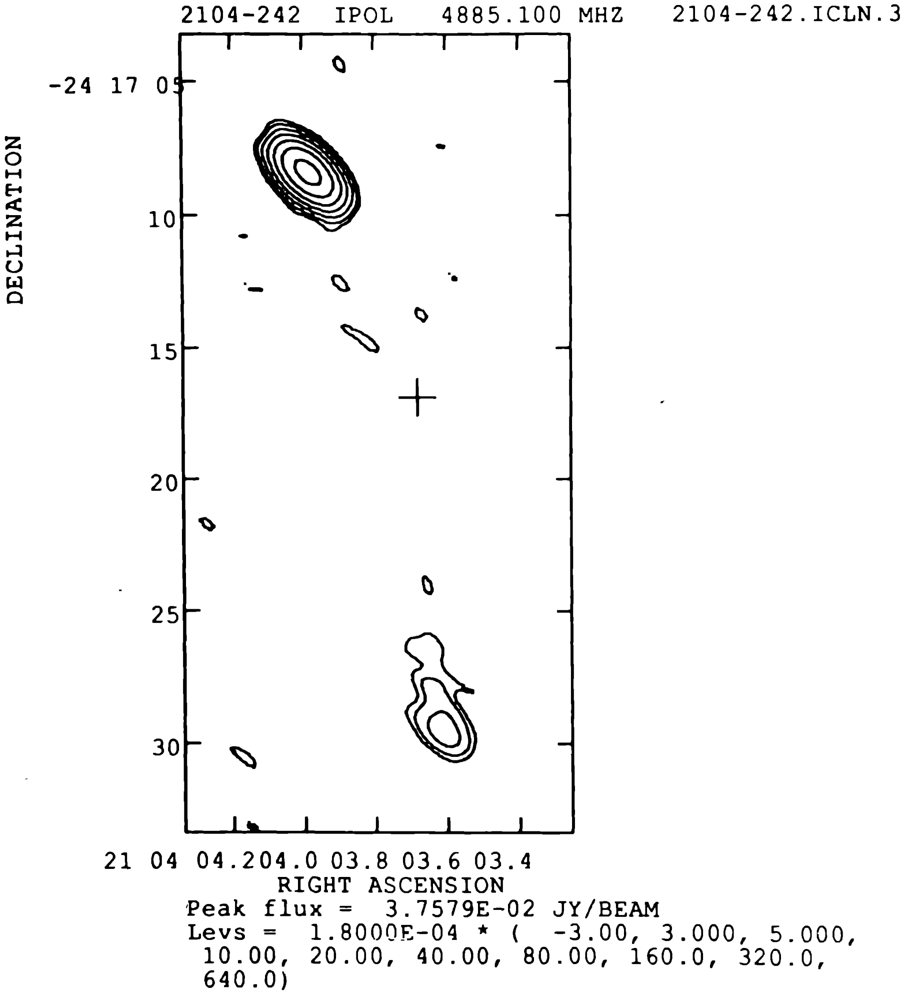

Figure 5.1c. The radio map of the galaxy 2104 – 242. Figures 5.1a, b, c has been adapted from McCarthy *et al.* (1990).

Palimaka *et al.* 1979). In figure 5.1 we show an example of this phenomenon studied by McCarthy *et al.* (1990), the radio galaxy 2104 - 242 at a redshift of 2.491. Figures 5.1a and 5.1b show the Lyman α and r band images of this galaxy which can be seen to be elongated in the same direction as the radio source shown in figure 5.1c. This 'alignment effect' has been the prime mover for much of the theoretical study of high z radio galaxies. So we begin in this section with a more detailed consideration of the alignment effect. The questions which arise for example are ; why is there a radio - optical alignment ? Where does the gas spread over such large volumes come from ? Since these high z radio galaxies are very different from their low z counterparts, do they represent young galaxies which have just formed the bulk of their stars ? We shall cosider tentative answers below.

A good review of the various mechanisms proposed to explain the alignment effect is given by Chambers & Miley (1990). The possible explanations include gravitational lensing (Lefevre *et al.* 1988a, 1988b,), electron and dust scattering of beamed optical emission from an underlying active nucleus (cf. Fabian 1989; De Alighery *et al.* 1989 respectively), and radio source induced star formation. Of these the last possibility appears at present to be the most likely explanation in the majority of the sources, although in some cases lensing or scattering of beamed optical emission may also play a role. Infact, there exists a possible nearby example of a starburst triggered by a radio source,the Minkowski's object (van Breugel *et al.* 1985; Brodie, Bowyer & McCarthy 1985). Chambers, Miley & van Breugel (1987) and McCarthy *et al.* (1987a) suggested in their very first papers which pointed out the alignment effect that such a process may also explain the radio - optical alignment. A number of authors have worked out in greater detail how this may happen (Rees 1989; De Young 1989; Subramanian 1989; Begelman & Cioffi 1989; Daly 1990). The basic idea is as follows :

Suppose a galaxy in the process of formation develops, a two phase medium ; clouds or filaments with a temperature $T \lesssim 10^4 K$ in pressure balance with a hotter medium at $T \gtrsim 10^6 K$. We shall discuss a little later how this may come about. The pressure felt by the clouds is then

$$p_0 \doteq nkT \approx 3 \times 10^{-11} (\frac{n}{10^{-2}\mathrm{cm}^{-3}})(\frac{T}{10^6 K})\mathrm{dyncm}^{-2} \qquad (5.12)$$

where n is the density of the hot inter cloud medium. Suppose such a galaxy also develops an active nucleus which squirts out twin jets of plasma to feed double radio lobes. As the lobe propogates into the gas, it will drive a bow shock ahead of it. There could also be shocks propogating transverse to the jet axis driven by the over pressured cocoon of plasma which gathers behind the radio lobes. The pressure of the shocked gas in front off the lobe is $\sim m_p v_h^2 4n$, where v_h is the velocity with which the head of the lobe advances into the gas. Assuming the lobe of relativistic plasma to be in rough pressure balance with the shocked gas in front of it, we can estimate the pressure of the plasma in the lobe to be

$$p_l \approx 7 \times 10^{-9} (\frac{v_h}{10^{-2}c})^2 (\frac{n}{10^{-2}\mathrm{cm}^{-3}})\mathrm{dyncm}^{-2} \qquad (5.13)$$

A similar radio lobe pressure is also estimated by Rees (1989) from applying the equipartition assumption to the high z radio galaxy 3C 368. Note that p_l is much larger than p_0, the pressure initially felt by the cool clouds. The pressure in the cocoon will be much smaller than p_l but still much larger than p_0, by a factor $\sim M^2$ where M is the mach number of the transverse shock. As the shocks produced by the radio jets propagates into the gas, the clouds wandering into the lobes or cocoon find themselves overpressured by a factor $p_l/p_0 \sim 10^4$ in the lobe for $v_h \sim 0.1c$ and perhaps a much smaller factor M^2 in the cocoon. The sudden jump in pressure, and the resulting compression would trigger the collapse of all clouds down to a fraction of order $(p_l/p_0)^{-1/2} \sim 0.01$ to $\sim 1/M$ of the previous Jeans mass, leading to a burst of star formation along the radio axis. Rees (1989) estimates that the rate of star formation along the radio axis maybe enhanced in this process by a factor $\sim 10-100$, comparable to the mach number of the shocks driven into the gas by the jets.

This explains the aligned optical continuum emission. In order to account for the line emission one needs also a source of ionizing photons. Such photons can come from hot stars which are co - existing with the cool gas or can arise from the active nucleus of the radio galaxy. It turns out that the line ratios seen in high z radio galaxies are best explained if the cool line emitting gas is photoionised by a power law continuum, expected from the active nucleus (cf. van Breugel & McCarthy 1990). In this case the alignment of the line emitting region with the radio axis requires the optical continuum from the nucleus to be beamed also along the radio axis. Interestingly, such beaming has been independently proposed in schemes to unify radio galaxies and quasars (Barthel 1989).

The above models for the alignment effect would not be complete without accounting for the gas spread over large volumes ($\sim$ 100kpc), which we postulated to be in the form of a two phase medium. This question has not perhaps received as much attention as it deserves. The most influential paper in this context has been the work of Fall & Rees (1985). Suppose one considers a collapsing protogalaxy. We saw in section 4.6 that for massive galaxies with $M \gtrsim 10^{12}M_{\odot}$, that the ratio τ of the gas cooling time to the collapse time initially exceeds unity (4.60). In this case as protogalaxy collapses most of the gas will get heated up initially to the virial temperature of the system, and try to settle into a pressure supported equilibrium. However as the gas cools, it will flow in, the density will rise and τ will decrease to unity. Fall & Rees (1985) argue that as this happens, the gas will become thermally unstable - slightly overdense regions will cool faster than their surroundings, be compressed, become denser, cool even faster and so on. The cool gas will start to condense out, thereby reducing the hot gas density and increasing τ. Fall and Rees argue in fact that the condensation of the cool gas depletes the hot gas in such a way that its cooling and collapse times scales remain comparable. It is from the cool - phase gas clouds that stars would form. Clouds massive enough to be Jeans unstable would contract, possibly initiating star formation within them. But clouds are expected with a broad spectrum of sizes, stretching well below the limit for Jeans instability. It is such clouds which may be pushed over the limit when they are engulfed by the radio source.

This picture nicely accounts for the fertile initial conditions which we postulated above, where gas clouds are waiting to be hit by a radio source to form stars. However it is not clear if it also accounts for the large extents of such gas. This is because thermal instability operating on small density peerturbations, initially leads only to power law growth of density contrasts (Fall & Rees 1985). If the initial density contrasts are $\sim 10\%$, then Fall and Rees show that cool clouds condense out only after the gas has collapsed by a factor ~ 10, from the radius when first $\tau \sim 1$, to say radii $\sim$ a few kpc. So to get cool gas at large radii $\sim$ 100 kpc, one has to postulate the protogalaxy to be highly inhomogeneous right at the start of its collapse. This may not be totally unnatural, if galaxies build up by the mergers of smaller masses. Another possibility which has been considered by Subramanian (1989) is that the protogalactic collapse is not spherical, but rather the protogalaxy initially collapses to a large pancake during which cool clouds are again produced due to thermal instability. Such pancakes may well be the absorbers producing the damped Lyman α absorbing systems (see the next section). Subramanian argues that the merger of a protogalactic pancake with another galaxy can disperse the cool gas to large volumes, and can also possibly switch on the radio source lighting up the merger.

Whatever the origin of the gas seen in high z radio galaxies, their very existence can strongly influence the properties of the radio source, for example the radio linear size. This argument can be turned around and one can ask whether observations of the linear sizes of high z radio galaxies, can be used to probe galaxy formation and evolution. Infact, observations already exist atleast up to a redshift $z \sim 1$, which show that the median linear size l, of radio sources with similar radio power, decreases steeply with redshift as $l \propto (1+z)^{-3}$ (Oort *et al.* 1987; Singal 1988; Kapahi 1989). This evolution may be understood if gaseous halos around these galaxies had a larger density in the past (Swarup 1988; Subramanian & Swarup 1990; Gopal Krishna & Wiita 1991). For forming galaxies, with most of the mass still in a gaseous form, the reduction in the radio linear size may be quite dramatic. For example, Subramanian & Swarup (1990) show that for forming galaxies, l may be reduced by factors $\sim 10-100$, for jets with power between $\sim 10^{46} - 10^{44}$ erg s^{-1} respectively, compared to values seen in local radio galaxies. So it may be possible to probe the epoch of vigorous galaxy formation by studying radio linear sizes as a function of redshift.

Perhaps the one spanner in the above works on explaining the radio - optical alignments of high z radio galaxies, is the fact that these alignments may persist even in the infrared waveband. The emission region in the infrared does not in general appear as elongated as in the optical, nevertheless for several high z radio galaxies, including 4C41.17, the radio galaxy with the largest z, it does appear to be aligned with the radio. (Chambers *et al.* 1988; Eisenhardt & Chokshi 1989). The case for infrared alignments is not yet as robust nor as universal as in the case of the optical emission (cf. Peacock 1990) and even the flux in the infrared may be much smaller than previously estimated for some objects (see Hammer, Lefevre & Proust 1991 in the case of 3C368 and Eisenhardt *et al.* 1990 for 0902 + 34). So at present one must be somewhat cautious in interpreting the infrared observations. We discuss below the possible origin of the infrared alignments, with the above caveat in mind.

The natural tendency to explain the infrared - radio alignments would be to extrapolate the picture for explaining the optical alignments and say that the infrared emission is also produced by stars formed due to the passage of the radio lobes. Since these stars have to be younger than the radio source which itself is believed to have a lifetime $\lesssim 10^8$yr, the observed infrared emission has to arise from a relatively young stellar population. However this has potential problems with two important sets of observations. These observations are also relevant to the question of the ages of these high redshift radio galaxies and therefore deserve discussion.

The first set of observations which is potentially embarassing for the above explanation of the infrared emission, is the observed colours of these galaxies. This is nicely quantified by plotting what is known as the spectral energy distribution (henceforth SED) of the galaxy, that is the dependence of flux density on the rest wavelength. Note that the infrared observations of a high z galaxy gives information about the the red to near infrared part of the spectrum in the rest frame of the galaxy. This part of the spectrum is generally dominated by emission from old red giant stars. On the other hand young massive stars generally contribute to the blue end of the spectrum. For an old elliptical galaxy in which there is very little ongoing star formation, the old stars dominate and the flux density is largest in the red part of the spectrum. So it is generally believed that a galaxy in its rest frame red wavelengths traces out more robustly the underlying stellar population. The SED for a number of high z radio galaxies have been determined. It turns out that the SED's are generally flat in the ultraviolet and show a rise in flux in the red. This is illustrated schematically in Figure 5.2. If this 'red bump' is due to an old stellar population then these stars atleast could not have been formed by the radio source. So in this case one should hardly see any elongation of the infrared (rest frame red) images along the radio axis.

In fact Lilly (1989, 1990) has taken this point of view. He fits the red bump with an old stellar population. In order to fit the flat ultraviolet part of the SED, he assumes that there is a second burst of star formation invoving only $\sim 4\%$ of the mass, lasting about 10^8yr and possibly associated with the radio activity. In order to fit the red bump of 4C41.17, it turns out that the older stellar population must atleast be $\sim 1.3G$yr old. Since the age of the universe at $z = 3.8$ in only ~ 1.25Gyr in a flat universe with $h = 1/2$, clearly Lilly's models if true have important implications for both galaxy formation and cosmology. Lilly expects that there would a be minimal alignment effect in the infrared, which is potentialy testable.

Taking the original observation of infrared alignments more seriously, a number of authors have tried to find ways of producing the red bump using only young stars. Chambers *et al.* (1988) and Bithell & Rees (1990) investigated whether the infrared continuum in the aligned radio galaxies is due to a large population of massive red supergiants. However it turns out these stars also spend some fraction of time being blue, and one only gets a red bump for a small fraction of the luminous phase of this population (Chambers & Miley 1990). Chambers & Charlot (1989) have come up recently with calculations which show that the observed SED's of high z radio galaxies can be reproduced even with young stellar populations. In their models the observed flat ultraviolet + red bump SED's can be produced with a normal initial mass function in $\lesssim 0.3$Gyr and persists for more than 0.6Gyr if most of the stars formed on timescales $\lesssim 0.1$Gyr. They estimate an age ~ 0.33Gyr for 4C 41.17 which implies a formation redshift of 4.9 for this galaxy in a flat universe with $h = 1/2$. Their evolutionary code is a modification of the Bruzual (1983) code, the main difference being in the treatment of the post main sequence evolution along the Asymptotic Giant Branch. It is far from clear to the non expert (and perhaps the experts) that the last word on models to fit the SED's of high z radio galaxies has been said as yet.

This brings us to the other important observational constraint on any model to explain the infrared emission from high z radio galaxies , the infrared Hubble diagram. A plot of the infrared K magnitude with redshift for esentially complete samples of 3C and 1 Jansky samples of radio galaxies shows a remarkably tight correlation between K and z (Lilly & Longair 1984; Lilly 1989). The K -z relation remains tight to the highest redshifts sampled, with a dispersion of ~ 0.4 magnitudes, constant over the redshift range $0 < z < 2$. The few systems known at $z > 3$ may also fall roughly on this relation. Any model for high z radio galaxies must be consistent with both

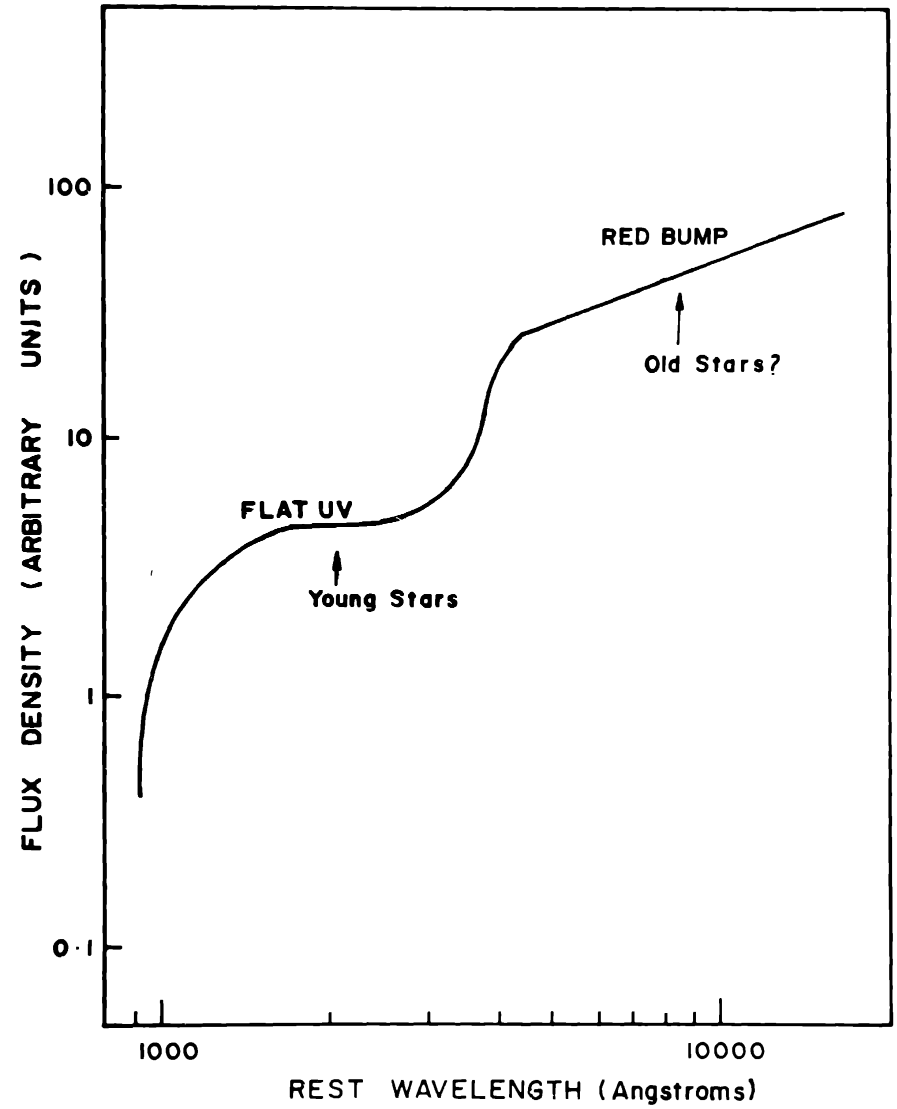

Figure 5.2. A scematic spectral energy distribution (SED) of a high redshift radio galaxy.

this small scatter and the fact that high z galaxies lie on the same relation defined by galaxies at lower redshift. As Lilly (1989) points out, the continuity and the small scatter in the infrared Hubble diagram obtains despite a factor of 6 difference in the radio luminosity between the two samples and a wide variation in the SED's at shorter wavelengths. These features can be easily understood if the K band light arises dominantly from a uniform 'old' population of stars. In models involving young stellar populations, their contribution to the light in the K band is expected to evolve on rather short timescales, say $\sim 10^8$yr. It is not then a priori clear whether the small scatter and the continuity in the infrared Hubble diagram would be maintained. However Chambers & Charlot (1989) claim that their models of the SED's of high z

radio galaxies, involving young stellar populations, is still consistent with the infrared Hubble diagram. They say that this is possible in their models because the different evolving components (main sequence, massive supergiants, AGB stars and red giants) conspire to maintain roughly constant visible to near infrared (in their rest frame) light from the time the galaxy is born up to $\gtrsim$ 1Gyr. As we said before we shall have to watch the future work on this question with interest.

The models to explain the SED's of high z radio galaxies also give as a byproduct, the total luminous mass and the redshift z_f, when the galaxy formed. For a cosmological model with $q_0 = 0.5$ and $h = 1/2$, Chambers and Charlot predict a characteristic luminous mass of $\sim 3 \times 10^{11} M_\odot$, for the high z radio galaxies that they model. If one takes into account the possible contribution of DM, the total mass of the galaxy could be an order of magnitude higher than the above value. The formation redshift z_f depends on the both the observed redshift and an estimate of the age of the galaxy from the model to account for its SED. For 4C 41.17 at a redshift of 3.8, Chambers and Charlot predict a luminous mass $\sim 5.7 \times 10^{11} M_\odot$ and a formation redshift $z_f \sim 4.9$. For lower q_0 models the formation redshifts are lower but the masses are significantly larger. Lilly (1989) concludes on the basis of his models invoving an old stellar population to explain the K - band light, similar large masses of stars $\sim 10^{12} M_\odot$ making up the high z radio galaxies. But his models imply that most of these radio galaxies formed well before their observed redshifts, with $z_f > 5 - 10$, the lower value being relevant for lower q_0. We can see from the above masses and formation redshifts, that the very existence of high z radio galaxies, just as in the case of quasars, will set interesting constraints on galaxy formation theories. In particular if many more radio galaxies are discovered at $z \gtrsim 3$, as seems inevitable at present, theories with late galaxy formation will be in trouble for the same reasons as outlined in the last section, in conection with high z quasars.

We have outlined in this section a variety of ways in which the existing observations of high z radio galaxies give clues the early evolution of at least some galaxies, which are perhaps exceptional in that they have learned how to make a radio source. We now consider the less special but equally interesting high z objects which are revealed as absorption line systems in the spectra of quasars.

5.4 Absorption lines and galaxy formation

The QSO absorption lines have been generally divided into several distinct categories. Of these the Lyman - α forest lines, the damped Lyman -α systems and the metal lines are the most interesting in the context of galaxy formation and evolution. Also of considerable importance is the constraint on the intergalactic medium (IGM) first discussed by Gunn & Peterson (1965), that is the absence of absorption by intergalactic hydrogen up to redshifts of atleast 4. In figure 5.3 we show schematically the spectrum of a QSO with the different types of absorption lines relevant to galaxy formation. The figure is exagerateed to show all the different types of apbsorption systems in the same spectrum. We begin this section with a discussion of the constraints imposed by the lack of the Gunn - Peterson dip in the spectra of QSOs.

Suppose the IGM were in the form of neutral hydrogen. It could then in principle be detected by examining the spectrum from a distant source, like a QSO. This is because neutral hydrogen atoms abosorb Lyman - α photons, of wavelenth $1216 A^0$,

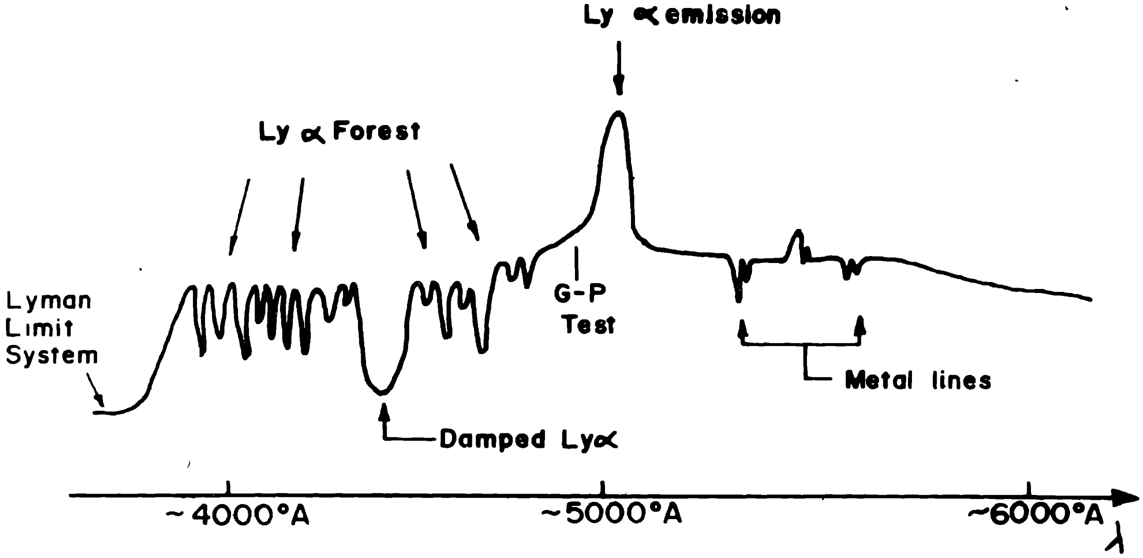

Figure 5.3. A schematic spectrum of a QSO exagerated to show the different absorption lines of interest for galaxy formation.

whose energy corresponds to the energy difference between the ground state and the first exited state of the atom. Because of cosmological redshift, the photons which are absorbed will have a shorter wavelength at the source and the signature of the absorption will be seen at longer wavelengths at the observer. So the spectrum of the source should show a dip at wavelengths on the blue side (shortwards) of the Lyman - α emission line if indeed neutral hydrogen is present in between the source and the observer. The magnitude of this dip of course depends on the neutral hydrogen density and can be quantified by calculating the optical depth, say p to such absorbtion.

Let n be the number density of hydrogen atoms and σ the absorption cross section. Then

$$p = \int n\sigma c dt \tag{5.14}$$

where the crossection for absorbing a Lyman - α photon is given by

$$\sigma(\nu) = \frac{\pi e^2}{mc} f g(\nu - \nu_\alpha). \tag{5.15}$$

Here e and m are the electron charge and mass respectively and $f = 0.416$ is the oscillator strength for the Ly α transition. The function g is sharply peaked at ν_α the frequency corresponding to the Lyman - α photon and its integral over ν is unity.

Converting the integral over cdt into an integral over z we then have

$$p = \int n(z) \times (\frac{\pi e^2 f g(\nu(1+z) - \nu_\alpha)}{mc}) \times (\frac{c}{H_0}) \frac{dz}{(1+z)^2(1+2q_0 z)^{1/2}}, \qquad (5.16)$$

In the above integral the observed frequency ν is fixed and corresponds to a photon of frequency $\nu(1+z)$ at the absorber. In the case when g is sharply peaked it behaves like a delta function and the main contribution to the integral is from absorbing atoms at a redshift $\bar{z}$ such that $(1+\bar{z}) = \nu_\alpha/\nu$ and we have

$$p = \frac{n(\bar{z})}{(1+\bar{z})(1+2q_0\bar{z})^{1/2}} \frac{\pi e^2}{mc\nu_\alpha} f \frac{c}{H_0} \approx 4.14 \times 10^{10} h^{-1} \frac{n(\bar{z})}{(1+\bar{z})(1+2q_0\bar{z})^{1/2}}. \qquad (5.17)$$

Gunn & Peterson (1965) did the above exercise and looked at the quasar 3C9 to see if there is any evidence for a dip in the spectrum shortward of the Lyman - α emission line. This quasar had a redshift of about 2 which brought this line into the visible part of the spectrum. There was hardly any dip noticable, which led them to put an upper limit to the density of neutral hydrogen in the IGM. Since then Steidel & Sargent (1987) have extended and improved this limit and get $p \lesssim 0.05$ for quasars in their sample with a mean redshift $\bar{z} \sim 2.6$. From (5.17) this lead them to derive an upper limit

$$n(z = 2.64) \lesssim 8.4 \times 10^{-12} h \mathrm{cm}^{-3} \qquad (5.18)$$

for a $q_0 = 1/2$ universe. This should be compared with the expected HI number density at this redshift

$$n_{HI} = 1.1 \times 10^{-5}(1+z)^3 \Omega_b h^2 \mathrm{cm}^{-3} \qquad (5.19)$$

where we have adopted a hydrogen mass fraction $X = 0.75$. If we take $\Omega_b = 0.026h^{-2}$ (see part 1) then at $z = 2.64$, $n_{HI} \approx 1.4 \times 10^{-5}$. So we see that n is much smaller than the expected HI density n_{HI}. This has been interpreted to mean that any hydrogen in the IGM must be almost completely ionised. There has been no evidence for the Gunn - Peterson dip in quasars at even a higher $z \sim 4$.

The lack of Gunn - Peterson dips upto redshifts of order 4 raises the question of what causes such ionisation of the IGM at these redshifts ? At present there is no clear cut answer to this question. Either the IGM has been heated to temperatures $\gtrsim 10^6 K$ or sufficient numbers of energetic photons have been produced by sources forming at large $z \gtrsim 4$ to completely photoionise the IGM. In either case strong constraints on epoch and the nature of galaxy formation are implied. The former possibility may obtain for example in scenarios of galaxy formation like the explosion picture of Ostriker & Cowie (1981). The later possibility has been extensively explored by Shapiro & Giroux (1987, 1989), who point out the inability of quasars to provide enough emissivity in ultraviolet photons. They also consider other possible sources of the required photons like young galaxies, primordial stars and decaying particles and do not find any of these to be entirely satisfactory. The lack of the Gunn - Peterson dips must clearly be understood as it may reveal some crucial aspect of galaxy formation.

Although high z quasars do not show any Gunn - Peterson dips, they do show many narrow absorption lines at wavelengths shortward of the Lyman α emission line. These are thought to be Lyman α absorption lines arising in clumps of gas with some fraction in neutral hydrogen. Because these lines are so numerous they are referred to as the Lyman α 'forest'. Infact, the line density is so high in many objects that it becomes difficult to find the continuum level to be used to determine any of the absorption line parameters. The parameters which can be measured are the wavelengths, the equivalent widths and where the spectum has sufficient resolution one can also determine the line profile or the doppler widths of the lines. Detailed studies indicate several important observational properties of these lines (cf. the review by Carswell (1989)):

Firstly, the mean number density of lines per unit redshift, say dN/dz, appears to evolve with epoch. For Lyα lines with rest equivalent widths $> 0.32\AA$ one has

$$\frac{dN}{dz} \approx k(1+z)^{\gamma}, \tag{5.20}$$

where $\gamma = 2.3 \pm 0.42$ for $1.5 \lesssim z \lesssim 3.8$ (Hunstead et. al. 1988). The constant k is more uncertain and is $\sim 3-5$ (cf. Bechtold 1987). For comparison suppose we calculate the number of absorption systems to be expected along the line of sight to a quasar from clouds of radius r_c and space density $n_c(z) = n_0(1+z)^3$. We get

$$\begin{aligned} \frac{dN_c}{dz} &\approx \pi r_c^2 \times n_0(1+z)^3 \times \frac{cdt}{dz} \\ &\approx 0.02(\frac{r_c}{10\text{kpc}})^2(\frac{n_0}{10^{-2}\text{Mpc}^{-3}})\frac{(1+z)}{(1+2q_0z)^{1/2}} \end{aligned} \tag{5.21}$$

where we have put in fiducial numbers for r_c and n_0 corresponding to nearby galaxies. Comparing (5.20) and (5.21) we see that the absorbing clouds have to be either larger than 10kpc and /or more abundant than nearby galaxies to match with the observed k. Moreover, it can be seen that the number density of lines increases much more rapidly with increasing z than if the comoving number density of absorbers were conserved. For example if $q_0 = 1/2$, one expects $dN/dz \propto (1+z)^{1/2}$, for constant comoving number density of clouds, where as observations indicate $dN/dz \propto (1+z)^{2.3}$. Any model for the Ly$\alpha$ forest has to explain this evolution.

The equivalent width and the profile of the Lyα absorption line can be used to derive the neutral hydrogen (HI) column density and the Doppler parameter of the absorbing cloud. The HI column densities for the Lyα forest systems range from 10^{13}cm^{-2} to 10^{15}cm^{-2} and the column density distribution in this range is well approximated by a power law of the type

$$f(\bar{N})d\bar{N} \propto \bar{N}^{-\beta} d\bar{N} \tag{5.22}$$

where f is the fraction of systems with column densities between $\bar{N}$ and $\bar{N}+d\bar{N}$, and $\beta \sim 1.75$. The above lower limit in the column density may not reflect the absence of such systems but rather arise because they are below the detection threshold of the present data. The exponent β also does not seem to strongly depend on redshift.

The doppler widths are usually quoted in terms of a parameter $b = \sqrt{2}\sigma$, where σ is the one dimensional velocity dispersion of the absorber. The doppler parameter shows a broad peak around $\sim 30 - 35\mathrm{kms}^{-1}$ and the mean value shows no strong dependence on redshift.

There has also been some interest in whether the Lyα systems are clustered. If the absorbers were associated with galaxies one would expect some degree of clustering. Interestingly, there is evidence for weak clustering over velocity scales of $\sim 150\mathrm{kms}^{-1}$ at intermediate $z \sim 2.5$, but for higher mean $z \sim 3.4$ it appears that the Lyα clouds are distributed uniformly. So there is a suggestion that clustering in the Lyα systems increases as the redshift decreases.

If we are to determine the physical parameters of the clouds causing the Lyα forest lines, we have to have more information than given above. The single key observation on which most models rely upon is the fact that the two images of the quasar Q2345+007 contain a number of common Lyα absorption lines. In order to intercept the line of sight to both the images the clouds then have to be atleast as large as the projected distance between the two images calculated at the absorption redshift. This quasar may be a case of multiple imaging by an intervening gravitational lens (cf. Weedman *et al.* 1982; Subramanian & Chitre 1984). In this case minimum cloud sizes $r_c \sim 5 - 25\mathrm{kpc}$ are implied (Foltz *et al.* 1984). Also some of the lines are not seen in both the spectra. So the cloud size cannot be much larger than the above estimate. Assuming a quasi spherical absorber one then gets an estimate of the HI density

$$n_{HI} \sim \frac{\bar{N}}{r_c} \sim 3 \times 10^{-8}(\frac{\bar{N}}{10^{15}\mathrm{cm}^{-2}})(\frac{r_c}{10\mathrm{kpc}})^{-1}\mathrm{cm}^{-3} \qquad (5.23)$$

To make further progress one has to know the neutral hydrogen fraction, since clouds with $\sigma \gtrsim 10\mathrm{kms}^{-1}$ or $T \gtrsim 10^4 K$ are likely to be highly ionised.

A popular assumption is that the clouds are photoionised, in which case the neutral hydrogen fraction depends on the background (or metagalactic) ionising flux. This in turn is estimated by noting how the number of Lyα lines changes as the redshift approaches that of the quasar. There is some indication that the line number density above a fixed equivalent width decreases as one approaches the quasar redshift (cf. Bajtlik, Duncan & Ostriker 1987). This effect, called the 'proximity effect', may be due to the increase in ionising flux from the quasar and the resulting decrease in HI column density, as one approaches the neighbourhood of the quasar. So the change in the line number density as one approaches the quasar will be governed by how the ratio of the quasar flux to the background ionising flux changes. This can then be used to estimate the metagalactic ionising flux. In their detailed study, Bajtlik, Duncan & Ostriker found that the metagalactic ionising flux at the Lyman limit, $J(\nu_T)$, is roughly constant over the redshift range $1.7 < z < 3.8$ and has a value $\sim 10^{-21.0\pm0.5}\mathrm{ergcm}^{-2}\mathrm{s}^{-1}\mathrm{Hz}^{-1}\mathrm{sr}^{-1}$.

The neutral hydrogen fraction can then be estimated from the equation governing the ionisation equilibrium

$$\alpha_H n^2 = \Gamma_H n_{HI}, \qquad (5.24)$$

where n is the total number density of the gas, assumed to be mostly hydrogen, α_H is the recombination rate coefficient and Γ_H the ionisation rate. Assuming that the ionisation flux J has a power law spectrum with spectral index of 1 above a thershold frequency for ionisation ν_T, the above equation can be rewritten as follows.

$$\begin{aligned}\alpha_H n^2 &= n_{HI} \int J(\nu)\sigma_H(\nu)d\nu \\ &= J(\nu_T) n_{HI} \int_{\nu_T}^{\nu_{max}} (\frac{\nu}{\nu_T})^{-1} \sigma_H(\nu)d\nu \\ &\equiv J(\nu_T) n_{HI} G_H\end{aligned} \tag{5.25}$$

Here σ_H is the crossection for ionisation, $\alpha_H = 4.36 \times 10^{-10}\, T^{-3/4}\mathrm{cm}^3\mathrm{s}^{-1}$ (Osterbrock 1974). And $G_H = 3.2 \times 10^9 \mathrm{erg}^{-1}\mathrm{cm}^2\mathrm{Hzsr}$ (Ostriker and Ikeuchi 1983). Using (5.25) and n_{HI} from (5.23) we then have

$$n \approx 4.9 \times 10^{-4} J_{21}^{1/2} \bar{N}_{15}^{1/2} T_4^{3/8} (\frac{r_c}{10\mathrm{kpc}})^{-1/2} \mathrm{cm}^{-3}. \tag{5.26}$$

One can also estimate the mass of the cloud from

$$M_c \sim \frac{4\pi}{3} n m_p r_c^3 \approx 4.6 \times 10^7 M_\odot J_{21}^{1/2} \bar{N}_{15}^{1/2} r_{10}^{5/2} T_4^{3/8}. \tag{5.27}$$

So cloud masses $\sim 10^7 - 10^8 M_\odot$ are typical.

What are these clouds and how are they produced ? For the above cloud parameters one can show that the ratio of the thermal energy to the gravitational energy of the gas in the cloud very much exceeds unity. So the clouds cannot be self gravitating and in any case self gravitating isothermal clouds tend to be unstable - if compressed they would collapse; if slightly inflated they go into free expansion (cf. Black 1981). If the cloud is not self gravitating then the gas must be confined in some other way. Two possible confining agents have been suggested ; the pressure of a hotter IGM or the gravity of DM.

Sargent *et al.* (1980) originally suggested the possibility that the Lyα clouds are primordial intergalactic clouds which are pressure confined by a surrounding hot IGM. Such a possibility naturally arises if the clouds form through thermal instabilities in shocks caused by either gravitational or hydrodynamical processes (cf. Ostriker 1988). The shock could be either the thermonuclear explosions associated with galaxy formation (Ostriker & Cowie 1981; Ostriker & Ikeuchi 1983; Ikeuchi & Ostriker 1986) or the pancaking which arises naturally in the hot DM models. In these models the Lyα clouds are those which are not massive enough to be Jeans unstable but at the same time of sufficient mass so that they are not evaporated by the surrounding hot IGM. This constrains not only the clouds but also the IGM. The observed evolution of the number density of lines given in (5.20) is explained by taking recourse to the evolution of the IGM (the density and temperature decrease) as the universe expands. To account for the range of column densities at a fixed z one has to postulate a range of masses for the clouds (Ikeuchi & Ostriker 1986) or variations in the pressure of the hot confining medium (Baron *et al.* 1989). The observed weak clustering of the lines is yet to find a natural place in this picture. Also of potential importance is the unsuccesful search for voids in the Lyα forest, which indicates that the pressure of any confining medium is the same to a factor ~ 2 in incipient voids as well as clusters (Carswell & Rees 1987).

Another possiblity which has been advocated by Rees (1986) and Ikeuchi (1986), is that the clouds are gravitationally confined by mini halos of cold DM. As we discussed earlier, in the cold DM theory and infact in any hierarchical clustering theory smaller mass objects form first then cluster and merge to form larger structures like galaxies. Suppose the IGM at these redshifts is photoionised and at a temperature $T \gtrsim 10^4 K$ (see the discussion on the Gunn - Peterson effect). For the smallest masses the potential well may be too shallow to capture any of this photoionised gas. On the other hand for larger masses comparable to the galactic mass the gas may cool efficiently, satisfying the $\tau < 1$ criteria of section 4.6, and sink to the centre of the DM halo. Rees pointed out there may be intermediate mass DM halos in which the captured gas may be stably confined being neither hot enouugh to escape nor cool enough to collapse, with heating by photoionisation balancing the cooling by radiative recombinations. Such mini halos have a virial velocity $\sim 30 \mathrm{kms}^{-1}$ and characteristic masses of $\sim 10^9 M_\odot$ with a 10% contribution to the mass from the baryons (Rees 1986). Rees points out that such gravitationally confined clouds are a neccessary consequence of the cold DM theory and had they not already been discovered one could have predicted such clouds.

In this model the observed evolution of the Lyα forest may arise due to several factors. Radiative cooling of the gas, a decrease in the background ionising flux with time and accretion of gas can all cause the gas to sink to the core and eventually fragment into stars. Rees infact hypothesises that dwarf galaxies result when this happens. Also the minihalos could get progressively destroyed if they get incorporated into larger clumps as galaxy formation proceeds. In the minihalo model, the distribution of column densities (5.22) may arise quite naturally from the fact that the lines of sight to sources will have a range of impact parameters. Suppose the baryon density follows an isothermal dark halo density law $n \propto r^{-2}$, then from (5.25) $n_{HI} \propto r^{-4}$ and $\bar{N} \propto r_p^{-3}$ where r_p is the impact parameter of the light rays. Assuming that r_p if randomly distributed one gets $f(\bar{N}) \propto \bar{N}^{-5/3}$, whhich is not too different from (5.22) (Rees 1988; Ikeuchi, Murakami & Rees 1990).

We have discussed two of the main models for the Lyα clouds. The state of flux of the field and the need to keep ones options open can be gauged from a recent controversy regarding some of the basic facts discussed above. Pettini *et al.* (1990) have claimed, from high spectral resolution ($6.5 \mathrm{kms}^{-1}$) observations of the QSO 2206 - 199N, that most Lyα lines in this object have $b \lesssim 22 \mathrm{kms}^{-1}$ and that lines with larger b have systematically larger column densities. The lines with the smallest b values they find can only be understood if the gas is much cooler than previously thought, with $T \sim 5000 - 10000 K$, and hence predominantly neutral. They conclude that all clouds must have this temperature and the $b - \bar{N}$ correlation arises because the larger $\bar{N}$ sytems involve many more clouds whose bulk motion produces a b larger than implied by the gas temperature. Their favored model is that the clouds are dense sheets or filaments with masses several orders smaller than those inferred above, possibly residing in forming galaxies. However high resolution observations ($\lesssim 9 \mathrm{kms}^{-1}$) by Carswell *et al.* (1991) of another quasar Q1100 - 204 does not show any correlation of b and $\bar{N}$. It is hoped that improvement in the signal to noise of the data may resolve the issue (cf. Peacock 1991).

We now turn to a class of absorbers called the damped Lyα systems, which are potentially a very important probe of galaxies at large z, since they may contain a significant fraction of the baryons at large redshifts.

The damped Lyman α systems were discovered and extensively studied by Wolfe and his collaborators. The original motivation was to see if one could detect the counterparts to the spiral disks at large redshift. Since these disks are expected to have a large column density in neutral hydrogen, it was hoped that they would produce strong Lyman α absorption lines which have been broadened by radiation damping. (Note that for large enough optical depth, the width of a line determined by the natural line-width dominates over that produced by thermal broadening. Since the classical analogue of this mechanism is radiation damping, the name 'damped' is used. The relative importance of such 'radiation damping' compared to thermal broadening, can be examined by considering the frequency dependence of the optical depth in the two cases. Doppler broadening leads to an optical depth $\tau = \tau_0 exp(-(\nu - \nu_0)^2/2\sigma^2)$, while the natural line width arising in quantum mechanics leads to a dependence $\tau = \tau_0/[1 + (\nu - \nu_0)^2/\Gamma^2]$. Here τ_0 is the optical depth at the frequency ν_0, corresponding to the line centre, σ the velocity dispersion in the absorber and Γ is the natural line width. Since the lorentzian profile falls of much more slowly than the gaussian profile it can be seen that for large τ_0, the frequency for which $\tau = 1$, say , and hence the line width, will be determined by the lorentzian profile. (see also Unsold 1977)). Wolfe *et al.* (1986) made a systematic study of the spectra of 68 QSOs to search for lines among the Lyα forest which were strong enough to be candidates for damped Lyα lines. Of the 47 candidates they found follow up spectroscopy confirmed 18 systems as damped Lyα systems (Wolfe *et al.* 1989, Turnshek *et al.* 1989). These systems have the following properties (Wolfe 1988, 1989) :

The redshifts of the detected systems lie between 1.8 and 2.8. The absorbers have an average HI column density $< \bar{N} >\sim 10^{21}\mathrm{cm}^{-2}$. By comparison the H_2 content is very low. Studies of two absorbers show that the H_2 mass fraction $\lesssim 10^{-5}$ and 10^{-4} respectively (Black *et al.* 1987, Lanzetta *et al.* 1988). In contrast this mass fraction is $\gtrsim 10^{-1}$ in the disk of our galaxy for lines of sight encountering comparable column densities in neutral hydrogen. A comparison of the optical continua of QSOs located behind the damped systems with a control sample shows that these systems may have some dust, between 1/20 to 1/4 that of the Galaxy (Fall, Pei & McMahon 1989). These absorbers also have some metals. Low ionisation states of carbon, silicon and iron (CII, SiII and FeII) are always detected whilst high ionisation state CIV and SiIV are less common. Pettini *et al.* (1989) find a metal abundance $Z \lesssim 1/10$ the solar abundance for the absorption system in PHL 957 with $z_{abs} = 2.3091$. The velocity dispersion revealed by the metal lines associated with the damped systems range from $10 - 100\mathrm{kms}^{-1}$. On the other hand associated 21cm absorption lines, in 7 of the damped systems, show the HI to be much more quiescent with a velocity dispersion generally $\lesssim 17\mathrm{kms}^{-1}$. This may be indicating that both a quiescent component producing HI absorption and a turbulent component producing the metal lines may be present in the absorber.

One of the most intriguing propertiees of these systems has to with their abundance. Firstly unlike in the case of the Lyα forest there is no positive evidence for redshift evolution of the damped Lyα systems ; their redshift distribution is consistent with the absorbers having a constant comoving number density and cros section. However the number of damped Lyα systems per unit redshift interval with $\bar{N} \gtrsim 10^{20.3}\mathrm{cm}^{-2}$ is

$$\frac{dN_{damp}}{dz} = 0.29 \pm 0.08; < z >= 2.4. \tag{5.28}$$

where $< z >$ refers to the average redshift of these systems. At this average redshift (5.21) gives $dN_c/dz \sim 0.05$ for galactic like absorbers. A more careful calculation for the number lines per unit redshift expected from spiral galaxies also gives $dN/dz \approx 0.05 \pm 0.03$. So the number density of the damped systems exceeds that expected from disks by a factor ~ 6. This can mean that either the number density or the cross section of the absorbers is larger than that associated with disk galaxies. Wolfe *et al.* have followed the later route and argued that disk galaxies at high z have radii $\gtrsim 3$ times the Holmberg radii of present day disks. We shall return to this question shortly.

Whatever the interpretation of the absorbers one can estimate the density parameter contributed by them from the observed dN_{damp}/dz, and using the fact they have a mean column density $< \bar{N} >= 10^{21}\mathrm{cm}^{-2}$. The mean mass density contributed by the damped Lyα systems at their mean redhift $< z >$ is

$$\begin{aligned} \rho_{damp}(< z >) &= \mu m_p < \bar{N} > \frac{dN}{cdt}|_{<z>} \\ &= \mu m_p < \bar{N} > \frac{dN}{dz}|_{<z>} \times \frac{H_0}{c}(1+ < z >)^2(1 + 2q_0 < z >)^{1/2} \end{aligned} \tag{5.29}$$

Here $\mu = 1.4$ is the mean molecular weight of the gas and m_p is the proton mass as before. As the universe expands this average mass density would have decreased by a factor $(1+ < z >)^3$. Comparing the resulting density with the present day critical density one gets the current density parameter of the HI making up the damped Lyα absorbers to be

$$\begin{aligned} \Omega_{damp} &= \mu m_p < \bar{N} > \frac{dN}{dz}|_{<z>} \frac{(1 + 2q_0 < z >)^{1/2}}{\rho_c(1+ < z >)} \\ &\approx \begin{cases} 1.2h^{-1} \times 10^{-3} & \text{for } q_0 = 0.05 \\ 2.3h^{-1} \times 10^{-3} & \text{for } q_0 = 1/2 \end{cases} \end{aligned} \tag{5.30}$$

where we have put in the appropriate numerical values from (5.28) to get Ω_{damp}. By comparison the mass density of the stars in disk galaxies makes up an $\Omega \sim 2h^{-1}\times 10^{-3}$. So we see that the damped Lyα systems may contain, at the redshift at which they are detected, a significant fraction of the baryons in the universe, comparable to that of luminous matter in galaxies. It is this that makes these absorbers such a crucial probe of the nature of protogalaxies at high z.

Based on the fact that to some extent they saw what they were searching for, Wolfe and collaborators put forward the suggestion that the damped Lyα systems arise in rotationally supported disks which are the progenitors of present day disk galaxies. However in order to explain the number of systems seen, they had to say, as we mentioned earlier, that these proto disks were several times larger in the past. The main problem is how to form rotationally supported disk galaxies with radii several times the Holmberg radius of present day disks ?

Recall our discussion of the formation of disk galaxies in section 4.8. We saw that at least in hierarchical theories, tidal torques give protogalaxies only about 5 – 10% of the angular momentum needed for rotational support. So to aquire rotational support the gas has to collapse by factors of ~ 10 in radius. Now if the Lyα disks are rotationally supported at radii of ~ 3 Holmberg radius, say at a radius ~ 30kpc, then this gas has to collapse from ~ 300kpc. Even if the gas turn around radius was this value (and not greater), it would take atleast ~ 5.5Gyr to collapse, using (4.18) with $M \sim 10^{12} M_{\odot}$, corresponding to a collapse $z \sim 1$ in a flat universe, which is too late to explain the damped population. One way out is to decrease the collapse factor. It is rather doubtful whether this can be done in the context of rotationally supported disks (but see Schiano, Wolfe & Chang 1991). On the other hand if one gives up the demand for rotational support, it is much easier to think of how such large sizes may arise by protogalactic collapse. One possiblity is that the damped Lyα absorbers are caustic sheets or pancakes of cool gas which have arisen from the general organised collapse or collapse in an asymmetric protogalactic potential. (Rees 1988, Hogan 1987, Subramanian 1988, 1989). One can then have large sizes for these systems even with modest collapse factors and reasonable timescales, since collapse has to occur only along the shortest axis of the system. Ofcourse caustic sheets and pancakes are also naturally expected in the explosion picture or the hot DM models.

Yet another suggestion about the nature of the damped Lyα systems is that they may represent a population of dwarf galaxies, which had smaller cross section but were much more abundant in the past (cf. Tyson 1988; Pettini *et al.* 1989) There is some indication that this may not be the case in atleast one object, the $z = 2.04$ absorption system in PKS 0458-02. For this object Briggs *et al.*(1988) find similar 21 cm line profiles towards different components of the extended radio structure of this source. This implies that the absorber must extend more than $8h^{-1}$kpc across the line of sight, and unlikely to be a dwarf galaxy at least in this case.

It is important to decide between the various possible explanations of the damped Lyα population since these systems have such a significant baryonic content.

Another class of absorption lines which may probe the evolution of galaxies particularly their gaseous halos at high z are the heavy element systems. These lines are seen as narrow absorption lines longward of Lyα emission. The commonly encountered lines are those corresponding to magnesium (MgII), carbon (CII, CIV), silicon (siIV) and iron (FeII). The typical neutral hydrogen column density is inferred to be $\sim 3 \times 10^{18}$cm^{-2} and the doppler parameter is in the range $5 - 25$kms^{-1} (cf. the review by Sargent 1988). There is a wide range of ionisation showing that the gas is probably being photoionised by a flat spectrum source (like for example the metagalctic flux). The heavy elements are somewhat underabundant with $Z \sim 0.1 Z_{\odot}$. The line depths show that the clouds producing the absorption cover the QSO emission region and so have a size $\gtrsim 10^{19}$cm. The two images of the gravitationally lensed quasar 0957 +561 show the same C IV absorption redshifts $z = 1.12$, but differences in detailed line profiles, implying that the absorbinng cloud has a size $\lesssim 10$kpc. It is thought that the absorption occurs in relatively small clouds embedded in much larger structures. A number of reasons seem to favour the hypothesis that the heavy element redshifts (with $z_{abs} < z_{em}$) arise in intervening galaxies.(cf. Weymann, Carswell & Smith 1981; Sargent 1988). Particularly encouraging are the discovery of galaxies at the same redshift as the absorption system in several cases, upto $z \sim 0.8$ (Bergeron 1988).

The observed frequency of absorption systems per unit redshift, for the metal lines, is of great interest. Since it may be used to infer the required mean cross sections of galactic halos, and how they evolve, via (5.21) . Sargent, Boksenberg & Steidel (1988) find in their survey of C IV absorption systems, that $< dN/dz >\sim 2.5$ at a mean $< z >\sim 2$ and that $dN/dz \propto (1+z)^{-1\ 2\pm0.7}$ for the lines in the redshift interval $1.3 < z < 3.4$. Since (5.21) gives $dN/dz \sim 0.035$ from galaxies with a crossection of 10kpc and at a $< z >= 2$, one infers galactic halo cross sections $\sim$ 85kpc to explain the observed mean dN/dz. Further the observed negative value for the evolution parameter γ (defined as in (5.20)) compared to an expected $\gamma \sim 0 - 1/2$ which would have obtained for a constant comoving density of absorbers, implies that the probability of absorption decreases in the past. One possible explanation for this is metal enrichment, that as galaxies evolve the abundance of heavy elements increases and the detectability of C IV absorption goes up with increasing time or decreasing z (Sargent, Boksenberg & Steidel 1988). Another possibility is halo expansion with time, possibly associated an earlier burst of star formation and the resulting energy injection into the gas (Ikeuchi 1990).

In contrast to the case for C IV, it appears that the low ionisation system Mg II has $dN/dz \propto (1+z)^{1.5\pm0.6}$ in the redshift interval $0.2 < z < 2.1$. Also the observed $< dN/dz >\sim 0.6$ at a mean $< z >\sim 1.1$, implying from (5.21) a cross section $\sim$ 45kpc for the absorbing halos. Ikeuchi (1990) hypothesises that this positive evolution may be due to a shrinking phase of the halo gas, which follows its expanding phase, due to gas cooling. The observations and the models of the evolution of metal lines are perhaps still in a preliminary stage. But as these improve one expects the metal line systems to become a very important probe of the evolving gaseous halos of forming galaxies.

This brings to an end our survey of high redshift objects which may shed some light and set constraints on theories of galaxy formation. The ideal high z objects to discover and study would have been galaxies in the process of forming. In the last section of this part we therefore consider such primeval galaxies. We must confess at the outset that at present there is no unambiguous detection of such objects, possibly due to the fact we mentioned earlier that we do not even know perhaps what to look for !

5.5 Primeval galaxies

Even the definition of what shoud be called a primeval galaxy is not universally agreed upon. Is a primeval galaxy one which has has just collapsed bringing $\sim 10^{11} M_\odot$ within a radius of say $\sim$ 10kpc ? Or is it an object which is in its most rapid star forming phase - an object which has completed forming half the stars seen in a luminous galaxy - a galaxy mass object which is still chemically young ? The first definition seems most sensible but such a primeval galaxy may or may not be the most easily observable.

The different definitions of primeval galaxies lead to differences in what one considers to be the epoch of galaxy formation. Peebles (1989) has reviewed some very simple constraints on this epoch, which deserve mention. The first arises from the sizes of galaxies. Suppose the number density of bright galaxies in the local universe is $n \sim 0.02h^3\mathrm{Mpc}^{-3}$, and each galaxy has an average size $r \sim 10\mathrm{kpc}$. The average separartion between galaxies today is $R \sim n^{1/3} = 4h^{-1}\mathrm{Mpc}$. In the past this separation would decrease roughly as $(1+z)$ and so galaxies would begin to overlap at a redshift earlier than $z_g \sim R/r$. If protogalaxies collapsed by a factor f_c in radius before they became the galaxies we see, then this redshift would be further reduced by the same factor. So we have

$$z_g \lesssim \frac{1}{f_c}(\frac{R}{r}) \sim \frac{200}{f_c} \qquad (5.31).$$

A somewhat more stringent constraint arises from the observed density of a galaxy, say ρ_{obs}. From the spherical model we roughly have

$$\frac{\rho_{obs}}{f_c^3} \sim \frac{9\pi^2}{16}\rho_b(t_m) \sim 5.6\Omega\rho_c(1+z_g)^3 \quad . \qquad (5.32)$$

For a luminous galaxy or cores of DM halos, of mass $\sim 10^{11}M_\odot$ within say a radius $r \sim 10\mathrm{kpc}$, we have $\rho_{obs}/\rho_c \sim 10^5$ and so from (5.32) the redshift

$$z_g \sim \frac{1}{f_c}(\frac{\rho_{obs}}{5.6\Omega\rho_c})^{1/3} \sim \frac{30}{f_c\Omega^{1/3}} \qquad (5.33)$$

If the collapse factor was much greater than unity, one has to allow for the time taken for collapse, which in the spherical model is roughly twice the turn around time. So the above z_g would have to be divided by a further factor of order $2^{2/3}$. What should we take for f_c in the above equations ? In case of dissipationless collapse we estimated on the $f_c = 2$ for the spherical model. On the other hand for the gas which goes to form the disk we may need larger collapse factors ~ 10 in order that it attains rotational support (see section 4.9). So, if gravitational instability is responsible for getting a galactic mass object within a radius of order 10kpc, the redshift of galaxy formation $z_g \lesssim 15$ for the halo cores and is probably less than ~ 3 for the formation of disks.

An interesting limit on the epoch by which galaxies can synthesise their metal content can be derived in the following way (Ostriker quoted in Peebles 1989). Note that heavy elements are made and distributed into the interstellar medium of a galaxy by reasonably massive stars, whose ages are $\sim 10^{7.5}$yr. Since several generations of such stars are needed to make the observed heavy elements, a limit on the time of formation of the bulk of the heavy elements is

$$t_e \approx \frac{2}{3\Omega^{1/2}H_0(1+z_e)^{3/2}} > 10^8\mathrm{yr}, \qquad (5.34)$$

which implies,

$$z_s < z_e < 20h^{-2/3}\Omega^{-1/3} \qquad (5.35)$$

Here z_s denotes the formation redshift of the bulk of the observed stars which are metal rich.

Finally if one wants preferred galaxy mass scales to be imprinted on cooling gas clouds in the way argued in section 4.6, then the redshift z_g and z_s has to be smaller than about 10.

What should primeval galaxies look like and how can one detect them ? The short answer to this question is that nobody knows ; otherwise they would probably have found them. Suppose we wish to search for a primeval galaxy, at say its most luminous phase. If for example this phase occurs at redshifts $z > 5$, then the infrared is the obvious place to look for them. This is because the dip in their spectral energy distribution due to absorption at the Lyman limit will move into the optical band at these redshifts, causing them to be invisible in optical. On the other hand the emission from young stars in the ultra violet to the visible band will be seen in the infrared for such z. So far searches in the infrared have not yielded any candidate primeval galaxy at large z (cf. Peacock 1991).

If primeval galaxies are at much lower redshifts then optical searches would be more suitable. The null results of such searches upto about 1985 has been reviewed by Koo (1986), who concluded that galaxies cannot form at redshifts below about 6. However Koo (1986) based his conclusions on models of galaxy formation which assumed rapid star formation rates, typically that a galaxy of stars formed in a dynamical time of the system (cf. Meier 1976). Baron & White (1987) re-examined this question and pointed out that there is a possible loophole in Koo's argument. They argued that in many galaxy formation theories stars could form over an extended timescale and in sub units which then merged to form the galaxy. They showed that a primeval galaxy could then have a low enough surface brightness to have escaped detection in the searches reviewed by Koo (1986). But is there any positive evidence for galaxy formation at low redshift ?

Recently it seemed that there was indeed such evidence, coming mainly fron the counts of faint galaxies. Such counts show a much higher galaxy number density than expected without evolution. For example Tyson (1988) in his counts of faint galaxies, finds a number density of order $3 \times 10^5 deg^{-2}$, for galaxies brighter than $B_J = 27$. On the other hand Koo (1989) calculates that, in a flat universe, the expected number of galaxies is only about $4.3 \times 10^4 deg^{-2}$ even upto a redshift 4. Here the comoving number density of galaxies has been taken to be $0.0015 \mathrm{Mpc}^{-3}$, corresponding to the density of average (L^*) galaxies in the local universe. One can increase this number by including fainter galaxies, but one has to go almost 5 magnitudes fainter and then it is doubtful if such a galaxy will be seen at all in the counts. It appears therefore that to explain the sheer number of galaxies seen in the faint galaxy counts, the galaxy luminosity function has to evolve with redshift.

Another reason why the faint galaxies are so interesting, is also because a significant fraction of them appear to be somewhat blue in colour (Tyson 1988, see also Lilly, Cowie & Gardner 1991, who however find that the faintest objects in their sample are not as blue as found in the survey of Tyson). The total flux from these faint blue galaxies may be so large that the associated star formation could account for a substantial fraction of present day metals (Cowie 1988). This has lead White & Frenk (1991), for example, to ask whether these faint galaxies are the long sought primeval galaxies. However redshift surveys at the bright end of this population show somewhat surprising results. Broadhurst *et al.*(1988) found that upto a limiting blue magnitude of 21.5 the median redshift of these galaxies in only ~ 0.2. Colless *et al.* (1990) extended these results to objects one magnitude fainter. Their results and also that of Lilly, Cowie & Gardner (1990) shows that even upto 24th B- magnitue where the number density of galaxies is already of order $10^4 deg^{-2}$, the median redshift is about 0.4.

Any model to account for the observed counts of galaxies, which takes recourse to some form of evolution in the galaxy luminosity function, must also be consistent with the low median redshifts discussed above. Rocca-Volmerange & Guiderdoni (1990) have for example examined a model where smaller galaxies progressively merge into bigger ones. Another possibility is that the faint end slope of the galaxy luminosity function steepens with redshifts, possibly due to increasing starburst activity of low luminossity galaxies at larger z (cf. Colless *et al.* 1990; Ellis 1990). However, as pointed out by Lilly, Cowie & Gardner (1991), this explanation may be in trouble in explaining the excess number of galaxies upto the faintest magnitudes. Since it would imply sustained star formation activity over cosmological timescales in these galaxies, which may give rise to a much larger fraction of metals than seen in the prsent day galaxies. The galaxy number counts in the K band may also help in distiguishing between the various possibilities. It appears that the K-band counts are consistent with a no evolution model. Since the K-band light is a more robust tracer of the mass of galaxies this makes models involving large scale merging less promising (cf. Peacock 1991) It appears therefore that the origin of the large excess of galaxies, seen in faint galaxy counts, still poses a challenge. Also if the blue galaxies are indeed low mass galaxies being made to undergo sporadic bursts of star formation, then it is not clear that they represent the major episode of massive galaxy formation, though they may be producing a significant fraction of present day metals.

We see from the above that any discussion about primeval galaxies at present is neccessarily incomplete. If high redshift and high star formation rates were the characteristics of a young galaxy, then the high z radio galaxies discussed in section 5.3 would be ideal candidates. However one has the uncomfortable feeling in this case that these radio galaxies may not be representative of the general population of galaxies. If the ability to produce a significant fraction of the present day metals were the criterion then the faint blue galaxies seen in deep galaxy counts could be called primeval, but as we just mentioned this is not totally satisfactory.

In Figure 5.4 we show schematically, in redshift space, the various high z objects that may be relevant to galaxy formation and which have been discussed in this part of the review. The highest redshift QSOs and radio galaxies show that some developed structure already exists at epochs corespponding to $z \sim 4 - 5$. This seems to be corraborated also by the fact that the IGM already appears to be highly ionised by a redshift of about 4. Coming down to smaller $z \sim 3 - 4$, we see evidence from the Lyα forest of the existence of possibly many intergalactic clouds or mini galaxies. The damped Lyα systems, which are also seen at present up to redshifts of about 3, indicates that a significant fraction of the baryons seen in present day galaxiess was in a neutral form at these redshifts. Whether these systems represent a primeval phase of galaxy formation is an open question. At somewhat lower $z \sim 2 - 2.5$, we

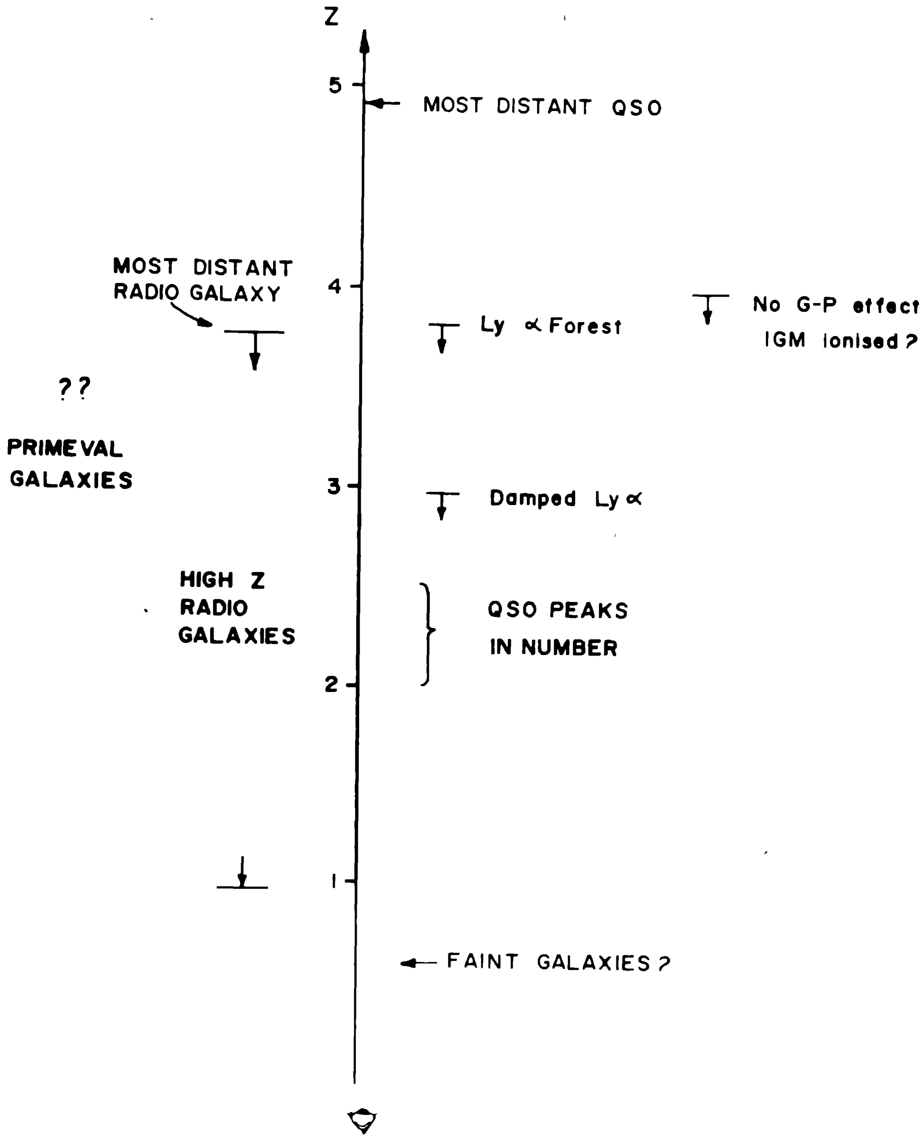

Figure 5.4. The high redshift universe.

have the very interesting epoch where the number density of QSOs and radio galaxies appears to peak. It is again not clear what this is implies for the epoch of galaxy formation. Further down in redshift we come to enigma presented by the large number of objects being revealed in the faint galaxy counts. Ofcourse of direct relevance to galaxy formation would be the discovery of a set of objects which could be clearly identified as primeval galaxies. This is shown as the big question mark in our figure!

We now turn to a consideration of how the perturbations which eventually grew into the structure we see may have arisen.

6: The origin of perturbations

6.1 The concept of inflation

The entire approach to structure formation, outlined so far, relies on the amplification of certain small perturbations which exist in the early universe. Such a picture can be considered complete (and satisfactory) only if we could produce these initial perturbations through some viable physical mechanism. Any attempt in this direction runs into two immediate difficulties:

(i) In the conventional big-bang model, there is no natural seed for these perturbations. Truly primordial perturbations e.g. those due to quantum gravitational effects at $t \approx t_P$ are likely to be of $\mathcal{O}(1)$ and will be difficult to estimate. Thus, the theory lacks predictive power.

(ii) The second difficulty is seen by the following argument. A linear perturbation at a characteristic physical scale λ_0 to-day would correspond to a proper length of $\lambda_0(a(t)/a_0) \propto t^n$ in the past, if we take $a(t) \propto t^n$. The characteristic expansion scale of the universe, on the other hand, is given by the Hubble radius $cH^{-1}(t) = c(\dot{a}/a)^{-1} = cn^{-1}t$. In realistic cosmological models, $n < 1$ and hence the ratio $[\lambda(t)/cH^{-1}(t)]$ increases as we go to the earlier epochs. *In other words $\lambda(t)$ would have been larger than the Hubble radius at sufficiently high redshifts.* Notice that a galactic mass perturbation was bigger than the Hubble radius for redshifts larger than a moderate value of about 10^6. This result leads to a major difficulty in conventional cosmology. It is usually assumed that physical processes can act coherently only over sizes smaller than the Hubble radius. Thus any physical processes leading to small density perturbations at some early epoch $t = t_i$ could have only operated at scales smaller than $cH^{-1}(t_i)$. But most of the relevant astrophysical scales (corresponding to clusters, groups, galaxies, etc.) were much bigger than $H^{-1}(t)$ for reasonably early epochs! Thus, if we want the seed perturbations to have originated in the very early universe, then it is difficult to understand how any physical process could have contributed to it.

It is possible to tackle some of these difficulties of the standard FRW models by modifying the dynamics of the very early universe. The trick lies in introducing a temporary phase during which the universe expanded *exponentially* as in the classical de Sitter model (Kazanas 1980; Sato 1981; Guth 1981). Such an exponentially expanding phase is called 'inflation'. The de Sitter model describes a universe with expansion caused by negative stresses due to the Λ-term. The inflationary universe also requires a Λ-term; but here it arises and is supposed to last only during the transient stage when the GUTs phase transition is taking place. We shall consider the actual mechanisms proposed for this purpose in the next section. Here we will outline the actual model that emerges and the way it can handle some of the awkward features of the standard model.

Consider a model for the universe, in which the universe was radiation dominated upto, say, $t = t_i$, but expanded exponentially in the interval $t_i < t < t_f$:

$$a(t) = a_i \exp H(t - t_i) \qquad t_i \leq t \leq t_f \tag{6.1}$$

For $t > t_f$, the evolution is again radiation dominated $[a(t) \propto t^{1/2}]$ until $t = t_{eq} \cong 4.36 \times 10^{10}(\Omega h^2)^{-2}$s. The evolution becomes matter dominated for $t_{eq} < t < t_{now} =$

t_0. Typical values for t_i, t_f, and H, suggested in the literature are:

$$t_i \approx 10^{-35}\text{s}; \quad H \approx 10^{10}\text{GeV}; \quad t_f \approx 70 H^{-1} \tag{6.2}$$

which give an overall 'inflation' of about $A \equiv \exp N \cong \exp(70) \approx 2.5 \times 10^{30}$ to the scale factor in the period $t_i < t < t_f$. At $t = t_i$, the temperature of the universe is about 10^{14}GeV. During this exponential inflation, the temperature drops drastically but the matter is expected to be reheated to the initial temperature $\sim 10^{14}$GeV at $t \approx t_f$. The reheating takes place when the phase transition is over and the energy released in the process is passed on to the radiation content of the universe. The situation is analogous to the reheating that takes place when supercooled steam condenses and releases its latent heat. Thus, inflation effectively changes the value of $S = T(t)a(t)$ by a factor $A = \exp(70) \approx 10^{30}$. Note that this quantity S is conserved during the non- inflationary phases of the expansion.

Such an evolution, if it can be implemented dynamically, has several interesting features. The most attractive feature of the inflationary model is probably the possibility of generating the seed perturbations which can grow to form the large scale structures (Bardeen *et al.* 1983; Guth & Pi 1982; Hawking 1982; Starobinsky 1982). This is realised in the following manner:

In the FRW models with $a(t) \propto t^n$ $(n < 1)$, the physical wavelengths (which grow as $\lambda \propto a \propto t^n$) will be far larger than the Hubble radius (which grows as $H(t)^{-1} \propto t$) in the early phases. This situation is drastically altered in an inflationary model. During inflation, physical wavelengths grow exponentially $[\lambda \propto a \propto \exp Ht]$ while the Hubble radius remains constant. Therefore, a given length scale has the possibility of crossing the Hubble radius twice in the inflationary models. Consider, for example, a wave length $\lambda_0 \sim 2$Mpc today (which contains a mass of a typical galaxy $1.2 \times 10^{12}(\Omega h^2)M_\odot$). This scale would have been

$$\lambda(t_f) = \lambda_0 \frac{a(t_f)}{a(t_0)} = 2\text{Mpc}\left(\frac{T_0}{T(t_f)}\right) \simeq 1.8 \times 10^{-2}\text{cm} \tag{6.3}$$

at the end of inflation. (This is, of course, much larger than the typical Hubble radius at that epoch, $cH^{-1} \approx 1.4 \times 10^{-24}$cm). But at the beginning of inflation, its proper length would have been

$$\lambda(t_i) = \lambda(t_f).\frac{a(t_i)}{a(t_f)} = A^{-1}\lambda(t_i) = 1.8 \times 10^{-32}\text{cm} \tag{6.4}$$

This is much smaller than the Hubble radius. This Hubble radius remains constant throughout the inflation while λ increases exponentially. In about $\Delta t = t - t_i \simeq 18H^{-1}$, λ will grow as big as the Hubble radius.

The situation is summarised in fig.(6.1). We see that the scales which are astrophysically relevant today were much smaller than the Hubble radius at the onset of inflation. (Therefore, causal processes could have operated at these scales). During inflation, the proper wavelength grows, and becomes equal to the Hubble radius cH^{-1} at some time $t = t_{\text{exit}}$. For a mode labeled by a wave vector $\mathbf{k}$, this happens at $t_{\text{exit}}(k)$ where,

$$\frac{2\pi}{k} a(t_{\text{exit}}) = cH^{-1} \tag{6.5}$$

That is, when $(kc/aH) = 2\pi$. In the radiation dominated era after the inflation, the proper length grows only as $t^{1/2}$ while the Hubble radius grows as t; Thus the Hubble radius "catches up" with the proper wavelength at some $t = t_{enter}(k)$. For $t > t_{\text{enter}}$, this wavelength will be completely within the Hubble radius.

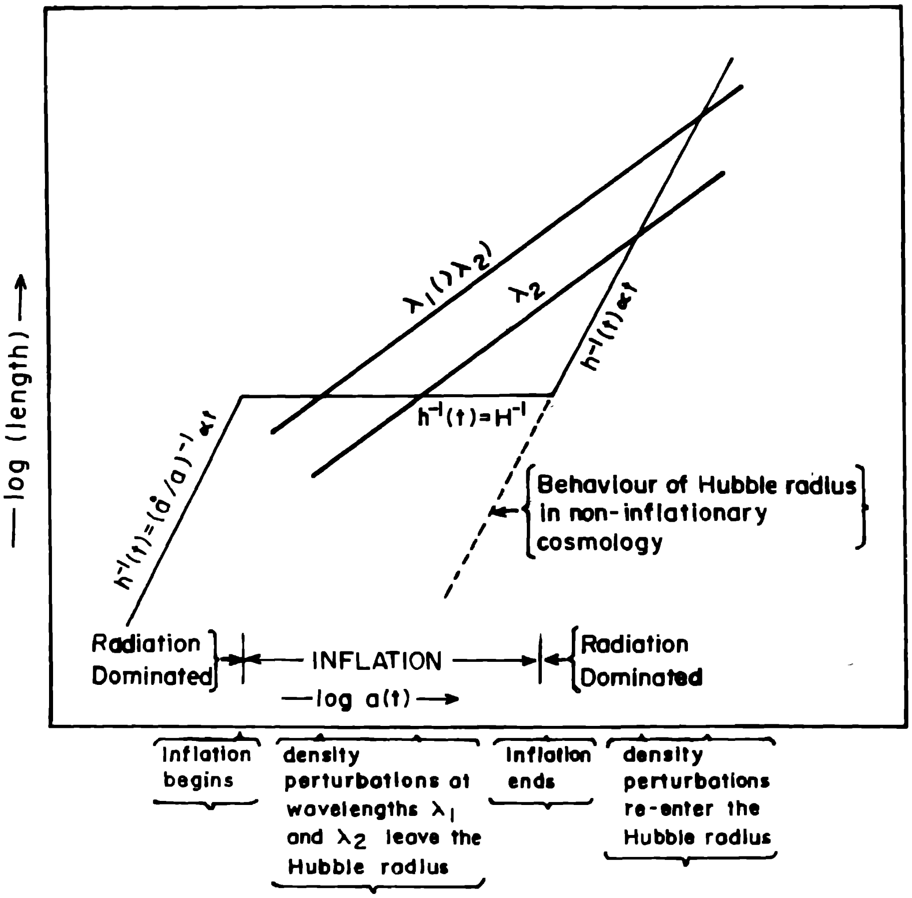
— log (length) →
$\lambda_1(()\lambda_2)$
λ_2
$h^{-1}(t) \propto t$
$h^{-1}(t)=(\dot{a}/a)^{-1} \propto t$
$h^{-1}(t)=H^{-1}$
Behaviour of Hubble radius in non-inflationary cosmology
Radiation Dominated
INFLATION
— log a(t) →
Radiation Dominated
Inflation begins
density perturbations at wavelengths λ_1 and λ_2 leave the Hubble radius
Inflation ends
density perturbations re-enter the Hubble radius

Figure 6.1. Wavelength and Hubble radius in the inflationary model.

Thus inflationary models allow λ to be less than cH^{-1} at *two* different epochs: an early phase, $t < t_{\text{exit}}(k)$ and a late phase, $t > t_{\text{enter}}(k)$. Any perturbation generated by physical processes at $t < t_{\text{exit}}$ can be preserved intact during $t_{\text{exit}} < t < t_{\text{enter}}$ and can lead to formation of structures at $t > t_{\text{enter}}$.

What can lead to perturbations at $t < t_{\text{exit}}$? Since the physical processes taking place in this epoch are quantum mechanical by nature, quantum fluctuations in matter fields are obvious candidates as seed perturbations. Therefore, in principle, we can now generate (and compute) the density inhomogeneities in the universe. This is indeed a major achievement of inflation.

Notice that all the above conclusions only depend on the scale factor growing rapidly (by a factor 10^{30} or so) in a short time. For example, if the energy density

$\epsilon(t)$ varies slowly during $t_i < t < t_f$, then one has near exponential expansion with

$$a(t_f) = a(t_i) \exp \int_{t_i}^{t_f} H(t)dt \equiv a(t_i) \exp N \tag{6.6}$$

where $H^2(t) = (8\pi G\epsilon(t)/3c^2)$. This provides a general definition of N.

6.2. The epicycles of Inflation

Since the inflationary idea seems to be quite attractive, several mechanisms were devised by which this idea can be implemented. Each of these models has some advantages and disadvantages and none of them is completely satisfactory. We will briefly summarise three different models.

For the universe to expand exponentially, the energy density should remain (at least approximately) constant. Various models of inflation differ in the process by which this is achieved. In most of them the 'quasi-constant' energy density $\epsilon(t)$ is derived from phase transition at the GUT epoch. Although in specific details one grand unified theory may differ from another almost all of them involve gauge theories with a mediating role played by the Higgs scalar field ϕ. We need not go into the intricacies of how ϕ is related to the other matter fields. The feature which is of interest to us is that the potential energy density V of the scalar field ϕ depends on the ambient temperature T.

At any given temperature T which is higher than a critical temperature T_c, the minimum value of V is found to be at the expected zero of ϕ. We may term this minimum at $\phi = 0$ as the 'vacuum state' of ϕ. As the temperature is lowered, however, it may happen that the minimum of V no longer remains at $\phi = 0$ but shifts to a finite value $\phi = \sigma$. This 'phase transition' occurs at $T = T_c$ and may be likened to the condensation of steam. Thus ϕ would tend to transit from $\phi = 0$ to $\phi = \sigma$.

If ϕ were to 'condense' immediately at T_c, all the excess energy could be released at once. However, in the more likely case of 'supercooling', ϕ may continue at $\phi = 0$ and move to the true minimum $\phi = \sigma$ later. During this transitional stage the state $\phi = 0$ is called the 'false vacuum' state since the 'true vacuum' is now at $\phi = \sigma$. The original model for inflation, due to Guth, invoked this temperature dependence of the potential energy of the Higgs field $V(\phi, T)$. Here $T_c \approx 10^{14}$GeV.

At temperatures $T \gg T_c$, the potential V has only one minimum (at $\phi = 0$) with $V(0) \approx (10^{14}\text{GeV})^4$. As the temperature is lowered to $T \sim T_c$, a second minimum appears at $\phi = \sigma$. For $T \ll T_c$, the $\phi = \sigma$ minimum is the 'true' minimum. [i.e. $V(\sigma) \approx 0 \ll V(0)$]. Now consider what happens in the early universe as matter cools through $T \approx T_c$. At $T \gg T_c$ the minimum configuration corresponds to $\phi = 0$ while for $T \sim T_c$ it is $\phi = \sigma$. But the matter in the universe does not instantaneously switch over from $\phi = 0$ to $\phi = \sigma$. The universe can get "stuck" at $\phi = 0$ (the "false vacuum"), with $V = V(0)$, even at $T < T_c$ and will expand exponentially because the dominant energy density driving the expansion is the constant $V(0) - V(\sigma) \approx V(0)$. Over the course of time thermal fluctuations and quantum tunneling will induce a transition from the 'false' vacuum $\phi = 0$ to the 'true' vacuum $\phi = \sigma$ ending the inflation in localized regions ("bubbles"). The phase transition is expected to be completed by the expanding 'bubbles' colliding, coalescing and reheating the matter.

Detailed analysis, however, shows that this model does not work (Guth & Weinberg 1983). In order to have sufficient amount of inflation, it is necessary to keep the "false" vacuum fairly stable. In such a case the bubble nucleation rate is small and even the resulting bubbles do not coalesce together efficiently. The final configuration is very inhomogeneous and quite different from the universe we need.

The original model was soon replaced by a version based on a very special form for $V(\phi)$ called the Coleman-Weinberg potential (Albrecht & Steinhardt 1982; Linde 1982 a,b). At zero temperature this potential is given by :

$$V(\phi) = \frac{1}{2}B\sigma^4 + B\phi^4[\ln\frac{\phi^2}{\sigma^2} - \frac{1}{2}]; \quad B \approx 10^{-3}; \quad \sigma \approx 2 \times 10^{15}\text{GeV}. \tag{6.7}$$

This potential is extremely flat for $\phi \lesssim \sigma$ and drops rapidly near $\phi \approx \sigma$. At finite temperatures, the potential picks up a small barrier near the origin [at $\phi \simeq \mathcal{O}(T)$] with height $\mathcal{O}(T^4)$, creating a local minimum at $\phi = 0$. This 'false' vacuum, however, is quite unstable when the temperature becomes $\mathcal{O}(10^9\text{GeV})$. The scalar field rapidly tunnels to $\phi \approx \phi_0 \approx \mathcal{O}(H)$, and starts 'rolling down' the gently sloped potential towards $\phi = \sigma$. Since the potential is nearly flat in this region, the energy density driving the universe is approximately constant and about $V(0) \approx (3 \times 10^{14}\text{GeV})^4$. The evolution of the scalar field in this 'slow roll-over' phase can be approximated as

$$\Box\phi + V'(\phi) = \ddot{\phi} + 3H\dot{\phi} + V'(\phi) \approx 3H\dot{\phi} + V'(\phi) = 0 \tag{6.8}$$

where we have ignored the $\ddot{\phi}$ term and $H = (4\pi BG\sigma^4/3c^2) \approx 2 \times 10^{10}\text{GeV}$ (in energy units). If the slow roll over lasts when ϕ varies from $\phi_{\text{start}} \simeq \mathcal{O}(H)$ to some $\phi_{\text{end}} \lesssim \mathcal{O}(\sigma)$ then

$$N \equiv \int_{t_i}^{t_f} Hdt = H\int_{\phi_s}^{\phi_e} \frac{d\phi}{|\dot{\phi}|} \approx 3\int_{\phi_s}^{\phi_e} \frac{H^2}{|V'(\phi)|}d\phi. \tag{6.9}$$

For the typical values of the Coleman-Weinberg potential this number can easily be about 10^2 ensuring sufficient inflation.

As ϕ_0 approaches σ, the field "falls down" the potential and oscillates around the minimum at $\phi = \sigma$ with the frequency $\omega^2 = V''(\sigma) \approx (2 \times 10^{14}\text{GeV})^2 \gg H^2$. These oscillations are damped by the decay of ϕ into other particles (with some decay time Γ^{-1}, say), and by the expansion of the universe. If $\Gamma^{-1} \ll H^{-1}$, the coherent field energy $(\frac{1}{2}\dot{\phi}^2 + V)$ will be converted into relativistic particles in a timescale $\Delta t_{\text{reheat}} \approx \Gamma^{-1} \ll H^{-1}$. This will allow the universe to be reheated to a temperature of about $T_{\text{reheat}} \approx \omega \approx 2 \times 10^{14}\text{GeV} \approx T_{\text{initial}}$. The decay width of several Coleman-Weinberg models can be about $\Gamma^{-1} \approx 10^{13}\text{GeV} \gg H$. This ensures good "reheating" of the universe (Albrecht *et al.* 1982; Dolgov & Linde 1982). Since the field has already tunneled out of the false vacuum before the onset of inflation, we do not face the problems which plagued the original inflation. Instead of several bubbles having to collide, coelesce and make up the whole observable universe of today, we have one huge bubble encompassing everything observable now.

Though an improvement on the original version, this model is also not free from problems. It turns out that the field should start its slow "roll over" from a value $\phi_s \approx H$ to ensure sufficient inflation. It can be shown that the quantum fluctuations in the scalar field are about $\Delta\phi \simeq (H/2\pi)$ (Linde 1982b; Vilenkin & Ford 1982). Since $\phi_s \sim \Delta\phi$, the entire analysis based on semiclassical $V(\phi)$ is of doubtful validity. The second - and more serious - difficulty stems from the calculation of density perturbations in this model: they turn out to be too large by a factor of about 10^6, unless the parameter B is artificially reduced by a factor 10^{12} or so!

The original model for inflation used a strongly first order phase transition while the second model may be considered to be using a weakly first order (or even second order) phase transition. It is possible to construct inflationary scenarios in which no phase transition is involved. The idea of "chaotic inflation", suggested by Linde falls in this class. (Linde 1983) In this model, the potential has a very simple form: $V(\phi) = \lambda\phi^4$. Inflation results because of the rather slow motion of ϕ from some initial value ϕ_0 towards the minimum. (The initial non-zero value of the ϕ_0 is supposed to be due to 'chaotic' initial conditions). This model can also lead to sufficient inflation but suffers from two other difficulties: (i) To obtain the correct value for the density perturbation, it is necessary to fine-tune λ to very small values: $\lambda \approx 4 \times 10^{-14}$. (ii) In order for the inflation to take place the kinetic energy of the scalar field has to be small compared to its potential energy. Detailed calculation shows that this requires the field to be uniform over sizes bigger than the Hubble radius! This is completely against the original spirit of inflation.

A further epicycle in the saga of inflation envisages a universe that is without a big bang origin. In this version the de Sitter type inflationary phase is *self reproducing* in a chaotic set up with the help of large scale quantum fluctuations of a scalar field φ. The bubbles of FRW models are nucleated in it at random points of space and time through quantum phase transitions.

A solution to the bubble nucleation and coelescence problem of the original Guth model (sometimes referred to as the 'graceful exit' problem) was proposed in yet another way (La & Steinhardt 1989; Steinhardt & Accetta 1990). In their 'extended inflationary cosmology' these authors used the Brans-Dicke theory of gravity instead of general relativity as the background theory for the early universe. The inflationary phase in this model has a power law type of expansion factor instead of the exponential one, thus making it possible for the inflationary phase to end gracefully through bubble nucleation.

Nevertheless this idea also ran into trouble with distortions of the MBR and was changed to 'hyper-extended inflation'. The background theory of gravity for this model differs from the Brans-Dicke theory through the inclusion of higher order couplings of the scalar field with gravity. In a rapidly changing subject in which the half-life of a theory is one year it is hard to pass judgement on the merits of this scenario.

The schemes and shortcomings discussed above are typical of several other models suggested in the literature. The most serious constraint on inflationary scenarios arises from the study of density perturbations. No single model for inflation, suggested so far, can be considered completely satisfactory.

6.3 The origin of scale-invariant spectrum

The most attractive feature of inflation, from the point of view of an astronomer, is the possibility that inflation may provide the seed perturbations which grow to form the structures we see today. In this section we will overview how this is achieved and what difficulties arise.

The most natural choice for the origin of seed perturbations, in the context of inflation, comes from the quantum fluctuations in the scalar field $\phi(t, \mathbf{x})$ driving the inflation. The computation of classical perturbations, generated by a quantum field is a difficult and technically involved issue. Several questions of principle are still unresolved in this calculation (see e.g. Padmanabhan, Seshadri &.Singh 1989). Since this review is primarily intended for the astronomer, we will avoid the technical aspects of the calculation and content ourselves by discussing the physical idea.

During inflation, the universe was assumed to be - on the average - in a FRW state with small inhomogeneties. This implies that the source - which is a *classical* scalar field $\Phi(t, \mathbf{x})$ - can be split as $\phi_0(t) + f(t, \mathbf{x})$ where $\phi_0(t)$ denotes the average, homogeneous, part and $f(t, \mathbf{x})$ represents the spatially dependent, fluctuating part. Since the energy density due to a scalar field is $\rho c^2 \cong (1/2)\dot{\phi}^2$, we get,

$$\delta\rho(t, \mathbf{x}) = \rho(\mathbf{x}, t) - \overline{\rho}(t) \cong \dot{\phi}_0(t)\dot{f}(t, \mathbf{x})/c^2 \tag{6.10}$$

(where $\overline{\rho}(t) = \frac{1}{2}\dot{\phi}_0(t)^2$ and we have assumed $f \ll \phi_0$). The Fourier transform will now give

$$\delta\rho(\mathbf{k}, t)c^2 \cong \dot{\phi}_0(t)\dot{Q}_k(t), \tag{6.11}$$

where we have put

$$f(t, \mathbf{x}) \equiv \int \frac{d^3\mathbf{k}}{(2\pi)^3} Q_k(t) e^{i\mathbf{k}\ \mathbf{x}}. \tag{6.12}$$

Since the average energy density during inflation is dominated by the constant term V_0, we have the density contrast

$$\delta(\mathbf{k}, t) = \frac{\delta\rho c^2}{V_0} = \frac{\dot{\phi}_0(t)\dot{Q}_k(t)}{V_0}. \tag{6.13}$$

It might now appear that all we have to do is to compute the quantities $\phi_0(t)$ and $Q_k(t)$ from the equation of motion for the scalar field. For $\phi_0(t)$ we can use the mean evolution of the scalar field during the slow roll-over phase and determine $\dot{\phi}_0(t)$ from the classical solution. The fluctuating field $f(t, \mathbf{x})$ is supposed to be some *classical* object mimicing the quantum fluctuations. Such a quantity is conceptually difficult to visualise, and justify. What is usually done is to choose some convenient quantum mechanical measure for fluctuations and *define* Q_k in terms of this quantity.

In quantum theory, the field $\hat{\phi}(t, \mathbf{x})$ and its Fourier coefficients $\hat{q}_k(t)$ will become operators related by

$$\hat{\phi}(t, \mathbf{x}) = \int \frac{d^3k}{(2\pi)^3} \hat{q}_k(t) e^{i\mathbf{k}.\mathbf{x}}. \tag{6.14}$$

The quantum state of the field can be specified by giving the quantum state $\psi_k(q_k, t)$ of each of the modes $\hat{q}_k$. (One can think of q_k as coordinates of a particle and $\psi_k(q_k, t)$ as the wavefunction describing this particle.)

The fluctuations in q_k can be characterised by the dispersion

$$\sigma_k^2(t) =< \psi|q_k^2(t)|\psi > - < \psi|q_k(t)|\psi >^2 =< \psi|q_k^2(t)|\psi > \tag{6.15}$$

in this quantum state. (The mean value of the scalar field operator $<\hat{\phi}(t,\mathbf{x})>=\phi_0(t)$ is homogeneous; therefore, we have set $<\hat{q}_k>$'s to zero in the above expression. Note that we are interested in only the $\mathbf{k}\neq 0$ modes). Expressing $\hat{q}_k$'s in terms of $\hat{\phi}(t,\mathbf{x})$ it is easy to see that

$$\sigma_k^2(t)=\int d^3\mathbf{x}<\psi|\hat{\phi}(t,\mathbf{x}+\mathbf{y})\hat{\phi}(t,\mathbf{y})|\psi>e^{i\mathbf{k}.\mathbf{x}}. \tag{6.16}$$

In other words, the 'power spectrum' of fluctuations σ_k^2 is related to the Fourier transform of the two-point-correlation function of the scalar field. Since $\sigma_k^2(t)$ appears to be a good measure of quantum fluctuations, we may attempt to *define* $Q_k(t)$ as

$$Q_k(t)=\sigma_k(t). \tag{6.17}$$

This is equivalent to *defining* the fluctuating classical field $f(t,\mathbf{x})$ to be

$$f(t,\mathbf{x})\equiv\int\frac{d^3k}{(2\pi)^3}\sigma_k(t)e^{i\mathbf{k}.\mathbf{x}}. \tag{6.18}$$

This leads to the result

$$\delta(\mathbf{k},t)=\frac{\dot{\phi}_0(t)}{V_0}\dot{\sigma}_k(t). \tag{6.19}$$

The procedure may be summarised as follows: (i) In quantum theory, the field $\hat{\phi}(t,\mathbf{x})$ and its Fourier coefficient $\hat{q}_k(t)$ become operators. In any quantum state, the variables will have a mean value and fluctuations around this mean value. (ii) Since the mean evolution of the scalar field is described by a *homogeneous* part $\phi_0(t)$, we expect the mean values of $\hat{q}_k$'s to vanish (for $\mathbf{k}\neq 0$); $<\psi|q_k(t)|\psi>=0$. However, the fluctuations around these mean values, characterised by $\sigma_k^2(t)=<\psi|q_k^2|\psi>$ do not vanish. (iii) We incorporate these quantum fluctuations in a semi-classical manner by taking the scalar field to be $\Phi(t,\mathbf{x})=\phi_0(t)+f(t,\mathbf{x})$ where $f(t,\mathbf{x})$ is related to $\sigma_k(t)$ by (6.18). (iv) The density perturbations are calculated by treating $\Phi(t,\mathbf{x})$ as a classical object.

The expression derived above gives the value of $\delta(\mathbf{k},t)$ in the inflationary phase: $t_i<t<t_f$. To compare this with observations, we need to know the value of $\delta(\mathbf{k},t)$ at $t=t_{\rm enter}(k)$ - that is when the perturbations enter the Hubble radius. Fortunately, there exists a (approximate) conservation law which relates the value $\delta(\mathbf{k},t_{\rm enter})$ with $\delta(\mathbf{k},t_{\rm exit})$ where $t_{\rm exit}(k)$ is the time at which the relevant perturbation 'leaves' the Hubble radius in the inflationary epoch (Bardeen *et al.* 1983; Frieman & Turner 1984). This law can be stated as

$$\frac{\delta(\mathbf{k},t_{\rm exit}(k))}{1+w(t_{\rm exit})}=\frac{\delta(\mathbf{k},t_{\rm enter}(\mathbf{k}))}{1+w(t_{\rm enter})} \tag{6.20}$$

where $w(t)$ is the ratio between pressure $p(t)$ and density $\rho(t)$ of the background (mean) medium: $w(t)=p(t)/\rho(t)$. In the inflationary phase with the scalar field,

$$p(t)=\frac{1}{2}\dot{\phi}_0^2-V_0;\quad \rho(t)=\frac{1}{2}\dot{\phi}_0^2+V_0;\quad 1+w(t)\cong\frac{\dot{\phi}_0^2}{V_0}, \tag{6.21}$$

where we have used the fact $\dot{\phi}_0^2 \ll V_0$. In the radiation dominated phase (at $t = t_{\text{enter}}$), $1 + w = 4/3$. Therefore

$$\delta(\mathbf{k}, t_{\text{enter}}) = \delta(\mathbf{k}, t_{\text{exit}}).\frac{4}{3}\left(\frac{V_0}{\dot{\phi}_0^2}\right); \tag{6.22}$$

or using (6.19),

$$\delta(\mathbf{k}, t_{\text{enter}}) = \frac{4}{3}\left(\frac{\sigma_k}{\dot{\phi}_0}\right)_{t=t_{\text{exit}}} \cong \left(\frac{\sigma_k}{\dot{\phi}_0}\right)_{t=t_{\text{exit}}} \tag{6.23}$$

This is the final result.

The problem now reduces to computing $\sigma_k(t)$ and $\phi_0(t)$, which can be done once the potential $V(\phi)$ is known. For a Coleman-Weinberg potential, detailed calculations (see e.g. Brandenberger 1985) give the result:

$$\delta(\mathbf{k}, t_{\text{enter}}) \approx \lambda^{1/2} N^{3/2} k^{-3/2} \approx 10^2 k^{-3/2} \tag{6.24}$$

where we have taken the effective e-folding time $N \approx 50$ and $\lambda \approx 0.1$. We see that the density perturbations have the scale-invariant spectrum but too high an amplitude. To bring it down to the acceptable value of about 10^{-4}, we need to take the dimensionless parameter λ to be about 10^{-13}! This requires an extreme finetuning for a dimensionless parameter especially since we have no other motivation for such a value.

This has been the most serious difficulty faced by all realistic inflationary models: they produce too large an inhomogeneity. The qualitative reason for this result can be found from (6.23). To obtain *slow* roll-over and sufficient inflation we need to keep $\dot{\phi}_0$ small which tends to increase the value of δ. We could have saved the situation if it were possible to keep σ_k arbitrarily small; unfortunately the inflationary phase induces a fluctuation of about $(H/2\pi)$ on any quantum field due to field theoretical reasons. This lower bound prevents us from getting sensible values for δ unless we fine-tune the dimensionless parameters of $V(\phi)$. Several 'solutions' have been suggested in the literature to overcome this difficulty but none of them appear to be very compelling. (see e.g. Ellis *et al.* 1985; Holman *et al.* 1984; Jensen *et al.* 1986; Padmanabhan 1988; Padmanabhan, Seshadri & Singh 1989).

Before concluding the discussion on inflation it is probably worth mentioning a definite prediction which emerges from inflationary models. It turns out that the same mechanism which produces the density inhomogenieties also produces gravitational wave perturbations. (see e.g. Abbot & Wise 1984; Allen 1988; Fabri & Pollock 1983; Rubakov *et al.* 1982; Yajnik 1990). These perturbations also have a scale invariant power spectrum and an r.m.s. amplitude of about $(H/10^{19}\text{GeV})$. The energy density of the gravitational waves contributes a fraction $\Omega_{\text{grav}} \approx 10^{-5}(H/m_P)^2 h^{-2}$ to the critical density. Such perturbations can induce a quadrupole anisotropy in the MBR background. The present bounds on this anisotropy ($\lesssim 10^{-4}$) suggest that $H < 10^{15}$GeV. The value of Ω_{grav} can be also restricted by the timing measurements of the millisecond pulsar; the present bound is $\Omega_{\text{grav}}(\lambda \sim 1pc) \leq 3 \times 10^{-7}$. A positive detection of quadrupole anisotropy in MBR or a direct detection of relic gravitational radiation will certainly go a long way in boosting confidence in inflation. [The

Laser Interferometer Gravity Wave Observatory (LIGO) and similar projects can, in principle, reach a sensitivity of $\Omega_{\rm grav} \sim 10^{-11}$].

7: Concluding remarks

We have emphasised in this review several general aspects of how gravitational instability may act on initially small density perturbations in the universe and lead to the development of the large scale structure and the galaxies that we see. We have also examined how several properties of galaxies may arise and the relevance of the high redshift universe to probe galaxy formation. We have not given a detailed discussion of particular theories, although the machinery that we have outlined particularly in parts 2 , 3 and the first half of part 4 can be used to examine any such theory.

At present it appears that no particular theory satisfactorily accounts for all the observed structure in the universe ; although many people would agree that the gravitational instability paradigm seems to offer the best hope of finding such a successful theory eventually. Peebles and Silk (1990) have compared the relative merits of several different theories in their cosmic book of odds. A glance at the list of various observational features that they feel any theory should account for, makes it apparent that galaxy formation theories are quite strongly constrained. It is no longer possible to speculate wildly when considering theories of galaxy formation ; one has to satisfy constraints from a wide range of phenomena.

One crucial test of the gravitational instability paradigm will be the detection of small fluctuations in the cosmic microwave background, especially on large scales where causal processes have not had time to act. That no fluctuation in the CMBR, other than the dipole, has been detected so far already provides a severe constraint on theories. It pushes the epoch of galaxy formation to low redshifts. At the same time the wealth of objects seen at high redshifts argues that the epoch of galaxy formation was not too recent. It remains to be seen how these observations firm up in the coming years and how any theory would eventually satisfy both constraints. If and when fluctuations in the CMBR are detected, they would provide an invaluable probe of the early universe. On the other hand a growing need to change the conventional paradigm would arise if the CMBR continues to show no fluctuations even as observations become more and more senstitive.

As far as the potential use of other types of observations, the recent efforts in conducting large redshift surveys is particularly interesting. It has brought to the fore another important probe of structure formation theories, the peculiar velocity field. We also keenly await radio observations of the redshifted 21 cm line from neutral hydrogen at high redshifts. This could be a powerful diagnostic of structure formation theories. In fact, recently Uson, Bagri and Cornwell (private communication) have claimed to have detected with the VLA the 21 cm line redshifted from $z = 3.4$, in absorption against a high redshift radio galaxy, and more interestingly in emission from a nearby region at the same redshift. The Giant Meterwave Radio Telescope being presently set up in India (Swarup 1984, 1990) will have a larger collecting area and is expected to perform better at meter wavelengths than the VLA. It should be possible, therefore with GMRT, to probe the neutral hydrogen phase of the universe at high redshift. We hope for exciting new discoveries from this latest window to the high redshift universe.

Acknowledgements

The idea for this review came up during a workshop held at TIFR during Sept.,90. We thank Rajesh Kochhar for encouraging us to produce such a review. We also thank Varun Sahni for sharing with us several figures prior to publication and Vijay Kapahi for the figures of the high redshift radio galaxy shown here.

Postscript

This review was completed in June, 91. Since then a number of exiting developments have taken place. The foremost is the discovery of fluctuations in the CMBR by the Cosmic Microwave Background Explorer (COBE) satellite. Smoot *et al.* (1992) have analysed the first year of the data from the Differential Microwave Radiometer (DMR) of COBE. After subtracting the mean and the dipole anisotropy of the CMBR, they find an rms sky variation in the CMBR temperature (smoothed over 10 degrees) of 30 ± 5 μK. The rms quadrapole amplitude is 13 ± 4 μK. These numbers translate into fluctuations in the gravitational potential $\sim 10^{-5}$. The data is also consistent with a Harrison - Zeldovich spectrum of density fluctuations. This is the first time one has obtained a direct view of the small fluctuations which could have seeded galaxy formation. The observed anisotropy also provides a way of unambiguosly normalising the power spectrum of density fluctuations of any given theory of structure formation ; once and for all ! A preliminary discussion of the implications of the COBE detection for structure formation theories has been given by Wright *et al.* (1992) and Padmanabhan & Narasimha (1992). The latter authors point out that the COBE result combined with data from galaxy surveys, severely constrains the shape of the density spectrum. There appears to be a sharp break in the density spectrum at about $50h^{-1}$ Mpc which may pose problems for existing theories of galaxy formation. Another important development, which we had referred to above, (and whose details are now in print) is the radio detection by Uson, Bagri & Cornwell (1992) of neutral hydrogen, in emission, from a redshift $z = 3.4$. The observed flux of redshifted 21 cm emission implies that the mass of HI in this object is $\sim 2 \times 10^{13}h^{-2}M_{\odot}$. This large HI mass has lead Uson *et al.* to identify the emitting object as the first example of a cluster sized Zeldovich pancake, of the kind expected in HDM theories. Subramanian and Swarup (1992) have focused, as an alternative, on theories where galaxies form before cluters. They suggest that the Uson *et al.* object may well be a large collection of HI rich protogalaxies, of the kind needed to explain the damped Lyman α absorption systems seen in quasar spectra. The detection of many more such objects promises to provide strigent constraints on theories of galaxy formation.

References

Aarseth, S.J. (1984) in *Methods of computational physics* (eds. J.U. Brackbill & B.I. Cohen) Academic.

Abbott, L. & Wise, M. (1984) *Nucl. Phys.* **B244**, 541.

Albrecht, A. & Steinhardt, P.J. (1982) *Phys. Rev. Lett.* **48**, 1220.

Albrecht, A., Steinhardt, P.J., Turner, M.S. & Wilczek, F. (1982) *Phys. Rev. Lett.* **48**, 1437.

Allen,B. (1988) *Phys. Rev.* **D37**, 2078.

Bachall, N.A. & Soneira, R. (1983) *Ap. J.* **270**, 20.

Bajtlik, S., Duncan, R. & Ostriker, J.P. (1987) *Ap. J.* **327**, 570.

Bardeen, J.M. (1980) *Phys. Rev. D.*, **22**, 1982.

Bardeen, J.M., Steinhardt, P.J. & Turner, M.S. (1983) *Phys. Rev.* **D28**, 679.

Bardeen, J.M., Bond, J.R., Kaiser, N. & Szalay, A.S., (1986) *Ap. J.* **300**, 15.

Barnes, J. & Efstathiou, G. (1987) *Ap. J.* **319**, 575.

Barnes, J. (1989) *Nature* **338**, 123.

Baron, E. & White, S.D.M. (1987) *Ap. J.* **322**, 585.

Baron, E., Carswell, R.F., Hogan, C.J. & Weymann, R.J. (1989) *Ap. J.* **337**, 609.

Barthel, P.D. (1989) *Ap. J.* **336**, 606.

Bechtold, J. (1987) in *High redshift and primeval galaxies* (eds. J. Bergeron, D. Kunth, B. Rocca-Volmerange & J. Tran Thanh Van) Frontiers; France, p.397.

Begelman, M.C. & Cioffi, D.F. (1989) *Ap. J. Lett.* **345**, L21.

Bergeron, J. (1988) in *QSO Absorption Lines* (eds. J.C. Blades, D. Turnshek & C.A. Norman) Cambridge Univ. Press, p.127.

Binney, J. (1978) *M.N.R.A.S.* **183**, 501.

Binney, J. (1977) *Ap. J.* **215**, 483.

Binney, J. & Tremaine, S. (1987) *Galactic dynamics*, Princeton Univ. Press.

Bithell, M. & Rees, M.J. (1990) *M.N.R.A.S.* **242**, 570.

Black, J.H. (1981) *M.N.R.A.S.* **197**, 553.

Black, J.H., Chaffe, F.H. & Foltz, C.B. (1987) *Ap. J.*, **317**, 442.

Blumenthal, G.R., Faber, S.M., Primack, J.R. & Rees, M.J. (1984) *Nature* **341**, 517.

Bond, J.R., Efstathiou, G. & Silk, J. (1980) *Phys. Rev. Lett.* **45**, 1980.

Bond, J.R. & Szalay, A.S. (1983) *Ap. J.* **274**, 443.

Bond, J.R., Cole, S., Kaiser, N. & Efstathion, G., (1991) *Ap.J.* **379**, 440.

Boyle, B.J. (1990) Talk at *Texas/ESO - CERN symposium on relativistic astrophysics, cosmology and fundamental physics*, Brighton, U.K.

Boyle, B.J., Fang, R., Shanks, T. & Peterson, B.A., (1987) *M.N.R.A.S.* **227**, 717.

Braun, E., Dekel, A. & Shapiro, P.R. (1988) *Ap. J.* **328**, 34.

Briggs, F.H., Wolfe, A.M., Liszt, H.S., Davis, M.M. & Turner, K.L. (1989) *Ap. J.* **341**, 650.

Broadhurst, T.J., Ellis, R.S. & Shanks, T. (1988) *M.N.R.A.S.* **235**, 827.

Brodie, J.P., Bowyer, S. & McCarthy, P.J. (1985) *Ap. J. Lett.*, **293**, L59.

Burgers, J.M. (1940) *Proc. R. Neth. Acad. Sci.* **43**, 2.

Burgers, J.M. (1974) *The Nonlinear Diffusion Equation*, Reidel.

Calberg, R.G., Couchman, H.M.P. & Thomas, P.A. (1990) *Ap. J.*, **352**, L29.

Carswell, R.F. & Rees, M.J. (1987) *M.N.R.A.S.* **224**, 13.

Carswell, R.F., 1989 in *The epoch of galaxy formation* (eds. C.S. Frenk, R.S. Ellis, T. Shanks, A.F. Heavens & J.A. Peacock) Kluwer, p.89.

Carswell, R.F., Lanzetta, K.M., Parnell, H.C. & Webb, J.K. (1991) *Ap. J.* **371**, 36.

Centrella, J. & Melott, A. (1982) *Nature* **305**, 196.

Centrella, J.M., Gallagher, J.S., Mellott, A.L. & Bushouse, H.A., (1988) *Ap. J.* **333**, 24.

Chambers, K.C. & Charlot, S. (1989) *Ap. J. Lett.* **348**, L1.

Chambers, K.C. & Miley, G.K. (1990) in *Evolution of the universe of galaxies* (ed. R.G. Kron) A S P Conf. Series **10**, p.373.

Chambers, K.C., Miley, G.K. & Joyce, R.R. (1988) *Ap. J.*, **329**, L75.

Chambers, K.C., Miley, G.K. & Van Breugel, W.J.M. (1987) *Nature* **329**, 604.

Chambers, K.C., Miley, G.K. & Van Breugel, W.J.M. (1990) *Ap. J.* **363**, 21.

Colless, M., Ellis, R.S., Taylor, K. & Hook, R.N. (1990) *M.N.R.A.S.* **244**, 408.

Collins, A., Joseph, R.D. & Robertson, N.A. (1986) *Nature* **320**, 506.

Cowie, L.L. (1988) in *The post recombination universe* (eds. N. Kaizer & A. Lansenby) Kluwer, p.1.

Cowsik, R. & McClelland, J. (1972) *Phys. Rev. Lett.* **29**, 669.

Daly, R. (1990) *Ap. J.* **355**, 416.

Davis, M., Efstathiou, G., Frenk, C.S. & White, S.D.M., (1985) *Ap. J.* **292**, 371.

Davis, M. & Peebles, P.J.E. (1983) *Ap. J.* **267**, 465.

Davis, M. & Peebles, P.J.E. (1983) *Ap. J.* **267**, 437.

De Lapparent, V., Geller, M.J. & Huchra, J.P. (1986) *Ap. J.* **302**, L11.

De Young, D. (1989) *Ap. J. Lett.* **342**, L59.

Di Serego Alighieri, S., Fosbury, R.A.E., Quinn, P.J. & Tadhunter, C.N. (1989) *Nature* **341**, 307.

Dekel, A. & Rees, M.J. (1987) *Nature* **326**, 455.

Dekel, A. & Silk, J. (1986) *Ap. J.* **303**, 39.

Dolgov, A. & Linde, A.D. (1982) *Phys. Letts.* **B116**, 329.

Doroshkevich, A.G. (1970) *Astrofisika* **6**, 581.

Doroshkevich, A.G. (1973) *Astrophys. Lett.* **14**, 11.

Doroshkevich, A.G., Kotok, E.V., Novikov, I.D., Polyudov, A.N., Shandarin, S.F. & Sigov, Yu. S. (1980) *M.N.R.A.S.* **192**, 321.

Dunlop, J.S. & Peacock, J.A. (1990) *M.N.R.A.S.* **247**, 19.

Efstathiou, G. & Rees, M.J. (1988) *M.N.R.A.S.* **230**, 5p.

Efstathiou, G. & Silk, J. (1983) *Fundam. Cosmic Phys.* **9**, 1.

Efstathiou, G., Davis, M., Frenk, C.S. & White, S.D.M. (1985) *Ap. J. Suppl.* **57**, 241.

Efstathiou, G. *et al.* (1990) *M.N.R.A.S.* **247**, 10p.

Eisenhardt, P. & Chokshi, A. (1990) *Ap. J. Lett.* **351**, L9.

Eisenhardt, P. et al. (1990) in *Evolution of the universe of galaxies* (ed. R.G. Kron) A S P Conf. Series **10**, p.371.

Ellis, J. *et al.* (1985) *Phys. Letts.***B152**, 175.

Ellis, R.S. (1990) in *Evolution of the universe of galaxies* (ed. R.G. Kron) A S P Conf. Series **10**, p.248.

Evrard, A. (1986) *Ap. J.* **310**, 1.

Faber, S.M. (1982) in *Astrophysical cosmology* (eds. H.A. Bruck, G.V. Coyne & M.S. Longair) Pontificia Acad. Sci. Vatican, p.219.

Fabri, R. & Pollock, M. (1983) Phys. Lett. B. **125**, 445.

Fabian, A.C. (1989) *M.N.R.A.S.* **238**, 41p.

Fall, S.M. (1979) *Nature* **281**, 200.

Fall, S.M. & Efstathiou, G. (1980) *M.N.R.A.S.* **193**, 189.

Fall, S.M. & Rees, M.J. (1985) *Ap. J.* **298**, 18.

Fall, S.M., Pei, Y.C. & McMahon, R.G. (1989) *Ap. J. Lett.* **341**, L5.

Foltz, C.B., Weymann, R.J., Roser, H.J. & Chaffe, F.H. (1984) *Ap. J.* **281**, L1.

Frenk, C.S., 1989 in *The epoch of galaxy formation* (eds. C.S. Frenk, R.S. Ellis, T. Shanks, A.F. Heavens & J.A. Peacock) Kluwer, p.257.

Frenk, C.S., White, S.D.M. & Davis, M. (1983) *Ap. J.*, **271**, 417.

Frenk, C.S., White, S.D.M., Efstathiou, G. & Davis, M. (1985) *Nature* **317**, 595.

Frenk, C.S., White, S.D.M., Efstathiou, G. & Davis, M. (1988) *Ap. J.* **327**, 507.

Frenk, C.S., White, S.D.M., Efstathiou, G. & Davis, M. (1990) *Ap. J.* **351**, 10.

Friedman, A. (1922), *Z. Phys.* **10**, 377.

Friedman, A. (1924), *Z. Phys.* **21**, 326.

Frieman, J.A. & Turner, M.S. (1984) *Phys. Rev.* **D30**, 265.

Giovanelli, R. & Haynes, M.P., (1982) *Ap. J.* **87**, 1355.

Gopal-Krishna and Wiita, P.J. (1991) *Ap. J.* **373**, 325.

Gott, J.R. (1977) *A. Rev. Astr. Ap.* **15**, 235.

Green, R.F. (1989) in *The epoch of galaxy formation* (eds. C.S. Frenk, R.S. Ellis, T. Shanks, A.F. Heavens & J.A. Peacock) Kluwer, p.121.

Gunn, J.E. & Peterson, B.A. (1965) *Ap. J.* **142**, 1633.

Gunn, J.E. (1982) in *Astrophysical cosmology* (eds. H.A. Bruck, G.V. Coyne & M.S. Longair) Pontificia Acad. Sci. Vatican, p.233.

Gurbatov, S.N., Saichev, A.I. & Shandarin, S.F. (1983) *Sov. Phys. Usp.* **26**, 857.

Gurbatov, S.N., Saichev, A.I. & Shandarin, S.F. (1985) *Sov. Phys. Doke.* **30**, 921.

Gurbatov, S.N., Saichev, A.I. & Shandarin, S.F. (1989) *M.N.R.A.S.* **236**, 385.

Guth, A.H. (1981) *Phys. Rev.* **D23**, 347.

Guth, A. & Pi, S.Y. (1982) *Phys. Rev. Lett.* **49**, 1110.

Guth, A.H. & Weinberg, E. (1983) *Nucl. Phys.* **B212**, 321.

Hammer, F., Le Fevre, O. & Proust, D. (1991) *Ap. J.* **374**, 91.

Harrison, E.R. (1970) *Phys. Rev.* **D1**, 2726.

Hart, L. & Davies, R.D. (1982) *Nature* **297**, 191.

Hawking, S.W. (1982) *Phys. Letts.* **B115**, 293.

Heavens, A.F. & Peacock, J.A. (1988) *M.N.R.A.S.* **232**, 339.

Hockney, R.W. & Eastwood, J.W. (1981) *Computer simulation using particles* McGraw Hill.

Hogan, C.J. (1987) *Ap. J. Lett.* **316**, L59.

Holman, R. et al. (1984) *Phys. Letts.* **B137**, 343.

Hoyle, F. (1949) in *Problems of cosmical aerodynamics* Central Air Documents office, Dayton Ohio, p.195.

Hoyle, F. (1953) *Ap. J.* **118**, 513.

Huchra, J., Davis, M., Latham, D. & Tonry, J., (1983) *Ap. J. Suppl.* **52**, 39.

Hunstead, R.W., Murdoch, H.S., Pettini, M. & Blades, J.C. (1988) *Ap. J.* **329**, 527.

Ikeuchi, S. & Ostriker, J.P. (1986) *Ap. J.* **301**, 552.

Ikeuchi, S. (1986) *Ap. Sp. Sci.* **118**, 509.

Ikeuchi, S., 1990 in *Dark matter in the universe* (eds. H. Sato & H. Kodama) Springer-Verlag, p.50.

Ikeuchi, S., Murakami, I. & Rees, M.J. (1988) *M.N.R.A.S.*, **236**, 21.

Irwin, M., Mcmahon, R.G. & Hazard, C. (1991) in *The space distribution of quasars* (ed. D. Crampton) A S P conf. series **21**, p.117.

James, P.A., Joseph, R.D. & Collins, C.A., (1987) *M.N.R.A.S.* **229**, 53.

Jensen, L. & Olive, K, (1986) *Nucl. Phys.* **B263**, 731.

Jones, B. & Wyse, R. (1985) *Astr. Ap.* **149**, 144.

Kaiser, N. (1985) *Ap. J.* **284**, L9.

Kaiser, N. & Lahav, O. (1989) *M.N.R.A.S.* **231**, 635.

Kapahi, V.K. (1989) A.J. **97**, 1.

Kashlinsky, A. & Jones, B.J.T. (1991) *Nature* **349**, 753.

Kazanas, D. (1980) *Ap. J.* **241**, L59.

Katz, N. & Gunn, J. E. (1991) *Ap. J.* **377**, 365.

Klypin, A.A. et al. (1987) *Sov. Astr. Lett.* **298**, L1.

Kofman, L.A., Pogosian, D. & Shandarin, S.F. (1990) *M.N.R.A.S.* **242**, 200.

Kolb, E.W., Turner, M.S. (1990) *The Early Universe* Addison-Wesley.

Koo, D. (1986) in *Spectral evolution of galaxies* (eds. C. Chiosi & A. Renzini) Reidel, p.419.

Koo, D. (1990) in *Evolution of the universe of galaxies* (ed. R.G. Kron) A S P Conf. Series **10**, p.268

Koo, D., Kron, R.G. & Szalay, A.S. (1986) *13th Texas Symp. Relativistic Astrophysics* (ed. P. Ulmer.) World Scientific, p.284.

Kormendy, J. (1989) *Ap. J. Lett.* **342**, L63.

Kotok, E.V. & Shandarin, S.F. (1989) *Soviet astr.*, **32**, 351.

Kreysa, E. & Chini, A. (1989) *Proc. 3rd ESO/CERN symp. astronomy, cosmology and fundamental particles* (eds. Caffo, M. *et al.*) Reidel.

La, D. & Steinhardt, P.J. (1989) *Phys. Rev. Letts.* **62**, 376.

Lanzetta, K.M., Wolfe, A.M. & Turnshek, D.A. (1989) *Ap. J.* **344**, 277.

Larson, R.B. (1975) *M.N.R.A.S.* **173**, 671.

Le Fevre, O. & Hammer, F., 1988c *Ap. J.* **333**, L37.

Le Fevre, O., Hammer, F., Nottale, L., Mazure, A., and Christian, C., 1988a *Ap. J.* **324**, L1.

Lilly, S.J. & Longair, M.S. (1984) *M.N.R.A.S.* **211**, 833.

Lilly, S.J. Cowie, L.L. & Gardner, J.P. (1991) *Ap. J.* **369**, 79.

Lilly, S.J. (1989) in *The epoch of galaxy formation* (eds. C.S. Frenk, R.S. Ellis, T. Shanks, A.F. Heavens & J.A. Peacock) Kluwer, p.63.

Lilly, S.J. (1989b) *Ap. J.* **340**, 77.

Lilly, S.J. (1990) in *Evolution of the universe of galaxies* (ed. R.G. Kron) A S P Conf. Series **10**, p.344.

Linde, A.D. (1982a) *Phys. Letts.* **B108**, 389.

Linde, A.D. (1982b) *Phys. Letts.* **B116**, 335.

Linde, A.D. (1983) *Phys. Letts,* **B129**, 177.

Lynden-Bell, D. (1967) *M.N.R.A.S.* **136**, 101.

Lynden-Bell, D. *et al.* (1988) *Ap. J.* **326**, 19.

Lyth, D.H. & Steward E.D. (1990) *Ap. J.* **361**, 343.

Maddox, S.J., Efstathiou, G., Sutherland, W.J. & Loveday, J., (1990) *M.N.R.A.S.* **242**, 43p.

Marx, G. & Szalay, A.S. (1972) *Proc. Neutrino* 72 **1**, 123.

McCarthy, P.J. (1989) Ph.D. thesis, Univ. of California, Berkeley.

McCarthy, P.J. *et al.* 1987a in *Cooling flows in clusters and galaxies* (ed. A.C. Fabian) Kluwer, p.325.

McCarthy, P.J., van Breugel, W., Spinrad, H. & Djorgovski, S., 1987b *Ap. J.* **321**, L29.

McCarthy, P.J., Kapahi, V.K., van Breugel, W. & Subrahmanya, C.R., (1990) A. J. **100**, 1014.

Meier, D.L. (1976) *Ap. J.* **207**, 343.

Mellot, A.L. & Shandarin, S.F. (1989) *Ap. J.* **343**, 26.

Mestel, L. (1963) *M.N.R.A.S.* **126**, 553.

Meszaros, P. (1975) *Astr. Ap.* **38**, 5.

Narlikar, J.V. (1983) *Introduction to Cosmology*, Jones and Bartlett.

Negropante, J. & White, S.D.M. (1983) *M.N.R.A.S.* **205**, 1009.

Olive, K., Sekel, D. & Vishniac, E. (1985) *Ap. J.* **292**, 1.

Oort, M.J.A., Katgert, P., Steeman, F.W.H. & Windhorst, R.A., (1987) *Astr. Ap.* **179**, 41.

Osterbrok, D.E. (1974) *Astrophysics of gaseous nebulae*, Freeman.

Ostriker, J.A. & Suto, Y. (1990) *Ap. J.* **348**, 378.

Ostriker, J.P. & Cowie, L.L. (1981) *Ap. J. Lett.* **243**, L127.

Ostriker, J.P. & Ikeuchi, S. (1983) *Ap. J. Lett.* **268**, L63.

Ostriker, J.P., 1988 in *QSO absorption lines* (eds. J.C. Blades, D. Turnshek & C.A. Norman) Cambridge Univ. Press, p.319.

Padmanabhan, T. (1988) *Phys. Rev. Lett.* **60**, 2229.

Padmanabhan, T., Seshadri, T.R. & Singh, T.P., (1989) *Phys. Rev.* **D39**, 2100.

Padmanabhan, T. (1990) unpublished.

Padmanabhan, T. & Vasanthi, M.M. (1987) *Ap. J.* **315**, 411.

Padmanabhan, T. & Narasimha, D. (1992), Tifr - Tap preprint - 3.

Palimaka, J.J., Bridle, A.H., Fomalont, E.B. & Brandie, G.W. (1979) *Ap. J.* **231**, L7.

Partridge, R.B. & Peebles, P.J.E. (1967) *Ap. J.* **147**, 868.

Peacock, J.A. & Heavens, A.F. (1990) *M.N.R.A.S.* **243**, 133.

Peacock, J.A. (1990) *Observational constraints on Galaxy Formation*,Edinburgh Astronomy Preprint No. 30/90.

Peacock, J.A. (1991) *Nature* **349**, 190.

Peebles, P.J.E. (1969) *Ap. J.* **155**, 393.

Peebles, P.J.E. (1980) *Large Scale Structure of the Universe*, Princeton Univ. Press.

Peebles, P.J.E. (1982) *Ap. J.* **258**, 415.

Peebles, P.J.E. (1986) *Nature* **321**, 27.

Peebles, P.J.E., 1989 in *The epoch of galaxy formation* (eds. C.S. Frenk, R.S. Ellis, T. Shanks, A.F. Heavens & J.A. Peacock) Kluwer, p.1.

Peebles, P.J.E., 1991 in *Observational tests of inflation*, Institute of Advanced Study, Princeton, Preprint.

Peebles, P.J.E. & Silk, J. (1990) *Nature* **346**, 233.

Pettini, M., Boksenberg, A. & Hunstead, R.W. (1989) *Ap. J.* **348**, 48.

Pettini, M., Hunstead, R.W., Smith, L.J. & Mar, D.P., (1990) *M.N.R.A.S.* **246**, 545.

Phinney, E.S. (1983) Ph.D. thesis, Univ. of Cambridge.

Press, W.H. & Schechter, P. (1974) *Ap. J.* **187**, 425.

Raymond, J.C., Cox, D.P. & Smith, B.W. (1976) *Ap. J.* **204**, 290.

Readhead, A.C.S. *et al.* (1989) *Ap. J.* **346**, 566.

Rees, M.J. & Ostriker, J.P. (1977) *M.N.R.A.S.* **179**, 541.

Rees, M.J. (1985) *M.N.R.A.S.* **213**, 75p.

Rees, M.J. (1986) *M.N.R.A.S.* **218**, 25.

Rees, M.J. (1988) in *QSO absorption lines* (eds. J.C. Blades, D. Turnshek & C.A. Norman) Cambridge Univ. Press, p.107.

Rees, M.J. (1989) *M.N.R.A.S.* **239**, 1p.

Rocca-Volmerange, B. & Guiderdoni, B. (1990) *M.N.R.A.S.*, **247**, 166.

Rowan-Robinson, M., et al. (1990) *M.N.R.A.S.* **247**, 1.

Rubin, V.C., Thonnard, N., Ford, W.K. & Roberts, M.S., (1976) *Ap. J.* **81**, 719.

Rubakov, V.A., Sazhin, M. & Veryaskin, A., (1982) *Phys. Lett.* **B115**, 189.

Sachs, R.K., Wolfe, A.M. (1967) *Ap. J.* **147**, 73.

Sahni, V. (1990) in *Texas/ESO - CERN symposium on relativistic astrophysics, cosmology and fundamental physics*, Brighton, U.K.

Sahni, V. (1991) in preparation.

Salpeter, E.E. (1964) *Ap. J.* **101**, 5.

Sargent, W.L.W. (1988) in *QSO absorption lines* (eds. J.C. Blades, D. Turnshek & C.A. Norman) Cambridge Univ. Press, p.1.

Sargent, W.L.W., Boksenberg, A. & Steidel, C.C. (1988) *Ap. J. Suppl.* **68**, 539.

Sargent, W.L.W., Young, P.J., Boksenberg, A. & Tytler, D. (1980) *Ap. J. Suppl.* **42**, 41.

Sato, K. (1981) *M.N.R.A.S.* **195**, 467.

Saunders, W. *et al.* (1991) *Nature* **349**, 32.

Schiano, A.V.R., Wolfe, A.M. & Chang, C.A. (1990) *Ap. J.*, **365**, 439.

Schneider, D.P., Schmidt, M. & Gunn, J.E., 1989a, *Astr. J.* **98**, 1507.

Schneider, D.P., Schmidt, M. & Gunn, J.E., 1989b, *Astr. J.* **98**, 1951.

Schneider, D.P., Schmidt, M. & Gunn, J.E. (1991) *Astr. J.* **102**, 837.

Schweizer, F. (1982) *Ap. J.*, 252, 455.

Schweizer, F. (1986) in *Nearly normal galaxies* (ed. S.M. Faber) Spinger Verlag, p.18.

Shandarin, S.F. & Zeldovich, Ya. B. (1990) *Rev. Mod. Phys.* **61** 185.

Shapiro, P.R. & Giroux, M.L. (1987) *Ap. J. Lett.* **321**, L107.

Shapiro, P.R. & Giroux, M.L., 1989 in *The epoch of galaxy formation* (eds. C.S. Frenk, R.S. Ellis, T. Shanks, A.F. Heavens & J.A. Peacock) Kluwer, p.153.

Silk, J. (1968) *Ap. J.* **151**, 459.

Silk, J. (1977) *Ap. J.* **211**, 638.

Silk, J. (1983) *Nature* **301**, 574.

Silk, J. (1985) *Ap. J.* **297**, 1.

Silk, J. & Norman, C. (1981) *Ap. J.* **247**, 59.

Singal, A.K. (1988) *M.N.R.A.S.* **233**, 87.

Smoot, G. F. *et al.* (1991) *Ap. J.* Letts. **371**, L1.

Smoot, G. F. *et al.* (1992) Preprint.

Starobinsky, A.A. (1982) Phys. Letts.B. **117**, 293.

Steidel, C.C. & Sargent, W.L.W. (1987) *Ap. J. Lett.* **318**, L11.

Steinhardt, P.J. & Accetta, F.S. (1990) *Phys. Rev. Letts.* **64**, 2740.

Strauss, M.A. & Davis, M., 1988. *Proc. Vatican Study week on Large-scale Motions in the Universe* (ed. G. Coyne & V. Rubin) Pontifical Sci. Acad. Vatican.

Strukov, I.A., Skulachev, D.P. & Klypin, A.A., (1987) *IAU Symp.***130**, 27.

Subramanian, K. (1988) *M.N.R.A.S.* **234**, 459.

Subramanian, K. (1989) *Invited talk at Symposium on 'Large Scale Structure and Evolution of the Universe'*, Ootacamund, April 20-21.

Subramanian, K. & Chitre, S.M. (1984) *Ap. J.* **276**, 440.

Subramanian, K. & Swarup, G. (1990) *M.N.R.A.S.* **247**, 237.

Subramanian, K. & Swarup, G. (1992) NCRA -TIFR preptint.

Sutherland, W. (1988) *M.N.R.A.S.* **234**, 159.

Swarup, G. (1984) *Gaint Meterwave Radio Telescope - A proposal*, Radio Astronomy Centre, TIFR, Ootacamund.

Swarup, G. (1988) *Proc. Indian Natl. Sci. Acad.* **54**, 853.

Swarup, G. (1990) *Indian J. Rad. Sp. Phys.* **19**, 493.

Teague, P.F., Carter, D. & Gray, P.M. (1990) *Ap. J. Suppl.* **72**, 715.

Toomre, A. (1977) in *Evolution of galaxies and stellar populations* (eds. B.M. Tinsley & R.B. Larson) Yale Univ. Obs., 401.

Trimble, V. (1988) *Contemp. phys.* **29**. 373.

Turner, M.S., Steigman , G. & Krauss, L., (1984) *Phys. Rev. Lett.* **52**, 2090.

Turner, E.L. (1991) *Astr. J.* **101**, 5.

Turnshek, D.A. *et al.* (1989) *Ap. J.* **344**, 567.

Tyson, J.A. (1988) A.J. **96**, 1.

Tyson, N.D. (1988) *Ap. J. Lett.* **329**, L57.

Unsold, A. (1977) *The new cosmos* Springer-Verlag.

Uson, J.M., Bagri, D.S. & Cornwell, T.J. (1992) *Phys. Rev. Lett.* **67**, 3328.

Van Breugel, W.J.M. & McCarthy, P.J. (1990) in *Evolution of the universe of galaxies* (ed. R.G. Kron) A S P Conf. Series **10**, p.359.

Van Breugel, W.J.M., Filippenko, A.V., Heckman, T.M. & Miley, G.K., (1985) *Ap. J.* **293**, 83.

Van den Bergh, S. (1990) in *Dynamics and interactions of galaxies* (ed. R. Weilen) Springer-Verlag, p.492.

Vilenkin, A. & Ford, L. (1982) *Phys. Rev.* **D26**, 1231.

Weedman, D.W., Weymann, R.J., Green, R.F. & Heckman, T.M. (1982) *Ap. J. Lett.* **255**, L5.

Weinberg, D.H. & Gunn, J.E. (1990a) *Ap. J.*, **352**, L25.

Weinberg, D.H. & Gunn. J.E., (1990b) *M.N.R.A.S.* **247**, 260.

Weinberg, S. (1972) *Gravitation and Cosmology*, Wiley.

West, M.J. & Richstone, D.O. (1988) *Ap. J.* **335**, 532.

Weymann, R.J., Carswell, R.F. & Smith, M.J. (1981) *Ann. Rev. Astr. Astrop.* **19**, 41.

White, S.D.M. & Rees, M.J. (1978) *M.N.R.A.S.* **183**, 341.

White, S.D.M. (1979) *M.N.R.A.S.* **189**, 831.

White, S.D.M. (1982) in *Morphology and dynamics of galaxies* (eds. L. Martinet & M. Mayer) Geneva Observatory, p.291.

White, S.D.M. (1984) *Ap. J.* **286**, 38.

White, S.D.M. (1986) in *Inner space/outer space* (eds. E.W. Kolb, M.S. Turner, D. Lindley, K. Olive & D. Seckell) Chicago Univ. Press, p.228.

White, S.D.M., Frenk, C.S. & Davis, M., 1983a *Ap. J.* **274**, L1.

White, S.D.M., Frenk, C.S., Davis, M. & Efstathiou, G., 1987a *Nature* **330**, 451.

White, S.D.M., Frenk, C.S., Davis, M. & Efstathiou, G., 1987b *Ap. J.* **313**, 505.

White, S.D.M. & Frenk, C.S. (1991) *Ap. J.* **379**, 52.

Williams, B.G., Heavens, A.F., Peacock, J.A. & Shandarin, S.F., (1991) *M.N.R.A.S.* **250**, 458.

Wolfe, A.M., 1988 in *QSO absorption lines* (eds. J.C. Blades, D. Turnshek & C.A. Norman) Cambridge Univ. Press, p.297.

Wolfe, A.M., 1989 in *The epoch of galaxy formation* (eds. C.S. Frenk, R.S. Ellis, T. Shanks, A.F. Heavens & J.A. Peacock) Kluwer, p.101.

Wolfe, A.M., Turnshek, D.A., Smith, H.E. & Cohen, R.D. (1986) *Ap. J. Suppl.* **61**, 249.

Wright, E.L. *et al.* (1992) Preprint.

Yagnik, U.A. (1990) Phys. Lett. B. **234**, 271.

Zeldovich, Ya. B. (1970a) *Astrofizika* **6**, 319.

Zeldovich, Ya. B. (1970b) *Astr. Ap.* **5**, 84.

Zeldovich, Ya. B. (1972) *M.N.R.A.S.* **160**, 1p.

Zeldovich, Ya. B., Novikov, I.D. (1983) *Relativistic Astrophysics*, Vol. 2.

Appendix 2.1. Newtonian perturbation theory

In the linear regime, each mode evolves independent of other modes. At any given time t, there will be modes such that, the proper wave length, $\lambda(t)$, is much smaller than the Hubble radius of the universe, $d_H(t)$, at that time. It should be possible to study such modes by Newtonian theory of gravitation; this may be done in the following manner. [The derivation in this appendix and the next are based on Padmanabhan 1990; Lyth & Stewart 1990].

There is a systematic procedure (involving expansion in powers of c^{-1}) which will allow one to determine the Newtonian limit of a given metric. For our purpose, this can be most easily done by transforming the FRW metric

$$ds^2 = c^2dt^2 - a^2(t)\left[dR^2 + R^2(d\theta^2 + \sin^2\theta d\phi^2)\right] \qquad (A1.1)$$

to a coordinate system which is locally inertial at the origin. It can be easily shown that, the metric in such a coordinate system is given by

$$ds^2 \cong c^2\left(1 + \frac{2\phi_b(t,\mathbf{x})}{c^2}\right)dt^2 - (dx^2 + dy^2 + dz^2) \qquad (A1.2)$$

where

$$\phi_b(t,\mathbf{x}) = -\frac{1}{2}\left(\frac{\ddot{a}}{a}\right)|\mathbf{x}|^2; \quad |\mathbf{x}| = a(t)|\mathbf{R}| \qquad (A1.3)$$

To the lowest non-trivial order, for $|\mathbf{x}| \ll d_H$, we may treat ϕ_b as an equivalent Newtonian potential due to the uniform, homogeneous background. The Newtonian limit of the matter flow equations ($T^{ab};b = 0$) will, in general, be

$$\dot{\rho} \equiv \frac{d\rho}{dt} \equiv \frac{\partial\rho}{\partial t} + (v^i\partial_i)\rho = -\rho(\nabla.\mathbf{v}) \qquad (A1.4)$$

$$\dot{v}^i = -\partial^i\phi - \rho^{-1}\partial^i P \qquad (A1.5)$$

which can be satisfied, for the potential in (A1.2), by the following ansatz: $P_b = 0$, $\rho_b(\mathbf{x},t) = \rho_b(t)$, $\mathbf{v}_b(t,\mathbf{x}) = f(t)\mathbf{x}$. Then we get,

$$\frac{\partial\rho_b}{\partial t} + 3\rho_b f(t) = 0 \qquad (A1.6)$$

$$\dot{f} + f^2(t) = \left(\frac{\ddot{a}}{a}\right) \qquad (A1.7)$$

The second equation integrates to give $f(t) = (\dot{a}/a) = H_b(t)$; substituting in the first equation we discover that $\rho_b \propto a^{-3}$. This set determines the Newtonian limit of FRW universe.

We now perturb this solution. As long as the perturbations are linear and have a scale length much smaller than d_H (so that the entire perturbed region can be covered by the region in which (A1.2) is valid), we can simply add the perturbed potential $\delta\phi$ (due to the perturbed density $\delta\rho$) to the background potential $\phi_b(t,\mathbf{x})$. Let $\phi, \mathbf{v}, P$ and ρ denote variables containing perturbed parts as well. By writing down the linearised versions of (A1.4) and (A1.5) one can easily obtain the perturbation equation for the variables like $\delta\rho$. We will, however, proceed in a (more complicated) manner which has the advantage that it can be easily adapted for a fully relativistic situation.

We begin by decomposing the gradient of the velocity field ($\partial_i v_j$) into an anti-symmetric part, symmetric traceless part and the trace by writing

$$\partial_j v_i = \omega_{ij} + \sigma_{ij} + H\delta_{ij} \qquad (A1.8)$$

where $2\omega_{ij} = (\partial_j v_i - \partial_i v_j)$; $2\sigma_{ij} = \partial_j v_i + \partial_i v_j - 2H\delta_{ij}$ and $H(t,\mathbf{x}) = [\partial_i v^i(\mathbf{x},t)/3]$ is the trace. [In the absence of perturbations $H(t,\mathbf{x}) = H_b(t)$ reduces to the Hubble constant of the background universe; in this sense we may consider $\nabla.\mathbf{v}$ to be proportional to the 'perturbed Hubble constant' $H(t,\mathbf{x}) = H_b(t) + \delta H(t,\mathbf{x})$]. It is trivial to verify that

$$(\partial_j v_i)(\partial^i v^j) = 3H^2 + 2(\sigma^2 - \omega^2) \tag{A1.9}$$

where $2\sigma^2 = \sigma_{ij}\sigma^{ij}$ and $2\omega^2 = \omega_{ij}\omega^{ij}$. Taking the divergence of the Euler equation

$$\frac{\partial v^i}{\partial t} + (v^j \partial_j)v^i + \partial^i \phi = -\rho^{-1}\partial^i P \tag{A1.10}$$

and using $3H(t,\mathbf{x}) = \partial_i v^i$ and (A1.9) we get

$$\frac{\partial}{\partial t}(\nabla.\mathbf{v}) + (v^j\partial_j)(\nabla.\mathbf{v}) + (\partial_i v^j)(\partial_j v^i) + \nabla^2\phi = -\partial_i(\rho^{-1}\partial^i P) \tag{A1.11}$$

or

$$3\dot{H} + 3H^2 + 2(\sigma^2 - \omega^2) + \nabla^2\phi = -\partial_i\left(\frac{\partial^i P}{\rho}\right) \tag{A1.12}$$

This equation is exact in the Newtonian limit; we now linearise it retaining only first order corrections to the background variables. We can ignore σ^2 and ω^2 since they are of quadratic order and replace ρ by ρ_b in the right hand side because P is essentially δP (since $P_b = 0$). We thus get

$$\begin{aligned} \delta\dot{H} &= -2H_b\delta H - \frac{1}{3}\nabla^2\delta\phi - \frac{1}{3}\frac{\nabla^2\delta P}{\rho_b} \\ &= -2H_b\delta H - \frac{4\pi G}{3}\delta\rho - \frac{1}{3}c_s^2\frac{\nabla^2\delta\rho}{\rho_b} \end{aligned} \tag{A1.13}$$

where we have set $\nabla^2\delta\phi = 4\pi G\delta\rho$ and $\delta P = c_s^2\delta\rho$ with c_s the sound velocity . The continuity equation

$$\dot{\rho} = -3H(t,\mathbf{x})\rho \tag{A1.14}$$

can be similarly linearised to give

$$\delta\dot{\rho} = -3H_b\delta\rho - 3\rho_b\delta H \tag{A1.15}$$

or, using the definition $\delta\rho \equiv \rho_b\delta$,

$$\delta H = -H_b\left(\frac{\delta\rho}{\rho_b}\right) - \frac{1}{3}\frac{\delta\dot{\rho}}{\rho_b} \equiv -H_b\delta - \frac{1}{3}\left(\frac{\dot{\rho}_b}{\rho_b}\delta + \dot{\delta}\right) = -\frac{1}{3}\dot{\delta} \tag{A1.16}$$

Suubstituting this into (A1.13), we obtain

$$\ddot{\delta} + 2H_b\dot{\delta} - 4\pi G\rho_b\delta - c_s^2\nabla^2\delta = 0 \tag{A1.17}$$

This looks very much like a perturbation equation; but notice that the 'overdots' in this equation stand for the operation $[(\partial/\partial t) + v^i\partial_i]$ and not just $(\partial/\partial t)$. However, when operating on δ, we only need to retain the zeroth order part of v^i which is just $Hx^i = (\dot{a}/a)x^i$. That is

$$\dot{\delta} \equiv \left(\frac{\partial \delta}{\partial t}\right)_x + v^i \partial_i \delta \simeq \left(\frac{\partial \delta}{\partial t}\right)_x + \left(\frac{\dot{a}}{a}\right) x^i \partial_i \delta \qquad (A1.18)$$

We will now reintroduce the FRW-coordinates (X, Y, Z) related to (x, y, z) by $x = a(t)X$ etc. Clearly, for any function $f(t, x)$

$$\begin{aligned} df &= \left(\frac{\partial f}{\partial t}\right)_x dt + \left(\frac{\partial f}{\partial x}\right)_t dx = \left(\frac{\partial f}{\partial t}\right)_x dt + \left(\frac{\partial f}{\partial x}\right)_t [\dot{a}X dt + a dX] \\ &= \left[\left(\frac{\partial f}{\partial t}\right)_x + \left(\frac{\partial f}{\partial x}\right)_t Hx\right] dt + \left(\frac{\partial f}{\partial x}\right)_t a dX \end{aligned} \qquad (A1.19)$$

showing

$$\left(\frac{\partial}{\partial t}\right)_x + Hx^i \left(\frac{\partial}{\partial x^i}\right)_t = \left(\frac{\partial}{\partial t}\right)_X \qquad (A1.20)$$

Thus, in the (t, X) coordinate system, the overdot merely means partial derivative with respect to t; however, $\nabla_x^2 = a^{-2}\nabla_X^2$. Thus we get the final equation

$$\ddot{\delta} + 2H_b\dot{\delta} - c_s^2 a^{-2}\nabla^2\delta = 4\pi G\rho_b\delta \qquad (A1.21)$$

This equation, set in the original FRW coordinates, describes the growth of perturbations in the Newtonian limit.

Appendix 2.2. General relativistic perturbation theory

The central equations of Newtonian perturbation theory disucssed in the previous Appendix 1 were for the perturbed density and perturbed Hubble constant:

$$\begin{aligned} \delta\dot{\rho} &= -3H_b\delta\rho - 3\rho_b\delta H \\ \delta\dot{H} &= -2H_b\delta H - \frac{4\pi G}{3}\delta\rho - \frac{1}{3}\frac{\nabla^2\delta P}{\rho_b} \end{aligned} \qquad (A2.1)$$

The correct, general relativistic, perturbation equations can be made to look identical to the above set except for the change of ρ_b to $(\rho_b + P_b)$. [Note that, in the Newtonian limit $P_b \ll \rho_b$.] Thus we get

$$\delta\dot{\rho} = -3H_b\delta\rho - 3(\rho_b + P_b)\delta H \qquad (A2.2)$$

$$\delta\dot{H} = -2H_b\delta H - \frac{4\pi G}{3}\delta\rho - \frac{1}{3}\frac{\nabla^2\delta P}{(\rho_b + P_b)} \qquad (A2.3)$$

For a single component model, we can set $\delta P = v^2\delta\rho$, eliminate δH and obtain a second order equation for $\delta\rho$. It is usual to use a Fourier transform with proper wavelength so that ∇^2 is equivalent to $a^{-2}k^2$. This will lead to the equation discussed in the main text. We will now derive (A2.2) and (A2.3) .

Let $a, b, \cdots$ denote spacetime indices and $i, j, k \cdots$ denote the space indices. At each event in spaccetime we choose an orthonormal basis such that the momentum density vanishes. In this frame, the four-velocity u^a has the components $u^o = u_0 = 1$ and $u^j = 0$. If D_a is the covariant derivative operator, then the 'overdot' will denote the directional derivative along u^a, i.e. the operator $u^a D_a$. Surfaces orthogonal to the comoving world lines (for which u^a are the tangent vectors) will be called comoving hypersurfaces; the projection tensor on to these surfaces will be $h_{ab} = g_{ab} - u_a u_b$ (in the co-moving basis, only nonzero components of this tensor will be $h_{ij} = g_{ij}$.) Using D_a and h_{ab} we can construct the natural derivative $h^b_a D_b$ on the comoving surfaces and the Laplacian $\nabla^2 = h^b_a D_b h^{ac} D_c$. In the absence of perturbations $\nabla^2 = a^{-2}\, \partial^i \partial_i$ because the comoving surfaces are flat (at any given t); in the presence of metric fluctuations, these surfaces show deviation from flatness but this deviation will only produce a second order effect when ∇^2 is applied to the perturbed quantity. Thus as far as perturbation equations are concerned, one can continue to use $\nabla^2 = a^{-2}\partial_i \partial^i$ even when the hypersurfaces are not flat.

With these preliminaries, we can start the derivation of the fluctuation equations. The relativistic analogues of continuity and Euler equations are

$$\dot\rho = -3H(\rho + P) \tag{A2.4}$$

$$\dot u_a = -\frac{h^b_a D_b P}{(\rho + P)} \tag{A2.5}$$

where all quantities refer to the background metric. We now proceed exactly in the same manner as before by starting with the relation

$$\begin{aligned} D_a\, \dot u^a &= D_a u^b D_b u^a = (D_a u^b)(D_b u^a) + u^b D_a D_b u^a \\ &= (D_a u^b)(D_b u^a) + u^b D_b D_a u^a + u^b [D_a, D_b] u^a \\ &= (D_a u^b)(D_b u^a) + u^b D_b (D_a u^a) + u^b R_{ab} u^a \end{aligned} \tag{A2.6}$$

Now, since u^a has constant norms $u^a D_b u_a = 0$ implying $D_b u^0 = 0$ in comoving basis. Therefore $(D_a u^b)(D_b u^a)$ and $D_a u^a$ reduce to purely spatial terms $(D_i u^j)\ (D_j u^i)$ and $(D_i u^i)$. We can now separate this tensor into σ_{ij}, w_{ij} and H exactly as before [with ∂_i replaced by D_i]. The last term in (A2.6) is

$$u^a R_{ab} u^b = R_{oo} = 8\pi G(T_{oo} - \frac{1}{2} T_{ab} g^{ab}) = 4\pi G(\rho + 3P) \tag{A2.7}$$

Using this result and

$$D_i u^i = 3H; \quad (D_i u^j)(D_j u^i) = 3H^2 + 2(\sigma^2 - w^2) \tag{A2.8}$$

we get

$$D_a \dot u^a = 3\dot H + 3H^2 + 2(\sigma^2 - w^2) + 4\pi G(\rho + 3P) \tag{A2.9}$$

The left hand side can be related to $\nabla^2 P$ by using the Euler equation. We have

$$\begin{aligned} D_a \dot u^a &= h^c_a D_c \dot u^a + u_a u^c D_c \dot u^a \\ &= h^c_a D_c \dot u^a + u^c D_c (u_a \dot u^a) - u^c \dot u^a D_c u_a \end{aligned} \tag{A2.10}$$

in which the middle term vanishes ($u_a \dot{u}^a = 0$) and the last term is of second order ($\dot{u}^a(u^c D_c u_a) = \dot{u}^a \dot{u}_a$). Using (A2.5), the first term can be written as

$$h^c_a D_c \dot{u}^a = -h^c_a D_c \left[\frac{h^{ab} D_b P}{(\rho + P)}\right] \approx -\frac{\nabla^2 P}{(\rho + P)} \qquad (A2.11)$$

which is correct to linear order. Therefore

$$D_a \dot{u}^a = -\frac{\nabla^2 P}{(\rho + P)} \qquad (A2.12)$$

Substituting back into (A2.9) we get

$$\dot{H} = -H^2 - \frac{4\pi G}{3}(\rho + 3P) - \frac{1}{3}\frac{\nabla^2 P}{(\rho + P)} \qquad (A2.13)$$

This equation, along with (A2.4) constitute the basic set. Before we separate out the quantities as $H = H_b + \delta H$ etc we have to take note of one additional complication. The interval $d\tau$ along a comoving world line between two adjacent comoving surfaces is position dependent on the surface and hence cannot be taken to be a valid label for the hypersurfaces. Suppose t stands for a valid ordering label for the hypersurfaces; then we can prove (we will do it at the end) that it can be chosen to give

$$\frac{d\tau}{dt} = 1 - \frac{\delta P}{\rho + P} \qquad (A2.14)$$

Given this relation, our equations can be recast with derivatives with respect to t. The continuity equation becomes

$$\begin{aligned}\frac{d\rho}{d\tau} &= \frac{d(\rho_b + \delta\rho)}{dt(1 - \delta P(\rho_b + P_b)^{-1})} \equiv (\dot{\rho}_b + \delta\dot{\rho})\left[1 + \frac{\delta P}{\rho_b + P_b}\right] \simeq \dot{\rho}_b + \delta\dot{\rho} + \frac{\dot{\rho}_b}{\rho_b + P_b}\delta P \\ &= -3H_b(\rho_b + P_b) - 3\delta H(\rho_b + P_b) - 3H_b(\delta\rho + \delta P)\end{aligned} \qquad (A2.15)$$

where we have kept only the linear terms and used the overdot symbol to denote derivatives with respect to t. Equating the zeroth order terms we get

$$\dot{\rho}_b = -3H_b(\rho_b + P_b) \qquad (A2.16)$$

Using this in the first order equations we get

$$\delta\dot{\rho} + \frac{\dot{\rho}_b}{(\rho_b + P_b)}\delta P = \delta\dot{\rho} - 3H_b\delta P = -3\delta H(\rho_b + P_b) - 3H_b\delta\rho - 3H_b\delta P \qquad (A2.17)$$

or

$$\delta\dot{\rho} = -3(\rho_b + P_b)\delta H - 3H_b\delta\rho \qquad (A2.18)$$

which is the advertised equation (A2.2) . The $\dot{H}$ equation proceeds along the same lines: We begin with :

$$\begin{aligned}\frac{dH}{d\tau} &= (\dot{H}_b + \delta\dot{H})\left(1 + \frac{\delta P}{\rho_b + P_b}\right) \\ &= -H_b^2 - 2H_b\delta H - \frac{4\pi G}{3}(\rho_b + 3P_b) - \frac{4\pi G}{3}(\delta\rho + 3\delta P) - \frac{1}{3}\frac{\nabla^2\delta P}{(\rho_b + P_b)}\end{aligned} \qquad (A2.19)$$

in which we have used the result $\nabla^2 P_b = 0$. The zeroth order term is

$$\dot{H}_b = -H_b^2 - \frac{4\pi G}{3}(\rho_b + 3P_b) \qquad (A2.20)$$

while the first order term is

$$\frac{\dot{H}_b}{(\rho_b + P_b)}\delta P + \delta\dot{H} = -2H_b\delta H - \frac{4\pi G}{3}\delta\rho - 4\pi G\delta P - \frac{1}{3}\frac{\nabla^2\delta P}{(\rho_b + P_b)} \qquad (A2.21)$$

Substituting the value of $\dot{H}_b$ in this, we get

$$\begin{aligned}\delta\dot{H} + 2H_b\delta H + \frac{4\pi G}{3}\delta\rho + \frac{1}{3}\frac{\nabla^2\delta P}{(\rho_b + P_b)} &= -\frac{\delta P}{(\rho_b + P_b)}\left[\dot{H}_b + 4\pi G(\rho_b + P_b)\right] \\ &= -\frac{\delta P}{(\rho_b + P_b)}\left[-H_b^2 - \frac{4\pi G}{3}(\rho_b + 3P_b) + 4\pi G(\rho_b + P_b)\right] \\ &= -\frac{\delta P}{(\rho_b + P_b)}\left[-H_b^2 + \frac{8\pi G}{3}\rho_b\right] = 0\end{aligned} \qquad (A2.22)$$

So we get the second of the advertised perturbation equation (A2.3)

$$\begin{aligned}\delta\dot{H} &= -2H_b\delta H - \frac{4\pi G}{3}\delta\rho - \frac{1}{3}\frac{\nabla^2\delta P}{(\rho_b + P_b)} \\ &= -2H_b\delta H - \frac{4\pi G}{3}\delta\rho - \frac{v^2}{3}\frac{\nabla^2\delta\rho}{(\rho_b + P_b)}.\end{aligned} \qquad (A2.23)$$

In arriving at the last equation we have used the definition $\delta P = v^2\delta\rho$. Also note that the 'dot' may be interpreted as $(\partial/\partial t)$ while acting on small deviations like δH, $\delta\rho$ etc. The rest of the analysis proceeds as in the non-relativistic case; rewriting the continuity equation as

$$\delta H = -\frac{1}{3(\rho_b + P_b)}\left[\delta\dot{\rho} + 3H\delta\rho]\right] = -\frac{1}{3(1+w)}\left[\dot{\delta} - 3Hw\delta\right] \qquad (A2.24)$$

and substituting into the $\delta\dot{H}$ equation, we get

$$\begin{aligned}\delta\dot{H} &= \frac{\dot{w}}{3(1+w)^2}\left[\dot{\delta} - 3Hw\delta\right] - \frac{1}{3(1+w)}\left[\ddot{\delta} - 3Hw\dot{\delta} - 3H\dot{w}\delta - 3\dot{H}w\delta\right] \\ &= -\frac{1}{3(1+w)}\ddot{\delta} + \dot{\delta}\left[\frac{\dot{w}}{3(1+w)^2} + \frac{Hw}{(1+w)}\right] - \delta\left[\frac{Hw\dot{w}}{(1+w)^2} - \frac{H\dot{w}}{(1+w)} - \frac{\dot{H}w}{(1+w)}\right]\end{aligned}$$
$$(A2.25)$$

It is easy to verify that

$$\begin{aligned} \dot{w} &= 3H_b(w - v^2)(1 + w) \\ \dot{H}_b &= -H_b^2 - \frac{1}{2}H_b^2(1 + 3w) = -\frac{3}{2}H_b^2 - \frac{3}{2}wH_b^2 = -\frac{3}{2}H_b^2(1 + w) \end{aligned} \qquad (A2.26)$$

Using these we can write

$$\delta\dot{H} = -\frac{1}{3(1+w)}\left\{\ddot{\delta} - 3H_b(2w - v^2)\dot{\delta} + \frac{9}{2}H_b^2\delta(2v^2 + w^2 - w)\right\} \qquad (A2.27)$$

Therefore the equation (A2.23) becomes

$$\begin{aligned} &\frac{1}{3(1+w)}\left\{\ddot{\delta} - 3H_b\dot{\delta}(2w - v^2) + \frac{9}{2}H^2\delta(2v^2 + w^2 - w)\right\} \\ &\qquad = -2H_b\frac{1}{3(1+w)}\left[\dot{\delta} - 3Hw\delta\right] + \frac{1}{2}H_b^2\delta + \frac{v^2}{3}\frac{1}{(1+w)}\nabla^2\delta \end{aligned} \qquad (A2.28)$$

Equivalently,

$$\ddot{\delta} + H_b\dot{\delta}(2 - 3(2w - v^2)) - \frac{3}{2}H_b^2\delta\left[1 - 6v^2 - 3w^2 + 8w\right] = -\left(\frac{kv}{a}\right)^2\delta \qquad (A2.29)$$

in which we have introduced the Fourier transform such that $\nabla^2\delta = -(k/a)^2\delta$. This was the equation which was used in the main text.

All that remains is the derivation of equation (A2.14). There are several ways of doing this, the easiest being the following:

Let $(t_1 x_i)$ be a set of coordinates with t labeling the comoving hypersurfaces; and let $(\lambda\mathbf{u}, \mathbf{e}_i)$ be a coordinate basis associated with this set, with $\lambda = (d\tau/dt)$. Obviously, $\mathbf{u}.\mathbf{e}_i = 0$ which implies that

$$\dot{\mathbf{u}}.\mathbf{e}_i = -\dot{\mathbf{e}}_i.\mathbf{u} = \dot{u}_i \qquad (A2.30)$$

On the other hand, since the basis vectors are coordinate induced, all Lie brackets vanish: $[\lambda\mathbf{u}, \mathbf{e}_i] = 0$. On expanding this, we get

$$(\lambda\dot{\mathbf{e}}_i)^a - D_i(\lambda\mathbf{u})^a = 0 \qquad (A2.31)$$

which is equivalent to

$$(\dot{\mathbf{e}}_i)^a - D_i u^a = \lambda^{-1}(\partial_i\lambda)u^a \qquad (A2.32)$$

Taking the dot-product with u_a, the second term vanishes; the first term gives $\dot{u}_i$ because of (A2.30). Thus we find

$$\dot{u}_i = \lambda^{-1}(\partial_i\lambda) \qquad (A2.33)$$

Combining this with (A2.5) and working upto first order, we find

$$\lambda^{-1}\partial_i\lambda = -\frac{\partial_i\delta P}{(\rho_b + P_b)} \qquad (A2.34)$$

giving

$$\lambda = \exp\left[-\frac{\delta P}{(\rho_b + P_b)}\right] \simeq 1 - \frac{\delta P}{(\rho_b + P_b)} = \frac{d\tau}{dt}, \qquad (A2.35)$$

This is the result to be proved.

Lastly, let us consider the case of several uncoupled fluids. [For example, radiation and dark matter]. In the Newtonian limit, this poses no special problems; we only have to take the H contributed by all matter and replace the driving term $4\pi G\delta\rho$ by $4\pi G \sum \delta\rho_i$. But there are some interesting subtleties in the relativistic case.

In a manner very similar to that of single component fluid, one can derive the following equations

$$\begin{aligned} \dot\rho_N &= -3H_N(\rho_N + P_N) \\ \frac{1}{3}D_a\dot u^a_N &= \dot H_N + H_N^2 + \frac{4\pi G}{3}(\rho + 3P) \\ D_a\dot u^a_N &= -\frac{\nabla^2 P_N}{\rho_N + P_N} + \frac{3(H - H_N)\dot P_N}{\rho_N + P_N} \end{aligned} \qquad (A2.36)$$

where $N = 1, 2, 3 \cdots$ denotes the various fluids (radiation, matter etc.) and

$$\rho = \sum_N \rho_N, \qquad P = \sum_N P_N, \qquad H = \sum_N H_N \qquad (A2.37)$$

etc. The last term in the third equation shows the main difference from the single component case. The mean values of all variables like H_N etc. are defined by averaging over each spacelike hypersurface. It, therefore follows that

$$< H_N >= H \qquad (A2.38)$$

giving

$$H - H_N = \delta H - \delta H_N \qquad (A2.39)$$

It is also easy to show that

$$(\rho + P)H = \sum(\rho_N + P_N)H_N \qquad (A2.40)$$

Combining all these equations, we can derive the perturbation equations

$$\begin{aligned} \delta\dot\rho_N &= -3(\rho_N + P_N)\delta H_N - 3H\delta\rho_N - 3H(\delta P_N - \theta_N\delta P) \\ \delta\dot H_N &= -2H\delta H_N - \frac{4\pi G}{3}\delta\rho - \frac{1}{3}\frac{\nabla^2\delta P_N}{(\rho_N + P_N)} + \frac{\dot P_N}{\rho_N + P_N}\left[\sum_M(\theta_M\delta H_M) - \delta H_N\right] \end{aligned} \qquad (A2.41)$$

where

$$\theta_N = \frac{\rho_N + P_N}{(\rho + P)} \qquad (A2.42)$$

To each component, we can associate a velocity dispersion

$$v_N^2 = \frac{\delta P_N}{\delta \rho_N} = \frac{\dot{P}_N}{\dot{\rho}_N} \tag{A2.43}$$

but this does not allow one to give a corresponding relation between the total δP and total $\delta\rho$. Because of this peculiarity, it is best to introduce a new variable which measures the deviation from adiabaticity.

Let us illustrate how it works for the case two uncoupled fluids. If $\delta_1 = (\delta\rho_1/\rho_1)$ and $\delta_2 = (\delta\rho_2/\rho_2)$ denote the fractional density contrast of each fluid, then the most convenient variables to use are the *total* density contrast

$$\delta = \frac{\delta\rho_1 + \delta\rho_2}{\rho_1 + \rho_2} = \frac{\rho_1\delta_1 + \rho_2\delta_2}{\rho_1 + \rho_2} \tag{A2.44}$$

and the 'non-adiabaticity' parameter

$$S = \frac{\delta_1}{1 + w_1} - \frac{\delta_2}{1 + w_2} \tag{A2.45}$$

[The name is due to the following fact: If the first fluid is dust ($w_1 = 0$) and the second is radiation ($w_2 = 1/3$), $S = \delta_1 - (3\delta_2/4)$; for adiabatic perturbations $\delta_1 = (3\delta_2/4)$ giving $S = 0$. Since the background model completely specifies the evolution of $\rho_1(t), \rho_2(t)$ etc, we can trade off the two unknowns δ_1 and δ_2 in favour of δ and S]. Using the general set of equations for the multicomponent medium, discussed above, we can write down the differential equations for δ and S. The procedure is straightforward but tedious. The final result is the set of equations:

$$\begin{aligned}\ddot{\delta} + [2 - 3(2w - v^2)]H\ddot{\delta} - \frac{3}{2}H^2(1 - 6v^2 + 8w - 3w^2)\delta \\ = -\frac{k^2}{a^2}(v^2\delta + w\eta)\end{aligned} \tag{A2.46}$$

$$\begin{aligned}\ddot{S} + (2 - 3u^2)H\dot{S} &= \frac{k^2}{a^2}[-u^2S + (v_2^2 - v_1^2)(1 + w)^{-1}\delta] \\ \eta &= \frac{(\rho_1 + P_1)(\rho_2 + P_2)}{(\rho + P)P}(v_1^2 - v_2^2)S\end{aligned} \tag{A2.47}$$

with

$$v^2 = \frac{\dot{P}}{\dot{\rho}}; \qquad v_A^2 = \frac{\dot{P}_A}{\dot{\rho}_A} \tag{A2.48}$$

$$u^2 = \frac{(\rho_1 + P_1)v_2^2 + (\rho_2 + P_2)v_1^2}{(\rho + P)} \tag{A2.49}$$

Quite clearly (A2.46)reduces to the density perturbation equation if $\eta = 0$ and there is only one component. When there exists more than one component δ and S are coupled; S generates δ and vice versa. In general, the solution to (A2.46) and (A2.47) needs to be obtained numerically.