Basics of Galaxy Formation

Asaf Pe'er[1]

March 15, 2017

This part of the course is based on Refs. [1] - [4].

1. Introduction

The discussion about perturbation growth in the early universe and its imprint on the cosmic microwave background (CMB), is aimed at addressing the basic question of galaxy formation. In other words, we need to understand how come our universe is populated by galaxies, rather than, say, a uniform sea of stars? Furthermore, the galaxies have typical scale, with luminosity $L_\star \simeq 3 \times 10^{10}\ L_\odot$.

The emergence of starts, galaxies, group and clusters of galaxies from the nearly homogeneous early universe is called **structure formation**. Due to lack of time, I will go only briefly about the basics of it.

1.1. Linear structure formation

In the discussion about perturbations (Equations 40,41 at that chapter), we introduced the dimensionless overdensity parameter,

$$\delta\rho(\vec{x}) \equiv \frac{\rho(\vec{x}) - \rho_0}{\rho_0} = \frac{\rho(\vec{x})}{\rho_0} - 1 \tag{1}$$

where ρ_0 denotes the average matter density over a volume V, that is sufficiently large so that the universe can be considered homogeneous. Further note that in many textbooks, the "ρ" is omitted, and one uses $\delta(\vec{x}) \equiv \delta\rho(\vec{x})$.

In the real universe, $\delta\rho(\vec{x})$ has a well defined value at each location, $\vec{x}$. However, for the mathematical treatment, we make some simplified assumption, such as that $\delta\rho(\vec{x})$ is continuous. In this case, the value it takes for two nearby points, $\vec{x}'$ and $\vec{x}' + \delta\vec{x}$ must be correlated. If we know $\delta\rho(\vec{x}')$, then the uncertainty in $\delta\rho(\vec{x}'+\delta\vec{x})$ must approach 0 as $\delta\vec{x} \to 0$. It is therefore costume to define a correlation function,

$$\xi(\delta\vec{x}) \equiv \langle \delta\rho(\vec{x})\delta\rho(\vec{x}' + \delta\vec{x})\rangle \tag{2}$$

[1]Physics Dep., University College Cork

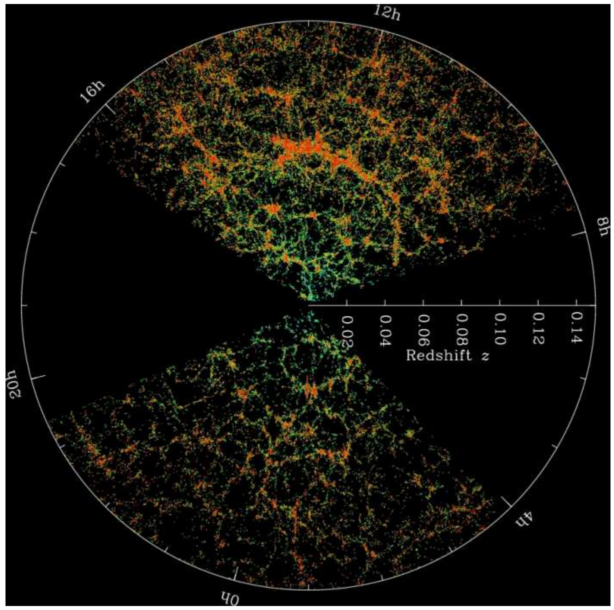

Fig. 1.— The Sloan Digital Sky Survey (SDSS) map of galaxies. Clearly, the universe is made of galaxies, rather than homogeneous sea of stars.

The brackets imply that we take the expectation value of this function.

For a universe that is isotropic, ξ does not depend on the direction of $\delta\vec{x}$, but only on its magnitude, $\delta x = |\delta\vec{x}|$. We thus have

$$\xi(\delta x) \equiv \langle \delta\rho(\vec{x})\delta\rho(\vec{x}' + \delta\vec{x})\rangle \tag{3}$$

In quantifying the fluctuations, we make several assumptions. First, we make a Fourier transform to write

$$\delta\rho(\vec{x}) = \frac{1}{V}\sum_{\vec{k}} \delta_{\vec{k}} e^{i\vec{k}\cdot\vec{x}}, \qquad \delta_{\vec{k}} = \int_V d^3x \delta\rho(\vec{x}) e^{-i\vec{k}\cdot\vec{x}}, \tag{4}$$

where $\vec{k} = 2\pi\vec{n}/V^{1/3}$ with $\vec{n} = (n_1, n_2, n_3)$ is a triple of integers, and the sum is over all $\vec{n}$. Since $\delta\rho(0)$ has zero mean by definition, we we must have $\delta_{\vec{k}=0} = 0$. Furthermore, since $\delta\rho(\vec{x})$ is a real number, we must have $\delta_{-\vec{k}} = \delta^\star_{\vec{k}}$.

Inserting the Fourier expansion into equation 3, we find

$$\xi(\delta x) = \frac{1}{V^2}\sum_{\vec{k}',\ \vec{k}} \langle \delta_{\vec{k}'}\delta_{\vec{k}}\rangle e^{i(\vec{k}'+\vec{k})\cdot\vec{x}'} e^{i\vec{k}\cdot\vec{x}} \tag{5}$$

Since the right hand side cannot depend on $\vec{x}'$, we have that

$$\langle \delta_{\vec{k}'} \delta_{\vec{k}} \rangle = 0 \;\; if \;\; \vec{k}' \neq -\vec{k}. \tag{6}$$

We thus find

$$\xi(\delta x) = \frac{1}{V^2} \sum_{\vec{k}} \langle \delta_{-\vec{k}} \delta_{\vec{k}} \rangle e^{i\vec{k}\cdot\vec{x}} = \frac{1}{V^2} \sum_{\vec{k}} \langle |\delta_{\vec{k}}|^2 \rangle e^{i\vec{k}\cdot\vec{x}} \equiv \frac{1}{V} \sum_{\vec{k}} P(\vec{k}) e^{i\vec{k}\cdot\vec{x}} \tag{7}$$

where $P(\vec{k}) \equiv \langle |\delta_{\vec{k}}|^2 \rangle / V$ is the **power spectrum** of the random field. Equation 7 thus states that the power spectrum is a Fourier transform of the correlation function. Since the universe is isotropic, the power spectrum can only depend on $k = |\vec{k}|$.

The variance of the overdensity is

$$\langle \delta\rho^2(\vec{x}) \rangle = \xi(0) = \frac{1}{V} \sum_{\vec{k}} P(k) \tag{8}$$

The second assumption we make is that $\delta_{\vec{k}}$ are independent random variables. When this is the case, we say that $\delta(\vec{x})$ is a **Gaussian random field**. For such fields, the power spectrum fully determines all their statistical properties [1].

The third thing which is common to do is **filtering**. Basically, we choose a **smoothing length**, L . What is done in practice is to convolve the overdensity field $\delta\rho(\vec{x})$ with a **window function** $W(\vec{x})$, to generate a smoothed field $\delta\rho_L(\vec{x})$ using

$$\delta\rho_L(\vec{x}) \equiv \int d^3x' W(\vec{x} - \vec{x}') \delta\rho(x'), \tag{9}$$

where $\int d^3x W(\vec{x}) = 1$, and $W(\vec{x}) \simeq 0$ for $|\vec{x}| \gtrsim L$.

A simple filter that is in common use is to simply cut all the modes with k greater then some value, K. This is known as **sharp K-space filter**. The variance of a an overdensity field that has been smoothes with such filter is

$$\sigma_K^2 \equiv \langle \delta\rho^2(\vec{x}) \rangle = \frac{1}{V} \sum_{|\vec{k}|<K} P(k). \tag{10}$$

In the limit $V \to \infty$, we get

$$\frac{1}{V} \sum_{\vec{k}} \to \frac{1}{V} \int d^3n = \frac{1}{(2\pi)^3} \int d^3k, \tag{11}$$

---

[1] this is a mathematical theorem that can be proven, though I do not bring the proof here.

and therefore

$$\sigma_K^2 = \frac{1}{2\pi^2}\int_0^K dk k^2 P(k). \tag{12}$$

It is typical to assume a power spectrum of the form $P(k) \propto k^n$, for which $\sigma_K^2 \propto K^{3+n}$. A very important model for power spectrum in cosmology is the **Harrison-Zeldovich** spectrum, which is also known as **scale invariant spectrum**. For this spectrum, we simply put $n = 1$. Its importance can be understood by the following argument: The RMS density fluctuations on scales K^{-1} is $\rho_{RMS}(K) \propto \sigma_K \propto K^{(3+n)/2}$. Thus, the corresponding RMS fluctuations in the mass is $M_{RMS}(K) \propto \rho_{RMS}(K)/K^3 \propto K^{(n-3)/2}$ (this can be understood as due to the use of the filter). The corresponding fluctuations in the gravitational potential is therefore $\Phi_{RMS}(K) \propto GM_{RMS}(K)K \propto K^{(n-1)/2}$. Thus, the RMS potential fluctuations is independent of scale if and only if $n = 1$. This spectrum therefore represents fluctuations that are independent on the scale.

Once the initial perturbation spectrum is set (or assumed) at $z \to \infty$, the perturbations grow with time, as discussed earlier. During most of the evolution of the universe, the perturbation growth was linear. As we showed in developing the theory of perturbation growth (eq. 41), in matter dominated era and non-relativistic fluid one finds that

$$\delta\rho \propto a \propto t^{2/3}. \tag{13}$$

In this case, we also have $\delta_{\vec{k}} \propto t^{2/3}$, which is correct as long as the universe is matter dominated. At short wavelength, on the other hand, Silk damping prevents the growth.

The results of the linear perturbation growth can be captured by defining the transfer function,

$$T^2(k) = \frac{\langle\delta_k^2\rangle_{z=0}}{\langle\delta_k^2\rangle_{z\to\infty}} / \frac{\langle\delta_o^2\rangle_{z=0}}{\langle\delta_0^2\rangle_{z\to\infty}}, \tag{14}$$

which is $T(k) \approx 1$ for large scales (small k) and $T(k) \propto (k\lambda_{\rm eq.})^{-2}$ for scales smaller than $x = \lambda_{\rm eq.}$. As already discussed, the exact transfer function depends on the values of the parameters defining the cosmology. Thus, observations of both the CMB and the large scale structure in fact provide valuable information about the cosmology.

If linear theory were valid at present, then current matter power spectrum $P(k)$ would be given by

$$P(k) \propto T^2(k) P_{\rm prim}(k) \tag{15}$$

where $P_{\rm prim}(k)$ is the primordial power spectrum. This is plotted in Figure 2 below, with primordial spectrum taken to be Harrison-Zeldovich.

The success of the theoretical models in predicting the observed power spectrum implies the following conclusions:

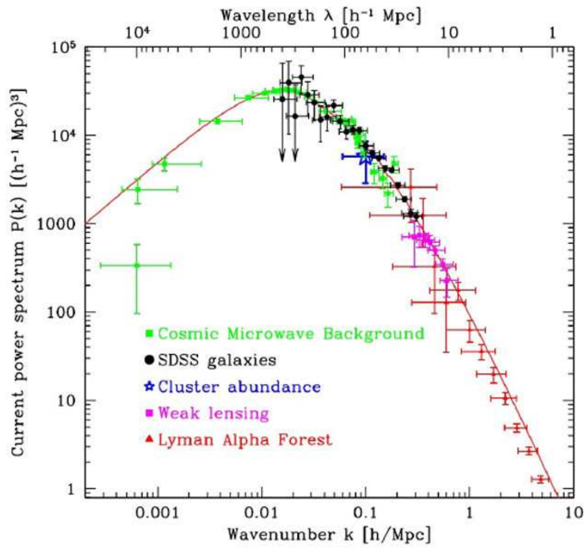

Fig. 2.— The matter power spectrum combined with CMB spectrum. The solid line is the theoretical spectrum for a flat FRW universe with a Harrison-Zeldovich initial spectrum.

1. On large scales, the universe is homogeneous and isotropic, described by FRW metric.
2. The geometry of the universe is flat (predicted by inflation)
3. The values of the cosmological parameters are set, $\Omega_{\lambda,0} = 0.763 \pm 0.034$, $\Omega_{m,0} = 0.237 \pm 0.034$, $\Omega_{b,0} = (0.0455 \pm 0.0015)h_{70}^{-2}$.
4. The dark matter is cold;
5. The initial density perturbations were small ($|\delta\rho| \ll 1$) and described by a Gaussian random field;
6. The initial power spectrum of the density fluctuations was (approximately) Harrison-Zeldovich spectrum.

These are known as **standard ΛCDM cosmology**.

2. Non-linear structure formation: spherical collapse

The typical density in a luminous galaxy at radius of $\sim$ few kpc is $\approx 10^5$ times larger than the critical density, ρ_c. This means, that in a certain point, the formation of galaxies must have involved non-linear density fluctuations.

Let us consider first the (too simplified) model of spherical collapse. For simplicity, we take the EdS cosmology (matter dominated and flat universe). This is a reasonably good approximation until relatively recent times.

Let us take, at some initial time t_i a (spherically symmetric) density fluctuations such as the (volume average) over density within a sphere of radius r_i is $\delta\rho(t_i) \equiv \delta\rho_i \ll 1$. Far outside the sphere, the density is that of FRW model, namely

$$\rho_m(t) = \frac{1}{6\pi G t^2} \tag{16}$$

The total mass enclosed in the sphere is therefore

$$M = \frac{4}{3}\pi(1+\delta_i)\rho_m(t_i) r_i^3. \tag{17}$$

Let us now follow the evolution of the radius $r(t)$ of material at initial radius r_i. The gravitational acceleration of this material is determined only by the total interior mass, M, which is constant (the matter is assumed to be cold).

We can use Newton's law to write

$$\frac{d^2 r(t)}{dt^2} = -\frac{GM}{r^2(t)} \tag{18}$$

This is an equation of motion of an object that is launched vertically from the surface of a spherical body of mass M. Of course, as we are interested in collapse, we assume that the object does not have enough energy to escape. The solution to Equation 18 is written in parametric form,

$$r(\eta) = A(1-\cos\eta); \quad t = \sqrt{\frac{A^3}{GM}}(\eta - \sin\eta) + t'. \tag{19}$$

In our analogy, the radius $r_{\rm max} = 2A$ is the turn-around radius, which occurs at $\eta = \pi$. By arguing that $r(t) = 0$ at $t = 0$, we can set $t' = 0$.

The average density inside the sphere is $\rho_s(M) \equiv M/\frac{4}{3}\pi r^3(t)$. We thus obtain the average overdensity inside $r(t)$ as

$$\delta\rho(t) \equiv \frac{\rho_s(t)}{\rho_m(t)} - 1 = \frac{9}{2}\frac{(\eta-\sin\eta)^2}{(1-\cos\eta)^3} - 1. \tag{20}$$

At turnaround, the overdensity is

$$\delta\rho_{\max} \equiv \delta(\eta = \pi) \quad \rightarrow \delta\rho_{\max} = \frac{9\pi^2}{16} - 1 = 4.55 \tag{21}$$

For small initial density contrast, we can further solve for A , and therefore $r_{\max}$ as follows. Taking $\eta \ll 1$ and Taylor expanding Equation 20, one gets $\delta\rho(t) \approx (3/20)\eta^2 + \mathcal{O}(\eta^4)$. Since the initial perturbations are small, $\delta\rho_i \ll 1$, we can write $\eta(t_i) = (20/3)\delta\rho_i$. Similar expansion of Equation 19 gives $A = 2r_i/\eta^2 \simeq (3/10)r_i/\delta\rho_i$. Applying this to equations 17, and using $r_{\max} = 2A$, one gets

$$r_{\max} = \left(\frac{243}{250}\right)^{1/3} \frac{(GMt_i^2)^{1/3}}{\delta\rho_i} \tag{22}$$

and the turnaround time is

$$t_{\max} = \pi\sqrt{\frac{A^3}{GM}} = \pi\left(\frac{243}{2000}\right)^{1/2} \frac{t_i}{\delta\rho_i^{3/2}} = 1.095\frac{t_i}{\delta\rho_i^{3/2}}. \tag{23}$$

This implies that **the larger the initial fluctuation, $\delta\rho_i$, the sooner the collapse commences.**

Note the following. If the collapse was linear (rather than non-linear), using Taylor expansion in Equation 19 we would get $t(\eta) \approx \sqrt{A^3/GM}\frac{\eta^3}{6}$. Using $\delta\rho(t) \simeq \frac{3}{20}\eta^2$, we could write

$$\delta\rho(t) = \frac{3}{20}(6\pi)^{2/3}\left(\frac{t}{t_{\max}}\right)^{2/3}.$$

In the non-linear model, the collapse is completed by $t = 2t_{\max}$. However, in the linear model, at this time we would get

$$\delta\rho(2t_{\max}) \equiv \delta\rho_c = \frac{3}{20}(12\pi)^{2/3} \simeq 1.686$$

This is known as the **critical density for collapse**. In modeling the resulting halos in the Press-Schechter theory, this critical density plays a role in determining the minimum mass of the forming halos.

In our simplified picture, the collapse reaches singular density at $\eta = 2\pi$, or $t = 2t_{\max}$. Does that mean that all matter collapses into a black hole?

Of course not. The collapsing dark matter will undergo phase mixing and (violent) relaxation, and settle into an equilibrium configuration known as **halo**. This process of collapse, mixing and relaxation is referred to as **virialization**.

As dark matter particles are only weakly interacting, the physical processes that lead to relaxation are NOT two body collision, but rather a combination of four other mechanisms:

1. **Phase mixing**: spread of neighboring points in phase-space due to the difference in frequencies between neighboring orbits.

2. **Chaotic mixing**: spread of neighboring points in phase-space due to the chaotic nature of their orbits

3. **Violent relaxation**: change in energy of individual particles due to the change in the overall potential.

4. **Landau damping**: damping of decay of perturbations due to decoherence between particles and waves (same as in free streaming).

Once the halo settled into an approximate equilibrium, we can use the virial theorem to estimate its radius. Recall that the virial theorem states that in a steady state,

$$E_k = -\frac{1}{2}E_p \tag{24}$$

where E_k is the kinetic energy and E_p is the potential energy of the system. At $r_{\rm max}$, the system has no kinetic energy ($E_{k,0} = 0$), and its potential energy is given by

$$E_{p,0} = -\frac{3}{5}\frac{GM}{r_{\rm max}} \tag{25}$$

Since the total energy is conserved, $E_k + E_p = E_{p,0}$, and we have $E_p = E_{p,0}/2$, from which we deduce that the radius of the halo is $r_h = r_{\rm max}/2$.

The density inside the halo can then be estimated using $\rho_h = M/(\frac{4}{3}\pi r_h^3)$, using equations 16 and 23 we find

$$\frac{\rho_h}{\rho_m(2t_{\rm max})} = \frac{M}{\frac{4}{3}\pi r_h^3} 6\pi G(2t_{\rm max})^2 = \frac{9}{4}\pi^2\left(\frac{r_{\rm max}}{r_h}\right)^3 \simeq 178 \tag{26}$$

Although the argument we based this derivate on are sketchy, it turns out that this results holds for various cosmologies (=cosmological parameters). It is therefore common to define the **virial radius**, denoted by r_{200} of a halo as the radius at which the matter density is equal to 200 times the critical density. Inside r_{200} the halo is assumed to be in virial equilibrium, and the mass inside r_{200} is a measure of the total mass of the halo.

2.1. Application: when did the galaxies form?

The analysis above enables us to estimate the time at which galaxies could have been formed. Looking at the Milky way galaxy, its average density is $\approx 10^5\rho_c$. Since it must be 200 times the density at the collapsing time, the matter density at the beginning of the collapse was ~ 200 times larger than the critical density at present time. Since $\Omega_m = 0.3$, the matter density at the beginning of collapse was $500/0.3 \simeq 1700$ times the present average density. However, the average density evolves with redshift as $(1+z)^3$, so the collapse began at

$$(1+z) \leq 1700^{1/3} \approx 12.$$

3. The cosmic web and Press-Schechter theory

(very briefly.) During the linear growth period, the over density field $\delta\rho(x)$ retains its shape; only its amplitude increases as $t^{2/3}$. However, once non-linear growth begins, regions that are slightly underdense expand more rapidly than the average, as gravity pulls less than the average. These underdense regions, or **voids** must therefore collide with each other and merge. These are separated by **sheets** of higher density matter. When these sheets meet, they form a **filament** of high-density matter. Thus, matter becomes more and more confined to a **cosmic web** of thin dense sheets that enclose voids (see Figure 3). When the filaments intersect, they form **nodes**, which then develop into the virial halos.

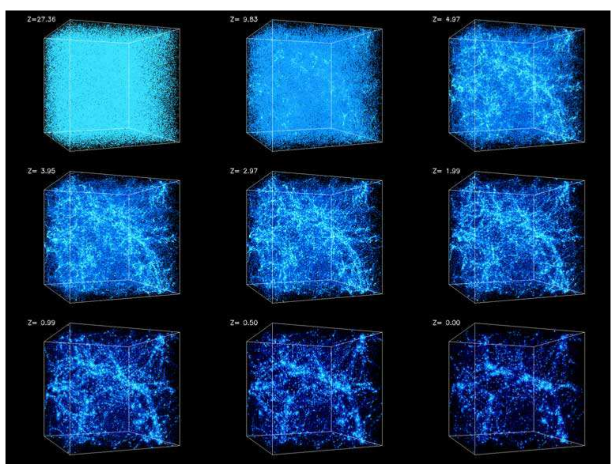

Fig. 3.— Formation of cosmic web and its growth over cosmological times. Halos and galaxies are formed at the nodes. Each cube have a size of ≈ 40 Mpc (comoving). Simulation results taken from the center for cosmological physics at University of Chicago, see http://cosmicweb.uchicago.edu/filaments.html

The halos are not static, but continuously evolve and merge. A basic question is what is the number of halos of a given mass per unit volume, $n(M,t)dM$; this is known as the **halo**

mass function. The answer to this question is given by the **extended Press-Schechter theory**, which we don't have time to discuss.

Furthermore, the halos themselves are dynamically evolving they are not necessarily spherical, but rather elliptical, they are rotating, and form of various sub-structures, which again, we will not consider here.

4. Collapse of gas, and formation of gaseous halos

So far, we discuss structure formation under the influence of gravity only. While this is relevant as long as we consider only dark matter (DM), when we consider baryons and electrons - the gas we see, we need to include the role of gas dynamics and radiative processes, in order to understand how the structure we observe form and evolve.

4.1. Hydrostatic equilibrium

As the gas collapses, it reaches a state of **hydrostatic equilibrium**. In this state, the gravitational forces are balanced by pressure gradients, so we have

$$\nabla p(\vec{r}) = -\rho(\vec{r})\nabla\Phi(\vec{r}). \tag{27}$$

Here, Φ is the gravitational potential, which, in the spherical case we can take as

$$\frac{d\Phi}{dr} = \frac{GM(r)}{r^2}. \tag{28}$$

For ideal gas, the pressure is given by $p = \rho k_B T/\mu m_p$, where ρ is the mass density, $\mu = \langle\rho/(n m_p)\rangle$ is the average molecular weight of the gas, and n is the number density of particles. Thus, we have for the spherical case,

$$\frac{dp}{dr} = \frac{k_B}{\mu m_p}\frac{d}{dr}(\rho T). \tag{29}$$

We can now solve equation 27, to get

$$M(r) = M_{\rm gas}(r) + M_{\rm DM}(r) = -\frac{k_B T(r) r}{\mu m_p G}\left[\frac{d\ln\rho_{\rm gas}}{d\ln r} + \frac{d\ln T}{d\ln r}\right]. \tag{30}$$

Thus, the hydrostatic equation (30) provides an estimate of the **total** mass of the halo (gas + DM) within some radius r, from measurements of the density and temperature profiles, $\rho(r)$

and $T(r)$, of the gas. This is a standard method for estimating the mass of X-ray clusters, for which both $\rho(r)$ and $T(r)$ can be obtained from their X-ray emission.

Equation 30 needs to be modified, if the pressure is not only determined by the thermal motion of particles ($p_{\rm thermal}$), but there are additional sources of pressure such as non-thermal turbulence, magnetic fields and/or cosmic rays. Denoting this pressure by $p_{\rm NT}$, Equation 30 reads

$$M(r) = M_{\rm gas}(r) + M_{\rm DM}(r) = -\frac{k_B T(r) r}{\mu m_p G}\left[\frac{d\ln\rho_{\rm gas}}{d\ln r} + \frac{d\ln T}{d\ln r} + \frac{p_{\rm NT}}{p_{\rm thermal}}\frac{d\ln p_{\rm NT}}{d\ln r}\right]. \tag{31}$$

Unfortunately, accurate measurements of $p_{\rm NT}(r)$ are extremely difficult to obtain...

Furthermore, in general, one cannot use hydrostatic equilibrium condition to fully specify the density distribution of the gas, since its temperature also depends on r, $T = T(r)$. One thus have to make further assumptions, such as:

- That the gas is isothermal, namely $T(r) = const$, and that $p_{\rm NT} = 0$;
- or that the gas is polytropic, namely $p(r) \propto \rho_{\rm gas}^{\Gamma}$/ An important example of a polytropic gas is isentropic gas, for which $\Gamma = \gamma$; for mono-atomic, isentropic gas, we have $\gamma = 5/3$, and thus $\rho(r) \propto T(r)^{3/2}$.

4.2. The virial temperature

The virial theorem provides a constraint on the global properties of any dynamical system. It reads:

$$\frac{1}{2}\frac{d^2 I}{dt^2} = 2E_k + E_p + \Sigma, \tag{32}$$

where I $(= \sum_k m_k r_k^2 = \int \rho \vec{r}^2 d^3x)$ is the moment of inertia of the system, $E_k = \frac{1}{2}\int \rho\langle\vec{v}^2\rangle d^3x$ is the kinetic energy, $E_p = -\int \rho\vec{x}\cdot\nabla\Phi d^3x$ is the potential energy, and $\Sigma = -\int \rho\langle v^2\rangle \vec{x}\cdot d\vec{S}$ is the work done by an external pressure (a surface pressure term). Note that this term is ignored when integrating over all space and taking $\rho \to 0$ at $\vec{x} \to \infty$, but this is true only for isolated system; in reality, a halo may be embedded in non-zero density field, yielding, in general, a non-zero surface pressure term.

The system is said to be in **virial equilibrium**, or static, if $d^2I/dt^2 = 0$, namely

$$2E_k + E_p + \Sigma = 0. \tag{33}$$

(If $d^2I/dt^2 > 0$ the system expands, $d^2I/dt^2 < 0$ it contracts).

Consider a uniform cloud of monoatomic gas ($\gamma = 5/3$) with constant temperature T. Recall that for such a gas we have: its energy density is $u = \rho c_v T$; its pressure is $p = \rho RT$, where $R = c_p - c_v$ is the gas constant; and thus $p = \frac{R}{c_v}u = (c_p/c_v - 1)u = (2/3)u$ (for adiabatic index $\gamma = 2/3$). The (static) virial theorem applied to this system then gives

$$2 \times \frac{3M_{\rm gas}k_B T}{2\mu m_p} - \frac{3GM_{\rm gas}M_{\rm vir}}{5r_{\rm vir}} - 4\pi r_{\rm vir}^3 p_{\rm ext} = 0, \tag{34}$$

where $r_{\rm vir}$ is the radius of the cloud (the **virial radius** when the cloud is in equilibrium), and $M_{\rm vir}$ is the total mass. If there is no external pressure, $p_{\rm ext} = 0$, the virial theorem defines a **virial temperature** via

$$T_{\rm vir} = \frac{\mu m_p GM}{5k_B r_{\rm vir}} = \frac{\mu m_p}{5k_B}V_{\rm vir}^2, \tag{35}$$

where $V_{\rm vir} = \sqrt{GM/r_{\rm vir}}$ is the halo circular velocity at the virial radius.

In fact, it is common to use a factor 5/2 of this result, and write the virial temperature as

$$T_{\rm vir} = \frac{\mu m_p}{2k_B}V_{\rm vir}^2 \simeq 3.6 \times 10^5 K \left(\frac{V_{\rm vir}}{100 \text{ km s}^{-1}}\right)^2 \tag{36}$$

Note, though, that in general, the gas will have a temperature profile, and therefore cannot be described by a single temperature. Nonetheless, the concept of virial temperature is useful as an order of magnitude estimate in the formation of gas halos.

4.3. Radiative cooling time

So far we ignored radiation, which can cause both heating and cooling. Let us see how the inclusion of radiation modifies the results developed so far.

We first define the total cooling and heating rates per unit volume by $[\mathcal{H}]$ and $[\mathcal{C}]$. Both have units of erg s^{-1} cm^{-3}. It is useful to define a **cooling function** by

$$\Lambda(T) \equiv \frac{\mathcal{C}}{n_H^2}, \tag{37}$$

where n_H is the number density of hydrogen atoms (both neutral and ionized). The units of $[\Lambda]$ are erg s^{-1} cm^{+3}.

We can further define the **cooling time**, as the time it takes the gas to radiate away its internal energy,

$$t_{\rm cool} = \frac{\rho\mathcal{E}}{\mathcal{C}} = \frac{\rho\mathcal{E}}{n_H^2\Lambda(T)}, \tag{38}$$

where $\mathcal{E}$ is the internal energy per unit mass and ρ is the mass density.

For an ideal, monoatomic gas, we have $\mathcal{E} = \frac{1}{\gamma - 1}\frac{k_B T}{\mu m_p}$ with $\gamma = 5/3$, and therefore

$$t_{\rm cool} = \frac{3nk_BT}{2n_H^2\Lambda(T)} \simeq 3.3\times10^9 \text{ yr} \left(\frac{T}{10^6\ K}\right)\left(\frac{n}{10^{-3}\ \text{cm}^{-3}}\right)^{-1}\left(\frac{\Lambda(T)}{10^{-23}\ \text{erg s}^{-1}\ \text{cm}^3}\right)^{-1} \tag{39}$$

In deriving this value, we have assumed that the number density of hydrogen atoms is $n_H = (4/9)n$ which is the case for fully ionized primordial gas, $n = \rho/(\mu m_p)$ is the number density of gas particles, and we took value for $\Lambda(T)$ which is typical for primordial gas (=gas produced at recombination) with low metallicity; see further discussion on this value below.

We further note that the cooling time $t_{\rm cool} \propto n/n_H^2 \propto n^{-1}$; thus, denser gas cools faster.

In order to estimate the importance of cooling on a system, we must compare it to the two other time-scales of the system:

- The age of the universe, which can be approximated by the Hubble time,

$$t_H = \frac{1}{H(z)} \propto \frac{1}{(G\bar{\rho})^{1/2}} \qquad \bar{\rho} = \Omega_m\rho_c \tag{40}$$

- the dynamical time, namely the free-fall time of the system,

$$t_{\rm ff} = \left(\frac{3\pi}{32G\bar{\rho}_{sys.}}\right)^{1/2} \propto \frac{1}{(G\bar{\rho}_{sys.})^{1/2}}. \tag{41}$$

  (Recall that this is the time on which a gas cloud will collapse in the absence of pressure, and also the time scale in which the system restores hydrodynamic equilibrium, if disrupted).

  Both the free fall time and the Hubble time are proportional to $\rho^{-1/2}$, and typically the free-fall time is much faster, $t_{\rm ff} \sim t_H/10$.

We can therefore discriminate between three cases:

1. $t_{\rm cool} > t_H$: cooling time is not important. Gas is in hydrostatic equilibrium, unless it was recently disturbed.

2. $t_{\rm ff} < t_{\rm cool} < t_H$: The system is in quasi-hydrostatic equilibrium. It evolves on cooling time scale. The gas contracts slowly as it cools, but the system has sufficient time to continue to re-establish hydrostatic equilibrium.

3. $t_{\rm cool} < t_{\rm ff}$: catastrophic cooling. Gas cannot respond fast enough to loss of pressure. Since cooling time decreases with increasing density, cooling proceeds faster and faster (=catastrophic). The gas basically falls to the center of the dynamic system on time scale comparable to the free fall time.

Further, note the following. Since $t_{\rm cool} \propto \rho_{\rm gas}^{-1} \propto (1+z)^{-3}$, while $t_{\rm ff} \propto \rho^{-1/2} \propto (1+z)^{-3/2}$, cooling is generally more efficient at higher redshifts.

4.4. Cooling processes

Let us now consider the physical cooling processes responsible for the cooling function. For galaxy formation, the relevant processes are **two-body radiative processes**, in which gas loses energy through the emission of photons as a consequence of two-body interactions.

There are mainly four processes which are particularly important:

1. Free-free emission (or bremsstrahlung): $e^- + X^+ \to e^- + X^+ + \gamma$ (X^+ is typically a proton, p^+). In this process, a free electron is accelerated by a heavy ion. Accelerating charge emits a photons.

2. Free-bound emission, or recombination: $e^- + X^+ \to X + \gamma$. In this process, a free electron recombines with an ion. The binding energy plus the free electron's kinetic energy are radiated away; only the free electron kinetic energy is counted as a loss (=cooling), since the binding energy is already accounted for as loss in collisional ionization process.

3. Bound-free process, or collisional ionization: $e^- + X \to X^+ + 2e^-$. The impact of a free electron ionizes an atom, releasing an initially bounded electron. This takes a kinetic energy from the free electron.

4. bound-bound interaction, or collisional excitation, $e^- + X \to e^- + X' \to e^- + X + \gamma$. Impact of a free electron knocks a bound electron to an excited state. As it decays, it emits a photon.

In order to calculate the cooling function, $\Lambda(T) \equiv \mathcal{C}/n_H^2$ for a certain gas, one needs to know the densities of the various ionic species. As a concrete example, we take a pure H/He mixture (which is the simplest, relevant case). For this gas, the relevant ions are $e^-, H_0, H^+, He_0, He^+, He^{++}$.

Assuming the total gas density is fixed, the densities of the different ions are governed by differential equation of the form

$$\frac{dn_{H0}}{dt} = \alpha_{H^+}(T) n_{H^+} n_e - \Gamma_{eH_0}(T) n_e n_{H0} - \Gamma_{\gamma H_0} n_{H0}, \tag{42}$$

where $\alpha_{H^+}(T)$ is the hydrogen recombination coefficient (cm^{3}s^{-1}); $\Gamma_{eH_0}(T)$ is the rate of collisional ionization (cm^{3}s^{-1}); and $\Gamma_{\gamma H_0}$ is the rate of photo-ionization (s^{-1}).

Without getting into the details of the calculations, a few key results are:

- The typical timescale for photo-ionization depends on the density of ionizing photons. For typical ionizing background, this time scale is

$$t_{\rm photo} \sim \frac{1}{\Gamma_{\gamma H_0}} \simeq 3 \times 10^4 \text{ yr}, \tag{43}$$

  which is much shorter then the dynamical time (the free-fall time, Equation 41).

- Even in the absence of photo-ionization, the density of hydrogen, n_{H_0} evolves on a time scale

$$\frac{1}{n_e(\alpha_{H^+} - \Gamma_{eH_0})} \sim 10^6 \text{ yr} \left(\frac{n_e}{10^{-5}\text{cm}^{-3}}\right)^{-1} \tag{44}$$

  which is also much shorter then the dynamical time.

Thus, in most cases, it is safe to assume that the system has equilibrated, and the destruction rate is equal to the creation rate of the various ions. This equilibrium is known as **ionization equilibrium.** If there is *no ionizing radiation*, namely photo-ionization is ignored but equilibrium is reached, the situation is called **collisional ionization equilibrium**, or CIE.

In ionization equilibrium, the ion abundances are determined by simple algebraic equations, which are much simpler than differential equations, such as

$$\begin{aligned} \Gamma_{eH_0} n_e n_{H_0} + \Gamma_{\gamma H_0} n_{H_0} &= \alpha_{H^+} n_e n_{H^+}; \\ n_{H^+} + n_{H_0} &= n_H; \\ n_{H^+} + n_{He^+} + 2n_{He^{++}} &= n_e; ... \end{aligned} \tag{45}$$

Once the gas is in ionizing equilibrium, it is relatively easy to determine its cooling function:

- At high temperatures, the gas is fully ionized. The only contribution to the cooling function is from bremsstrahlung, for which $\Lambda \propto T^{1/2}$.

- At $T < 10^4$ K, all the gas is neutral. There are no ions, and therefore there is no bremsstrahlung emission. Furthermore, residual free electrons do not have enough energy to excite the hydrogen to its first excited state (minimum energy required is 10.2 eV).

- If $T > few \times 10^4$ K, all hydrogen is ionized. Thus, the hydrogen no longer contributes to the cooling, causing a local drop in $\Lambda(T)$. At around $T \sim 10^5$ K, a second peak in $\Lambda(T)$ is attributed to helium.

- When metals are present, many new cooling channels are available, mainly between 10^4 K and 10^7 K, greatly increasing $\Lambda(T)$. For $Z = Z_\odot$, the cooling rate at 10^6 K is about a factor 100 larger than the cooling rate of primordial gas.

A plot of the cooling function appears in Figures 4 and 5.

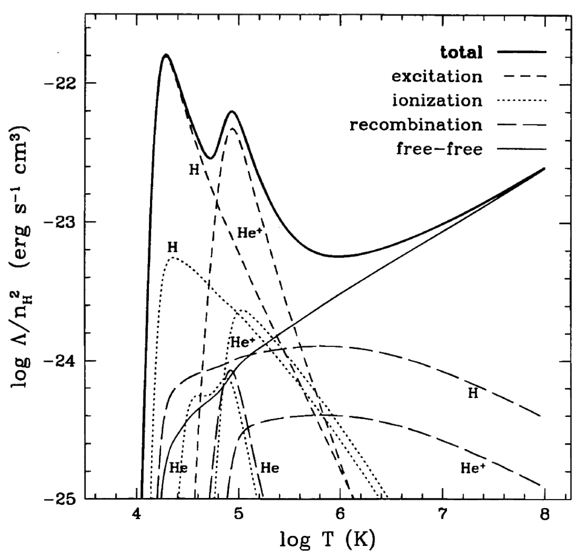

Fig. 4.— Cooling rate as a function of temperature for a primordial composition of gas. At high temperature, the cooling rate increases as $\Lambda \propto T^{1/2}$. Source: Katz, Weinberg & Hernquist, 1996, ApJS, 105, 19

Cooling is essential in the formation of galaxies. Under the assumption of ionization equilibrium, one can calculate the ratio $t_{\rm cool}/t_{\rm ff}$, which is a function of the density, n, temperature, T and metallicity, Z. This function is plotted in Figure 6.

It is clear from the figure that halos with $M_{\rm vir} < 10^8\ M_\odot$ (or $V_{\rm vir} < 20$ km/s) can't cool their gas. (Recall that $M_{\rm gas} \simeq 0.15 M_{\rm vir}$). Similarly, halos with $M_{\rm vir} > 10^{12}\ M_\odot$ ($Z = 0$)

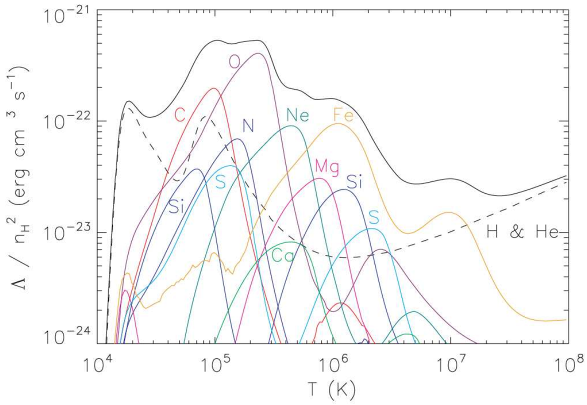

Fig. 5.— Cooling rate as a function of temperature for solar-abundances composition of gas at collisionless ionization equilibrium. Source: Wiersma, Schaye & Smith, 2009, MNRAS, 393, 99

or $M_{\rm vir} > 10^{13}\ M_\odot$ ($Z = Z_\odot$). This is why clusters and group of galaxies at present time usually contain large amount of hot gas.

The critical mass for effective cooling is similar to that of most massive galaxies, suggesting that the physics of cooling plays an important role in limiting the mass of galaxies. In other words, more massive galaxies can't form because they can't efficiently cool their gas.

This argument, though, is applicable only to galaxies that form directly from gas cooling in dark matter halos. It does not apply in cases where massive galaxies are formed due to merger of smaller galaxies, which is possibly an important channel of formation of giant elliptical galaxies. Furthermore, realistically halos have both density and temperature profiles, which complicates the calculations. In fact, in 1978, While & Rees showed that when mergers of dark matter halos are accounted for, most of the gas should have already cooled and formed stars; this model, thus in fact over-predicts the number density of faint galaxies, known as the **over-cooling** problem. This model, thus, requires some extra process that **heats** the gas during the formation of galaxies !.

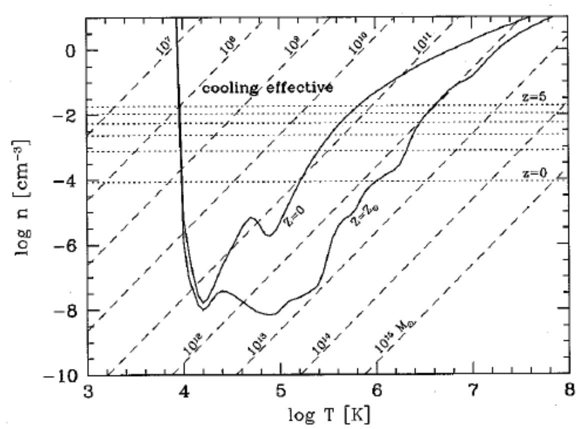

Fig. 6.— Cooling diagram showing the locus $t_{\rm ff} = t_{\rm cool}$ in the $n - T$ plane. The upper and lower curves correspond to gas with zero and solar metallicity, respectively. The tilted dashed lines are lines of constant total mass. Cooling is effective for clouds with n and T above the locus. Source: [1], Figure 8.6.

A. The Virial Theorem

The collisionless Bolzmann Equation (CBE) that describes conservation of phase-space density around each point is

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x_i} v_i + \frac{\partial f}{\partial v_i} dv_i = 0 \tag{A1}$$

When the velocity is derived from a potential, $dv_i = -\frac{\partial \Phi}{\partial x_i}$, Equation A1 takes the form

$$\frac{\partial f}{\partial t} + \frac{\partial f}{\partial x_i} v_i - \frac{\partial \Phi}{\partial x_i}\frac{\partial f}{\partial v_i} = 0 \tag{A2}$$

(This is known as Vlasov equation).

Recall that the concept of distribution function f enables to write the density and the average velocity in the i-th direction as

$$\rho = \int f d^3v, \quad \rho\langle v_i\rangle = \int f v_i d^3v. \tag{A3}$$

As a concrete example, if we integrate Equation A2 over d^3v, recalling that v_i is independent on x_i (hence $\partial v_i/\partial x_i = 0$), Equation A2 becomes

$$\rho + \frac{\partial \rho\langle v_i\rangle}{\partial x_i} - \frac{\partial \Phi}{\partial x_i}\int \frac{\partial f}{\partial v_i} d^3v = 0 \tag{A4}$$

The last term can be expanded using the divergence theorem, $\int \frac{\partial f}{\partial v_i} d^3v = \int f \cdot d^2S = 0$, where we used the fact that $f \to 0$ as $|v| \to \infty$. We are thus left with

$$\rho + \frac{\partial \rho \langle v_i \rangle}{\partial x_i} = 0, \tag{A5}$$

which is nothing but the continuity equation.

Returning to Equation A2, multiplying it by v_j and integrating over d^3v, we get

$$\frac{\partial \rho \langle v_j \rangle}{\partial t} + \frac{\partial \rho \langle v_i v_j \rangle}{\partial x_i} - \frac{\partial \Phi}{\partial x_i} \int v_j \frac{\partial f}{\partial v_i} d^3v = 0 \tag{A6}$$

The last term can be integrated by parts, to give

$$\int v_j \frac{\partial f}{\partial v_i} d^3v = \int \frac{\partial v_j f}{\partial v_i} d^3v - \int \frac{\partial v_j}{\partial v_i} f d^3v = \int v_j f \cdot d^2S - \int \delta_{ij} f d^3v = -\delta_{ij}\rho.$$

We thus finally obtain

$$\frac{\partial \rho \langle v_j \rangle}{\partial t} + \frac{\partial \rho \langle v_i v_j \rangle}{\partial x_i} + \rho \frac{\partial \Phi}{\partial x_j} = 0. \tag{A7}$$

These are called **momentum equations** (note that there are three equations, for $j = 1, 2, 3$).

Multiplying Equation A7 with x_k and integrating over d^3x (real space) gives

$$\frac{\partial}{\partial t} \int \rho x_k \langle v_j \rangle d^3x = -\int x_k \frac{\partial \rho \langle v_i v_j \rangle}{\partial x_i} d^3x - \int \rho x_k \frac{\partial \Phi}{\partial x_j} \tag{A8}$$

The first term on the right hand side can be put in a simpler form, using integration by parts:

$$\begin{aligned} \int x_k \frac{\partial \rho \langle v_i v_j \rangle}{\partial x_i} d^3x &= \int \frac{\partial \rho x_k \langle v_i v_j \rangle}{\partial x_i} d^3x - \int \rho \langle v_i v_j \rangle \frac{\partial x_k}{\partial x_i} d^3x \\ &= -\int \delta_{ki} \rho \langle v_i v_j \rangle d^3x \\ &= -\int \rho \langle v_k v_j \rangle d^3x = -2K_{kj} \end{aligned} \tag{A9}$$

where in the second line we used the assumption that ρ vanishes at large radii.

Here, we defined the **kinetic energy tensor**

$$K_{ij} \equiv \frac{1}{2} \int \rho \langle v_i v_j \rangle d^3x. \tag{A10}$$

It is customary to split the kinetic energy tensor into contributions from *ordered* and *random* motions:

$$K_{ij} \equiv T_{ij} + \frac{1}{2}\Pi_{ij} \tag{A11}$$

where

$$T_{ij} \equiv \frac{1}{2}\int \rho\langle v_i\rangle\langle v_j\rangle d^3x \; ; \quad \Pi_{ij} \equiv \int \rho\sigma_{ij}^2 d^3x \tag{A12}$$

where we defined the **stress tensor**,

$$\sigma_{ij}^2 \equiv \langle v_i v_j\rangle - \langle v_i\rangle\langle v_j\rangle. \tag{A13}$$

(the stress tensor measured the random motion of objects (e.g., stars) around the streaming part, $\langle v_i\rangle\langle v_j\rangle$; it describes the anisotropic pressure).

In addition to K_{ij}, we cefine the **potential energy tensor**,

$$W_{ij} \equiv -\int \rho x_i \frac{\partial\Phi}{\partial x_j}. \tag{A14}$$

Using these definitions in Equation A8, we get

$$\frac{\partial}{\partial t}\int \rho x_k\langle v_j\rangle d^3x = 2K_{kj} + W_{kj}. \tag{A15}$$

Since both K_{kj} and W_{kj} are symmetric ($K_{kj} = K_{jk}$, $W_{kj} = W_{jk}$), we can write

$$\frac{1}{2}\frac{d}{dt}\int \rho\left[x_k\langle v_j\rangle + x_j\langle v_k\rangle\right] d^3x = 2K_{jk} + W_{jk}. \tag{A16}$$

We finally define the **moment of inertia tensor**,

$$I_{ij} \equiv \int \rho x_i x_j d^3x. \tag{A17}$$

Differentiating with respect to time, and using the continuity equation (A5), we get

$$\begin{aligned} \frac{dI_{jk}}{dt} &= \int \frac{\partial\rho}{\partial t} x_j x_k d^3x \\ &= -\int \frac{\partial\rho\langle v_i\rangle}{\partial x_i} x_j x_k d^3x \\ &= -\int \frac{\partial\rho\langle v_i\rangle x_j x_k}{\partial x_i} d^3x + \int \rho\langle v_i\rangle \frac{\partial(x_j x_k)}{\partial x_i} d^3x \\ &= \int \rho\langle v_i\rangle\left[x_j\delta_{ik} + x_k\delta_{ij}\right] d^3x \\ &= \int \rho\left[x_j\langle v_k\rangle + x_k\langle v_j\rangle\right] d^3x \end{aligned} \tag{A18}$$

from which we deduce that the left hand side of Equation A16 can be written as

$$\frac{1}{2}\frac{d}{dt}\int \rho\left[x_k\langle v_j\rangle + x_j\langle v_k\rangle\right] d^3x = \frac{1}{2}\frac{d^2 I_{jk}}{dt^2} \tag{A19}$$

Equation A16 can thus be written as the **tensor virial theorem**,

$$\frac{1}{2}\frac{d^2 I_{jk}}{dt^2} = 2T_{jk} + \Pi_{jk} + W_{jk}. \tag{A20}$$

If the system is in a steady state, the moment of inertia tensor is stationary, and the tensor virial theorem reduces to

$$2K_{ij} + W_{ij} = 0. \tag{A21}$$

Of particular interest is the **trace** of the tensor virial theorem, which related the **total kinetic energy**, $E_k = \frac{1}{2}M\langle v^2\rangle$ to the **total potential energy**, $W = \frac{1}{2}\int \rho(\vec{x})\Phi(\vec{x})d^3x$. We get:

$$\begin{aligned} Tr(K_{ij}) \equiv \sum_{i=1}^{3} K_{ii} &= \frac{1}{2}\int \rho(\vec{x})\left[\langle v_1^2\rangle(\vec{x}) + \langle v_2^2\rangle(\vec{x}) + \langle v_3^2\rangle(\vec{x})\right] d^3x \\ &= \frac{1}{2}\int \rho(\vec{x})\langle v^2\rangle(\vec{x})d^3x \\ &= \frac{1}{2}M\langle v^2\rangle = E_k \end{aligned} \tag{A22}$$

where we have used

$$\langle v^2\rangle = \frac{1}{M}\int \rho(\vec{x})\langle v^2\rangle(\vec{x})d^3x.$$

Similarly, the trace of the potential energy tensor is equal to the total potential energy,

$$Tr(W_{ij}) = E_p = \frac{1}{2}\int \rho(\vec{x})\Phi(\vec{x})d^3x \tag{A23}$$

(this result is derived using $\Phi(\vec{x}) = -G\int \frac{\rho(\vec{x})}{|\vec{x}' - \vec{x}|}d^3x$). We thus obtain the familiar **scalar virial theorem**,

$$2E_k + E_p = 0. \tag{A24}$$

REFERENCES

[1] H. Mo, F. van den Bosch and S. White, *Galaxy Formation and Evolution* (Cambridge), chapters 7,8.

[2] J. Binney and S. Tremaine, *Galactic Dynamics* (Princeton), chapter 9

[3] L.S. Sparke and J.S. Gallagher, *Galaxies in the Universe* (Cambridge), chapter 8

[4] T. Padmanabhan, *Structure Formation in the Universe* (Cambridge), chapter 8.